A Stereo Vision Lip Tracking Algorithm and Subsequent Statistical Analyses of the Audio-Video Correlation in Australian English

Roland Goecke

A thesis submitted for the degree of Doctor of Philosophy of
The Australian National University

30 January 2004

Research School of Information Sciences and Engineering
The Australian National University
Canberra, Australia

Declaration

This thesis describes the results of research undertaken in the Computer Sciences Laboratory, Research School of Information Sciences and Engineering, The Australian National University, Canberra. This research was supported by scholarships from The Australian National University and the Cooperative Research Centre for Advanced Computational Systems, Canberra.

The results and analyses presented in this thesis are my own original work, accomplished under the supervision of Doctor J Bruce Millar, Professor Alexander Zelinsky, and Doctor Jordi Robert-Ribes (SingTel Optus Pty Limited, Sydney), except where otherwise acknowledged. This thesis has not been submitted for any other degree.

Roland Goecke
Computer Sciences Laboratory
Research School of Information Sciences and Engineering
The Australian National University
Canberra, Australia
30 January 2004

Acknowledgments

First of all I would like to thank the members of my supervisory panel Bruce Millar, Alex Zelinsky, and Jordi Robert-Ribes. Without their help throughout the project, this thesis would not have been possible. They have added invaluable insight from their respective areas which has been a big help in this multi-disciplinary project. I enjoyed the many discussions we had and would like to thank them wholeheartedly for guiding me. Bruce, thank you for being an excellent supervisor and mentor and believing in me. Alex, I would like to thank you for your constructive criticism and for making me a 'part' of the Robotic Systems Laboratory. Jordi, a big thank you for the time and effort you have put in to keep in contact from Sydney and for the support you have given me throughout the project.

Secondly, I would like to express my gratitude to the ANU Statistical Consulting Unit who played an instrumental role in explaining the statistical analyses that form such a big part of this PhD project. In particular, I would like to thank Ann Cowling, John Maindonald, and Jeff Wood. In addition, I want to thank Stéphane Dray, Universite Lyon 1 (France), for his expertise on canonical correlation analysis and coinertia analysis. My gratitude also goes to Jim Ramsay, McGill University (Canada) for the many discussions we had on the usage of Functional Data Analysis. I would also like to thank Cheol-Woo Jo, Changwon National University (Korea), and Lionel Revéret, Institut de la Communication Parlée, Grenoble (France), for their help in the initial stages of this project.

Next, this thesis and the whole PhD project would not have been possible without the support of the staff and fellow students at the Research School of Information Sciences and Engineering. Thank you to all of you! In particular, I would like to thank the administrative and support staff of the Computer Sciences Laboratory and the Department of Systems Engineering. Special thanks go to Dave Davis, who helped me tremendously during the time after the hard disk had crashed and I had lost all of my data.

Although on the sidelines of this project, my gratitude goes to the Human Language Technologies Department at IBM for giving me the opportunity of a research internship at the TJ Watson Research Center in Yorktown Heights, New York (USA). It was a great experience and it helped to get a broader view of this complex field and the many applications of audio-visual speech processing. I want to thank Gerasimos 'Makis' Potamianos, Chalapathy Neti, and Giri Iyengar.

No list of thankyous would be complete without mentioning the friends I have met during my time in Canberra and who have helped me to keep my sanity. There is unfortunately not enough space to mention all their names but a few. So thank you Janine, Kris, Megan, and Nina. A special thankyou goes to Mika for introducing me to bushwalking in Australia. It has made my Australian experience so much more complete and what started with the bushwalks, turned into a great friendship. Last but not least, there are the numerous friends at the ANU Mountaineering Club and the ANU Graduate House. I hope I could give you some of that back, of what you gave me.

Ein sehr großes Dankeschön geht an meine Eltern Barbara und Reinhard für ihre unendliche Liebe und Unterstützung, nicht nur während der Doktorarbeit, sondern während meines ganzen Lebens. Mit Eurer Hilfe und dem Wissen, daß Ihr immer für mich da seid, bin ich zu der Person geworden, die ich heute bin. Ihr seid einfach die besten Eltern, die ich mir wünschen könnte. Danke!

Almost finally, I would like to thank my partner Nicole with all my heart for being in my life! You had been a good friend already and that you now want to share your life with me, is as much an honour as it is fun. I look forward to us taking the next steps together and starting a new chapter in our lives, now that we both have finished our PhDs. You are the love of my life!

And really the last thank you goes to all those countless, wonderful Australians who I have met during my four years in Canberra and elsewhere and who have helped to make this such a great experience for me. You have welcomed me unconditionally and with an openness and heartiness, that is rare to find these days. Australia will always have a special place in my heart!

Abstract

Human perception of the world is inherently multi-sensory because the information provided is multimodal. The perception of spoken language is no exception. Beside the auditory information, there is visual speech information as well, provided by the facial movements as a result of moving the articulators during speech production. Visual speech information contributes to speech perception in all kinds of audio conditions, but its effect is perhaps most readily noticed in noisy audio conditions. Various research groups around the world have studied the effects of incorporating visual speech information in automatic speech recognition (ASR) systems in recent years. They have found that audio-video (AV) ASR systems result in an improved recognition rate compared to audio-only systems, in particular in noisy audio conditions. Exactly how to incorporate the additional visual speech information best is still not known.

This study aims to extend our knowledge of relationships between audio and video speech parameters. It investigates ways of describing such relationships using statistical analyses and their application to the example of Australian English (AuE). The work described in this thesis is multi-disciplinary. Apart from the statistical analyses, it also required algorithms to extract speech parameters and a corpus of AV speech sequences, which were not readily available.

A novel non-intrusive automatic lip tracking algorithm is presented, which uses a stereo camera system to enable accurate 3D measurements of facial feature points without the need for artificial markers on the face. Due to the lack of an AV speech corpus for AuE, a new modular framework for AV speech corpora was developed and followed in a newly created corpus for AuE.

Equipped in such ways, it was possible to test the hypothesis that combinations of audio and video speech parameters are related, rather than single parameters, and that these combinations are phoneme-specific. Based on articulatory theory, it is clear that the audio and video domain are related in some way and to some

extent because the visible speech articulators form a part of the whole set of articulators. However, it also means that not all of the information contained in the audio modality has equivalent information in the video modality. The set of audio speech parameters was formed by voice source excitation frequency F_0, formant frequencies F_1, F_2, F_3, and RMS energy. Mouth width, mouth height, protrusion of upper and lower lip, and the novel teeth visibility measure *relative teeth count* formed the video speech parameter set.

Extensive univariate and multivariate statistical analyses, such as pairwise linear correlation analysis, principal component analysis, statistical shape analysis, canonical correlation analysis, and coinertia analysis, were performed to explore the AV relationships in AuE. The AV relationships found by this study support the hypothesis that linear combinations of parameters correlate well ($r = 0.5{-}0.8$) across the two modalities and that their composition is phoneme-specific. The results show that with the given parameter sets, between one fifth and one third of the variance in either modality can be recovered from the other modality. For visible speech information purely based on the lips, this agrees with studies on human speech perception found in current literature. Further investigations are required to test the stability of the found relationships and their suitability for a rule-based AV ASR system.

Publications

During the course of this study, the following refereed conference and journal papers were published.

- C.-W. Jo, R. Goecke, and J.B. Millar. *Collection of Korean Audio-Video Speech Data*. In Proceedings of the Second International Workshop on East-Asian Language Resources and Evaluation (Oriental COCOSDA-99), pages 73–76, Taipei, Taiwan, May 1999.

- C.-W. Jo, R. Goecke, and J.B. Millar. *Design and Collection of Korean Audio-Video Speech Data*. In Proceedings of the International Conference on Speech Processing ICSP'99, pages 519–523, Seoul, Korea, August 1999.

- C.-W. Jo, R. Goecke, and J.B. Millar. *Collection of Korean Audio-Video Speech Data*. Speech Sciences, Vol.7(1), pages 5–15, March 2000.

- R. Goecke, J.B. Millar, A. Zelinsky, and J. Robert-Ribes. *Automatic Extraction of Lip Feature Points*. In Proceedings of the Australian Conference on Robotics and Automation ACRA2000, pages 31–36, Melbourne, Australia, August 2000.

- R. Goecke, Q.N. Tran, J.B. Millar, A. Zelinsky, and J. Robert-Ribes. *Validation of an Automatic Lip-Tracking Algorithm and Design of a Database for Audio-Video Speech Processing*. In Proceedings of the 8th Australian International Conference on Speech Science and Technology SST2000, pages 92–97, Canberra, Australia, December 2000. Australian Speech Science and Technology Association (ASSTA).

- G. Loy, R. Goecke, S. Rougeaux, and A. Zelinsky. *Stereo 3D Lip Tracking*. In Proceedings of the Sixth International Conference on Control, Automation,

Robotics and Computer Vision ICARCV2000, CD-ROM, Singapore, December 2000.

- R. Goecke, J.B. Millar, A. Zelinsky, and J. Robert-Ribes. *Stereo Vision Lip-Tracking for Audio-Video Speech Processing.* In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP 2001, Student Forum, CD-ROM, Salt Lake City (UT), USA, May 2001. IEEE.

- R. Goecke, J.B. Millar, A. Zelinsky, and J. Robert-Ribes. *Analysis of Audio-Video Correlation in Vowels in Australian English.* In Proceedings of the International Conference on Auditory-Visual Speech Processing AVSP2001, pages 115–120, Aalborg, Denmark, September 2001.

- R. Goecke and J.B. Millar. *Statistical Analysis of the Relationship between Audio and Video Speech Parameters for Australian English.* In Proceedings of the ISCA Tutorial and Research Workshop on Auditory-Visual Speech Processing AVSP2003, pages 133–138, St Jorioz, France, September 2003.

Abbreviations

AAM	Active appearance model
ANDOSL	Australian national database of spoken language
ANOVA	Analysis of variance
ASM	Active shape model
ASR	Automatic speech recognition
AuE	Australian English
AV	Often audio-visual but more correctly audio-video or auditory-visual
AVOZES	Audio-video Australian English speech data corpus
AVSP	Audio-video speech processing
CANCOR	Canonical correlation analysis
CCA	Canonical correspondence analysis
CMP	Categorical model of perception
COIA	Coinertia analysis
CV	Canonical variate
CVC	Consonant-vowel-consonant syllable
DCT	Discrete cosine transform
DFT	Discreet Fourier transform
DI	Direct identification model
DMC	Discriminative model combination
DR	Dominant recoding model
DTW	Dynamic time warping
DV	Digital video
ESPS	Entropic signal processing software
F_0	Voice source excitation frequency
F_1	First formant frequency
F_2	Second formant frequency

F_3	Third formant frequency
FDA	Functional data analysis
FFT	Fast Fourier transform
FIR	Finite impulse response filter
FLMP	Fuzzy logical model of perception
HiLDA	Hierarchical LDA
HLAC	Higher order local autocorrelation
HMM	Hidden Markov model
HSI	Colour model consisting of hue, saturation, and intensity samples
HSV	Colour model consisting of hue, saturation, and value samples (similar to HSI)
IIR	Infinite impulse response filter
LDA	Linear discriminant analysis
LPC	Linear predictive coding
MANOVA	Multivariate analysis of variance
MH	Mouth height (vertical distance of mouth opening)
MLLT	Maximum likelihood linear transform
MLP	Multi-layer perceptron
MR	Motor space recoding model
MSA	Multiscale spatial analysis
MS-TDNN	Multi-state time-delay neural network
MUX	Multiplexer
MW	Mouth width (horizontal distance of mouth opening)
NCC	Normalised cross-correlation
NTSC	TV and video coding standard
PC	Principal component
PCA	Principal component analysis
PCM	Pulse Code Modulation
PLL	Protrusion of lower lip midpoint

PLP	Perceptual linear predictive analysis
PUL	Protrusion of upper lip midpoint
R	Statistical software package similar to S-PLUS
RDA	Redundancy analysis
RGB	Colour model consisting of red, green, and blue samples
RMS	Root mean square energy
RTC	Relative teeth count (measure of teeth visibility)
RV	Measure of similarity between two sets of variables
SI	Separate identification model
SNR	Signal-to-noise ratio
VCV	Vowel-consonant-vowel syllable
YUV	Colour model consisting of one intensity and two colour difference samples (used in NTSC)
2D	Two-dimensional
3D	Three-dimensional

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The Australian often speaks without obviously opening his lips at all, through an immobile slit, and in extreme cases through closed teeth.

Hector Dinning

The Australian Scene, Sydney, 1939

With the rapid advances in computer technology over the last two decades and its many obvious and hidden uses in today's life, the way people interact with computer systems has become an important aspect and has received significantly more attention in recent years. As computer systems become commonplace, they are used more and more by non-experts, so that user-friendly systems are required. Traditional ways of interaction through the usage of keyboards, mice, and monitor displays are often cumbersome or simply impractical in many application areas outside the laboratory or office environment. Human-computer interaction is inherently and unavoidably social. People often respond to computers as if they were human. The social and emotional aspects of that interaction form an important part of the field of human-computer interaction and the future direction is undoubtedly towards more human-like interactions with computer systems.

Automatic speech processing has long been regarded as an important means of human-computer interaction because of its naturalness, but only recent advances in technology, combined with a significant reduction in cost, have made the widespread

use of speech processing technology viable. For example, telephone companies such as Telstra[1] employ automatic speech recognition tools in their directory services. PC-based automatic speech recognition systems like IBM's ViaVoice are available for dictating letters. Similarly, the synthesis of voices has progressed and is utilised in many areas, the film industry being but one example.

Technology developments in recent years have made continuous speech recognition possible in reasonably good acoustic environments such as the office. However, these systems can fail unpredictably if the conditions are not ideal, for example, when faced with acoustic noise, changes in the rate of speaking, or certain speaker characteristics which cause no problem to human perceivers.

Human perception of the environment is inherently multi-sensory because the information provided is multimodal. Humans use single senses only rarely, usually the different senses are employed in a coordinated way. When we touch an object, we also see it with our eyes and we might also smell its odour. The information from the object is transferred through different modalities and different sensors receive the information, but our mind combines the various information channels again and produces a coherent understanding of the object's properties.

Human perception of spoken language is no exception to this multi-sensory, multimodal perception of the environment. To the naive observer, speech perception is often considered to be a unimodal process, purely based on the audio modality. However, human beings make use of visual information, provided by the facial movements during speech production, as well. As Burnham and Sekiyama [Burnham 02] point out, visual speech information contributes to speech perception not only when the acoustic information is degraded by noise or when the listener is hearing-impaired, but also in clear audio conditions as can be seen in the McGurk effect [McGurk 76, MacDonald 78] (see the literature review in Chapter 2 for details). Hence, human speech perception is really a multi-sensory, multimodal process.

In a similar fashion to human speech processing, the incorporation of additional visual information, extracted from facial movements during speech production, has

[1] An Australian telecommunications company.

been shown to overcome some of the limitations of audio-only automatic speech recognition systems and to improve the recognition rate, particularly in conditions where the auditory information is degraded by noise (see the literature review in Chapter 2 for details). This combination of auditory and visual speech information leads us to the relatively new, but fast growing field of Audio-Video (or Auditory-Visual) Speech Processing (AVSP) and the research in this field will bring us closer to natural, human-like interactions with computers.

1.1 Motivation and Aim

The aim of the work described in this thesis was to enhance the understanding of some of the many complex aspects of AVSP for automatic speech recognition. Of particular interest were a scientific understanding of the interplay of the auditory and the visual modality of the speech signal. Although studies by various research groups around the world have shown that adding visual speech information in the recognition process is advantageous, little is known on how these two modalities interact and how this interaction can be exploited best.

To gain such an understanding, the relationships of various parameters of the audio and video speech signals were statistically analysed on the sequences of a purpose-built, yet general and comprehensive, new audio-video speech data corpus of Australian English (AVOZES). No such data corpus has previously been publicly available for Australian English (AuE). While 'standard' parameters describing the audio speech signal are known, the visual speech parameters most useful for automatic speech recognition are still debated. This study investigated the empirical relationship of visual speech parameters, derived from geometric features such as the lip corners, with auditory speech parameters like formants. These parameters were chosen because their more direct relation to the vocal tract and the articulators, compared to other parameters, is expected to facilitate the interpretation of the results. The central theme of this study was the investigation of the presence and the nature of the AV relationships in AuE. Along the way, the suitability of the chosen visual speech parameters to explain variance in the auditory speech

parameters (and vice versa) was explored.

Moreover, this study presents new methodology for the extraction of visual speech parameters. Recently developed methods for fast and reliable facial feature tracking, which is a pre-requisite for accurate parameter measurement, were applied and tested. In particular, the use of a stereo vision system, new to the field of AVSP, was expected to deliver more accurate and more reliable results than a monocular system can, because of the superior capability of 3D reconstruction. This means the speaker is free to move the head in a natural way during speech production, while the system still provides accurate measurements of the location of facial feature points in 3D. A novel lip tracking system extending the stereo vision face tracking system is presented, which does not require the use of any artificial markers.

Furthermore, the relatively young multivariate statistical technique of coinertia analysis was introduced to the field of AVSP. Coinertia analysis was developed for ecological studies, where small sample sizes and large parameter sets often conflict. This problem also occurs frequently in spoken language studies. Traditional multivariate statistical analyses can result in stability problems. Coinertia analysis has the advantage that its results are independent of the sample size, which made it very suitable for this investigation.

From Hector Dinning's quote at the beginning of this chapter, one could expect AVSP to be a lost cause for speakers of AuE. AuE is traditionally said to be lazy, nasal, drawling, not clear, similar to Cockney, monotonous, flat, and marred by lip-laziness. However, Mitchell and Delbridge [Mitchell 65] comprehensively refute most of these views. AuE has its own characteristic rhythm and intonation as well as a shift in vocalic sounds, particularly diphthongs, that make it stand out when compared to other regional dialects of English but perhaps surprisingly for such a large country, regional differences are very small. Only one AuE dialect exists, but different pronunciation varieties occur. An interesting side aspect was thus to compare the AV relationships of AuE within its varieties, although such a comparison could only be a starting point for a more comprehensive study on this aspect due to the limited sample size. If lip-laziness was indeed prominent in AuE, AV relationships were expected, which are not as strong as they are for strong lip

movements.

The AuE 'speech varieties' are usually categorised as broad, general, and cultivated but categories are not discrete entities, rather, they span a continuum with considerable phonetic overlap. Perhaps broad AuE, with its characteristic vocalic pronunciations and consonantal unclarity, is the variety that Dinning had in mind when making the above mentioned critcisms. However, speech production is an auditory-visual event in any case and is perceived as such. Hence, an investigation of the statistical relationship between auditory and visual speech parameters is an interesting topic for AuE, and certainly in general anyway, and one that has not received much attention so far.

1.2 Chapter Outline

Chapter 2 contains a comprehensive literature review. First, AVSP by humans is reviewed, including the McGurk effect, general aspects of human speech production and perception, and models of the integration of the audio and the video modality. Secondly, the characteristics of AuE are presented. Thirdly, the area of AVSP by machines is visited which includes topics such as the fundamentals of audio-only and video-only automatic speech recognition, the methods of automatic extraction of video speech parameters, the ways of integrating the two modalities in AV automatic speech recognition, and AV speech data corpora.

A novel real-time lip tracking system is presented in Chapter 3. This system extends a real-time stereo vision face tracking system to deliver accurate video speech parameters. The way the face and lip tracking systems work in measuring facial feature points in 3D is explained in detail. The accuracy of the lip tracking system is validated in an experiment and the results of this validation are given.

In Chapter 4, the design principles of the new AV speech data corpus for Australian English (AVOZES) are introduced. A general framework for modular, extendable AV speech corpora is proposed and this framework is followed in the AVOZES corpus. In addition, the chapter details the experimental setup and the recording environment for the creation of the data corpus and the subsequent sta-

tistical analysis of the relationship of audio and video speech parameters.

Chapter 5 describes the methods of analysis. This chapter is structured in two parts. Firstly, methods for the separate analysis of the audio and the video modality for extracting parameters that describe the speech signal in these modalities are discussed. Secondly, the theoretical background of the performed statistical analyses is presented. This study focussed on the analysis of the originally measured (static) parameters and left the analysis of derived (dynamic) parameters to future work. Both traditional methods of univariate and multivariate statistical analysis as well the relatively new methods of coinertia analysis and functional data analysis were used.

In Chapter 6, the results of the statistical analyses are presented and discussed. This chapter starts with remarks about the extracted parameters by discussing observations made on the measurements and by using linear discriminant analysis. Next, each parameter set is tested for redundancies by applying principal component analysis and linear correlation analysis to each set. This is followed by a statistical shape analysis of the parameter curves for each phoneme under investigation to determine patterns in the temporal domain. To test the hypothesis of combinations of parameters being related across the two modalities, the multivariate statistical analyses of canonical correlation analysis and coinertia analysis were performed and their results discussed in detail. Also, an example is given on how curve registration using functional data analysis can aid the analysis.

Finally, Chapter 7 presents a summary and the conclusions of the work described in this thesis as well as an outlook on open issues for future work.

Chapter 2

Literature Review

In this chapter, a comprehensive overview of the related literature is given to set the scene for the remaining parts of this study. Audio-video speech processing (AVSP) is a complex field which draws from many other disciplines, for example, linguistics, psychology, machine learning, and computer vision. Such a multi-disciplinary field offers different angles on the same topic and research can be taken in different directions. This chapter provides an overview of the most important areas in AVSP, keeping in mind the aim of this study, to investigate the statistical relationship of audio and video speech parameters.

This chapter consists of four main sections: AVSP by humans, characteristics of Australian English (AuE), AVSP by machines, and statistical analyses of audio-video (AV) relationships. Section 2.1 gives a detailed overview of human speech production and perception mechanisms from an AVSP angle. It includes discussions of the McGurk effect, of the sources of visual speech information, and of models of the integration of the audio and the video modality. Section 2.2 describes the characteristics of AuE in general and the continuum of accent variation spanned by the varieties of AuE in more detail. Thirdly, Section 2.3 outlines the methods used in audio-only, video-only, and AV automatic speech recognition (ASR) systems. This includes a discussion of visual feature extraction methods as well as of ways to integrate the two modalities. The section ends with an overview of AV speech data corpora. Finally, Section 2.4 discusses previous analyses of AV relationships.

2.1 Audio-Video Speech Processing by Humans

Extensive psychological and linguistic research has shown that human speech perception does not only involve the processing of auditory information but also the processing of visual speech information. Humans use single senses only rarely. The different senses are employed in a coordinated way and the world is perceived multimodally. Inspection of an object by one sense also leads to expectations and predictions about what will be experienced by other senses. For example, simply by looking at a surface, one's mind creates an expectation of whether the surface is likely to be smooth or rough, soft or hard, and so on, when touched.

This section starts with looking at the McGurk effect which generated a surge of activity in the field of AVSP (Section 2.1.1). Section 2.1.2 introduces the important concepts of phonemes and visemes, before background information on speech production and perception mechanisms (Sections 2.1.3 – 2.1.5) are provided. Next, the sources of visual speech information are discussed in Section 2.1.6. Finally, models of integrating auditory and visual speech information are presented in Section 2.1.7.

2.1.1 The McGurk Effect

In many situations in which spoken communication between humans occurs, the listener cannot only hear the speaker, but also see them. Although speech processing is often regarded as merely an auditory process, it is influenced by vision as well. McGurk and MacDonald showed this effect in their seminal work [McGurk 76]. In their experiments, sequences of /ba-ba/, /ga-ga/, /pa-pa/, and /ka-ka/ were recorded. Audio and video signals were rearranged to create sequences such as audio-ba + video-ga, audio-ga + video-ba, audio-pa + video-ka, and audio-ka + video-pa. These sequences were shown to subjects from different age groups who were asked to repeat what they had just 'heard'.

Two types of responses were found: *fused* and *combined*. Audio-ba + video-ga and audio-pa + video-ka resulted in fused responses of /da-da/ and /ta-ta/, respectively. In other words, the subjects perceived something that was not present in either modality — the so-called *McGurk effect*. Audio-ga + video-ba and audio-ka

+ video-pa resulted in combined responses, such as /bagba/, /gabga/, and /kapka/, /pakpa/ etc. These are composite responses comprising relatively unmodified elements from each modality. These observations suggest that people with normal hearing under good listening conditions employed lip reading skills. The information received from the eye influenced the perceptual process in such a way, that they were 'hearing' something different from what was presented directly to their ears. Even with objective knowledge about the McGurk effect, the illusions do not disappear. A previously auditory-visually perceived /da/ is heard correctly as /ba/ by simply closing the eyes, only to become /da/ again after opening the eyes again.

These findings were confirmed in a further study by MacDonald and McGurk [MacDonald 78]. The authors proposed the theory that manner of articulation (voiced or voiceless, oral or nasal, stopped or continuant, etc.) is detected by ear, while place of articulation (e.g. front, central, or back articulation) is detected by eye. Similar to the first study, audio and video recordings of /pa, ba, ta, da, ka, ga, ma, na/ were combined in all possible ways and presented to subjects whose responses were recorded. Their results showed that combinations within the group of labials /p b m/ and within the group of non-labials /t d k g n/ produced very few perception errors. In contrast, combinations of labial audio and non-labial video (and vice versa) led to a considerable number of fused and combined responses as predicted by their manner-place theory. However, the nature of these errors found in responses to non-labial audio and labial video differed from what was expected.

McGurk effects exist for speakers of many languages. For example, they were found for German and Spanish speakers [Fuster-Duran 96], Japanese speakers [Sekiyama 98], Dutch speakers [Vroomen 92], French speakers [Colin 98], Finnish speakers [Sams 97], and Thai speakers [Burnham 96].

Age Matters

An interesting result of McGurk's initial experiments is that speech perception of adults is more strongly influenced by visual stimuli than is that of children. Speech perception is subject to age-related changes. When speech perception is dominated

by a single modality, this tends to be the auditory for children and the visual for adults [McGurk 76]. With increasing age and acquired language knowledge, the awareness of the relationship between speech sounds and their associated visible articulations grows. However, infants as young as four months have been shown to perceive the McGurk effect [Burnham 96]. Through long experience with natural conversation, humans have a strong expectation for lips and voices to convey the same speech information and they seem to have an implicit understanding of the constraints placed upon speech production by the activity of the visible articulators. For example, the plosive /b/ is rarely perceived unless one sees the lips closing.

2.1.2 Phonemes and Visemes

Before giving an overview of the processes involved in speech production and processing in the following sections, two important concepts shall be introduced: *phonemes* and *visemes*. A phoneme is a member of the set of auditory speech sounds (in any given language) that serves to distinguish the meaning of one word from another. For example, /p/ and /b/ are separate phonemes[1] in English because they distinguish words such as *pet* and *bet*. It may consist of several phonetically distinct articulations (*allophones*), which are regarded as identical by native speakers, since one articulation may be substituted for another without any change of meaning.

By analogy, a viseme is a member of the set of visually distinguishable articulations [Fisher 68]. Taking the example of /p/ and /b/ again, they belong to the same viseme category as they are both bilabial sounds. /p/ and /b/ are more readily separable auditorily than visually. In contrast, /θ/ and /f/ are auditorily very similar but visually distinguishable by the visibility of the tongue and the teeth, respectively. Thus, the set of phonemes is at least partially different from the set of visemes and, hence, there is no 1–1 mapping between the two. The difference can be used to resolve ambiguities evolving in one modality but not in the other.

[1] The alphabet of the International Phonetic Association (IPA) is used throughout this thesis [IPA 99].

No 1–1 Mapping between Phonemes and Visemes

Although visible speech articulation, particulary lip motion, is potentially informative, it is also inherently ambiguous. Spoken AuE consists of 44 separate phonemes (24 consonants and 20 vowels/diphthongs) and their various combinations [Bernard 81, Mitchell 46, Woodward 60]. On the other hand, there are only 11 distinguishable visemes[2] in AuE, which means each viseme has to accommodate more than one phoneme [Plant 77, Plant 80]. The phonemes and visemes of AuE are shown in Tables 4.1, 4.2, and 4.3 in Chapter 4.

Moreover, a study by Kricos [Kricos 96] shows that viseme categories vary across speakers both in the number of visemes and in their constitution. This fact is related to the ease with which talkers can be speechread, that is how clear their visible articulation is. The visual expressiveness and distinctiveness of a speaker has a significant effect on the speechreader's perception. According to the same study, most people are likely to be presenting significantly fewer than 11 or 12 visemes. Factors affecting the number and constitution of perceived visemes are coarticulation effects of accompanying sounds, environmental conditions such as lighting and the angle at which the perceiver watches the speaker's face, and articulatory differences among the speakers. These factors must also be addressed in AV ASR systems.

Auditory Intelligibility $\neq$ Visual Intelligibility

Kricos [Kricos 96] found also that normal auditory intelligibility of a speaker does not ensure high visual intelligibility because the visible movements of the articulators, needed for successful speechreading, are not needed to produce auditorily intelligible speech. In addition, a given speech sound can be produced in different ways by various talkers and still sound the same. For example, the inter-dental phonemes /θ/ and /ð/ can be satisfactorily produced by both protruding the tongue through the teeth, or placing the tongue behind the teeth of the upper jaw. While the auditory consequences are very much the same, the visual consequences are clearly different. The tongue will be visible in the first case but not in the latter.

[2] 12 visemes if /aʊ/ is taken as a separate viseme.

2.1.3 The Speech Chain

Speech production and speech perception in humans are related to each other. In speech production, there are linguistic, physiological, and physical (acoustical) stages, the order of which is reversed in speech perception. This has been called the *speech chain* [Denes 93]. The very purpose of uttering speech is that it is received and understood by a listener. An idea is transmitted from the mind of the speaker to the mind of the listener. Speech production and speech perception both involve processing in the brain to formulate or understand a message, to transfer to or from a language code, and to control certain muscle groups to produce sounds or receive signals from the hearing organs in the ear. An understanding of these processes can help create good solutions for the difficult task of processing speech by machines.

After a general overview, details of speech production and speech perception, that are considered useful background information and relevant to a study of AV relationships, are given in Sections 2.1.4 and 2.1.5. For a detailed explanation of all processes, see Denes and Pinson [Denes 93]. Also, most textbooks on speech processing or phonetics give an overview of the processes involved (for example, Ainsworth [Ainsworth 76], Clark and Yallop [Clark 95], Fry [Fry 76, Fry 79], Furui [Furui 89, Furui 00], Ince [Ince 92], Rabiner and Juang [Rabiner 93]).

General Overview

Although not all details of every step have been discovered yet, the generally accepted view of the processes is as follows. Human speech communication starts with the formulation of a message that the speaker wants to convey to a listener. This message is converted via a linguistic structure into language code. The conversion involves choosing words from a dictionary, ordering them in the appropriate order according to grammatical rules, adding prosodic information (specifying pitch, loudness, and duration of sequential segments), and stringing together the phonemes.

In the next step, the message is encoded in the physical properties of the speech sounds. The speaker produces neuromuscular (motor nerve) commands which are executed by the muscles of the vocal organs. These are broadly the muscles in the

chest and abdomen used for breathing (which as a by-product produce the energy to generate sound waves), the muscles of the larynx used in phonation, and the muscles of the vocal tract which take part in the articulation of speech sounds. The resulting movements of the vocal organs generate and shape continuous sound waves (pressure waves) which travel from the speaker's vocal operators in all directions.

The propagation of the sound waves with respect to velocity and energy loss (damping) depends on the physical properties of the transmission medium (usually air). Some of the sound waves travel to the listener's ear as well as to the speaker's ear allowing continuous control of the vocal organs by this feedback. When the sound waves reach the ear, the acoustic energy contained in the waves causes the eardrum to vibrate, with the ear canal working as an acoustic resonator. The ossicles of the middle ear connect the eardrum with the inner ear. The middle ear also amplifies the acoustic energy delivered to the inner ear and protects it from loud sounds. In the hair cells of the organ of Corti, contained within the cochlea, the pressure waves are transduced into electrical signals, which are sent to the speech centres of the brain via the nerves connecting the ear and the brain. In a way that is as yet not well understood, the neural activity is converted into a language code in the speech centre and the sequence of language units, originally formulated in the speaker's mind, is reconstructed to achieve understanding of the message.

2.1.4 Speech Production

In this section, the processes involved in speech production are described. The generation and shaping of sound waves in the vocal organs are detailed because the physical properties of these waves are measured and analysed later (see Chapter 5).

When the abdominal muscles force the diaphragm up, air is pushed up and out of the lungs. The airflow passes through the trachea and the glottis — the opening between the vocal cords — in the larynx and from there through the pharynx past the velum to the oral and nasal cavities. The upper part beginning with the larynx is called the *vocal tract* which ends at the lips and nostrils. The vocal tract is an air-filled tube of about 15–17cm length (in adults). No two vocal tracts are of

exactly the same size or shape. Human speech communication relies on the fact that we learn to disregard the effects due to different vocal tracts and instead pay attention to the effects of articulatory changes to the shape of the vocal tract. The shape can be varied by moving the so-called *articulators* which are the tongue, the lips, the jaw, and the velum.

The vocal cords act as an adjustable barrier to the airflow. They open and close rapidly when under tension during speech production, turning the air stream into a series of volume-velocity pulses. The intermittent airflow is called the *glottal source* (or *source of speech* or *voice source*). In normal vocal cord vibration, the vocal cords are first drawn together, so that the subglottal pressure builds up. When the pressure becomes too large, the vocal cords are blown apart by the sudden release of air. According to the myoelastic-aerodynamic theory, the Bernoulli effect assists the closure of the vocal cords and the cycle is repeated [van den Berg 58]. The process of vocal cord vibration is called *phonation*. The pressure wave generated is quasi-periodic and the shape of the waves is approximately asymmetric-triangular. The frequency of the vocal cord vibration is commonly referred to as the *fundamental frequency* of the glottal source or the *voice source excitation frequency* F_0.[3] The range of F_0 frequencies used in normal speech extends from about 60Hz to 350Hz [Denes 93]. It should be noted that the sound wave generated by the glottal source is a complex wave consisting of fundamental and harmonic components [Fry 79].

Acoustics of the Vocal Tract

The term *acoustics* refers to the scientific study of sound. Sound results from vibration of one kind or another. In order to generate audible sound, a propagating medium is required through which the sound can travel. In addition, the frequency of the vibrations must be within the sensitivity range of the ear, which is usually considered to be from 20Hz to 20kHz, although this varies among individuals. Fur-

[3] Fundamental frequency and pitch are often used synonymously although it is important not to confuse the two. Fundamental frequency is a physical property of the sound source, while pitch is the sensation that this frequency gives rise to [Fry 79].

thermore, the vibration must have an amplitude large enough to be audible.[4] From amplitude, we can derive a property called *intensity* which corresponds to power per unit area. Power itself is a measure of the rate of energy being used and hence intensity is also a measure of energy. Intensity is proportional to the square of pressure, so from samples of the speech waveform the measure known as *Root Mean Square (or RMS) Value*[5] can be derived.

There are four main processes in sound generation and the acoustic consequences of these processes can be considered independently [Fant 60]. The first process involves the creation of sound waves at the glottis and / or in some turbulent airstream in the vocal tract. Secondly, the shape of the vocal tract modifies these waves and functions as a frequency-selective filter (more on articulation in the next subsections). These two processes together are often considered as the *source-filter model of speech* [Fant 60]. As a third process, energy losses due to the damping effect of the vocal tract walls affect the acoustics. The final process involves the radiation of the sound waves from the lips and / or the nostrils. The first two processes are described in more detail in the following paragraphs. For more details on all four processes, see Harrington and Cassidy [Harrington 99], for example.

In *voiced* sound production, the air in the vocal tract — acting as a resonator — is set into vibration by the harmonic rich glottal source with the fundamental frequency F_0. The fundamental frequency is the lowest in a harmonic series, a series in which the harmonic frequencies (or simply *harmonics*) are multiples of the fundamental frequency. The vocal tract responds more strongly to those harmonics which coincide with its *natural* or *resonant frequencies*. The resonances of the vocal tract are called *formants* and their frequencies *formant frequencies*, which are commonly numbered F_1, F_2, F_3, and so on, starting from the lowest. The lowest three or four are considered to be the most useful in describing speech sounds. These frequencies determine the frequency spectrum of the sound waves radiating from the lips and nostrils. The formant frequencies of women are higher than those of men and chil-

[4] Amplitude refers to what is commonly called loudness, although these are not the same because the former is a physical property and the latter a sensation.

[5] The terms RMS intensity and RMS energy are often used synonymously.

dren's formant frequencies are higher again. Changing the shape of the vocal tract results in a change of the formants and thus creates different sounds. Resonances only depend on the shape of the vocal tract, not on the fundamental frequency of the sound source. Source and filter are independent in first approximation.

In *unvoiced* sound production, air turbulences in the vocal tract are the sound source. This is achieved by constricting the airflow in the vocal tract with one or more articulator(s), thereby changing the shape of the vocal tract and possibly creating a front and a back cavity. Doing so creates resonances and anti-resonances, the latter having a cancelling effect on frequencies. There is some evidence in the literature (e.g. [Fant 60, Flanagan 72]) that the length of the front cavity accounts for many frequency differences between unvoiced sounds with different places of articulation. Shorter front cavities result in higher resonance frequencies. However, in general, the relationship between vocal tract shape and sound production is less understood for excitation forward of the glottis than it is for excitation at the glottis. Formants are still present but are often considerably less distinct.

Vocal Tract Models

Vocal tract (or articulatory or speech production) models relate articulatory properties to acoustic (spectral) properties. The vocal tract is assumed to be a tube with adjustable shape (cross-sectional area, length) [Maeda 79, Maeda 82, Maeda 88]. The most simplistic model is a circular tube with constant cross-sectional area and no energy losses.[6] Despite such large simplifications — the vocal tract is at best ellipsoidal near the glottis and of more complex shape elsewhere — this single-tube model is sufficient for the approximation of the central vowel /ə/, which corresponds to a virtually unrestricted vocal tract. The vocal tract can be approximated more accurately by a series of short tube sections of fixed length (≈5–10mm) and variable cross-sectional area which can be adjusted to reflect the vocal tract shape for a certain sound. The effective frequency response of the vocal tract can then be computed and the corresponding formant data be obtained [Dunn 50, Fant 60].

[6] This is equivalent to a single Helmholtz resonator.

Such a model is accurate but computationally expensive. Stevens and House [Stevens 55, Stevens 56] as well as Fant [Fant 60] developed much simpler and less computationally expensive, approximate models with circular tubes and only four sections. These sections form a back cavity, a tongue constriction, a front cavity, and the lip region.[7] Apart from the neutral /ə/ and /h/, all sounds are produced by creating a constriction of some kind and some degree with the help of the articulators. This view is supported by X-ray studies of the vocal tract in speech production ([Fant 60, Wood 79]). Fant showed in his study that such a model can calculate the formant pattern of voiced sounds and unvoiced sounds, although the latter require more complex calculations, because they involve turbulent airstreams as sound sources which are more difficult to model. In the special case of nasal consonants, the nasal cavity is coupled with the totally obstructed oral cavity as a branch resonator. Nasal formants can be found and they are relatively stable, because the nasal cavity system cannot be varied in a systematic way like the oral cavity by means of some articulators (e.g. [Fant 60, Harrington 99]).

Attempts have also been made to construct *acoustic-to-articulator models* but the difficulty is that more than one articulator configuration can produce the required speech sounds. For example, intelligible speech can be produced when holding a bite block between the teeth [Gay 81, Hoole 87, Jones 03]. More details on acoustic-to-articulator models can be found in the literature, for example, see Harshman and Ladefoged [Harshman 77, Ladefoged 78, Ladefoged 79].

Stevens' Quantal Theory

Stevens [Stevens 72, Stevens 89] suggests in his *quantal theory of speech production* that there are regions in the vocal tract for which even relatively large changes in the degree of constriction by articulators result only in relatively small changes to the acoustic output. On the other hand, a small shift beyond the boundaries of such a region produces a large change in the generated speech signal. Thus, the relationship between vocal tract shape and frequency spectrum is not linear

[7] Such a system with front and back cavities is equivalent to a double Helmholtz resonator.

but rather 'step-wise' or *quantal*. These comparatively large changes contribute to phonological distinctiveness. Stevens suggests that articulation is organised in such a way, that optimal use of the vocal tract to produce distinctive sounds is made. Examples are the vowel triangle of /i/, /a/, and /u/ [Boë 94], or the sudden change from laminar to turbulent airflow when a constriction reaches a critical point and the sound changes suddenly from vocalic to fricative [Harrington 99].

Simulations with vocal tract tube models support the quantal theory. It can be shown that if the location of the constriction is changed slightly within a quantal region, the formants change little [Stevens 72]. So even when speakers are imprecise in the actual positioning of the tongue constriction, the acoustic properties of the produced sound can be recognisable by the listener. The quantal theory is also supported by evidence from X-ray studies such as Wood's [Wood 79].

Acoustical Consequences of Articulatory Movements

Lindblom and Sundberg [Lindblom 71] studied the acoustic effects of articulatory movements using a vocal tract model similar to Fant's [Fant 60] but with different articulatory parameters. These were the larynx height, position and shape of the tongue body, and the area of the mouth opening depending on jaw (mandible) position and lip configuration which are interdependent to some degree. For example, lowering the jaw leads to increased mouth opening even when the lips are in a neutral position. The tongue body position describes the location of the vocal tract constriction in the range from palatal to pharyngeal, while tongue body shape relates to the degree of constriction. The larynx height determines the length of the pharyngeal cavity. As mentioned before in "Acoustics of the Vocal Tract" in Section 2.1.4, the creation of front and back cavities of variable length changes the resonance frequencies of the vocal tract and thereby enables the production of different sounds. The formant frequencies from any combination of these parameters can then be determined through area function analogs as in previous work [Dunn 50, Fant 60, Lindblom 71].

The findings of Lindblom and Sundberg with respect to the relationship between

articulatory and acoustic features can be summarised as follows [Clark 95]:

- Jaw movement affects the acoustic properties of vowels. With all other parameters being constant, jaw movement alone causes F_1 to rise markedly. F_2 rises when the tongue constriction is in a velar location and this effect is stronger for spread lips. F_3 rises slowly for small to moderate jaw openings but may show a sharp increase when the tongue constriction is in palatal position.

- Movement of the tongue body position from front to back causes a moderate increase in F_1 but a large decrease in F_2, most of which occurs between the frontal and central positions. For small jaw openings, F_2 rises again slightly for central to back positions, while it continues to fall for large jaw openings. F_3 decreases sharply for frontal tongue position and spread lips, but then rises again slowly as the tongue moves backwards for all lip configurations.

- The shape of the tongue body — determining the degree of constriction — has little effect on F_1, except for a slight decrease when the tongue is in a frontal position and constriction is at a maximum. The effect of tongue body shape is primarily in F_2. It decreases significantly with increasing constriction when the tongue is in neutral or back position. It rises with increasing constriction when the tongue is in frontal position. F_3 shows little effect.

- Lip rounding has the effect of lowering all formant frequencies due to the increase in vocal tract length and the decrease in the size of the mouth opening. The effect is modest in F_1 but quite significant for F_2 and F_3. Tongue position as well as the degree of jaw opening determine how strong the effect is.

- A lowering of the larynx increases the vocal tract length and results in a lowering of all the formant frequencies, with the largest changes in F_2 and F_4.

The main conclusion of Lindblom and Sundberg [Lindblom 71] is that the position and degree of tongue constriction are more suitable to characterise vowels in the articulator-to-acoustic transformation than traditional ways by tongue height.

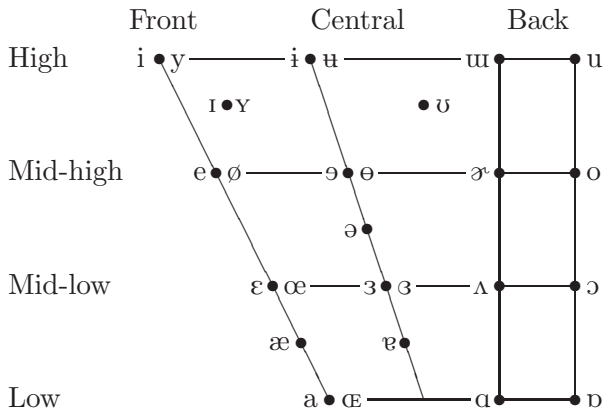


Figure 2.1: Vowel quadrilateral: Vowel classification according to tongue position [IPA 99]. In pairs of symbols, the right symbol represents a rounded vowel.

The Articulation of Vowels and Consonants in English

English vowels are generally voiced sounds in normal speech (not when whispering). A relatively stable vocal tract shape is maintained during their production and quasi-periodic pulses of air are generated at the glottis. One common way of classifying vowels is in terms of their articulatory configuration in a so-called *vowel quadrilateral* (Figure 2.1) where the corners represent the extreme cases of tongue position and a vowel's place is determined approximately according to its tongue position [Jones 17]. Tongue position in this context means the location of the highest part of the tongue body. On the vertical axis, the vertical tongue position or degree of constriction of the vocal tract is shown. A low tongue position corresponds to a low degree of constriction, a high tongue position to a high degree of constriction. On the horizontal axis, the tongue position with respect to fronting or backing is depicted. The literature (e.g. [Joos 48, Clark 95, Harrington 99]) illustrates that such a schematic vowel classification corresponds well to acoustic properties of the formant frequencies F_1 and F_2 (see previous subsection).

The vowel quadrilateral only reflects the tongue position but not the lip configuration. Basically, any lip configuration could be used with any tongue position and differences exist between languages. In English, front vowels are usually produced with spread lips and back vowels with rounded lips. The lower the tongue position, the more the lips tend to become open and the back vowels less rounded.

Vowels can also be distinguished by duration. While short and long vowels are produced by similar articulator positions, long vowels are distinguished from short vowels by the amount of time they are pronounced.

The production of consonants is more complex. First, in addition to the vocal cord vibration, two other methods are used to make the airstream from the lungs audible. These are the mechanisms to produce *fricatives* and *oral stops (plosives)*. Fricatives, e.g. /f θ s h/, are noise-like sounds which are generated by constricting the airflow in the vocal tract by tongue or lips (in combination with teeth) which creates turbulence. Each fricative corresponds to a fairly precisely located constriction. Oral stops, e.g. /p t k/, are produced by stopping the airflow altogether by blocking the vocal tract with the tongue or lips. The production of fricatives and oral stops is independent of vocal cord vibration. When the vocal cords vibrate at the same time, voiced oral stops and fricatives are produced. Otherwise, they are voiceless.

Other consonant groups are *affricates*, *nasals*, and *liquids and glides* (or *semivowels* or *approximants)*. Affricates (/tʃ dʒ/) consist of a brief oral stop shortly followed by a fricative which merge into a new sound. In the production of nasals (/m n ŋ/), the velum is lowered and as a result, the nasal cavity is coupled with the pharynx and oral cavity to become part of the vocal tract. The airflow through the mouth is blocked and diverted to the nasal cavity.[8] Liquids and glides (/l r w j/) are produced in a similar way as vowels. They are voiced consonants. An articulator moves quickly towards another, thereby constricting the airflow to some extent but not blocking it completely, nor causing turbulence sufficient for a fricative.

Consonants are commonly classified by the place and manner of articulation (for example, see the consonant chart of the International Phonetic Alphabet [IPA 99]). Place of articulation describes the location of the constriction in the vocal tract produced by teeth, tongue, or lips. Place of articulation classes correspond well to the viseme classes (see Table 4.3 in Section 4.2.4). Manner of articulation describes the differences in articulatory methods for the production of oral stops, fricatives, affricates and so on. These methods correspond well to the phoneme classes (Table

[8] Nasalised vowels are produced in the same way but are not phonemically distinctive in the English language.

4.1 in Section 4.2.4).

In continuous speech, vowels and consonants, voiced and unvoiced sounds alternate (not necessarily 1–1). Therefore, formant transitions play an important role, particularly in carrying consonant-related information. The formant transitions reflect the rapid, yet continuous shift of articulatory position between consonants and vowels, i.e. between a constricted and an unobstructed resonant system. For more details on the aspects involved, see Heinz and Stevens [Heinz 61] for an acoustic perspective, Liberman *et al.* [Liberman 76] for a perception perspective, and Clark and Yallop [Clark 95] for a phonetic perspective.

Coarticulation

Speech sounds are normally not produced in isolation, rather they are influenced by the context in which they occur. Although we can think of spoken language as a string of separate words made up of letters on a linguistic level, this is not the case on an acoustic level. The acoustic pattern of a particular phoneme can vary considerably depending on the neighbouring phonemes. Such an overlap of phonetic features from phoneme to phoneme is called *coarticulation* and it is important to also take its effects into account in automatic speech processing. The articulators constantly position themselves for the phonemes in the sequence and in that process they move ahead of time towards a position appropriate for the next phoneme before the position for the current sound has been fully reached. As soon as the target position of the current phoneme has been approached closely enough to be intelligible for the listener, the articulators move towards the next target position.

Coarticulation is a necessary, natural part of speech production (Harrington and Cassidy [Harrington 99]). It allows for a higher rate of speech because sounds are transmitted partially in parallel. Coarticulation helps minimise the effort required to produce speech because articulators do not have to move the full distance.

The nature and degree of coarticulation are affected by many factors. Different allophones of the same phoneme can have different acoustic properties. For example, the articulation of the phoneme /k/ depends strongly on the following vowel

[Harrington 99]. If followed by a front vowel (e.g. *keep*), it is produced in a post-palatal position. If followed by a central vowel (e.g. *curd*), it is realised as a velar stop. Before back vowels (e.g. *caught*), articulation is in a postvelar or preuvular position. Furthermore, coarticulation patterns vary across languages, dialects within languages, and speakers. The prosody of an utterance also influences coarticulation. At faster rates of speech, the degree of coarticulation between sounds increases. The phonetic overlap is also greater for stressed syllables than for unstressed syllables.

2.1.5 Theories of Speech Perception

In the general overview given in Section 2.1.3, it was discussed briefly how acoustic signals are detected by the ears, transduced, and then sent to the brain for processing. While some of the processes involved are not yet known, some theories of speech perception exist and a brief overview is given at this point. The interested reader is referred to the cited literature for more detailed information as well as [Klatt 89] for a review.

The recognition of linguistic units (syllables, words, phrases) depends on a number of factors which include the acoustic structure of the speech sounds, the context, the familiarity with the speaker, and the expectations as a listener. A lot of the understanding of continuous speech involves 'top-down' linguistic processing, which draws on the personal knowledge base of the listener. A segmental processing of the acoustics of the speech signal is not necessarily needed to determine the phonological structure and achieve understanding of a message's meaning [Clark 95].

One common, although not unchallenged [Repp 84], theory is that of *Categorical Perception of Speech* [Liberman 57], formalised by Massaro in the *Categorical Model of Perception* [Massaro 87]. It suggests that phonemes are used as perceptual categories. Listeners do not change their opinion gradually as the signal changes but make a sudden, categorical change in the perception of a speech segment. It can also be shown that sounds from the same (phoneme) category, e.g. oral stops, are harder to distinguish from each other than sounds from a different category. The categorical perception fits well as a counterpart to Stevens' quantal theory (see Sec-

tion 2.1.4) and is often considered to be further evidence of the interwoven nature of speech production and perception. However, the categorical model of perception does not handle the fusion of auditory and visual speech information well, which gives rise to the *Fuzzy Logical Model of Perception* (see Section 2.1.7 below).

A widely acknowledged, but also controversial, theory is the *Motor Theory* [Liberman 67, Liberman 85]. It suggests that humans decode the perceived acoustic signal in terms of articulatory patterns and compare these patterns with those stored for articulation of their own messages. The advantage of this theory is that the listener's mental neuromuscular planning compensates for coarticulation effects. How exactly the model works in detail and how the storage and accessing of articulator patterns function remains unclear.

The *Analysis by Synthesis* theory by Stevens and Halle is, in some aspects, similar [Stevens 67]. Listeners perform a spectral analysis of the acoustic signal, decoding it into features and parameters. This information is then further analysed to establish an estimate of the phonological structure of the speech signal. A phonological rule system compares this estimate with an appropriate neural representation of the analysed input and only if the match is good, it is accepted. Otherwise the process is iterated until a good match is found.

Another theory is Klatt's *Lexical Access From Spectra (LAFS)* [Klatt 79, Klatt 81]. According to this theory, spectral templates of all familiar words are stored in the listener's memory. No segmental representation or analysis is required and thus it avoids problems with context-dependent coarticulation. The spectrum of the incoming signal is evaluated against a number of competing spectral templates and the one closest to the input is chosen. This theory assumes the availability of very powerful processes for storage, analysis, access, and decision.

Finally, the *Trace* theory by McClelland and Elman [McClelland 86, Elman 97] was inspired by work on connectionist models[9] of cognition. In this theory, spectral slices of the acoustic signal are generated every 5ms and form the input to a connectionist model. The nodes of the first level of the model act as feature detec-

[9] Also known as neural networks, parallel distributed computing, or neuro-computing.

tors. They are connected to segmental detection nodes which identify a particular speech segment. The outputs of the segmental nodes are connected to a set of word connection nodes. Connections between nodes generally work by activation and inhibition and the structure of the network is typically learned in a training phase. For a good review on connectionist models see [Medler 76] and [Christiansen 99].

2.1.6 Sources of Visual Speech Information

One of the central issues in AVSP is the question of which part or feature of the face humans rely on as source of visible speech information. Silent speechreading is different from the processing of AV speech stimuli [Campbell 96, Smeele 96]. They are performed by different parts of the brain and involve different processes.

While visual evidence of speech articulation can be found everywhere on the face, the lower half of the face carries the vast majority of information [Smeele 96]. The major factors are the visible articulators consisting of the mouth region including lips, teeth and tongue, as well as the lower jaw. From these, the following parameters have typically been derived and analysed:

- the width and height of mouth opening,
- the area of mouth opening,
- the protrusion (or rounding) of upper and lower lip and lip contact point, and
- the vertical distance of the lower jaw from the upper lip or nose.

These parameters are not independent of each other. Cosi and Caldognetto [Cosi 96] found a negative correlation between the width of the mouth opening and the protrusion parameters as well as a positive correlation between the height of the mouth opening and the jaw distance. Rounded and protruded lips lead to a small width of the mouth opening, while the opposite is true for spread lips. Also, moving the jaw up and down results, from a certain point on, in an opening and closing of the mouth because the upper lip cannot move up and down much and the lower lip is directly linked to the jaw by skin (see also Bailly *et al.* [Bailly 98]).

Benoît *et al.* [Benoît 96] tested the intelligibility of AV speech with real and synthetic faces in comparison to audio-only speech. The display of moving lips alone restored about one third of the missing information irrespective of the level of noise on the acoustic signal. Displaying movements of the jaw resulted in even higher intelligibility. Nevertheless, the best results were obtained with a full face display. The question remains whether this is due to additional information carried on the rest of the face or simply because of the unnatural look of lips and jaw moving without a face. The results were similar for both real and synthetic faces.

Smeele [Smeele 96] also investigated which parts of the face influence speech processing most by determining the number of times that McGurk effects were encountered. The visual information from the lips and the oral cavity together were sufficient to influence auditory speech processing. Additional facial parts added only little to that effect. Jaw movements alone were not sufficient. This result is in accordance with a study by Plant and Macrae [Plant 77]. They found that the movement of the lower lip was the most important visible factor in vowel and diphthong articulation while upper lip and jaw movements did not provide much information to differentiate vowels. Yakel *et al.* [Yakel 95] as well as Green [Green 94, Green 96] showed that the visibility of facial features has a strong influence on auditory speech perception regardless of the face orientation (e.g. upside down or colour inverted).

Static vs. Dynamic Parameters

Beside static parameters, studies by Cosi and Caldognetto [Cosi 96], Campbell [Campbell 96], Cathiard *et al.* [Cathiard 96], Goldschen *et al.* [Goldschen 96], Green [Green 96], and Vatikiotis-Bateson *et al.* [Vatikiotis-Bateson 96] also investigated the dynamic patterns — the first (velocity) and second (acceleration) derivatives — of these parameters. In fact, some consider face kinematics as more useful than shape parameters. This view is supported by Rosenblum and Saldaña [Rosenblum 96] who used a moving light model. A speaker with 28 point lights fixed on his face (lips, teeth, tongue tip, chin, cheeks, jaw, nose tip) was filmed in the dark. The face was not recognised from static video frames but McGurk effects

were reported for dynamic presentations (video sequences). However, the issue of whether shape or motion is more important in AVSP has not yet been resolved and contradicting evidence can be found in the literature. Beside the possibility that shape is recovered from motion, it is also possible that a combination of both is actually used. Based on a study of various visual speech parameters for ASR systems, Scanlon [Scanlon 01] suggested that a certain amount of static information is required as a base before dynamic parameters improve the recognition process. Human speech perception might be similar.

Although some studies have attempted to measure protrusion parameters, almost all have only looked at the effects of visual speech in a frontal view of the face. An exception is the work by Cathiard *et al.* [Cathiard 96] who performed experiments for frontal and profile views with both static images and dynamic image sequences. For their experiments limited to the French vowels /i/ and /y/, which cause rounded lips, the profile view gave comparable results to the frontal view. On the other hand, the information to be gained from any particular view angle appears to depend on the viseme, too, as some are more readily speechread from a frontal view and others from a profile view. A 45° angle might be the solution to make best use of the total visual speech information.

Visual Cues Enhance Speech Sound Detection

Visual speech cues do not only enhance the intelligibility of spoken language, but also improve the detection of speech sounds, in particular in noisy conditions. Grant and Seitz [Grant 00, Grant 01a] recently showed that as the visible speech articulators move to be in position for the articulation of the next speech sound, these cues help the perceiver to detect speech sounds and to filter them in a noisy acoustic environment. These results have been confirmed by Bernstein [Bernstein 03], Kim [Kim 01, Kim 03], and Schwartz [Schwartz 02, Schwartz 03]. Visual speech cues typically precede auditory cues by 50ms and more (see [Kohlrausch 00] for a review) and perceivers tolerate AV asynchronies where the video leads the audio more than where the audio leads the video, which suggests that they are more used

to visual evidence preceding auditory evidence [Conrey 03, Grant 01b, Grant 03]. Girin *et al.* [Girin 01, Sodoyer 03] showed that visual speech cues can also help to separate the sources in the case of multiple audio speech signals.

2.1.7 The Integration of the Two Modalities

A final important issue in human AVSP is when and how the information from the two modalities is integrated and processed. As crucial as the answer to this issue is for the understanding of how humans process AV speech information, as controversially it is debated in the field. Intersensory fusion and intermodal transfer is a well-studied area of psychology. For more information, see for example [Stein 93]. In speech perception, four categories of fusion models are commonly considered (based on [Summerfield 87], adapted by [Robert-Ribes 96]):

- Direct Identification (DI) based on the *Lexical Access From Spectra* theory [Klatt 79, Klatt 81],
- Separate Identification (SI) as in the *Vision Place Auditory Mode* model [McGurk 76] or the *Fuzzy Logical Model of Perception* [Massaro 87],
- Recoding in the Dominant (auditory) modality (DR),
- Recoding in the Motor space (MR) based on the *Motor Theory* [Liberman 67, Liberman 85].

Robert-Ribes *et al.* [Robert-Ribes 96] present a taxonomy of these sensor-fusion models (Figure 2.2). The taxonomy is based on three basic questions:

1. Is there a common, intermediate representation of the audio and video stimuli?
2. When does the integration happen: early or late?
3. What is the nature of the common representation?

The two extremes in terms of sensor-fusion models are the DI model and the SI model. In the DI model, both input signals go directly into a bimodal classifier. In

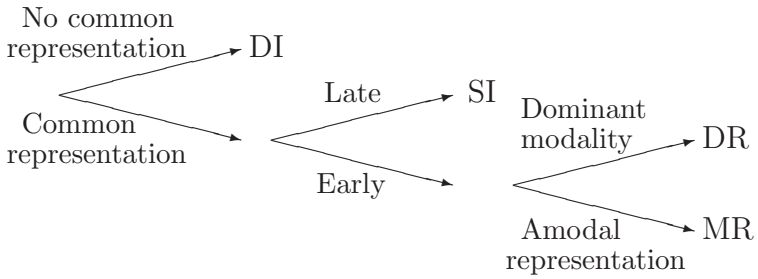


Figure 2.2: Taxonomy of four basic AV integration models by Robert-Ribes *et al.*

the SI model, two separate recognition processes for the auditory and visual information run in parallel. The results of each recognition process are then integrated to yield the final outcome. Robert-Ribes *et al.* [Robert-Ribes 94, Robert-Ribes 95a, Robert-Ribes 95b] suggest two additional models. The DR model considers the auditory modality as dominant and the visual information is recoded into a representation of the auditory modality. This model is considered as unlikely by Robert-Ribes *et al.* In the fourth model, the MR model, both inputs are transformed into an amodal common space and then fused in that space before classification.

Robert-Ribes *et al.* draws the following conclusions. Based on studies that used conflicting auditory and visual stimuli, the direct identification model was rejected. Subjects were able to detect AV incompatibilities, to estimate a perceptual distance between the two stimuli, and still fuse both inputs. Hence, the stimuli can be compared in a common space before fusion.

The issue of early versus late integration is one of the most controversial ones in the field of AVSP. The literature presents conflicting evidence for both claims, although a majority seems to favour an early integration model (e.g. Green [Green 96], Robert-Ribes *et al.* [Robert-Ribes 96]). One argument in favour of an early integration model is the fact that humans are able to extract temporal coordinations between the audio and video signals. This would not be the case in a late integration model where the separate classifiers lose the temporal coordination information.

The third question distinguishes the DR model and the MR model. The DR model was rejected by Robert-Ribes *et al.* because studies on a particular French vowel had shown that it was perceived as rounded when presented auditorily but

judged as unrounded when presented visually or auditory-visually. It is hard to account for this fact in a model where the visual input is recoded into an auditory representation. Robert-Ribes *et al.* conclude that the MR model is the only model compatible with all experimental data, but that a more complex hybrid model as a combination of DI and SI models could also explain the experimental data.

The Modified Fuzzy Logical Model of Perception

Robert-Ribes' conclusions are supported by a bimodal speech perception model developed by Massaro [Oden 78, Massaro 83, Massaro 87, Massaro 96]. Its theoretical framework is based on an information processing approach which assumes that there is a sequence of processing stages in spoken language understanding. The *Modified Fuzzy Logical Model of Perception (FLMP)* assumes that, firstly, both audio and video sources of information support speech perception and, secondly, continuously-valued features — hence a fuzzy logical model — are evaluated, integrated, and matched against prototype representations in memory (Figure 2.3).

The central point of the model is the independent evaluation of auditory (A_i) and visual features (V_j) before integration. For example, the degree of visible mouth opening at the beginning of a syllable can be evaluated independently of whether auditory information is available. In addition, bimodal information (B_{ij}) about the temporal asynchrony between auditory and visual information is also evaluated. The evaluation stage transforms the sources of information into psychological values (a_i, v_j, b_{ij}) which are then integrated to give an overall degree of support for a given representation in memory. Finally, the decision stage maps this overall value into some response, R_k, such as a discrete decision or a rating. Hence, according to the FLMP, humans have information about the degree to which a given alternative is present rather than just the information about which alternatives are present.

The original FLMP is mathematically equivalent to Bayes' Theorem and hence is optimal for combining multiple sources of information. As such, the FLMP is more appropriate to describe multimodal speech perception than the theory of categorical perception (see Section 2.1.5). The CMP is equivalent to a SI model

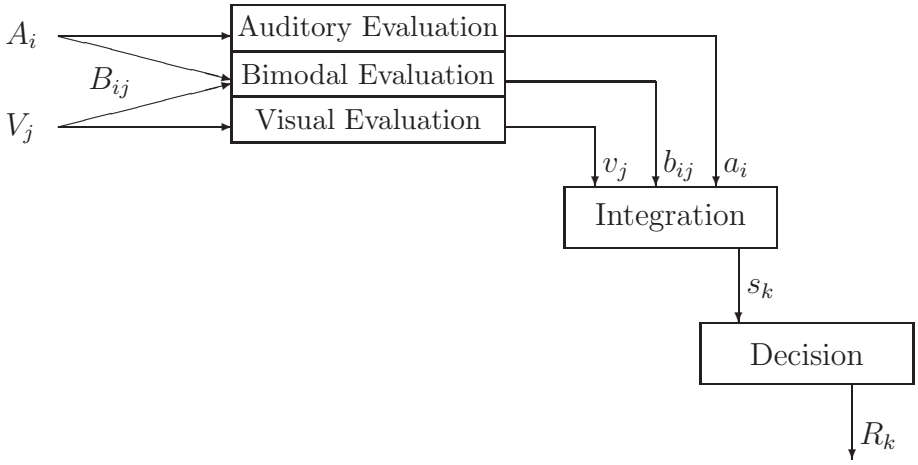


Figure 2.3: The three stages of the modified Fuzzy Logical Model of Perception.

in the taxonomy by Robert-Ribes *et al.* [Robert-Ribes 96]. The CMP is as fit to describe the results from the individual sources as the FLMP, but it fails on the description of the overall results, i.e. on the integration of the different sources of information. This supports the view by Robert-Ribes that early integration is more likely than late integration. The FLMP predicts that two sources of information can be more informative than just one. It gives a good description of the results not only in speech perception but also in reading, object recognition, sentence interpretation, recognition of affect, memory, and decision making.

Information $\neq$ Information Processing

Massaro [Massaro 92] distinguishes between information and information processing. One component of information corresponds to the outcome of the evaluation stage (Figure 2.3), i.e. how much does a particular stimulus presented to a given input channel support the various alternatives? On the contrary, one component of information processing corresponds to the process of integrating the various sources of information. There are significant differences in the information value of audible and visible speech as a function of age, but no differences in the information processing. These processes appear to exist at age 3 and remain constant for the rest of life. Pre-school children still acquire language knowledge and speech perception skills and therefore do not speechread as well as adults [Massaro 96]. On the other hand, ageing decreases the resolution of the sensory systems which results in less

accurate speech perception but the availability of, and the ability to process, multiple sources of information appears to compensate for that effect. For example, some older adults report they 'hear' the TV better with their glasses on [Massaro 96].

Massaro [Massaro 96] also rejects the frequent claim that speakers of different linguistic backgrounds are influenced differently by visible speech. Even in the often stated example of Japanese, who are not used to watching the speaker's face because it is considered to be impolite, Massaro found that perceivers are similarly affected by visible speech as perceivers from other cultural backgrounds. The information processing in the integration and decision making is identical across languages, but the information made available by the evaluation stage differs. This view is in contrast to studies by Sekiyama (e.g. [Sekiyama 93]) which showed that Japanese speakers are less influenced by visible speech information than speakers of English.

The Influence of Experience

Linguistic experience also influences human speech processing. Exposure to a specific language, in particular the native language, results in stored representations of language units and also alters the person's phonetic perception [Kuhl 92]. Of what kind these representations in the brain are, remains unclear. It has been suggested that cross-language effects, i.e. when utterances spoken by a foreign speaker in the native language of the perceiver are misunderstood, are due to the fact that listeners employ the stored information about auditory and visual characteristics of their native language during speech perception and that the productions of foreign talkers fail to match these stored representations [Kuhl 94]. Similar results have been experienced for persons perceiving AV speech in a foreign language [Fuster-Duran 96]. When facing incongruent AV speech, as in the study by McGurk and MacDonald [McGurk 76], subjects make use of the knowledge and experience of their own language to find the best fit using both auditory and visual information. The interlanguage differences, however, lead to misperceptions [Fuster-Duran 96].

2.2 Characteristics of Australian English

Australian English can be described as a regional dialect of English. It is spoken by people who are born in Australia, or who arrive in Australia at a linguistically impressionable age, and who grow up in an AuE speaking peer group [Bernard 81]. Despite there being only one AuE dialect, different pronunciations exist. These so-called 'speech varieties' are usually categorised as: *broad*, *general*, and *cultivated* [Mitchell 65].[10] AuE is characterised by specific vowel (and diphthong) pronunciations, intonation patterns, lexical items, and various paralinguistic features [Clark 89, Cochrane 89, Harrington 97, Mitchell 46, Mitchell 65]. All the varieties share characteristically Australian intonation, rhythm and stress patterns.

Regional variation in AuE is minor and is usually limited to a small number of words or phonemes. Unlike dialects in places like Britain or the USA, varieties of AuE do not differ in the number or disposition of phonemes to a significant extent, but only in the pronunciations which span a continuous range. It is therefore fair to speak of only one AuE dialect. Or in the words of Bernard [Bernard 81]: *"The picture is of a widespread homogeneity stretching from Cairns to Hobart, from Sydney to Perth, a uniformity of pronunciation extending over a wider expanse than anywhere else in the world."* This is often attributed to the fact that Australia is a migrant country with immigration originally from all parts of the British Isles and, since the end of World War II, from other parts of Europe and later Asia. Hence, AuE can be considered as a 'mixing-bowl' for English from various backgrounds with a tendency towards British English as the former colonial homeland [Bernard 81]

On the other hand, an AuE speaker's accent is much more influenced by socio-economic factors as well as by age and gender (see [Harrington 97] for references to various studies on these factors). Broad AuE has long been associated with the working class, less educated part of the population, whereas cultivated AuE has been associated with the educated and verbally more skilled population. Harrington

[10] Bernard [Bernard 81] adds *modified* AuE as another category, which is more 'cultivated' than *cultivated* AuE. It is spoken by a numerically insignificant part of the Australian population and therefore not further considered in this study.

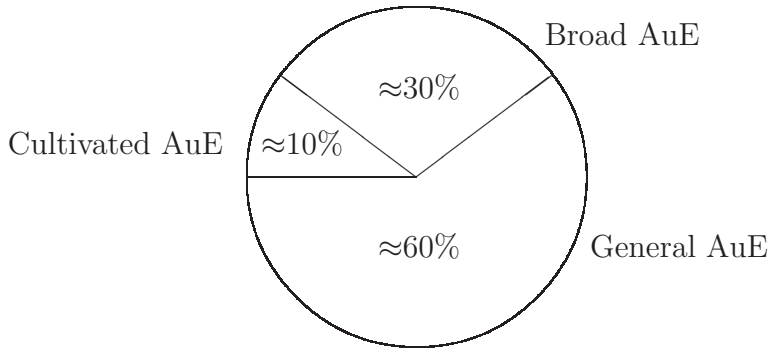


Figure 2.4: Distribution of AuE varieties in percent.

et al. [Harrington 97] point to a larger proportion of the older population as well as female population speaking cultivated AuE than the young and male populations.

A Continuum of Accent Variation

It should be noted that the AuE varieties are not discrete entities but rather span a continuum in the order *broad-general-cultivated* with considerable phonetic overlap. Speakers from one variety may well use, for example, a particular pronunciation of a diphthong from another variety, either habitually or temporarily because it seems appropriate under certain social pressure. Speakers may also change their position within the spectrum of AuE over the span of their lives.

Figure 2.4 shows the percentages of occurence for the three varieties [Bernard 81, Harrington 97]. At one end of the continuum is broad AuE which has some vowel features similar to London Cockney English [Cochrane 89]. Vowels and in particular diphthongs are given their own characteristic pronunciation, while consonant pronunciation is similar to that in other forms of English. However, speakers of broad AuE are not noted for consonantal clarity and make frequent use of assimilation and elision. At the other end of the continuum of variation, the 'prestige' form cultivated AuE most closely approximates Southern British English (or Received Pronunciation of British English). General AuE lies between these two varieties. It is spoken by the majority of the population and some evidence suggests that it is the most rapidly expanding of the three categories [Blair 93].

Differences between the Varieties of Australian English

Harrington *et al.* [Harrington 97] studied the phonetics of the three accent varieties in AuE by analysing the formant frequencies F_1, F_2, and F_3. Their findings are consistent with earlier work by Mitchell and Delbridge [Mitchell 65]. The main phonetic difference between broad, general, and cultivated AuE lies in the rising diphthongs /aɪ aʊ/. In /aɪ/, the effect is that F_1 and F_2 are significantly lower in the first vocalic target for broad AuE than for cultivated AuE, with general AuE between these two extremes. In /aʊ/, broad AuE exhibits a lower F_1 but a higher F_2 in the first target than cultivated AuE, general AuE again being between these two. To a lesser extent, there are also differences in the rising diphthongs /eɪ oʊ/. For /eɪ/, broad and general AuE exhibit higher F_1 values and lower F_2 values for the first target compared to cultivated AuE. In /oʊ/, the accent effect is in F_2 of the first target which is raised for broad AuE speakers which indicates fronting. Harrington *et al.* found considerably fewer differences between accent varieties of AuE in the second targets of the diphthongs.

As far as the falling diphthongs /ɪə ɛə ʊə/ are concerned, the same study by Harrington *et al.* showed some accent effects. For /ɪə ɛə/, broad AuE shows higher F_2 values in the first and second target and higher F_1 values in the second target only than both general and cultivated AuE. In the case of /ʊə/, there are no significant differences in the first target but broad AuE has clearly higher F_2 values in the second target. While mostly produced as diphthongs, the falling diphthongs can also be found as a long monophthong or two syllables. In fact, some variation is phonemic. For example, /ʊə/ as in words like 'tour' is produced as /ɔː/ by about a quarter of the population. This can also be found in Southern British English. In any case, the context in which these sounds are produced appears to have an impact on how they are pronounced [Bernard 81, Harrington 97].

The same study [Harrington 97] also showed that accent variations of monophthong vowels are much smaller than for diphthongs. They are mostly confined to /uː/ which is found to have higher F_2 values for broad AuE than for cultivated AuE, with general AuE being between these two extremes. A similar effect can

be found for /ɜː/ in female speakers. In both /ɪ/ and /ɛ/, F_2 is higher for broad AuE speakers, compared to both general and cultivated AuE. In addition, broad AuE has a longer onglide in /iː/ and /uː/ than general and cultivated AuE. The F_2 onglide is lowest in frequency for broad /iː/ and the F_2 and F_3 onglides are highest for broad /uː/. The degree of onglide in /iː/ also varies with age, where it is less marked for younger speakers compared to older speakers.

In summary, the acoustic differences between the varieties of AuE seem to be mostly in F_1 and F_2. Based on the discussion of the acoustic consequences of articulator movements in Section 2.1.4, it is possible to hypothesise that the differences between the varieties are mostly a result of differences in the position and shape of the tongue. If this hypothesis holds, significant differences in the visible speech articulation are not expected. A discussion of this issue for the speakers in the AVOZES data corpus used in this study, can be found in Section 6.1.11.

2.3 Audio-Video Speech Processing by Machines

A number of AVSP systems have been developed over the past two decades, for various application areas such as automatic speech recognition (ASR), speaker identification / verification, and speech synthesis. This section gives an overview of the most prominent system architectures for ASR, because the investigations in this study are from an ASR angle, but many issues apply equally to other AVSP application areas. By no means can this section account for every system that has been built. The reader is referred to Hennecke *et al.* [Hennecke 96] for an extensive comparison of different automatic AV speech recognition systems, and to a general review of AVSP by Chen [Chen 01].

Firstly, a brief overview of methods used in audio-only ASR systems is given in Section 2.3.1, followed by the fundamentals in visual-only ASR systems (Section 2.3.2). An oveview of facial feature extraction is presented in Section 2.3.3 and more details on lip feature extraction methods are given in Section 2.3.4. Next, combined AV ASR systems are described in Section 2.3.5, including the issue of integration of the two modalities. Finally, this section ends with an overview of AV

speech data corpora (or databases) in Section 2.3.6.

2.3.1 Fundamentals of Audio Automatic Speech Recognition

Audio-only ASR has been an ongoing research topic for decades and hence a plethora of publications in the literature exists, describing various approaches at all levels of detail. The discussion here focuses on a general overview and the interested reader is referred to books like [Rabiner 93, Furui 00] for more details.

Following Rabiner and Juang [Rabiner 93], approaches to audio ASR can be categorised as:

1. the rule-based acoustic-phonetic approach,
2. the data driven pattern recognition approach, and
3. the artificial intelligence approach, which is a mixture of the first two.

In addition, connectionist models, or artificial neural networks, can be seen as a fourth approach to audio ASR, or they can be regarded as an implementational technique (just as Hidden Markov Models (HMM)) used in any of the other three approaches. A review of the use of artificial neural networks in speech recognition can be found in Section 2.5.4 of [Rabiner 93]. The other approaches are discussed in the following sections.

Acoustic-Phonetic Approach

This approach is based on the theory of acoustic phonetics. According to this theory, finite, distinctive phonetic units exist and they are characterised by the frequency spectrum of the speech signal over time. It is assumed that these properties of the phonetic units can be learned and applied readily. However, the properties are highly variable among speakers as well as depending on the phonetic context (coarticulation, see Section 2.1.3), which is one of the problems in implementing a reliable ASR system based on this approach.

Acoustic-phonetic ASR systems typically consist of a four-step process. First, like in other ASR systems, a speech analysis method is required that measures certain features which describe the speech signal (or usually its frequency spectrum) appropriately over time. Common methods of spectral analysis are filter bank analysis, linear predictive coding (LPC) analysis, cepstral analysis and discrete Fourier transform (DFT) analysis (see Section 5.1.1 for an overview of the first two and, for example, [Furui 00, Harrington 99] for more details on all methods).

Secondly, in the feature-detection step, the measured features are converted to another set of features which describe the acoustic properties of the various phonetic units. Features can be continuous, such as formant frequencies and energy, or binary, as in voiced-unvoiced, nasality, and frication classifications.

In the segmentation and labelling step, the ASR system attempts to find stable or salient regions and then label these with matching phonetic units. This step is the core step of the acoustic-phonetic approach, and the most difficult one. One way of accomplishing the labelling task is by classifying each speech segment into one of several broad classes (e.g. unvoiced stop, voiced fricative, etc.) based on predetermined rules. However, this method is error-prone.

As a result of the segmentation and labelling step, a phoneme lattice is created from which a lexical access procedure determines, in the final step, the best matching word (if we assume word recognition for the moment) as the output of the recogniser.

Pattern Recognition Approach

Good reviews of this approach are given by Padmanabhan and Picheny [Padmanabhan 02] and Ney [Ney 03]. Typically, the audio signal is sampled every 10ms and feature vectors are formed using similar methods of spectral analysis as in the acoustic-phonetic approach [Juang 00]. Widely used are filter bank methods which simulate the human auditory system (see also Section 5.1.1). The sensitivity to the energy in each filter follows a logarithmic relationship, where the ratio of the centre frequencies of adjacent filters is constant and the filter bandwidth is proportional to the centre frequencies. Differences exist in modelling perceptual aspects of

the frequency scales, such as mel and bark frequency scales. Temporal information is captured through first and second derivatives of these features.

In the recognition phase, the approach seeks the word sequence with the highest likelihood, given the measured feature vectors and the trained models. Using Bayes' theorem and ignoring the denominator term, this is equivalent to maximising the product of the probability of the measured features given the word sequence and the probability of the word sequence itself. Such a system uses a lexicon of all possible words, each represented as a sequence of phonemes, a language model which models the linguistic structure, and an acoustic model which models the relationship between the feature vector and the phonemes. Language and acoustic models are learned in the training phase requiring a large amount of data to create good models. Language models are often based on word trigrams, assuming that the probability of a word only depends on the previous two words, to reduce the complexity. Acoustic models typically use HMMs whose parameters are commonly estimated by a maximum likelihood estimation process.[11]

The performance depends on many factors like vocabulary size, language model perplexity, background noise, speech spontaneity, sampling rate, and the amount of training data [Padmanabhan 02]. Speaker-dependent systems perform better than speaker-independent ones but require acoustic model adaptation, for example vocal tract normalisation [Wegmann 96]. Many commercially available ASR systems use the pattern recognition approach, which has proven to perform well. It is robust and invariant to differences in vocabularies, speakers, feature sets, pattern classification algorithms and decision algorithms, because no speech-specific knowledge is used. It is also insensitive to the recognition unit (e.g. subword units, whole words, phrases) and hence the basic techniques are applicable to a wide range of applications.

Artificial Intelligence Approach

This approach can be considered to lie somewhere between the other two approaches, because concepts of both approaches are used. The central concept is the

[11] Millar and Davis [Millar 99] suggest to build better acoustic models by representing time relative to the acoustic-phonetic structure rather than physical time.

use of knowledge from a variety of sources to improve performance. In particular, expert systems for the difficult task of segmentation and labelling in the acoustic-phonetic approach have proven to be useful as they combine acoustic knowledge with lexical, syntactic, semantic and pragmatic knowledge. The higher-level knowledge is capable of correcting incorrectly chosen speech units by the lower-level stages before a decision is finally made on the measured features. Another advantage of the artificial intelligence approach is the ability to learn and adapt over time, which corresponds well to the idea that knowledge is both static and dynamic, and that expert systems must adapt to those dynamic changes.

An interesting problem is how to integrate the knowledge. Rabiner and Juang [Rabiner 93] report three different approaches. These are the bottom-up approach, the top-down approach, and the blackboard approach. In the bottom-up approach, lower-level processes precede higher-level processes in a sequential way, which means that each stage is constrained as little as possible. In the top-down approach, the language model generates word hypotheses that are compared to the speech signal. Syntactically correct and semantically meaningful sentences are then generated on the basis of the similarity scores of the tested word hypotheses. In the blackboard approach, all knowledge sources are considered to be independent. A hypothesis-and-test paradigm communicates between the knowledge sources which compare the speech signal patterns with stored representations individually. An overall rating policy is used to combine the results from the knowledge sources.

2.3.2 Fundamentals of Visual Automatic Speech Recognition

Fundamental to visual speech recognition are the abilities to, first of all, automatically find the face in an image and track it over a sequence of video frames, and secondly, to extract useful parameters that describe the visible speech-related movements of the articulators. In terms of speech recognition, similar approaches are taken as in the case of audio-only ASR described in the previous section.

Finding the Face

A fundamental requirement is the ability to find the mouth region in an image. This equates to first determining the face position and then, within the face, the position of the mouth region, which carries the largest amount of visible speech information (cf. Section 2.1.6). A large number of face tracking systems is described in the literature. Common methods are discussed below and examples given.

Using Colour. A simple face tracking method is based on finding *artificial markers* attached to the face in the image. Revéret [Revéret 98] developed an elaborate system, in which a 3D lip model is fitted to the image data for lip tracking, speech recognition and visual speech animation purposes, but the face tracking is done by a marker on the nose. Bothe [Bothe 96] uses colour patches on the forehead and the nose to find the face.

A common method is *skin colour detection*. In normal RGB colour space, the skin colour is overlayed with highlights (reflections of light sources). Simple thresholding does not work well. Yang *et al.* [Yang 98] developed an adaptive stochastic model of the skin colour distribution in RGB colour space using colour histograms. A colour histogram characterises the distribution of colours in the colour space. Human skin colours cluster in a small region in RGB colour space. They vary more in intensity than in colour. The skin colour distribution of each individual is a multivariate normal distribution, with the parameters of the distribution accounting for differences among people and lighting conditions [Yang 96].

Yang's adaptive approach transforms the original skin colour model into the new environment of viewing conditions, which can be done in real-time because the Gaussian model has only few parameters (mean vector and covariance matrix). In addition, a motion model is used to estimate the speaker's motion and to predict the location of the search window in the next frame. Inside the found face area, a search for the pupils is started by looking for two dark regions that satisfy certain face constraints. The approximate positions of the lip corners are then predicted from the position of the eyes and the face orientation in the previous frame. A window

containing the mouth region is extracted for further processing. Using horizontal and vertical integral projection within the window, the lip corners are determined.

Senior [Senior 99] presented an approach based on the Fisher discriminant and eigenspaces. An image pyramid over a range of scales is used to search for face candidates. Each candidate is given a score based on several features like skin tone, proximity to face space, and the Fisher discriminant. Iyengar *et al.* [Iyengar 01a, Iyengar 01b] add a verification step by using trained Gaussian mixture models to differentiate correctly found mouth regions from incorrectly found other face regions.

Normalised RGB colour offers another solution [Graf 96, Wark 98] but a transformation into a colour space such as the HSI space (sometimes also called HLS colour space [Foley 96]) is better still. In the HSI space, hue and saturation are separated from intensity [Vogt 96]. The hue and saturation values can be used to find skin colour regions in the image. These values are also surprisingly consistent across human skin colour types [Kjeldsen 96].

Petajan and Graf [Petajan 96] used a face tracking system based on morphological filtering of single frames to find the relative positions of eyes, nose and mouth. The processing was done on colour images with colour thresholds used to distinguish skin colour and non-skin colour parts. The results (head position, scale due to distance from camera, and tilt) were used to initialise a nostril tracking algorithm from which the position of the mouth was derived. With a camera placed slightly below the face, the nostrils were claimed to be practically always visible and never obscured, not even by facial hair.

Applying Geometrical Constraints. To make the face tracking system more robust, typically a model enforcing geometrical constraints is applied to the skin colour blob. This often takes the form of positional information (e.g. the eyes are above the mouth) and relative distances (e.g. the horizontal distance between the eyes is roughly the same as the vertical distance from the eyes to the mouth). Such information is either based on heuristics or gathered when the face model for a specific speaker is built. The first automatic speechreading system, developed by Petajan [Petajan 84], used the information from tracking the nostrils to identify

the mouth region. Another example is the face tracking system of the Robotic Systems Laboratory at the Australian National University [Heinzmann 97, Newman 00, Zelinsky 99]. In order to make the tracking more robust, the tracked facial features are interconnected as a rigid structure which pulls badly tracked features to their correct position. Furthermore, the position of the features in the next frame is estimated from the motion information during the change from the previous frame to the current frame using a Kalman filter. Other motion-prediction methods use optical flow techniques (e.g. McKenna and Gong [McKenna 96]).

Other Approaches. Sobottka and Pitas [Sobottka 96] applied an active shape model (see 'Explicit Lip Feature Extraction' in Section 2.3.3) after the skin colour segmentation in HSV colour space (similar to the HSI colour space [Foley 96]). The mouth and other facial features were found using morphological filtering and geometrical models. Although the face and mouth were found in the test images, the documented results appeared to be rather rough and not very accurate.

Last but not least, systems that learn to track a face shall be mentioned here. Features (or landmarks) are selected — often manually — and their geometrical distribution is learned by a statistical model. The models can be statistical feature models (Cootes and Taylor [Cootes 96]), face graphs (Maurer and von der Malsburg [Maurer 96]), maximum likelihood models (Colmenarez and Huang [Colmenarez 96]), artificial neural networks (Reinders *et al.* [Reinders 96]), or Gaussian mixture models (Iyengar and Neti [Iyengar 01a]).

Practical Issues

For real-world applications, face tracking systems must be able to cope with multiple faces in the scene. The systems mentioned above almost all assumed that there was only one face in the scene. Zelinsky *et al.* [Zelinsky 99] used a face tracker based on skin colour segmentation to find face candidates in the image. The candidates were then tested for certain geometrical constraints such as the relative position of the eyes and the mouth. The largest skin colour blob fulfilling the constraints was taken as the head to be tracked. A multi-person system needs to test all face

candidates and, if found to be valid, track them as well [Krumm 00, Nakadai 01].

A final issue is the head pose towards the camera. Most systems simply assume the face to be in a (near) frontal position but for real-world applications, it is questionable how realistic such an assumption is. One solution is the use of a stereo camera system, as was done by Newman *et al.* [Newman 00], which recovers depth information and thus offers truly 3D information independent of the exact face position towards the cameras. This approach is also used in the work described in this thesis. Another solution was presented by Holden *et al.* [Holden 00a], where a 3D head pose model, based on the outer corners of the eyes and the corner of one nostril, is projected into 2D image space, compared to the measured feature locations, and its 3D position adapted to reflect the actual head pose.

2.3.3 Automatic Facial Feature Extraction — An Overview

While well-established parameters exist for the audio modality of the speech signal, it is not clear which parameters best describe the visual speech information. Facial features must be extracted on which video speech parameters can be based. Two main streams of feature extraction can be identified: *implicit feature extraction* and *explicit feature extraction*. These are discussed in the following sections. Note, that combinations of implicit and explicit features have also been proposed, for example by Chan [Chan 01] who used geometric and appearance features. For the sake of providing a concise overview, such approaches are not considered further here.

Implicit Feature Extraction

A part of the image data which contains the mouth area is taken as is, and the pixel values are used as input of the recognition engine (e.g. HMM, artificial neural network). Thus, the recogniser learns the typical pixel patterns associated with certain lip movements. A principal component analysis (PCA) or linear discriminant analysis (LDA) can be employed to reduce the dimensionality of the input vector and to define the main directions of variation. Only a few principal components are typically required to account for almost all variation. Examples of such implicit feature

extraction systems are the ones by Meier *et al.* [Meier 96, Meier 00], Movellan and Chadderdon [Movellan 96], and Potamianos *et al.* [Potamianos 00, Potamianos 01].

Holden and Owens [Holden 00b] used shift-invariant, computationally inexpensive higher order local autocorrelation (HLAC) features extracted from cepstral images. The cepstral images were generated by applying LPC to image sequences and by converting the LPC coefficients into cepstral coefficients, which were then taken as pixel values to form the cepstral images.

Optical flow techniques also fall into the category of implicit feature extraction methods [Horn 81]. Mase and Pentland suggested an automatic speechreading system on the basis of optical flow [Mase 91]. Motion rather than shape was extracted which follows the earlier argument that the dynamic patterns of the visible speech articulation are perhaps more important to the recognition than the static lip shapes. The flow field was computed from two adjacent frames and used for recognition (the velocity components comprising the flow fields, to be more exact).

Implicit feature extraction avoids explicitly finding facial feature points and preserves both shape and appearance information [Scanlon 01]. The disadvantages are that without a PCA or similar technique, the dimensionality of the input vector becomes very large (e.g. a 20×15 pixels window results in a vector with 300 elements!) and some effort must be made to compensate illumination changes (either by having a well-illuminated face or by using a normalised colour space at least). Most important of all, however, is the fact that the systems can only be trained for one specific angle of the face towards the camera and can thus not cope with a freely moving head, unless several recognisers were trained for different head poses and some sort of interpolation between these were used. Alternatively, some head pose compensation method based on image warping could be employed.

Explicit Feature Extraction

Here, image processing techniques are used to extract the position of certain facial features, such as eye corners or nostrils in general, or the mouth features in the case of AVSP. Mouth features are certain points on the lips (e.g. lip corners) as well as

the internal and external lip contour line or the position of the teeth. Parameters describing the shape and the movements in the mouth region are then derived from the positions of these features in the image data. The effect of the overall head movement must be eliminated in the set of parameters to be extracted. Only components comprising mouth region movements are wanted. Parameter sets are often based on studies on what human perceivers appear to use for AV speech perception (see Section 2.1.6 and [Cathiard 96, Cosi 96, Smeele 96]). The mouth region with the visible articulators carries most of the visual speech information. Other facial features (e.g. lower jaw) contribute as well but they are harder to track automatically in a non-intrusive way and are therefore not considered in this study.

Methods for the extraction of mouth features are described in more detail below, as this approach is followed in this study. Using explicit features has the advantage that they can be chosen to be directly related to the visible speech articulators which facilitates the interpretation of the results of a statistical analysis of the relationship between audio and video speech parameters (see Chapter 6). Generally, measurement problems can arise from a number of observed features. Firstly, the lip corners are often in a shady area, if normal illumination is used in the scene. As a result, the internal and external lip contour lines are hard to distinguish at the lip corners. Secondly, the lip colour can be very similar to the surrounding skin colour, so that it is hard to find the external lip contour line [Eveno 01]. The same is true for the tongue, which will affect the extraction of the internal contour line. Thirdly, lips move very quickly while speaking, so any method must be able to adjust rapidly to different shapes in the area surrounding the feature points.

2.3.4 Automatic Explicit Lip Feature Extraction

Once the approximate location of the mouth region has been determined, the position of the lip features needed to calculate the above mentioned parameters must be found. Methods range from *image-based methods* to *model-based approaches*. An overview of these methods is given now. The lip tracking algorithm used in this study is presented in detail in Chapter 3.

Image-based Methods

A simple method on greyscale images is the *integral projection*, in which the grey-values of each row or column are summed up and yield an intensity distribution curve [Yang 98]. Another simple image processing method is *thresholding*, specifically targeted to the extraction of the lips [Petajan 84, Petajan 96, Wojdel 01b]. However, it suffers from the fact that intensity information from light source reflections (highlights) is superimposed on the skin values. The use of colour adds a lot more information to the process. Colour spaces that separate hue and saturation from intensity work best [Kjeldsen 96, Wark 01, Wojdel 01a]. Petajan and Graf [Petajan 96] used the results of their nostril tracking system to find the mouth region and then applied colour thresholding to determine the inner lip contour. Integral projection and thresholding work reasonably well for the extraction of the inner lip contour because the dark mouth opening contrasts well to lips and skin. However, they cannot reliably distinguish the external lip contour from the surrounding skin because the values both in greyscale and colour images are too close.

A third image-based method is *edge detection* by a suitable convolution filter, for example a Sobel or Canny filter [Russ 95]. Edge detection methods typically work well for the middle parts of the lip contours but often fail for the lip corners because they lie in a shady area with little contrast. Since edge detectors are essentially contrast enhancers, their failure to detect lip corners does not surprise.

Recently, Holden and Owens ([Holden 02b], *personal communication*, Robyn Owens, University of Western Australia, Western Australia, Australia) showed that wavelets can be used to efficiently represent facial points and their local surrounding features. Facial feature points similar to a sample wavelet response of a facial point were used to find candidate points which were then tested for compliance with certain geometric relationships between facial feature points.

Many of the problems with image-based methods arise from the lack of sufficient contrast in the mouth region. Moreover, the contrast of the lips to the surrounding skin is illumination-dependent. While there is good contrast in well-illuminated areas, there is very little in shady areas. Benoît *et al.* [Benoît 96] and Bothe

[Bothe 96] use artificially coloured lips to enhance the contrast and to facilitate the above mentioned image-based methods for feature extraction. Although that is a valid way of simplifying the feature extraction problem, wearing blue lipstick is a considerable step away from a natural, non-intrusive system, which is the application scenario in mind for this study. Artificial facial markers, for example infrared LEDs tracked by an infrared camera system (OPTOTRAK [Vatikiotis-Bateson 95], Qualisys [Nordstrand 03]), offer another way of extracting feature points with high accuracy, but again it is an intrusive system which requires familiarisation for the speaker and is a step away from practical applications.

Template Matching

A step further towards model-based approaches is the use of template matching algorithms [Russ 95]. They are based on the cross-correlation of images which are taken as 2D functions. Some part of an image — the template — is moved across the target image and the correlation values are calculated for each position. The position with the highest correlation value is the one with the highest degree of similarity. If the template was taken from the same image, then the exact position is found. The idea of template matching is, however, to find a feature of interest in a different image. Noise as well as shape or pose differences will thus affect the matching process. Since the shape of the lips changes quickly and quite significantly while speaking, static image templates do not work very well for the mouth region.

Deformable 2D Models

This led to the development of deformable 2D models or templates. Yuille *et al.* [Yuille 92] developed a *2D deformable mouth template* consisting of a mouth-closed template and a mouth-open template. The mouth-closed template is attracted by the deep intensity valley corresponding to the dark shadow line between upper and lower lip. The mouth-open template uses the presence of teeth in addition. Energy potentials are calculated to determine the goodness of fit.

Kass *et al.* [Kass 88] propose *active contour models* or *snakes*. They are energy-

minimising splines guided by external constraint forces and influenced by image forces that pull it toward features such as lines and edges. However, if the features move too rapidly, the snake can lose them and be attracted by a different feature. Given that lips can move very rapidly, this is a potential risk unless the video frame rate is so high that inter-frame differences are small. Nevertheless, snakes have been applied to lip tracking in ASR systems [Kass 88]. The relative position of the control points of the splines serves as input to the recogniser. Such a system is thus closer to an implicit feature extraction system. Barnard *et al.* [Holden 02a] combined snakes with 2D template matching, so that the snakes were driven by matched templates of lip contour points, thus safe-guarding the lip tracking process.

A way to make active contour models more robust is to imply constraints on the shape of the snakes. Such a method was introduced by Cootes *et al.* [Cootes 95] with their *active shape models (ASM)* and subsequently used by Luettin *et al.* [Luettin 96] to track the internal and external lip contour for automatic speechreading. An ASM can only deform in ways characteristic to the class of objects it represents. These characteristics are learned from a set of training images and stored in a point distribution model. Whereas deformable templates and active contour models align to strong gradients for locating the object, ASMs learn the typical shape deformation and use it during feature search. The set of points comprising the active shape undergo a PCA to obtain the main modes of variation which are also used as input in the automatic speechreading system. Any normalised lip shape can be approximated using the learned mean shape and the first few principal modes of variation. A similar approach was used by Dalton *et al.* [Dalton 96]. The possible motion patterns of a snake were learned from training image sequences and a Kalman filter was used to predict the motion during tracking.

In order to make the ASM even more robust, Matthews *et al.* [Matthews 98] employed an *active appearance model (AAM)* which is an extension of the ASM. It combines an ASM with a statistical model of the grey-values in the region around each point of the ASM. By iteratively minimising the difference of the grey-values of the model and the image, the parameters of the shape model can be updated to fit the model better to the object's shape (the lips in this instance) in the image.

Model-based Methods

Finally, fully model-based approaches are described. Based on the experience with the previous system by Benoît and Adjoudani [Adjoudani 96], Revéret [Revéret 98] developed a system in which a *3D lip model* is fitted to image data for lip tracking, speech recognition and visual speech animation purposes. The lip model consists of a 3D polynomial surface model controlled by three articulatory-oriented parameters learned on the speaker. The surface is defined by three 3D contour curves which are exactly interpolated by 10 control points, corresponding to geometrical features such as the lip corners, each. Thus, a total number of 30 points control the surface.

During the training of the system, a graphical user interface is used to fit the model onto the lip image. A combination of calibrated front and profile views is employed to fit the 3D model. The XYZ positions of the control points form the visual feature vector. From a viseme analysis, ten key lip shapes are defined and a PCA on these is performed. The first three principal components account for 94% of the total shape variance and are thus subsequently used as direct control parameters of the model. During lip tracking, a 2D projection of the 3D model is calculated at each time step and the three control parameters are changed until the projection resembles the lip contour in the image.

The lip tracking was tested on a single phonemically-balanced sentence. The results show that internal width and height as well as external height are recovered well, whereas the external width is harder to measure due to shadows around the lip corners which make an exact extraction difficult.

A similar system based on the *backprojection of a 3D model* into 2D image space and adjusting the model parameters until the model fits the mouth shape in the image was developed by Basu *et al.* [Basu 98]. The model is built on physical and statistical information about permissible mouth shapes from training image sequences. Colour information is used to determine the similarity of the lip model and the shape in the image data.

In summary, image processing methods work well, if appropriately chosen. More robust feature extraction can be achieved with model-based approaches but the

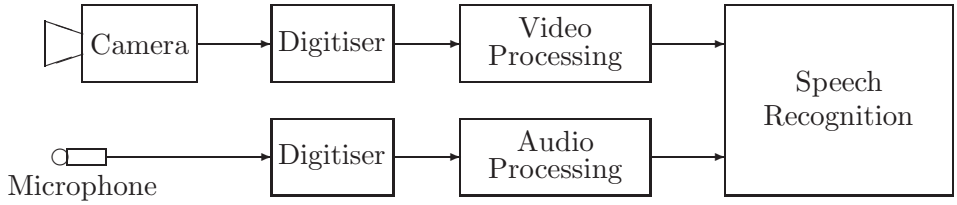


Figure 2.5: Schematic representation of an AVSP system for speech recognition.

increased computational complexity affects the real-time performance. An extra level of speaker dependency is added through the model learning (training) phase.

2.3.5 Audio-Video Automatic Speech Recognition and Integration

In this section, the various approaches to AV automatic speech recognition (ASR) are summarised and categorised in terms of the method used for the visual feature extraction. Illustrating examples from the literature are given and the approaches to AV integration are discussed and classified using the taxonomy by Robert-Ribes *et al.* [Robert-Ribes 96] (Section 2.1.7), where information was available.

Figure 2.5 shows a general schematic representation of an AVSP system for ASR. It consists of a video subsystem and an audio subsystem. The video system captures images of the scene including one or more talkers. Using an analogue camera, the images first need to be digitised, e.g. by a framegrabber. This step can be omitted if a digital camera is used. The images are then processed to determine the location of the talking face (if any) and to extract parameters based on relevant facial features for the speech recognition system. Such parameters describe the location, shape, or motion of facial features such as the lips and the jaw.

Parallel to the video subsystem, the audio subsystem records the acoustic signal through a microphone. The signal is subsequently also digitised and processed to extract relevant features for the recognition task. The synchronisation of the audio and video signals plays an important role if the additional visual information is to be of any use. According to Hennecke *et al.* [Hennecke 96], an AV ASR system is not much different from an audio-only recognition system in the sense that the available

input is digitised and processed before it is fed into some kind of statistically-based recognition engine. The recogniser responsible for the AV processing is often the same as in an audio-only system. As discussed in the Section 2.1.7, the integration of the audio and video information, both in terms of how and when, is still an open research issue for both human beings and machines.

Using Image-based Methods for Visual Feature Extraction

The first automatic speechreading system to be combined with an acoustic ASR system was developed by Petajan in 1984 [Petajan 84]. The system consisted of two separate recognisers for the audio and video input. Audio processing was done by a commercial discrete utterance recogniser (Voterm) which output the two most likely recognition candidates together with their recognition scores. The video processing was based on grey-value thresholding and contour coding to detect the nostrils which were used for face tracking by a single camera. The mouth was assumed at a fixed distance below the nostrils. Next, parameters of the mouth region such as area, perimeter, width and height of the mouth opening were derived. Since the acoustic information is a time series, the video recogniser was trained with the difference in the parameter values from frame to frame for each utterance. During recognition, the parameter template with the closest set of parameters was chosen.

The acoustic recogniser dominated the system. The word candidates from the acoustic recogniser were the input into the video recogniser which chose the one with the best lipreading score as the final recognition result. As such, the system employed a combination of the SI and DR models of AV integration. The system was tested on single-word utterances and showed an improvement in recognition scores in comparison to the audio-only recognition. Apparently, no particular effort was made to control the acoustic environment, which is only stated as *"moderately noisy due to cooling fans and air conditioners"* in [Petajan 84]. The speaker's face was positioned directly in front of the camera with two light sources, pointing at 45° to the speaker, at either side of the camera.

While this early system required the user to maintain a steady, frontal pose to-

wards the camera, a later system by Petajan and Graf [Petajan 96] incorporated a face tracking system which allowed the user some head movement (see 'Finding the Face' in Section 2.3.2). The face tracking system initialised a nostril tracking procedure from which the mouth region and then the inner lip contour were determined. The visibility of the teeth was detected by a colour search within the inner lip contour area. Petajan and Graf refer to improved applications in AV ASR systems without giving details about them or the way the two modalities were integrated.

Using Implicit Methods for Visual Feature Extraction

Yang's real-time face tracking system [Yang 98] was used in an AV ASR system by Meier *et al.* [Meier 96, Meier 00]. Using a modular multi-state time-delay neural network (MS-TDNN) architecture, an acoustic and a visual TDNN were trained separately. The position of the lip corners was used to determine a 24×18 pixels 'lip window'. The grey-values of all pixels in it formed the visual input vector into the network. The dimensionality of that vector was reduced using Linear Discriminant Analysis (LDA). Explicit feature extraction was thus avoided which made the algorithm more robust but slower due to processing redundant information. The acoustic input consisted of 16 mel-scale cepstral coefficients.

Different levels of combination of the two signals were tried: on a phonemic level (SI model), on the input level (DI model), and on the hidden layer level (MR model). The recognition task was the speaker-dependent continuous spelling of German letter strings in different noise scenarios. All combination methods resulted in an improved recognition performance over the audio-only recognition, particularly with high background noise. The best results were obtained using a combination on the phonemic level (SI model) [Meier 96].

A similar approach was taken by Movellan and Chadderdon [Movellan 96]. Again, instead of explicitly extracting facial features, a preprocessed part of the image showing the lips was used as input vector, so that the recognition system developed the feature detectors that best solved the task. The speakers had to centre and align their lips to the camera; no further face tracking was incorporated.

The image frames were symmetrised along the vertical axis, the temporal difference to the previous frame was obtained, then low-pass filtered and soft-thresholded. The acoustic input features came from a standard LPC/cepstral analysis. Hidden Markov Models (HMMs) were used as recognition engines.

The auditory and visual features were either processed by separate banks of HMMs and then their results were integrated (SI model), or by a common bank of HMMs (DI model). The training and recognition was done with the Tulips1 database [Movellan 95] containing 12 speakers uttering the digits from "one" to "four" twice. Assuming conditional independence in Bayesian analysis, as for the late integration (SI) model, yielded marginally better results than the early integration (DI) model. However, the results must be treated with care because of the small size of the database. Four isolated words per speaker cover only few phonemes and visemes. Nevertheless, the results showed a clear improvement by using audio and video signals over the audio-only speech recognition results.

Hierarchical LDA. Also in this category of systems is the IBM AV ASR system [Potamianos 01]. In this system, first an LDA was applied separately to both input signals to discriminantly reduce the dimensionality of the feature vector. This was followed by a maximum likelihood linear transform (MLLT) which maximises the observation data likelihood in the original feature space under the assumption of diagonal data covariance in the transformed space.[12] In a second stage, LDA and MLLT were applied again, this time to the concatenated AV feature vector. This two-stage process was therefore called *Hierarchical LDA (HiLDA)*. It corresponds to the recoding integration models but recoding is neither done in the motor space, nor in the auditory space, but in a different space.

The audio and video feature vectors were concatenated 'static' features from consecutive sampling points, as a way of incorporating dynamic information in the ASR process. The static audio features were 24 mean normalised mel-cepstral coefficients. The static video features were the 24 highest-energy coefficients of a

[12] Diagonal covariances are typically assumed in ASR when modelling the observation class probability distribution.

discrete cosine transform (DCT) applied to the mouth region which was found using a statistical face tracking algorithm [Potamianos 00, Senior 99].

Recognition experiments were done with the IBM AV large vocabulary continuous speech database (see Section 2.3.6) [Neti 00, Neti 01]. Various other integration methods were tested as described in [Luettin 01, Glotin 01]. Multi-stream HMMs were shown to be able to handle asynchrony between the audio and video streams which may be important as visual speech activity usually precedes acoustic activity (e.g. [Kohlrausch 00, Massaro 98]). The best results were achieved for multi-stream HMMs when the training of the HMMs was done jointly. Slight decreases in the word error rate were reported for clean audio conditions, but significant decreases found for noisy audio conditions.

Improved results were achieved when applying adaptive weights to the audio and video input depending on an estimate of the audio reliability (cf. [Rogozan 97]), which can be seen as a SI model of AV integration. Both clean and noisy audio AV recognition improved considerably. In the discriminative model combination (DMC) [Beyerlein 98], the audio and video streams were used independently to train models which were then combined with a language model, with weights optimised to minimise the word error rate on a held out training set. DMC aims at an optimal integration of independent sources of information. Some improvements of recognition performance were reported for clean audio conditions, while no experiments were done on noisy audio conditions. As a general comment, it should be noted that the DMC approach suffers from the lack of synchronisation between the two streams, as they are used independently.

Using Model-based Methods for Visual Feature Extraction

Bregler and Omohundro [Bregler 94b] presented a system based on *active contour models* which performs a PCA on the points of the snake to determine the major components responsible for shape variations. The term 'eigenlips' was coined for these components [Bregler 94a] and they were used as input into a hybrid connectionist MLP/HMM speech recognition system. Bregler and Omohundro used a

DI model to integrate the visual and acoustic input features already on the input level. The MLP was trained to estimate the likelihoods that a 'specific phone' (sic - [Bregler 94b]) was related to the current bimodal input vector and these likelihoods were used as input for the HMM. The speech material came from a multi-speaker spelling task database (6 speakers, 2955 connected letters in total). The system was tested in different levels of acoustic noise. The results showed that the additional visual information helped to reduce the word error rate by up to 20%.

Active Shape Models. A system based on *active shape models (ASM)* was introduced by Luettin *et al.* [Luettin 96]. ASMs have been discussed in 'Deformable 2D Models' in Section 2.3.3. To restrict the model to only deform to shapes similar to the ones in the training set, the shape parameters of each principal mode of variation were constrained to stay within ± 3 standard deviations. The system assumed that the mouth region had been extracted by another face tracking algorithm which was not specified in the article. Movellan's Tulips1 database [Movellan 95] was used for the experiments. However, the results presented state only the outcome of the lip extraction stage of the system. They showed that ASMs were able to represent deformable objects such as lips with an acceptable degree of accuracy, although the correctness of the lip extraction process was only judged visually. The advantage of ASMs over snakes is that the deformation is purely governed by statistics learned from a training set and is therefore neither too constrained nor too flexible.

Active Appearance Models. Matthews *et al.* [Matthews 98, Matthews 02] compared three different methods of lip shape extraction for AV ASR systems: model-based *ASMs*, intermediate *active appearance models (AAMs)*, and image-based *multiscale spatial analysis (MSA)*. The first two techniques have been described in Section 2.1.6. MSA is a pixel-based method, which decomposes an image into a granularity domain using a nonlinear scale-space decomposition, which is a mathematical morphology serial filter (Matthews *et al.* [Matthews 02]). The granules are the extrema which are progressively removed from the input signal by using the filter with increasing scale. Scale histograms show the distribution of features over

scale and are used as visual input in the recognition experiments.

Classification was done with HMMs. The acoustic features were not further specified. The AV speech material consisted of the letters A to Z spelled by 10 speakers three times. Separate recognisers were applied to the audio and video signals. The results were combined (SI model) using a confidence measure based on the uncertainty of the acoustic recogniser about a word at a given SNR. In all three methods, the AV recognition performed better than the audio-only one. AAMs and MSAs performed similarly well but ASMs showed poorer results.

Lip Models. Dalton *et al.* [Dalton 96] used a dynamic contour tracker to track the lips. The use of blue lipstick was required to enhance the contrast. The 2D outline of the lips was parameterised by quadratic B-splines. Tracking was achieved by generating estimates of the B-spline control points to match the lip contour. Lip motions were described in terms of deformations to an average mouth shape, controlled by some shape constraints. A PCA on the set of training data discovered the main modes of variation of the model, with the first six modes expressing 99% of the variation. The acoustic features were 8 mel-scale cepstral coefficients. Bimodal feature vectors were formed which relates to a DI model. Recognition experiments were conducted using an AV dynamic time warping (DTW) isolated-word recogniser. The performance was evaluated on a single speaker and 40 words vocabulary with different levels of acoustic noise. The results showed an improved performance for the combined input signals at all levels of noise.

The AV ASR system by Benoît and Adjoudani [Adjoudani 96, Benoît 96] consisted of two calibrated cameras recording the carefully made up (blue lipstick) mouth region from a frontal and a profile view, and a chroma-key system which converts the blue lips into saturated black colour to ease edge-detection. Parameters measured included internal and external lip contour width and height, the lip area, the area of the oral cavity, as well as the protrusion of the upper lip, the lower lip, and the lip contact point measured to a vertical ruler mounted on the speaker's glasses. These protrusion parameters were not used in the recognition experiments because they would be difficult to extract in a real application. 12

cepstral coefficients were used as auditory features. The speech material consisted of nine repetitions of 54 isolated non-sense words spoken by one speaker. The two modalities were integrated in a HMM. Both a DI and SI models of AV integration were tested. In the latter case, the decision to rely more strongly on either the audio or the video modality was based on the uncertainty of the acoustic recogniser reflecting the signal-to-noise ratio (SNR). The experiments showed that the SI model outperformed the DI model. In both cases, the combined AV recognition results were better than the audio-only recognition ones.

2.3.6 Audio-Video Speech Data Corpora

For testing and comparing results published by various research groups in the field of AVSP, a common basis in the form of a comprehensive, systematically designed AV speech data corpus would be of great value. Such a publicly available 'benchmark' AV speech data corpus still does not exist, despite a number of corpora having been produced over the last few years. Many corpora appear to have been designed with a specific application in mind, rather than being based on a general phonemic and visemic analysis. Some corpora have already been mentioned in previous sections. This section only discusses some major corpora for the English language.

A good overview of existing AV speech corpora is given by Chibelushi *et al.* [Chibelushi 96a]. Their study led to the creation of the well-designed DAVID corpus [Chibelushi 96b] which consists of four different subcorpora, each addressing a particular research issue. The first subcorpus addresses the issue of facial image segmentation under different conditions, including variable illumination, variable backgrounds, and facial distractors such as glasses and hats. This subcorpus has 6 subjects. The second subcorpus is designed for research in the area of automatic speech and person recognition and contains recordings of 31 clients and 92 impostors. A subset of 9 subjects has highlighted lips (blue make-up) to facilitate the lip extraction process. Both the first and second subcorpus use the set of digits from 0 to 9 as speech material. Subcorpus 3 is intended for speech-assisted video

compression and the synthesis of talking heads. VCVCV[13] utterances of 5 subjects were recorded. The fourth subcorpus is concerned with automatic speech and person recognition with application in video-conferencing systems. Hence, it contains sentences from a business control set spoken by 31 clients and 92 impostors. All recordings show a frontal and profile view, achieved by a mirror construction and a single camera, together with the associated synchronous audio.

A well-established AV speech data corpus is the M2VTS database and its successor XM2VTSDB [Messer 98, Messer 99]. Whereas the M2VTS database contains 37 speakers, the XM2VTSDB database comprises recordings of 295 speakers. Four sessions were recorded to account for natural changes in appearance of the speakers. During each session, an AV speech recording was made as well as a head rotation sequence. The speech material recorded consists of three sequences, two of which contain the digits from 0 to 9 in different order. The third sequence is *"Joe took father's green shoe bench out."* which was designed to maximise visible articulatory movements. It contains all phoneme and viseme categories (but not all phonemes). The XM2VTSDB is currently the largest publicly available AV corpus in terms of numbers of speakers but suffers from the small number of different sequences for each speaker with respect to a complete phonemic and visemic analysis.

The Tulips1 data corpus recorded by Movellan [Movellan 95, Movellan 96] contains the four digits 'one', 'two', 'three', and 'four' repeated twice by 9 male and 3 female subjects. This speech material was chosen with a phone number spelling task in mind. Only frontal views are recorded. As such, the corpus is rather small and application-driven.

Other AV speech databases have been recorded by various research groups but are not publicly available. One such proprietary data corpus is the IBM LVCSR[14] AV corpus [Neti 00], which contains continuously spoken utterances of the IBM ViaVoice training set from more than 290 American English speakers in different environments (office, car). The video stream is compressed using MPEG-2.

Recently, the CUAVE corpus was introduced by Patterson *et al.* [Patterson 02].

[13] VCVCV = Vowel-Consonant-Vowel-Consonant-Vowel

[14] LVCSR = Large Vocabulary Continuous Speech Recognition

It contains recordings from about 50 American English speakers, uttering connected and isolated digits. The sequences are stored as MPEG-2 files. The data is fully labelled at a millisecond level.

Although a comprehensive and systematically-designed audio data corpus exists for AuE (ANDOSL [Millar 94]), no AV speech corpora exist. As a result, a new AV corpus for AuE has been created for this study (see Chapter 4). The AVOZES data corpus is systematically designed to contain the phonemes and visemes of AuE and comprises utterances from 20 speakers.

2.4 Statistical Analyses of Audio-Video Relationships

The literature is scarce with respect to statistical analyses of the relationships between audio and video speech parameters. Yehia *et al.* [Yehia 97, Yehia 98] presented a study with one speaker of American English and one of Japanese, in which they investigated the degrees of correlation between vocal-tract, facial, and acoustic parameters. Vocal-tract motion was tracked electro-magnetically using small transducers placed mid-sagittally on tongue, lips, and lower teeth. Facial motion was captured as 3D trajectories of infrared LEDs placed on the lower face, including the lips. RMS amplitude and line spectrum pairs (LSP) derived from linear prediction coefficients were used as acoustic parameters. Vocal-tract data accounted for 91% of the variance in the facial data, with the latter accounting for 80% of the variance in the vocal-tract data. The acoustic data accounted for ≈60–70% of the variance in both the vocal-tract data as well as the facial data. Surprisingly, the acoustic data was also well estimated (≈75% correct) by the facial data alone.

A similar study of the correlation between facial movements, tongue movements, and speech acoustics in American English was performed by Jiang *et al.* [Jiang 02]. An optical tracking system (Qualisys) tracked the position of markers on the entire lower face, from which relative distances were computed as video speech parameters. Tongue movements were captured by an electro-magnetic midsaggital artic-

ulography system. The audio speech parameters were LSP and RMS parameters. Averaged across the four speakers, ≈69% of the information in the video speech parameters was accounted for by the audio speech parameters and 47% of the acoustic information could be recovered from the video speech parameters. Tongue movements were well (≈75% correlation) predicted from audio or video speech parameters, but in the opposite direction, the correlation values decreased to 52% for the audio speech parameters and to 66% for the video speech parameters. Chin movements were easiest to recover, then lip motion, and finally cheek movements.

Barker and Berthommier [Barker 99] tested both linear and non-linear models for AV relationships in French, using similar techniques to Yehia *et al.* [Yehia 98]. Their facial parameters were purely related to lip and jaw movements, not the entire lower face, measured using a chroma-key technique on blue made-up lips (see Section 2.3.4). The acoustic parameters were again LSP and RMS parameters. Acoustic data accounted for ≈75% of the variance in the facial data but the opposite way was only correlated at 55%, which is less than what Yehia *et al.* reported. The difference was attributed to measuring the lip and jaw movement only, not the lower face. Yehia *et al.* reported strong correlations between small movements of the cheeks and the horizontal position of the tongue. Such movements could not be measured by lip and jaw parameters. Barker and Berthommier demonstrated that non-linear models are able to represent the AV relationships better than linear models, because of the non-linear relationship between vocal tract shape, acoustics, and visible speech articulation, but linear models provide a good first approximation.

2.5 Chapter Summary

This chapter has given a broad overview of methods presented in the literature on the various aspects of audio-video speech processing (AVSP). First, the processes involved in AVSP by humans have been described. Most notable is the so-called McGurk effect which clearly shows that human speech processing is not only affected by the acoustics, but also by visual speech information, which can be found in particular on the lower face half. A review of the processes involved in speech

production and speech perception from an AVSP angle has been given. This has included the acoustics of the vocal tract as well as models which connect the vocal tract shape with the acoustic properties, e.g. formant frequencies, of the speech sounds generated. Of the articulators, only the lips and jaw (mandible) are fully visible and the tongue and teeth are visible at times. Movements of these articulators lead to a change in the vocal tract geometry which results in changes to the produced speech sound. An overview has been presented on theories of speech perception and investigations into which face parts are most relevant for the visual speech cues. Models of how the two modalities are integrated have been discussed. Integration in a common, amodal space (motor space recoding) or in a hybrid model (e.g. FLMP) appeared to be the models best explaining the integration.

Next, the characteristics of AuE, which shows little dialect variation due to regional differences but more variation due to socio-economic factors, have been discussed. Speakers of AuE are typically classified into broad, general, and culitvated pronunciation with the changes from one class to another being continuous rather than discrete. The strongest differences are found for vowels and diphthongs.

Finally, literature on AVSP by machines has been presented. The discussion has included an overview of fundamental techniques in audio- and visual-only speech recognition systems. While well-established parameters exist for the audio modality (in form of features from a spectral analysis), it is not clear which features describe the visual speech information best. Implicit and explicit feature extraction methods have been described. Methods of the explicit lip feature extraction approach followed in this study have been discussed in more detail. Next, some AVSP ASR systems based on the various approaches have been compared. It should be noted that common data corpora to enable close comparison of the various results reported are still not available in the field of AVSP, with the XM2VTSDB corpus coming closest to the requirements among the publicly available corpora.

In conclusion, the literature review has shown that several areas of AVSP need further investigation. Among them are the non-intrusive extraction of visual speech information from the face as well as a thorough investigation of the AV relationships of extracted audio and video speech parameters. These two areas are investigated in

this study. A novel non-intrusive lip tracking algorithm is presented which is based on a stereo vision system. Such a system increases the naturalness of utterances spoken by test subjects because they are not restricted by artificial markers on their face. Modern statistical methodology offers interesting ways to investigate relationships of AV speech parameters. The recent method of coinertia analysis is introduced to the field of AVSP and applied to characterise the AV relationships for AuE. As suitable AV speech corpora do not exist for AuE, the new AVOZES corpus is created based on a proposed new design framework for AV speech corpora.

Chapter 3

Lip Tracking Using Stereo Vision

To derive parameters describing the visible speech movements in the mouth region, it is essential to track these movements. This chapter describes a set of techniques to achieve this motion tracking. First of all, a method to find a human head (or face) in the video data and to determine its pose was required. This is referred to as the face tracking system. It is described in detail in Section 3.1. The 3D lip model used in this study is presented in Section 3.2. Secondly, a way to localise the mouth region and to track the motions in this region was needed. As discussed in the literature review (Sections 2.3.3 and 2.3.4), feature extraction for the description of movements in the mouth region is either implicit, i.e. using pixel-based techniques, or explicit, i.e. using geometric techniques. In this study, an explicit geometric feature extraction approach was followed, because it was judged to facilitate the interpretation of the results of the statistical analyses of the relationship between audio and video speech parameters. A novel real-time lip tracking algorithm based on stereo vision was developed, which is able to track certain lip feature points accurately without requiring manual adaption to the speaker (Section 3.3). Stereo vision has the advantage that 3D coordinates of facial points can be measured irrespectively of the head pose, while monocular systems measure only 2D image coordinates without separating head pose-related effects from facial movements. Finally, Section 3.4 describes an experiment to validate the accuracy of the new lip tracking algorithm.

3.1 Real-Time Stereo Vision Face Tracking

Before a lip tracking system can be applied, it is first necessary to establish the position and orientation of the human face in the video frames. Several systems that achieved this were discussed in the literature review (Section 2.3.2). The face tracking system used in this study was developed by Newman [Newman 99a, Newman 00]. It is based on earlier work by Matsumoto[1] and Heinzmann [Heinzmann 98, Heinzmann 99, Matsumoto 99, Matsumoto 00, Zelinsky 99]. A detailed overview of the Robotic Systems Laboratory (RSL) face tracking system is given below, as it forms the basis for the subsequent lip tracking system.

3.1.1 System Outline

The face tracking system is based on real-time stereo vision processing. A stereo vision system has the advantage that depth information (distance from cameras to object) can be recovered from the stereo disparity, if the cameras are calibrated. A calibrated monocular camera system can only estimate depth — or the object's 3D position in general — if the object dimensions and its orientation are known. This is obviously not the case in unrestricted face tracking. The RSL face tracking system allows for a non-intrusive way of tracking facial features. No markers or special make-up are required, yet the system achieves a high degree of accuracy. These properties are highly desirable in an analysis of AV relationships, because any artificial tracking aids might inhibit the speaker from speaking naturally. Hence, there would be a risk of generalising the results to normal speech.

Figure 3.1 illustrates the system configuration. The two video cameras are standard, colour analogue NTSC cameras mounted side by side on a rig. The video output signals from the cameras are multiplexed into a single channel using field multiplexing [Matsumoto 97]. In this technique, a device containing a video switching integrated circuit selects the signal from one video stream as the odd field of the video output, while the signal from the other video stream becomes the even

[1] Nara Institute of Science and Technology, Japan. The work mentioned here was carried out during a visit to the Robotic Systems Laboratory, RSISE, ANU.

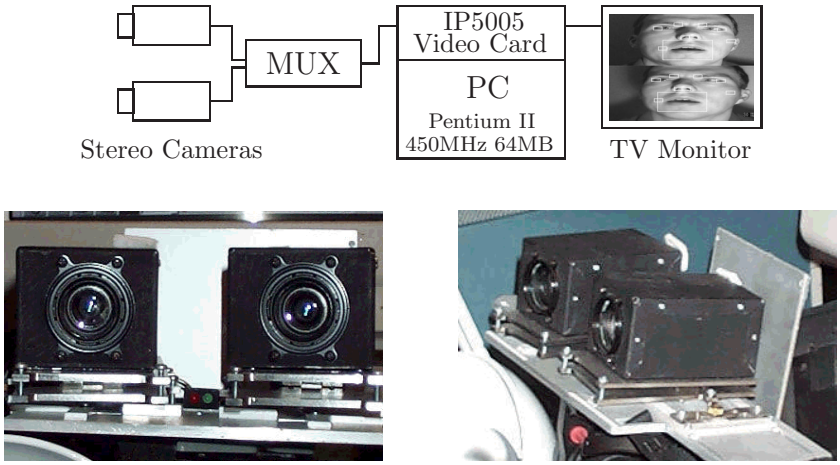


Figure 3.1: Top: Configuration of the stereo vision face tracking system. Bottom: Front and side view of the stereo camera rig.

field. This requires to first de-interleave the odd-even fields of the video frames from each camera. Multiplexing video signals in the analogue phase has the advantage that it can be applied to virtually any video hardware system. Images from two cameras can be stored in a single video frame. Stereo image processing can be performed within the computer's memory using only one image processing board. Single video stream processing is thus transformed into stereo vision processing.

A weakness of the field multiplexing technique is that only half the vertical resolution of the original video frame from each camera is available, as two video streams are compressed into a single frame. However, this disadvantage is more than outweighed by the ability to perform stereo vision processing with a single video card. Nevertheless, it would be worthwhile in future studies to consider rotating the stereo cameras by 90°, so that the halved resolution is in the horizontal direction rather than the vertical direction, which is potentially the more informative axis in visible speech articulation. Both face and lip tracking algorithms would have to be adjusted. Based on the results of the validation experiments (see Section 3.4), it was judged that the current setup was sufficiently accurate.

If displayed directly on a TV monitor, the multiplexed video output looks strange to the human eye, as the images from the two video streams alternate every other line. From an image processing point of view, this is no problem because

corresponding lines can easily be put back together to form two separate images (at half the original vertical resolution) again. One other weakness is the delay of 16.6ms between the images from the two video streams, which is inherent in the NTSC standard, as it is an interlaced video/TV standard. That is, first all the lines of one field, let's say the odd lines, are processed, then all the lines of the other field. The field frequency is 60Hz in the NTSC standard, or 30Hz frame frequency, and hence there is a 16ms delay between fields. This delay poses no problem for a face tracking application, but it is a potential error source for lip tracking (see Section 3.1.4 for a discussion of error sources).

The multiplexed video frames are sent to a Hitachi IP5005 video card for further processing. This video card was designed to perform a variety of fast image processing functions in real-time, for example, filtering, smoothing, erosion, convolution, normalised correlation. The card itself is a PCI-bus card running under the Linux operating system on a PC with a 450MHz Pentium II CPU and 64MB RAM. Image processing is done by the IP5005 video card hardware, while stereo reconstruction and head tracking[2] are performed in software in the PC memory. These steps are described in Sections 3.1.2 – 3.1.5 below. Once tracking information has been updated, the video output with overlayed tracking information is sent to a TV monitor (see Figure 3.1).

3.1.2 From 2D to 3D — Stereo Reconstruction

The standard pinhole camera model was used in the face tracking system, because any non-linear camera effects (radial and tangential lens distortions) were relatively small compared to the errors due to noise and stereo matching inaccuracies. In this model, the camera performs a linear perspective projection of an object point onto a pixel in the image plane through the camera centre. The camera arrangement and world coordinate system are shown in Figure 3.2. The cameras' centres, $\vec{c}_l$ and $\vec{c}_r$, are located equidistantly (about 55mm in the experiments) from the origin of

[2] The RSL face tracking system is also capable of gaze direction estimation. However, details of this estimation are omitted here, as they have no relevance to this work.

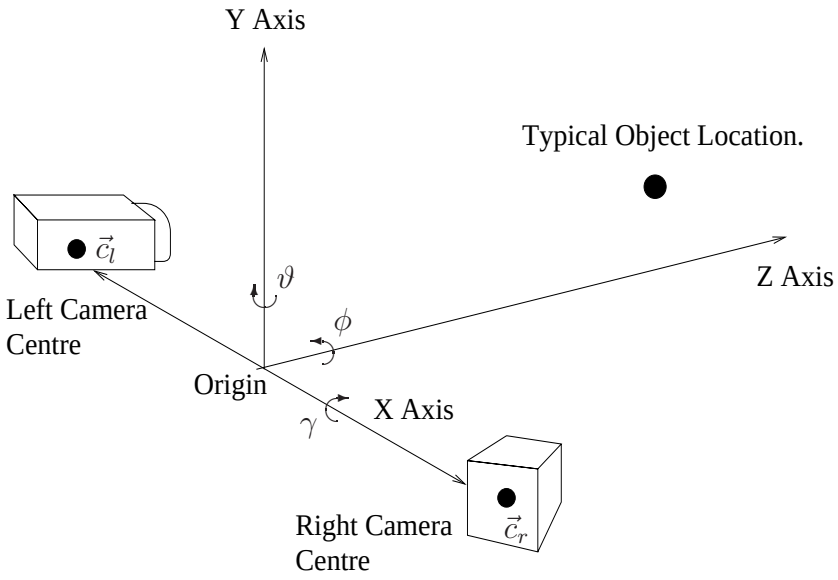


Figure 3.2: Stereo camera arrangement and stereo world coordinate system.

the world coordinate system on the x axis. The y axis is vertically upwards and the z axis points horizontally out into the scene.

It is important to understand and distinguish the various coordinate systems that will be referred to in the following. First of all, there is the *image coordinate system of each camera*. This is a 2D coordinate system, which is represented by (u, v) coordinates for the left camera and (r, s) coordinates for the right camera, respectively. Secondly, there is the *world coordinate system of each camera*. These are 3D coordinate systems with the origin (= centre of projection) in the camera centre. Finally, there is the *stereo world coordinate system*, depicted in Figure 3.2, with its origin halfway between the two camera centres.

Epipolar Geometry

Figure 3.3 shows the *epipolar geometry*, which is the basic constraint arising from having two cameras (at different locations = viewpoints) looking at the same scene. A very good introduction into epipolar geometry can be found in [Xu 96]. The line through the two camera centres, $\vec{c}_l$ and $\vec{c}_r$, projects to a point $\vec{e}_l$ in the left image plane and $\vec{e}_r$ in the right image plane. The points $\vec{e}_l$ and $\vec{e}_r$ are called *epipoles*. The camera centres $\vec{c}_l$, $\vec{c}_r$ and point $\vec{m}$ form a plane — the *epipolar plane* for the

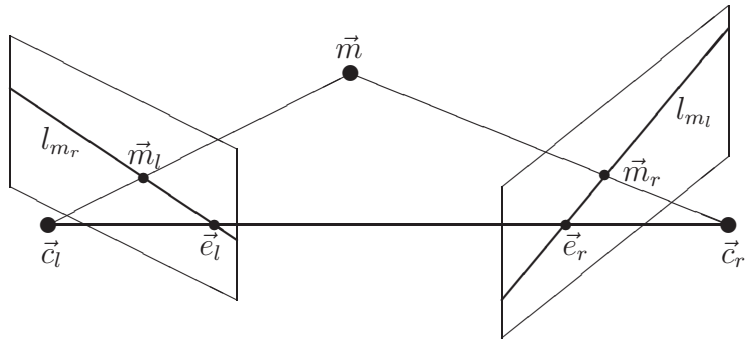


Figure 3.3: Epipolar Geometry.

point $\vec{m}$. The image points, $\vec{m}_l$ and $\vec{m}_r$, must lie on the *epipolar lines* l_{m_l} and l_{m_r}, respectively. These epipolar lines are defined by the intersection of the epipolar plane with the image planes of the cameras and must therefore, by definition, go through the epipoles.

An algorithm for computing the 3D structure of a scene from a pair of perspective projections, where the spatial relationship between the two views is unknown, was first presented by Longuet-Higgins [Longuet-Higgins 81]. He showed that if a scene contains at least eight corresponding points in the images from both views, the relative orientation of the two projections and the structure of the scene can be computed by solving a set of simultaneous linear equations based on the eight sets of image coordinates. This only accounts for extrinsic camera parameters, i.e. rotation and translation (see Section 3.1.3 for an explanation of camera parameters). The relationship between corresponding image points in the two camera images is described in the *Essential matrix* $\mathbf{E}$ — a 3×3 matrix — and satisfies

$$\vec{m}_r^T \, \mathbf{E} \, \vec{m}_l = 0 \quad . \tag{3.1}$$

Luong and Faugeras [Luong 96] generalised Longuet-Higgins' algorithm to also include intrinsic camera parameters (see Section 3.1.3). The relationship between corresponding image points is expressed in the 3×3 *Fundamental matrix* $\mathbf{F}$, which can be computed from coordinates of corresponding points in uncalibrated images, see [Luong 93, Luong 96] for details. The Fundamental matrix satisfies

$$\vec{m}_r^T \, \mathbf{F} \, \vec{m}_l = 0 \quad . \tag{3.2}$$

Let us denote the (2D) image point of an object point $\vec{m}_i$ in the left and right image planes respectively by

$$\vec{m}_i^l = \begin{pmatrix} x_i^l \\ y_i^l \\ z_i^l \end{pmatrix} \quad \text{and} \quad \vec{m}_i^r = \begin{pmatrix} x_i^r \\ y_i^r \\ z_i^r \end{pmatrix} \tag{3.3}$$

with the z_i element representing the distance of the image plane from the camera centre. It is generally more convenient to use homogeneous coordinates, which can be established by dividing the vector elements by the element in the third row

$$u_i = \frac{x_i^l}{z_i^l}, \quad v_i = \frac{y_i^l}{z_i^l}, \quad r_i = \frac{x_i^r}{z_i^r}, \quad s_i = \frac{y_i^r}{z_i^r}, \tag{3.4}$$

$$\vec{m}_i^l = \begin{pmatrix} u_i \\ v_i \\ 1 \end{pmatrix} \quad \text{and} \quad \vec{m}_i^r = \begin{pmatrix} r_i \\ s_i \\ 1 \end{pmatrix} \quad . \tag{3.5}$$

(Almost every textbook on computer graphics or computer vision will discuss the use of homogeneous coordinates in detail, for example, consult [Foley 96].)

The *perspective transformation matrix* (or *camera calibration matrix*) defines the transformation from image coordinates to camera world coordinates. It is determined during camera calibration as described in Section 3.1.3. If matrices for both cameras are known, $\vec{m}_i^l$ and $\vec{m}_i^r$ can be transformed into vectors in camera world coordinates. Not considering non-linear camera effects, this matrix represents a rotation as well as a translation. If the centre of projection coincides with the camera centre (and origin of each camera's world coordinate system), then the translational component equals 0.

The resulting vectors $\vec{p}_i$ and $\vec{q}_i$ represent directions from the camera centres, through the respective point on the image plane, to the object point in the scene

$$\vec{p}_i = R_y(\vartheta_l)\ R_z(\phi_l)\ R_x(\gamma_l)\ f_l\ \vec{m}_i^l \tag{3.6}$$

$$\vec{q}_i = R_y(\vartheta_r)\ R_z(\phi_r)\ R_x(\gamma_r)\ f_r\ \vec{m}_i^r \tag{3.7}$$

The scalars f_l and f_r are the focal lengths of the left and right cameras, respectively. R_x, R_y, and R_z are rotations around the x, y, and z axes, respectively. As mentioned

in the previous subsection, the cameras in this project were mounted on a rig with a baseplate in the xz plane which allows one to verge the cameras around the y axis, but limits rotation around the other two axes. In this study, the angles were $\vartheta_l \approx -5°$, $\vartheta_r \approx 5°$, and $\phi_l \approx \phi_r \approx \gamma_l \approx \gamma_r \approx 0$.

Under ideal conditions, the vectors $\vec{p}_i$ and $\vec{q}_i$ intersect at the 3D point $\vec{m}_i = (x, y, z)^T$. However, since $\vec{p}_i$ and $\vec{q}_i$ are likely to be corrupted by noise (lens distortion, point correspondence), $\vec{m}_i$ is determined by minimising the error term

$$E_i = \|\vec{p}_i\, s_i \;+\; \vec{c}_l \;-\; \vec{m}_i\|^2 \;+\; \|\vec{q}_i\, t_i \;+\; \vec{c}_r \;-\; \vec{m}_i\|^2 \tag{3.8}$$

with respect to the three coordinates of $\vec{m}_i$ and the two scalars s_i and t_i. If stereo matching fails, i.e. if the image points $\vec{m}_i$ are found incorrectly, minimising the error term E_i will not determine the coordinates of $\vec{m}_i$ correctly. Finding matching image points — solving the 'correspondence problem' — is therefore of great importance.

The final step in 3D reconstruction is the stereo triangulation, which leads to the coordinates of $\vec{m}_i$ in stereo world coordinates. If the orientation and distance of each camera to the origin of the stereo world coordinate system is known, then, together with the perspective transformations of each camera, the relative orientation of the two cameras to each other and the stereo world coordinates of a point viewed in both camera images can be calculated.

Setting the partial derivatives of E_i to zero gives the following solution for $\vec{m}_i$

$$P_i \;=\; \frac{\vec{p}_i \cdot \vec{p}_i^{\,T}}{\|\vec{p}_i\|^2} \;-\; \mathbf{I} \qquad Q_i \;=\; \frac{\vec{q}_i \cdot \vec{q}_i^{\,T}}{\|\vec{q}_i\|^2} \;-\; \mathbf{I} \tag{3.9}$$

$$(P_i \;+\; Q_i)\, \vec{m}_i \;=\; P_i\, \vec{c}_l + Q_i\, \vec{c}_r \quad . \tag{3.10}$$

Inverting the matrix coefficient $(P_i + Q_i)$ yields the three coordinates of $\vec{m}_i$.

3.1.3 Camera Calibration

Camera calibration is the process of relating the camera's image (pixel) coordinates to the world coordinates. The relationship between the coordinate systems is described in the perspective transformation matrix. In the most general case, neither the intrinsic nor the extrinsic camera parameters are known. Intrinsic parameters define the perspective transformation from 3D object coordinates in the

camera world coordinate system to the 2D camera image coordinate system. These parameters are

- f: focal length (or distance from image plane to centre of projection),
- κ_1, κ_2: lens distortion coefficients for both directions in image plane,
- s_x: uncertainty scale factor due to camera scanning and acquisition timing error,
- (u_O, v_O): coordinates of origin of image coordinate system in image plane.

Extrinsic parameters define the transformation from the 3D object world coordinate system to the 3D camera world coordinate system. In detail, these parameters are

- γ, ϑ, ϕ: rotation angles,
- $T = (t_x, t_y, t_z)^T$: elements of the translation vector.

Tsai [Tsai 86] developed a camera calibration technique, for both a single camera system as well as stereo camera systems, that takes all of these 12 camera parameters into account. It is common to not calibrate the camera(s) for some parameters to simplify (and speed up) the calibration process by reducing the number of corresponding image points required. For example, if a perfect linear perspective transformation and no lens distortion are assumed, then the intrinsic parameters κ_1 and κ_2 can be omitted. Faugeras and Toscani [Faugeras 86] presented another approach to the calibration problem in stereo camera systems that assumes such a perfect perspective transformation.

Newman's Two-Step Process

However, the approach taken by Newman [Newman 99a] in the RSL face tracking system is slightly different in that each camera is calibrated separately but using the same algorithm. As mentioned in the previous paragraph, a linear perspective transformation is assumed and non-linear camera effects (lens distortion) are not considered. Camera calibration is achieved in a 2-step process

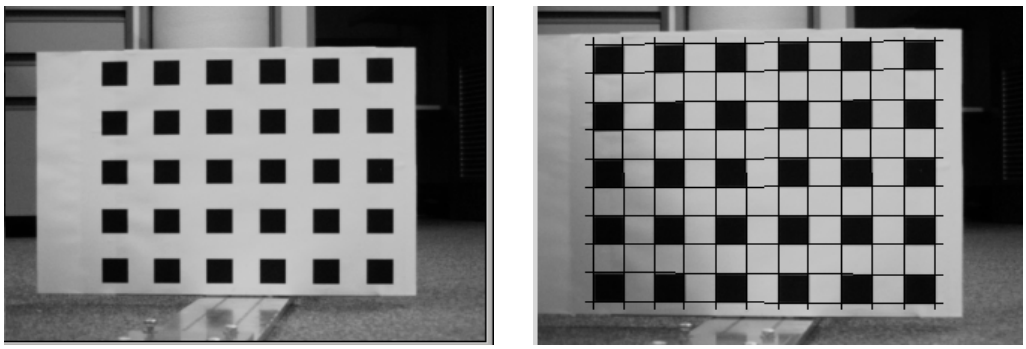


Figure 3.4: The calibration pattern: normal (left) and after edge detection (right).

1. Define a set of known 3D points in the scene and determine their image coordinates in the image plane.

2. Determine the perspective transformation matrix which maps the 3D object points onto their 2D image points.

In the first step, the stereo camera rig is placed on one end of an exactly measured calibration rig. The cameras observe an object plane parallel to the plane defined by the x and y coordinate axes. It features a rectangular 5×6 grid of 30 black rectangles on a white background similar to the grid used by Tsai [Tsai 86] (Figure 3.4). The object plane is placed at various distances from the stereo camera rig, which are known exactly from the process of manufacturing the calibration rig.

The four corners of each rectangle are semi-automatically (the user has to click the mouse pointer on the corner rectangles to start the process) detected using edge detection in snapshots from both cameras. This procedure is repeated for all five positions (650–850mm) in which the object plane is placed. In total, this gives 30 rectangles × 4 corners × 5 positions = 600 corresponding image points for determining the 10 intrinsic and extrinsic camera parameters. The procedure takes only a few minutes and can be done offline, before using the stereo camera system for face tracking or any other application.

In the second step, a minimisation procedure is usually necessary because of errors introduced by image noise and incorrectly located corresponding image points. As described in Section 3.1.2, the error between the measured and predicted 2D positions is minimised. Many papers in the literature describe general non-linear

minimisation techniques (see [Ganapathy 84, Tsai 86] for good overviews). Instead, a direct method, proposed in [Ganapathy 84] and described in [Trucco 98], is used here, because it is more accurate. Here, the perspective transformation matrix is determined by finding a matrix A such that for all i

$$A = \begin{pmatrix} a_1^T \\ a_2^T \\ a_3^T \end{pmatrix} \qquad u_i = \frac{a_1^T \cdot \vec{m}_i}{a_3^T \cdot \vec{m}_i} \qquad v_i = \frac{a_2^T \cdot \vec{m}_i}{a_3^T \cdot \vec{m}_i} \tag{3.11}$$

where $(u_i,\, v_i)$ are the image coordinates of the i^{th} calibration point with world coordinates $\vec{m}_i = (x_i,\, y_i,\, z_i)^T$. Errors in the measured image points $(u_i,\, v_i)$ make it practically impossible to satisfy these equations exactly, so A is found by minimising

$$E \;=\; \sum_i \left((a_3^T \cdot \vec{m}_i)\, u_i - a_1^T \cdot \vec{m}_i \right)^2 \;+\; \left((a_3^T \cdot \vec{m}_i)\, v_i - a_2^T \cdot \vec{m}_i \right)^2 \quad . \tag{3.12}$$

The camera parameters, and hence the transformation matrix, can then be extracted from the elements of A. However, because of noise, the rotational parameters will not necessarily form an exact rotation matrix. Choosing the closest rotation matrix may not minimise the error E any more. Instead of employing an iterative non-linear minimisation procedure, a more precise algorithm was developed by Newman *et al.* [Newman 00]. It can be shown that along each ordinate representing a rotation angle, the differential of E is quartic in that angle's cosine. Similarly, E is quadratic along the ordinates of all other camera parameters. Since closed form solutions of any quadratic, as well as the roots of any quartic, can be obtained, E can be minimised precisely. A small number of iterations (≈ 5) is sufficient to find the minimum. The resulting calibration is accurate to within 1mm over the range ($x = \pm 200$mm, $y = \pm 200$mm, $z = 600 \pm 300$mm) [Newman 99b].

3.1.4 Discussion of Error Sources in Camera Calibration

In the previous subsections, two main error sources were identified. Firstly, camera lenses can show some radial and tangential distortions, which can be accounted for by the lens distortion coefficients κ_1 and κ_2. As a result of such lens distortions, the epipolar constraint may not hold. However, a perfect linear transformation is often

assumed and the effects of lens distortions are neglected, because the effects are considered small and omitting the determination of κ_1 and κ_2 simplifies the calibration process. Secondly, errors occur during the determination of corresponding points in the stereo images. This can be due to inaccuracies in an automatic determination (depending on the method(s) used), incorrectly manually chosen points, image quantisation, the delay between left and right images in the stereo vision system used in this project, and the fact that the cameras view the scene from different angles. The last point presents no problem for salient image features (the corner of a cube, for example), but may lead to incorrectly chosen correspondences for points on smooth surfaces. By using a specific calibration pattern of exactly known dimensions, as was done in this study, the problems of finding corresponding points in the stereo images can be avoided or at least reduced to a negligible level. For the remainder of this thesis, correctly calibrated cameras were assumed.

3.1.5 The Tracking Procedure

Before the actual face tracking can start, the RSL system requires the creation of a face model, which is stored in a 'face model' file. This model has three components

1. an edge map of the entire face,

2. template images of the facial features, and

3. 3D coordinates of the facial features.

The edge map of the whole face is used in the searching mode to find the approximate position of the face in the stereo images. The facial features to be used for tracking are identified interactively by the user during the face model creation. Up to 32 features in total can be selected, but usually about 15–20 features are sufficient for accurate face tracking. In principle, these facial features can be any landmarks on the face but distinctive regions, for example the eye corners, the eyebrow ends, the nose region, the ears, and moles, give the best results. In addition, the system requires the user to select the corners of the eyes as well as the mouth corners.

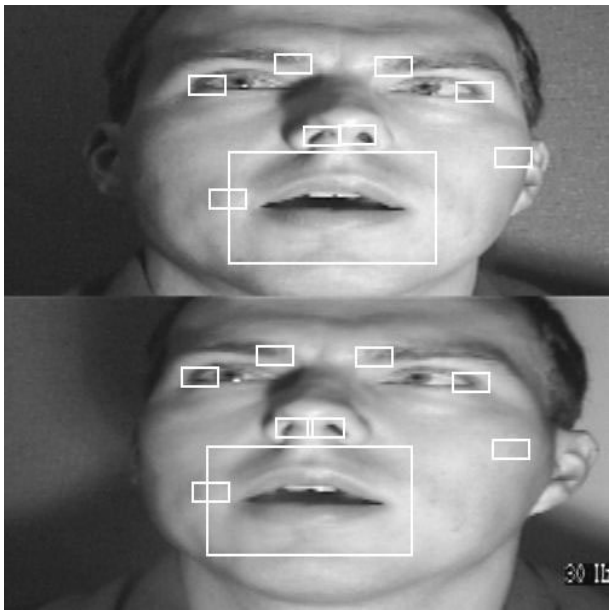


Figure 3.5: Small rectangles: Feature templates selected for face tracking. Large rectangles: Automatically selected mouth region for lip tracking.

These are the face model points used in determining the head pose. They also define the area of the image for the whole face edge map.

An image of the entire face is taken in full frontal position, as well as for views of the face on a 45° angle to the left and right, for feature selection and subsequent extraction of template images. By selecting the position of each feature in both stereo images, their 3D coordinates can be computed as described in Section 3.1.2. Figure 3.5 shows an example of a frontal view with features selected for tracking.

The face tracking system is always in one of two modes. In the searching mode, the algorithm needs to find the face in the video stream. This is obviously necessary at the beginning of the face tracking, but also every time the face is 'lost', due to the face not being in the images or because of a tracking failure. In the tracking mode, the algorithm performs the actual tracking of the selected facial features.

Searching Mode

The searching procedure firstly obtains a multiplexed stereo image and then splits it into left and right views. Both images are smoothed and subsampled to reduce them to quarter size. Next, the edge map of each image is computed using standard edge

detection methods. Using the face model edge map, the left image is searched firstly for instances of the area surrounding the bridge of the nose. Then, a second search region located below the best result of the first search is defined and searched for instances of the region around the mouth. This 2-stage search procedure introduces additional scale and rotation invariance.

In the next step, templates of the nose bridge and mouth regions are extracted from the best matches found in the left image. The search is then repeated with these new templates in the right image. Template matching is done using hardware-implemented normalised cross-correlation (NCC) in the Hitachi IP5005 video card. NCC matching is more robust to changes in illumination than other methods. If the combined certainty of the NCC matches exceeds the search threshold, the face is considered to be found. Initial estimates of the rotation and translation parameters of the head pose are calculated and the proper face tracking process started.

Tracking Mode

Based on the head pose from the previous video frame, an estimate of the pose in the current frame is predicted from Kalman filters. The head is considered to be a rigid object for this purpose, so that only rotation and translation parameters of the current head pose need to be estimated. For each feature, 2D search regions are computed for both left and right view using the uncertainty of the predicted pose estimate to determine the size. In other words, search regions can be smaller, when the head pose estimate is good and they need to be larger, when the estimate is not so good. Of course, the size of the search region has an immediate impact on the speed of processing. Next, the best match of each feature template from the face model is found in the individual search regions of the left image. Then, a template is extracted from each position of the best match in the left image and used to find the best match in the corresponding search regions of the right image.

Once the best matches have been found in both images, the 3D coordinates of each feature are computed using stereo reconstruction (see Section 3.1.2). Finally, the optimal head pose is determined by finding the rotation and translation param-

eters (R, T) that map the face model points to the observed 3D coordinates. This mapping is a mapping from the average of the model coordinates — the 'model centre point' $\vec{m}_O$ — to the weighted average of the 3D feature coordinates $\vec{x}_c$. The uncertainty of each template correlation is used as the weight w_i in computing $\vec{x}_c$

$$\vec{x}_c \;=\; \frac{\sum_{i=0}^{n} w_i \, \vec{x}_i}{n} \tag{3.13}$$

where n is the number of observed points $\vec{x}_i$. Because of the noise inherent in any observation of the transformed model, the transformation parameters (R, T) are found in a least squares minimisation of the error term E with respect to R and T

$$\min E \;=\; \|\vec{x}_c - R\,\vec{m}_O - T\|^2 \quad . \tag{3.14}$$

Well-tracked features have high correlation values. The optimal head pose found is, thus, biased towards these points, which makes the system more robust to occlusions, perspective distortions, and image noise. If the overall tracking certainty is above the tracking threshold, the Kalman filters that predict the head pose and template search regions in the next frame are updated with the pose parameters, and tracking continues. Otherwise, the face searching mode is started again.

The version of the RSL face tracking system used in this study is implemented in the Java programming language. Depending on the number of features to be tracked, face tracking was typically performed at 15–20Hz. Fewer features increased the frame rate up to 30Hz, but led to inaccuracies in the head pose estimate. Accurate real-time face tracking often requires a trade-off between speed and accuracy. However, no such trade-off was necessary in this study, because offline processing was available and the statistical analyses were done on a pre-recorded data corpus.

3.2 A Model of Lip Movements in 3D

In order to understand the lip parameters measured in the experiments, it is helpful to define a lip model. A 3D model is more realistic and more accurate than a 2D projection. Given that the stereo vision face tracking system enables the measurement of the 3D coordinates of object points, such a 3D lip model is desirable and was, therefore, used in this study.

King *et al.* [King 00] presented an anatomically-based 3D parametric lip model for synchronised speech and facial animation. Its parameter set corresponds to 21 muscles around the mouth that control the movement of the lips. These 21 parameters are mapped to a B-Spline surface. While this lip model is very realistic, it is, in many aspects, too complex and more importantly, it is difficult to measure muscle movements without using artificial markers or electromyographic equipment.

The 3D lip model used in our research was inspired by previous work on a lip model for speech synthesis at the Institut de la Communication Parlée (ICP) in Grenoble (France), mainly by Adjoudani [Adjoudani 93] and Guiard-Marigny [Guiard-Marigny 94, Guiard-Marigny 97, Guiard-Marigny 96]. The ICP lip model defines the inner and outer lip contours in 3D as a set of 12 curve segments, six for the inner lip contour and six for the outer lip contour. Left and right half of the lips are considered symmetrical, so that only six curve segments need to be calculated. The parameters of the curve segments are controlled through ten polynomial equations. An iterative process allows the prediction of the coefficients of these equations from five control parameters, shown in Figure 3.6. They are

- the internal width w of the mouth opening,
- the internal height h of the mouth opening,
- the protrusion c of the lip contact point,
- the protrusion u of the midpoint of the upper lip, and
- the protrusion l of the midpoint of the lower lip.

The protrusion parameters are measured as distances to an imaginary vertical plane — assuming the face to be in an upright position — behind the mouth. Mathematically, these distances are equivalent to the length of the normal vector of that plane to the respective lip contour points. Guiard-Marigny placed a mirror next to the speaker's face at an angle of 45°, so that both a frontal and a side view of the face could be taken by a single camera concurrently. A vertical line marker attached to the face served as an indication of the position of the vertical plane.

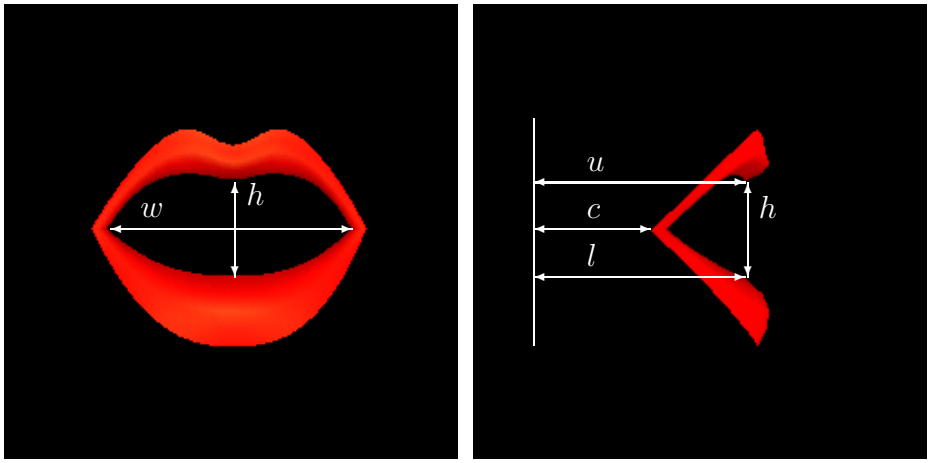


Figure 3.6: Frontal (left) and side (right) view of the ICP 3D lip model.

The lip model in our research followed the same basic principle. Measuring the 3D coordinates of certain feature points on the inner lip contour leads to a variety of parameters describing the shape of the lips in 3D. From just 4 feature points — the lip corners as well as the midpoints of upper and lower lip — 3D measures such as mouth width, mouth height, and lip protrusion can easily be determined. The inner lip contour was preferred over the outer lip contour for a number of reasons. Firstly, people differ in the generic shape of their lips.[3] Some people have thicker lips than others, some have stronger protrusion (in the rest state) than others. Extracting the outer lip contour would mean that such personal characteristics influence the measurements, while the inner lip contour can truly be considered as the final boundary of the vocal tract. Hence, inner lip contour measurements are, in my opinion, better suited for the investigation of relationships between audio and video speech parameters. Secondly, the difference between lip colour and the surrounding facial skin can be quite small. Many lip tracking methods described in the literature review have difficulty in coping with this lack of contrast, if employed on tracking the outer lip contour. Given the different appearance of the oral cavity, the inner lip contour does not suffer from these problems.

[3] In this context, lip shape means the 'basic structure' of the lip area between inner and outer lip contour.

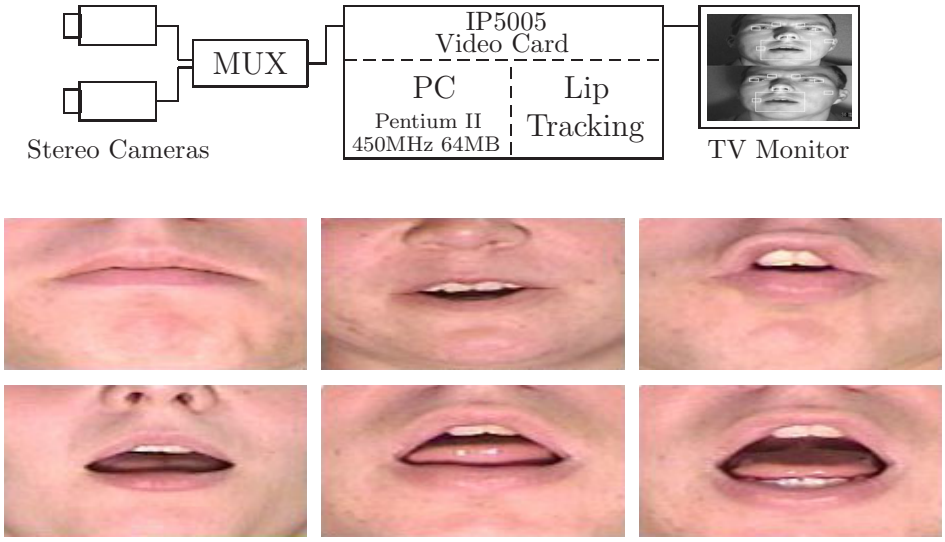


Figure 3.7: Top: Outline of the combined stereo vision face and lip tracking system. Bottom: Different degrees of mouth openness as well as teeth and tongue visibility.

3.3 Lip Tracking in 3D

3.3.1 Overview

The requirements of a lip tracking algorithm generally depend on the application. In the case of AV speech processing, an algorithm that is both fast and accurate is needed. Lip movements during speech production can be very quick and changes in mouth shape (mouth closed, mouth partially open, mouth wide open, lips rounded, lips spread etc.) can take place in a time span as short as 10ms. This highlights the need for a real-time algorithm which tracks the lip movements continuously. At the same time, accuracy is of great importance for the results of the statistical analyses described in Chapter 6 to have any meaningful value. It is particularly important to distinguish apparent distortions in mouth shape due to head pose (rotation) from speech production-related mouth deformations (see Section 2.3.3). Furthermore, the algorithm must be able to cope with different personally characteristic lip shapes as well as various mouth shapes ranging from a completely closed mouth to a fully open mouth in which upper and lower teeth as well as the tongue may or may not be visible (Figure 3.7 bottom). Finally, a lip tracking algorithm must take the level of illumination in the lower face half into account.

As discussed in Section 2.3.3, lip tracking can be either implicit or explicit. Implicit lip tracking analyses the statistical behaviour of feature vectors representing the pixels of the mouth area. Explicit lip tracking, on the other hand, attempts to fit a 2D or 3D lip model to the observations by locating facial feature points that define the model. Such an explicit approach was followed in this work because it was expected to facilitate the interpretation of the analysis of the AV relationships.

Extracting the Mouth Region

The newly developed lip tracking algorithm builds on the real-time stereo vision face tracking system described in Section 3.1 (Figure 3.7). It assumes that the face has been located in the video stream from the two cameras and that this face is being tracked. The face tracking system estimates the head pose and from it, the locations of eye and mouth corners according to the 3D face model. The face model considers the position of the mouth corners to be static which is, of course, not the case during speech production. It is merely an approximate estimate of the mouth's position in the images. However, by applying some heuristics, a suitably sized part in each camera image containing the mouth area can be chosen automatically. The heuristics used to determine the size of this *mouth window* in pixels are

$$X = L_x - 30 \ldots R_x + 30 \tag{3.15}$$

$$Y = L_y - 40 \ldots L_y + 40 \quad (\text{or } R_y \pm 40) \tag{3.16}$$

where (L_x, L_y) and (R_x, R_y) are the estimated image coordinates of the lip corners.

The mouth window is a rectangular part of the image containing the lips, oral cavity, and some of the surrounding facial skin as shown in Figure 3.8. To account for differences in the cameras, the image data in the mouth windows are mean-normalised. The orientation of the mouth will obviously follow the general head pose. In a general lip tracking algorithm, the angle of rotation around the z axis (deviation from the horizontal xy plane) ϕ should be considered when the mouth window is extracted. An image processing technique such as warping can then be applied to the mouth window to realign the mouth horizontally. If the speaker's head is approximately in upright position ($\pm 20°$), then the influence of ϕ can be

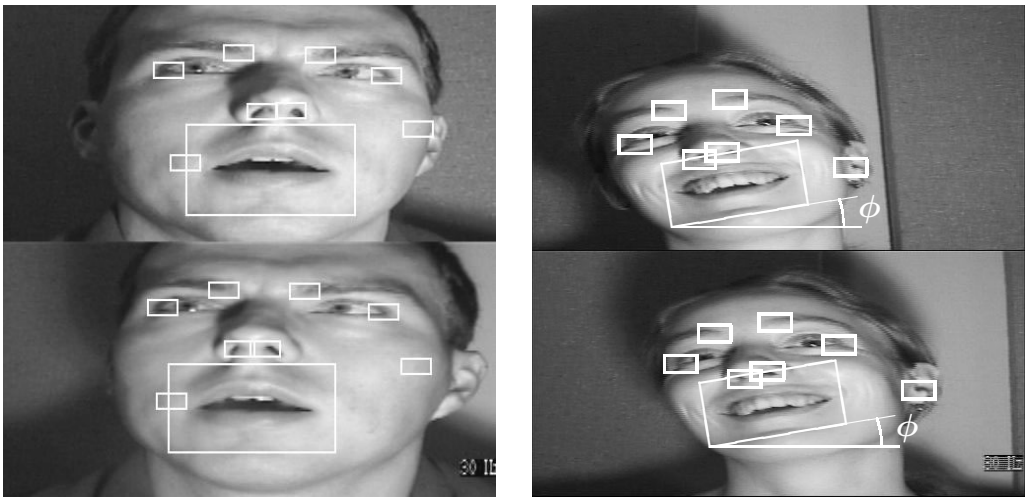


Figure 3.8: Extracting the mouth region: Large rectangles enclose automatically selected mouth windows.

neglected and the axes of the mouth window rectangle can be aligned with the image coordinate axes, as was done in the left picture in Figure 3.8. Lip tracking is performed on the mouth windows rather than the entire image, which reduces the amount of processing that is required and limits the error due to a lip tracking failure. The lip tracking algorithm is too complex to be done on the IP5005 video card. The mouth windows are therefore uploaded into the PC's main memory and processed there, although doing so has a negative impact on the frame rate achieved by the system, because of the limited bus speed of the video card.

Solving the Point Correspondence Problem

In order to take advantage of having a stereo vision system and thereby being able to make measurements in 3D, it is necessary to solve the point correspondence problem. If corresponding points in the two camera views are identified incorrectly, incorrect 3D coordinates of the point and subsequently erroneous lip parameters will be computed. It is therefore of utmost importance for accurate results to find the 2D image coordinates of corresponding points reliably.

Applying the same technique — static template matching — used in the face tracking system might initially appear as a logical choice. However, static template matching requires that the object, from which the template images were taken in

the first place, fulfills the requirement of being a rigid object. Although this is a reasonable assumption for many parts of the face, the assumption does not hold for the mouth area during speech production, when rapid movements and changes in the shape of the lips and surrounding facial skin are common. A way to deal with this situation is to use adaptable (or dynamic) template matching, in which the template image is updated regularly. Such an approach was described by Loy *et al.* in [Loy 00a]. There, the updated template $T_i[k]$ for the k^{th} video frame was a weighted average of the initial template $T_i[0]$ and the best match of the previous template image in frame $k-1$

$$T_i[k] \;=\; \beta\, T_i[0] \;+\; (1-\beta)\, T_i[k-1], \quad \beta \in [0,1]. \tag{3.17}$$

The weighting factor β determined the contribution of the initial template image. 'Grounding' the template image was necessary, because fully updated templates had the tendency to 'drift' over the image after some time, due to the quantisation error and possible mismatches [Loy 00b]. A different approach was used in the work described in this thesis which does not use templates for the lip tracking and, thus, avoids the mentioned problems. Instead, it uses image processing techniques based on a combination of colour and structural information (see next subsection).

Generally, the amount of processing required to find corresponding points in stereo images can be reduced drastically by taking advantage of epipolar geometry, as described in Section 3.1.2 and depicted in Figure 3.3. There, it was shown that the corresponding right camera image point $\vec{m}_r$ of left camera image point $\vec{m}_l$ has to lie on its epipolar line l_{m_l}. Hence, image points not lying on the epipolar line can be discarded in the search. However, the difficulty lies in determining the epipolar geometry correctly [Luong 93, Luong 96].

Combining Colour and Structure

It was, therefore, decided to take a different approach in this study [Goecke 00a]. The lip tracking algorithm was designed to find the coordinates of the four lip feature points necessary to define the parameter set of our lip model, i.e. the two lip corners plus the midpoints of upper and lower lip. The algorithm combines colour

information from the images with knowledge about the structure of the mouth area. Colour information is a powerful cue in facial feature detection. However, the YUV (also known as YIQ) signal from the NTSC cameras alone is of little use, because the image signal is encoded into an intensity (Y) signal and two colour difference signals (U, V). The YUV signal can be transformed into a standard computer RGB signal. However, images in the RGB colour space are affected by changes in illumination, i.e. the RGB colour space is intensity-sensitive, which can be a real problem for image processing or object recognition. Using intensity-normalised RGB values can overcome this problem, but such colour information does not separate lip and surrounding skin well, as was discussed earlier in this chapter and Section 2.3.4, and thus does not assist lip feature point extraction. A better choice is the HSV colour space which separates hue (H) and saturation (S) from intensity (V) [Foley 96]. Colour space transformation takes place when the mouth windows are uploaded from the IP5005 video card into main memory.

Our original idea of using the hue value to separate the lips and surrounding skin flesh from the oral cavity did not work, because the hue of the oral cavity is still very close to the hue ('red') of other face parts, although a human observer hardly perceives it that way. However, there is a clear difference in the saturation values between skin/lips, teeth, and oral cavity. The dark oral cavity exhibits the largest saturation values and the teeth the smallest, while the skin values lie between these two extremes. A combination of intensity (Y) and saturation (S) values is, therefore, used throughout the algorithm.

The lip tracking algorithm is a three-stage process (Figure 3.9) outlined in the following paragraphs and described in detail in Sections 3.3.2 – 3.3.6. The first step determines the general degree of mouth openness. As discussed previously, the lip tracking algorithm must be able to handle mouth shapes ranging from a completely closed mouth to a wide open mouth. Using image processing techniques for lip tracking, as was done here, no single technique would give good results for all possible mouth shapes. However, by pre-classifying mouth shapes into one of three categories based on mouth openness, specific techniques individually targeted at each category can then be applied to give better results. These categories are

- closed mouth,
- partially open mouth, and
- wide open mouth.

In the second step, the lip corners are found. Here, the *a priori* knowledge about the structure of the mouth area becomes useful. For example, if the mouth is closed, teeth will not be visible, so the shadow line between upper and lower lip is the outstanding feature. Various definitions of what constitutes the inner lip contour of a closed mouth are possible. In this study, the shadow line between the lips was considered to be part of the inner lip contour. Therefore, the algorithm looks for this line. When the mouth is open, it is very likely that either or both the upper and lower teeth are visible, so the algorithm looks for them as well as for the oral cavity. By tailoring the algorithm in this way to fit a particular situation, more accurate results can be obtained than from a general-purpose, 'one-size-fits-all' algorithm. The first and second steps are applied separately to both the left and right camera images. Once the 2D image positions of the lip corners in both views are known, their 3D positions can be calculated. This result is then used in the third and final step, in which the positions of the lip midpoints are determined.

The face tracking and lip tracking systems together achieved a real-time frame rate of 5–8Hz, depending on the processing required for finding the desired lip feature points. Such frame rates are sufficient for tracking the basic mouth shapes during speech production. However, if the detection of quick changes is needed, for example for a detailed statistical analysis of lip movements, the tracking needs to be done at a higher frame rate. Higher frame rates could generally be realised by porting the system from Java to C/C++ or by using a fully digital video system which offers higher frame rates. Alternatively, offline processing of recorded sequences, as was done in the work described in this thesis, avoids the limitations set by the hardware and software, and is then only limited by the properties of the recording equipment. In this study, the limiting factor was the NTSC frame rate of 30Hz determined by the analogue cameras used. That means that one video frame was taken every 33ms, which captured a lot of detail of lip movements, but information

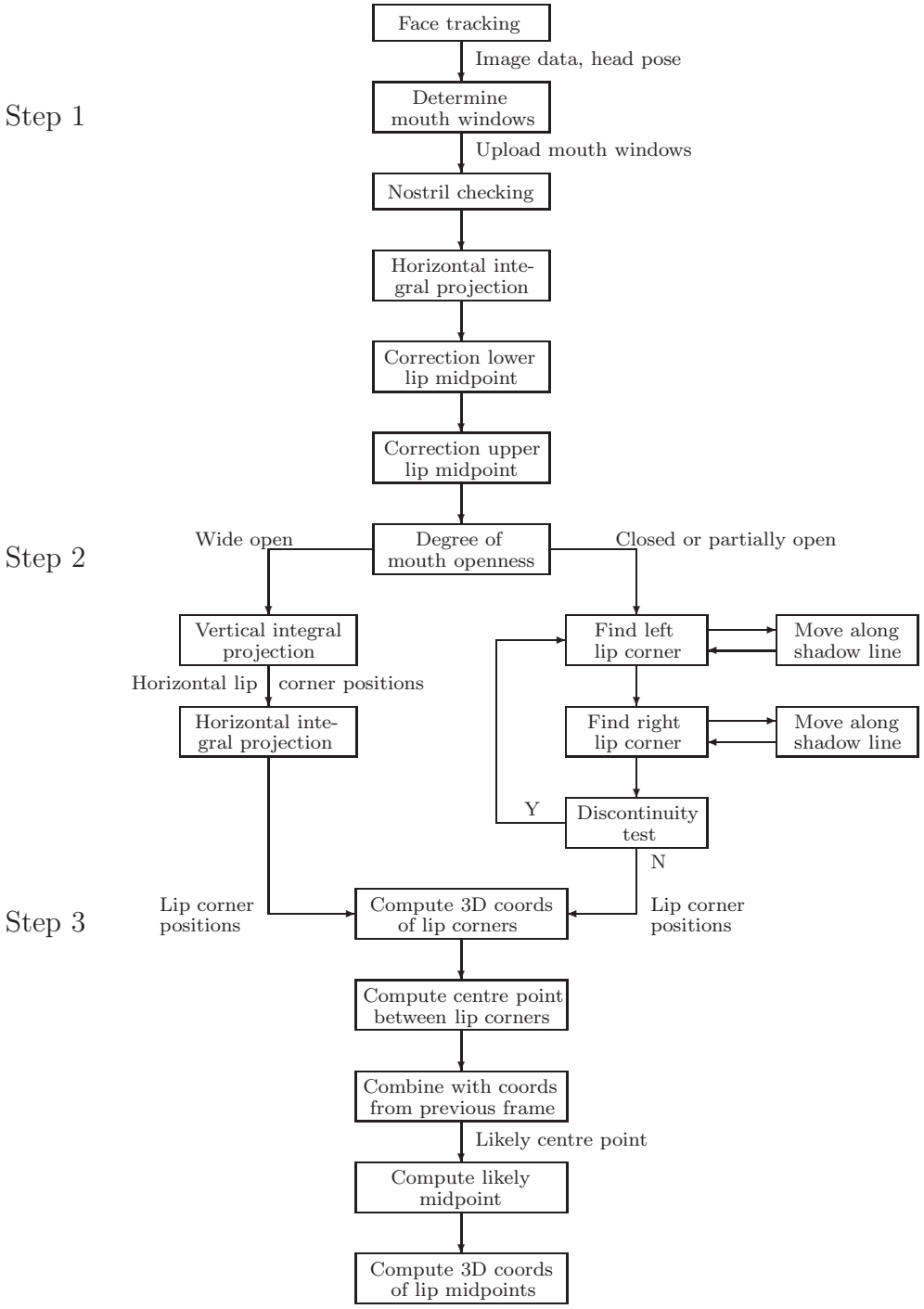


Figure 3.9: Lip tracking algorithm.

about faster lip motion was lost. It is suggested that new studies following this current one should use video equipment which offers higher frame rates.

3.3.2 Algorithm Techniques

Various techniques are used several times in the lip tracking algorithm and are, therefore, discussed here. Firstly, there is *dynamic thresholding*. Whenever a threshold operation takes place, the threshold is determined dynamically at that time, instead of using hard-coded threshold values, to improve robustness. That is, the starting value of the threshold is chosen to be overly conservative, so that no pixel value in the area of interest will pass it. The threshold is then iteratively changed until the value has been found, at which pixel values start passing the threshold. The algorithm then continues to use this threshold value for the rest of the processing of that frame. In this way, the algorithm adapts itself to changes in illumination and different skin tones.

Secondly, whenever a particular pixel position is tested, not only the pixel value of that position is checked, but also of up to n (empirically set to 9) other pixels in the neighbourhood (but some pixel positions away). Selection is done by pixel masks like this one

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & X & 0 & 1 & 0 & 1 \end{bmatrix} \tag{3.18}$$

where X indicates the current pixel position, a 1 relates to other checked pixel positions and a 0 to pixel positions ignored in the test. *Voting* takes place for each such pixel X, and only when a certain number of positive votes is reached, does it indicate that a threshold has been passed and that position is accepted as being correct. Voting turns weak cues into strong ones. It can be either 'hard' or 'soft'. The former requires a majority of votes, for example two thirds, and is used for finding lip contours. The latter only requires a few positive votes. It indicates the presence of a particular feature. Soft voting is only used for detecting visible teeth.

Thirdly, two modules which are used numerous times in the algorithm, test for the *shadow line* between the lips and for the *visibility of teeth* in the image data. The shadow line is detected by the typical high saturation and low intensity values. Teeth can be distinguished from other parts of the face by their characteristic low saturation and high intensity values. However, since some skin parts can show similar values, the teeth check must also pass an edge detection test, which looks

for the horizontal edge between the lip and the teeth. Edge detection is usually done by convolving the image with a filter [Pratt 78, Russ 95]. In our algorithm, a 3×3 vertical gradient filter is applied

$$K = \begin{bmatrix} +1 & +1 & +1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix} \quad . \tag{3.19}$$

Compared to other edge detection filters (e.g. Sobel, Kirsch, Canny, Laplacian [Pratt 78, Russ 95]), it requires very few computations and therefore allows fast processing, while the results are sufficient for our purpose.

Finally, *integral projection* is a technique, in which all the pixel values of one row or column of pixels are summed. This technique can be extremely effective in locating facial features, if the rectangular image part, on which it is performed, is chosen suitably, as was shown in Kanade's pioneering work [Kanade 73]. In his work, he used integral projection on binary images, whereas Brunelli and Poggio later applied this technique to grey-level images [Brunelli 93]. Let $I(x, y)$ denote the pixel value, e.g. intensity or saturation, at the coordinates (x, y). Horizontal and vertical integral projection are then respectively defined as

$$H(y) = \sum_{x=x_1}^{x_2} I(x, y) \tag{3.20}$$

$$V(x) = \sum_{y=y_1}^{y_2} I(x, y) \tag{3.21}$$

where x_1, x_2, y_1, y_2 denote the boundary coordinates of the image part under investigation. Horizontal integral projection is useful for detecting vertical gradients and vertical integral projection similarly for horizontal gradients.

3.3.3 Step 1: Determine Mouth Openness

To determine the degree of mouth openness, the vertical positions of the lip midpoints must be determined. For the following calculations, the horizontal position of the lip midpoints is (temporarily) considered to be at the middle between the left and right boundaries of the mouth window. This estimate is close enough to

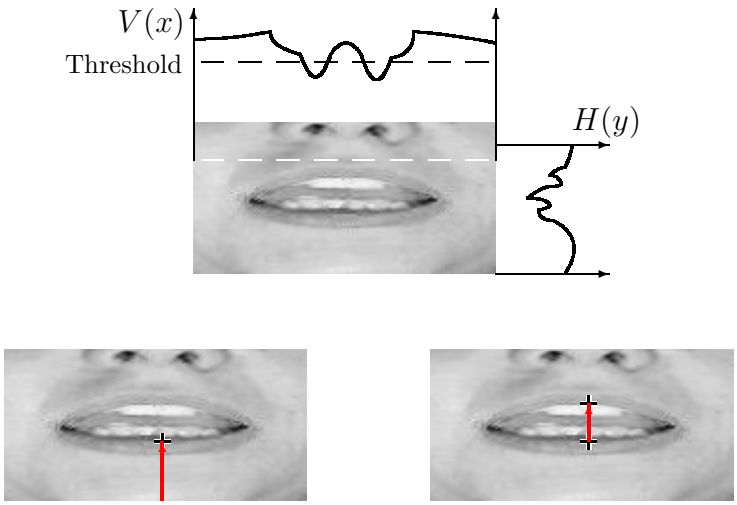


Figure 3.10: Step 1 - Top: Nostril detection by vertical integral projection in the top quarter of the mouth window. Horizontal integral projection to find vertical position of lip midpoints. Bottom: Possible correction of lower and upper lip midpoints.

the true position to start the algorithm for finding the lip corners (Step 2), but the position is recalculated (Step 3) after the lip corner positions are found.

Horizontal integral projection on the intensity values of the mouth window pixels is used to find a starting estimate of the vertical positions of the lip midpoints (Figure 3.10 top). These sometimes rough estimates need to be refined. If the mouth is closed, either the shadow line between the lips (correct) or the external lip contour (incorrect) is found. If the mouth is open, either or both the upper and lower teeth are visible and the horizontal integral projection detects either the edge between lip flesh and teeth (correct), or between teeth and oral cavity (incorrect), or in rare cases the outer lip contour (incorrect).

Let us first look at correcting the midpoint of the lower lip. To test if a correction is necessary, an imaginary vertical line from the lower boundary of the mouth window to the midpoint of the upper lip at the estimated horizontal position is followed upwards (Figure 3.10 bottom left). Necessarily, the lower lip midpoint cannot lie above the upper lip midpoint. While walking along the line, the algorithm checks for either the shadow line between the lips or for the appearance of teeth. If either is found and the position is different from the one obtained from the previous

horizontal integral projection, the position is updated. This vertical position might be just off the lip contour, so in a final step, the algorithm adjusts the vertical position of the lower lip midpoint to the lip pixel bordering the oral cavity.

Secondly, the vertical position of the midpoint of the upper lip is corrected, if necessary. The algorithm first tests for the appearance of teeth above the position of the upper lip midpoint estimated from the horizontal integral projection. If found, it means that the edge between oral cavity and teeth was detected and, hence, the position is moved upwards until no further teeth pixels are detected above the current position. Subsequently, a second test for teeth, this time below the current position, is performed. If the edge between teeth and upper lip was found correctly by either of the steps before, there are teeth pixels below the current position. However, if the outer lip contour was detected, no teeth pixels are found below. In that case, the algorithm starts just above the lower lip midpoint and moves upwards until the edge between teeth and upper lip is found (Figure 3.10 bottom right). Again, the vertical position of the upper lip midpoint is finally accurately placed on the lip pixel forming the edge to the oral cavity. If neither of these tests indicates any necessary changes in the lip midpoint positions, then the coordinates found in the horizontal integral projection step are retained.

Nostril Detection

The size of the mouth window is chosen to be sufficiently large to contain the mouth area under all circumstances. This means that the nostrils could sometimes be included in the mouth window (Figure 3.7). Since intensity values are used in the horizontal integral projection step, the nostrils' low intensity values potentially lead to incorrect results for the position of the lip midpoints. Therefore, a *nostril detection* algorithm was also developed. The top quarter of the mouth window is scanned for the minimum and maximum pixel column using vertical integral projection on the intensity values. A threshold to determine nostril candidates is set by

$$T \;=\; \mathrm{Min} \;+\; \frac{\mathrm{Max} - \mathrm{Min}}{3} \quad . \tag{3.22}$$

The horizontal position of nostril candidates is determined by testing the pixel column sums with this threshold. Values below the threshold are candidates, but they are only confirmed as horizontal nostril positions, if they extend at least 3 pixels wide horizontally. The lower edge of the nostril is then found using edge detection and the vertical position closest to the upper lip ('lowest' nostril position in the image) is taken as the new upper boundary of the mouth window used for determining the mouth openness described before.

3.3.4 Step 2: Find Lip Corners

So far, the vertical positions of the lip midpoints have been established at their temporary horizontal positions. The 2D distance between the lip midpoints defines the following steps in the algorithm. If the distance is less than 15 pixels, the mouth is either fully closed or only partially open. Otherwise, the mouth is considered to be wide open. The threshold of 15 pixels is an experimentally determined heuristic. It is equivalent to about 15mm in 3D space for an object at a distance of about 600mm. It is worthwhile to remember that the lip tracking algorithm deals with parts of camera images of half the original vertical resolution (see Section 3.1.1). So 15 pixels here correspond to 30 pixels in a full resolution video frame.

Mouth Closed or Partially Open

If the mouth is fully closed or only partially open, a vertical integral projection would not yield enough information to find the lip corners reliably. Thus, starting from the current position of the lower lip midpoint, a search along the shadow line to either side is performed through a cycle of tests (Figure 3.11 top left).

Let us consider the speaker's right lip corner, noting that the speaker's left lip corner is found similarly, except for moving in the opposite direction. Starting from the midpoint (x, y) of the lower lip, the algorithm moves left in image space. This line of pixels is marked (1) in Figure 3.11, bottom left. Testing the five pixel positions $(x-1, y), \ldots, (x-5, y)$ enables the algorithm to jump over pixels, where the shadow line is discontinuous due to image noise. If one of the pixels indicates

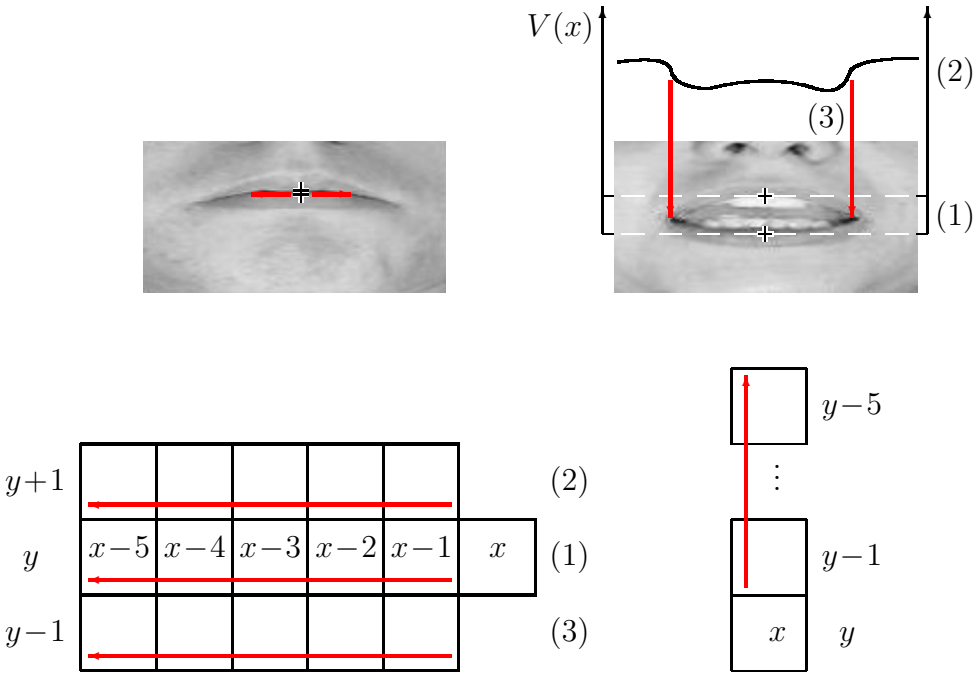


Figure 3.11: Step 2 - Top left: Moving along the shadow line for closed or partially open mouth. Top right: Vertical integral projection for wide open mouth. Bottom left: Checking for discontinuities in the shadow line. Bottom right: Testing for shadow line pixels above current position.

the continuation of the shadow line, the current position is moved to $(x-1, y)$. If not, the test is repeated first for $(x-1, y+1), \ldots, (x-5, y+1)$ (pixel line (2) in Figure 3.11, bottom left) and then for $(x-1, y-1), \ldots, (x-5, y-1)$ (pixel line (3) in Figure 3.11, bottom left). For positive tests, the current position is moved to $(x-1, y+1)$ and $(x-1, y-1)$, respectively. The reason for testing different vertical positions y is that the inner lip contour, which corresponds to the shadow line, of the lower lip is not necessarily a straight line but can be curved up or down, depending on the generic lip shape of the speaker and the mouth shape during speech production. If there are no more shadow line pixels ahead, the algorithm tests the five pixels $(x, y-1), \ldots, (x, y-5)$ above the current position (x, y) (Figure 3.11, bottom right). If a shadow line pixel is found, the current position is moved to $(x, y-1)$ and the test cycle is repeated. Otherwise, the lip corner has been found.

In rare cases, the shadow line is discontinuous for more than five pixels. There-

fore, the found lip corner positions are checked to be at least 25 pixels away from the midpoint of the lower lip. Otherwise, the search along the shadow line is restarted from a point 25 pixels away from the lip midpoint. For an object at a distance of 600mm from the camera, 25 pixels are equivalent to 10–12mm in Euclidean space. It is important to note that there is no reduction in horizontal resolution in the images. A distance of at least 10mm to either side of the lip midpoint, or at least 20mm mouth width in total, has been found to be a reasonable lower bound for any mouth shape experienced during speech production.

Mouth Wide Open

If the mouth is wide open, vertical integral projection on the intensity values of the image gives reliable estimates of the horizontal positions of the lip corners (Figure 3.11 top right). The vertical positions of the midpoints of upper and lower lip (1), determined in Step 1, define the vertical range for the integral projection (2). The largest changes in the resulting values determine the horizontal positions of the lip corners (3). Once these have been found approximately, a search along the (vertical) pixel columns through these horizontal positions looks for the pixels with the lowest intensity value and the highest saturation value which corresponds to the internal lip contour in the lip corner. The resulting pixel positions from the intensity and saturation searches are averaged to yield an estimate of the vertical position, which makes the algorithm more robust against misleading pixel values. Given that the accuracy of the results from both searches is unknown, averaging offers a way of most likely reducing any error. Finally, the found positions are refined by using the search technique along the shadow line described above for the closed or partially open mouth, but with the current estimated positions as starting points.

3.3.5 Step 3: Find Lip Midpoints

Now that the lip corner coordinates are established, the horizontal position of the lip midpoints, which has so far simply been the middle between left and right boundaries of the mouth window, needs to be recalculated. From the 2D image

coordinates of the lip corners in the left and right stereo images, their 3D coordinates $\vec{l}$ and $\vec{r}$ are calculated using the known camera parameters. Based on these 3D coordinates, the centre point $\vec{c}$ between these two points is computed as

$$\vec{c} = \frac{\vec{l} + \vec{r}}{2} \quad . \tag{3.23}$$

Since the lip corner coordinates could be wrong, a linear combination of the previous midpoint estimates and the newly computed centre point $\vec{c}$ is used to determine the likely centre point $\vec{c}\,'$

$$\vec{c}\,'[k] \;=\; \alpha\,\vec{c}[k] \;+\; (1-\alpha)\,\vec{c}\,'[k-1] \tag{3.24}$$

where k is the frame number. The linear factor α is the width confidence measure described in the next subsection. Information about the head pose from the general face tracker is used to define a normal vector perpendicular to an imaginary face plane and pointing away from the face. Then, the likely centre point $\vec{c}\,'$ is moved 5mm along this vector (Figure 3.12). An analysis of test video data had shown that the lip midpoints protrude about 5mm more than the lip corners. The likely lip midpoint $\vec{m}\,'$ is then back projected into image space and the x coordinate of that point taken as the horizontal position of the lip midpoints in each of the two images. After this, small adjustments to the vertical positions determined in Step 1 are likely and can be made in the same way, as when finding the exact lip contour at the end of Step 1. Finally, the 2D coordinates of the lip midpoints in the stereo image pair are combined to give their respective 3D coordinates.

3.3.6 Confidence Measures

Despite the generally accurate results of the lip tracking algorithm, no such algorithm is likely to be always 100% correct in practice. It is, therefore, important to define a confidence measure, which estimates the accuracy of the measurements. A novel confidence measure is proposed based on the difference between the corresponding 2D mouth width and mouth height distances in the left and right stereo images, respectively. Disagreeing 2D distances in both images are an indication that the lip tracking process has failed. As it is not known if the algorithm failed in

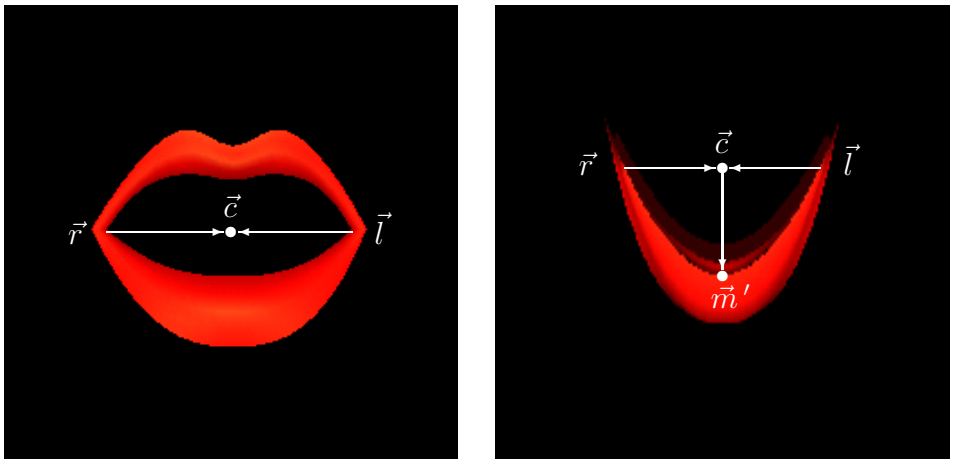


Figure 3.12: Step 3 - Finding the horizontal position of the lip midpoints (viewed from front and above).

only one or in both images, the results can only be marked as unreliable. However, the chances that the lip tracking algorithm fails in the same way in both images are small (see Section 3.4 for a validation of the accuracy of the lip tracking algorithm).

Two confidence measures are defined following this method: α_w for the mouth width and α_h for the mouth height. They are computed as

$$\alpha_w = 1 - \frac{|W_l - W_r|}{W_{max}} \tag{3.25}$$

$$\alpha_h = 1 - \frac{|H_l - H_r|}{H_{max}} \tag{3.26}$$

where W_l, W_r denote the 2D mouth width in the left and right image, respectively, H_l, H_r denote the 2D mouth height in the stereo image pair, and W_{max}, H_{max} are the width and height of the mouth window, respectively. It is $\alpha \in [0, 1]$, with $\alpha > 0.9$ indicating reliable lip tracking. The value of the confidence measure decreases linearly as the differences in the 2D distances become larger.

3.4 Validation of the Lip Tracking Algorithm

Visual inspection of the extracted lip feature positions showed a high degree of accuracy. The algorithm failed only in very few frames, which were well detected by the confidence measures and rejected. Figure 3.13 shows some correct and in-

Figure 3.13: Correct and incorrect feature positions. The speaker's left lip corner in the lower right image was not found correctly.

correct results. However, visual inspection alone can be deceptive. In order to quantify the error, a ground-truth would be required, but could not be obtained for practical reasons. Instead, a software tool was developed by the author of this thesis to compare automatically extracted lip feature points with the results from a manual extraction, in which the experimenter selects the feature positions by mouse click [Goecke 00b, Tran 00]. Although this process would be tedious for long video sequences and potentially introduces a new source of error — experimenter skill — it gave a clear indication of the accuracy of the lip tracking algorithm and was the best validation that could be achieved under the circumstances. Despite the importance of such validation experiments, the literature on lip tracking algorithms rarely describes any, so that it is not known, if validation experiments were never performed or if they were performed but not reported. In my opinion, validation experiments are essential to confirm the accuracy of any claimed results.

The validation of our lip tracking algorithm was performed for three speakers from the AVOZES data corpus (see Chapter 4). Three sequences were selected from the corpus for each speaker. These were the calibration sequences "ba ba ba ..." (/bɑː .../) and "e o e o e o ..." (/iː ɔː .../), and the first sentence of the continuous speech examples "Joe took father's green shoe bench out." The first sequence maximises vertical lip movement (opening and closing), while the second sequence emphasises horizontal lip movement (rounding and stretching). Fifty consecutive video frames were selected from each sequence. The mouse pointer was clicked by

the experimenter on the position of each of the four lip feature points on a frame-by-frame basis. Then, the parameter set (see Section 5.2) of our lip model was computed from the coordinates of these four points in exactly the same way as in the automatic case. For each lip feature point and for each parameter, the average absolute difference ('error') $\bar{d}$ and the standard deviation σ between automatic and manual feature extraction were computed (Table 3.1).

The comparison showed that the manual and automatic lip tracking procedures yielded similar results, sufficiently accurate for the purpose of AV speech processing. The results only differed at about 1–2mm absolute error for mouth width and mouth height. This appears to be a very accurate result given that totally non-intrusive face and lip tracking algorithms with no artificial markers or made-up lips were used. The standard deviation was 1–3mm except for subject 2 in sequence 1 and subject 1 in sequence 2. Their larger standard deviations resulted from outliers where the algorithm failed to find a feature position close to the 'true' position selected by the user. However, these tracking failures were detected by the confidence measures. When excluded from the analysis, these two sequences also showed similar results to the ones from the other sequences.

The difference between manual and automatic lip feature point extraction was larger for the protrusion parameters. This was due to two problems. Firstly, corresponding points were not always located correctly in the stereo images. Inaccuracies in the stereo matching result in incorrect depth (z) values, which are obviously of great importance for measuring protrusion. Secondly, using a point on the internal lip contour is perhaps not ideal for measuring protrusion because small changes, for example due to image quantisation, in the vertical position can have a considerable impact on the recovered depth value (Figure 3.14 left). Given that protrusion parameters only exhibit a range of about 10mm, the relative effect is considerable. Choosing a point on the external lip contour would be more suited, but also faces the difficulty of exactly identifying the same point in both camera views.

Analysing the feature points separately revealed that the differences in vertical position were less than one pixel. However, the horizontal difference was 2–3 pixels on average. Such stereo mismatches lead to less accurate 3D coordinates in the

Sequence	Speaker	Mouth Width $\bar{d} \pm \sigma$	Mouth Height $\bar{d} \pm \sigma$	Protrusion Upper Lip $\bar{d} \pm \sigma$	Protrusion Lower Lip $\bar{d} \pm \sigma$
ba ba ba	1	1.7 ± 2.1	1.9 ± 2.4	7.8 ± 9.5	6.2 ± 7.8
...	2	2.8 ± 10.2	1.9 ± 2.6	7.1 ± 8.4	6.0 ± 7.6
	3	0.8 ± 1.0	0.8 ± 1.0	6.7 ± 8.6	4.8 ± 6.2
e o e o e o	1	2.6 ± 4.0	1.4 ± 1.8	3.5 ± 4.5	5.9 ± 7.3
...	2	1.6 ± 2.9	1.1 ± 1.4	6.0 ± 7.6	5.3 ± 7.4
	3	1.7 ± 2.5	1.9 ± 2.6	4.2 ± 6.0	4.8 ± 7.0
Joe took...	1	2.3 ± 3.0	1.7 ± 2.3	6.6 ± 8.5	5.8 ± 7.5
	2	1.2 ± 1.6	2.2 ± 2.7	4.9 ± 6.5	5.0 ± 6.6
	3	1.9 ± 2.8	1.2 ± 1.6	4.4 ± 5.7	3.4 ± 4.5

Table 3.1: Average absolute difference $\bar{d}$ and standard deviation σ (both in mm) between automatic and manual feature extraction for three sequences.

stereo reconstruction process, in particular in the z coordinate, which affects the accuracy of the parameter set. In an effort to find out where the algorithm failed for those misplaced feature points, the different steps of the algorithm were examined separately. Two error sources became evident. Firstly, while locating the horizontal position of the lip corners for the wide open and the fully closed mouth revealed no problems, the partially open mouth showed differences in the manual and automatic extraction process. This was due to the difficulty of deciding the location of the lip corners on the inner lip contour, as shown in Figure 3.14 (right). Various interpretations are possible, but for this study it was defined that the shadow line between upper and lower lip always forms part of the inner lip contour. Secondly, the horizontal placement of the lip midpoints was more accurate in the automatic feature point extraction, where it was computed as the protruded centre point of the 3D line between the lip corners, than in the manual one. Thus, a corresponding horizontal position in both images was ensured in the automatic feature point extraction, whereas the experimenter could move the mouse pointer freely. A similarly guided approach in the manual extraction would improve the results.

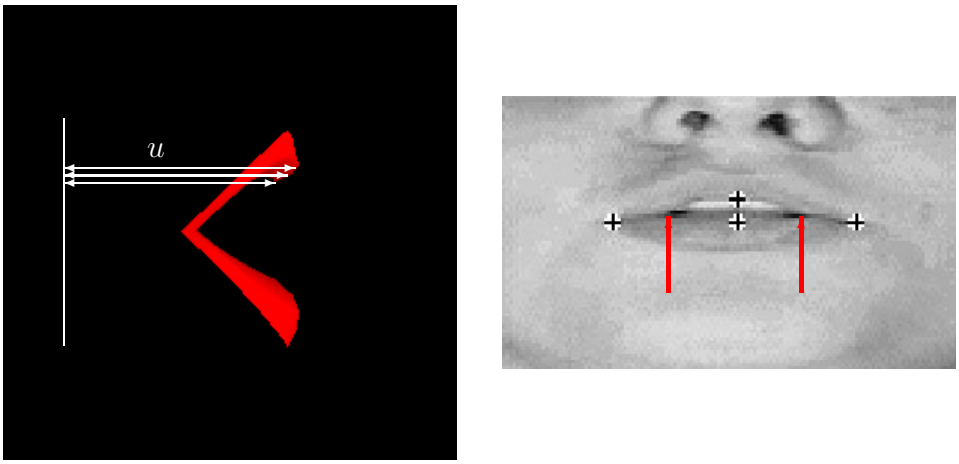


Figure 3.14: Lip tracking problems: Left: Small vertical changes can have a drastic effect on protrusion parameters. Right: Lip corner position depends on definition of internal lip contour — position of crosses or of red arrows?

3.5 Chapter Summary

Two essential systems used in this project have been described and discussed in this chapter. The RSL stereo vision face tracking system allows non-intrusive real-time face tracking. By using stereo vision in calibrated cameras, the 3D (world) coordinates of any object point can be recovered. The fact that no artificial markers are required is important for the analysis of AV relationships, and in general AV speech processing, because it allows people to act and speak naturally. The face tracking system combines both hardware and software systems of a 450MHz PC to achieve real-time processing at a frame rate of 15–20Hz. The details of the systems have been described, in particular the stereo reconstruction, the camera calibration, and the face tracking procedure.

The lip tracking system was newly developed in the course of this study. It is a software system that extends the face tracking system. It is based on combining colour information with knowledge about the structure of the mouth region to tailor the applied algorithms to different mouth shapes. No single cue is good enough by itself, but combining them in a multi-cue system creates a powerful system. To the best of my knowledge, this is the first lip tracking system that applies stereo vision in the field of AVSP. Unlike other systems, it does not require artificial

markers or made-up lips, thus facilitating natural speech. A 3D lip model has been defined in which the structure of the entrance to the mouth is described by a set of parameters. These parameters are mouth width, mouth height, protrusion of upper lip, and protrusion of lower lip. The parameter values can easily be derived from the 3D coordinates of four lip feature points: the lip corners and the midpoints of upper and lower lip. The steps of the lip tracking algorithm have been described in detail. A new confidence measure has been designed to judge the reliability of the measured parameters. As a control, the accuracy of the lip tracking system has been validated in a comparative experiment with manual feature point selection. The mouth width and mouth height parameters were generally found to be very accurate, while problems in the measurement of the protrusion parameters have been discussed. It is important to perform such validation experiments to confirm the accuracy of any results based on a lip tracking system.

The combined face and lip tracking systems achieved a real-time frame rate of 5–8Hz. Fast and accurate processing is a prerequisite for real-time AV speech processing. Higher frame rates can be realised by offline processing, which is only limited by the properties of the recording equipment. The current offline frame rate of 30Hz was a result of the analogue NTSC cameras. The face and lip tracking systems can be applied to both real-time as well as offline processing. Offline processing was used for the statistical analyses in this study (Chapter 5). For fast lip movements, an even higher frame rate than 30Hz would be desirable, but could not be achieved with the current equipment. Recent systems with digital cameras achieve 60Hz (see for example the *faceLab* product by Seeing Machines at http://www.seeingmachines.com). However, the offline performance of the current system was judged to be sufficient for the statistical analyses of AV relationships.

Comparing the reported results of non-intrusive 2D face and lip tracking methods (see Sections 2.3.2 - 2.3.4), the results presented here are encouraging to believe that the distances measured in the face plane by using a stereo lip-tracking system are more accurate than those gained from 2D systems. Of course, such judgement can only be subjective at this point in time, as a true comparison can only be made if the systems are applied to the same input data.

Chapter 4

Audio-Video Speech Data Corpus of Australian English

To achieve the goal of this project, the investigation of the relationship between audio and video speech parameters for Australian English, an audio-video speech data corpus was needed. In this chapter, the new *A*udio-*V*ideo *OZ*stralian *E*nglish *S*peech (*AVOZES*) data corpus, used in the experiments described in Chapters 5 and 6, is presented. It was necessary to establish a new corpus because no comprehensive corpus of AV speech existed for AuE. Despite a number of small and large AV speech data corpora having been produced over the last ten years, the field of AVSP still lacks a 'benchmark' data corpus for testing and comparing the various results published on a common basis (cf. Section 2.3.6). Moreover, it seems that no agreed framework on the design of a 'standard' data corpus for various languages exists. A new extensible framework is proposed in Section 4.1.

The design of the AVOZES data corpus followed that proposed framework, covering all visemes and phonemes of AuE (Section 4.2). It is thereby the most comprehensive AV speech data corpus of AuE to date. In addition, it is the first AV speech data corpus to use a stereo vision system. As mentioned previously in Section 3.1, a stereo vision system has the advantage over monocular systems that 3D coordinates can be recovered accurately. Thus, 3D distances can be measured, not just distances in 2D image coordinates, which makes the measurements robust

against rotations of the face. Other speech corpora do not include a video component (ANDOSL [Millar 94, Millar 97]), are either not designed for AuE (DAVID [Chibelushi 96b]), not comprehensive enough in terms of the number of speakers or sequences/phonemes (Tulips1 [Movellan 95]), or not publicly available (IBM [Neti 01]). An overview of AV speech data corpora was given in the literature review in Chapter 2, Section 2.3.6.

The experimental setup used for the recordings of the AVOZES data corpus is described in Section 4.3. Details are given about the recording media, the recording room and the layout of the components, as well as the recording equipment used. Finally, Section 4.4 describes the recording process and summarises information about the speakers recorded in the AVOZES data corpus.

4.1 A Framework for the Design of Audio-Video Speech Corpora

A survey by Chibelushi *et al.* [Chibelushi 96a] examined existing AV speech data corpora as well as which features researchers would like to see in such corpora. The latter was established by a questionnaire-style survey. Although only five questionnaires were received in response, the conclusions that were made by the authors match the observations made by the author of this thesis. The usual limitations of existing AV speech corpora are:

- a small number of speakers,
- a small number of phonemes and visemes covered, and
- isolated words such as digits or letters of the alphabet rather than embedded (carrier phrase) or continuous speech.

Most existing AV speech corpora were clearly designed for a particular research project and not as a publicly available corpus for the comparison of methods developed by various research groups around the world. The limited size of many

corpora is clearly related to the time- and resource-consuming effort required in the creation of a corpus.

The features that researchers would like to see in a benchmark corpus were:

- a large number of speakers for statistical significance,
- a broad coverage of phonemes and visemes,
- different levels of acoustic noise starting with 'clean speech' (no noise),
- whole face images in colour,
- short words and continuous speech with transcription, and
- extensibility.

Chibelushi *et al.* [Chibelushi 96b] presented a number of ideas on how to design an AV speech data corpus that covers a variety of experimental themes. Similar considerations on design issues were made by Öhman [Öhman 98] for a Swedish language AV speech data corpus. Generally, various factors play a role in the design of such corpora, and these are briefly discussed in the next section.

4.1.1 Factors in AV Speech Corpus Design

Data Collection Factors

These factors relate to the corpus recording process. One can argue that recordings made in laboratories do not mirror exactly the conditions in the real world. However, in terms of facilitating the interpretation of experimental results, it is an advantage to be able to control the experimental conditions. These conditions include the recording equipment, the possible use of markers, the layout of the recording room (e.g. background), the sitting arrangement, the illumination arrangement, and the level of acoustic noise. Going through all possible combinations of these conditions in a systematic way would result in an exponential growth of the corpus and quickly become impractical. It is suggested here to leave all conditions but one constant at a time, and to study the effects of changing that condition, rather than mixing the effects of various changing conditions in one recording.

Speaker Factors

Speaker-related factors can be categorised into language background, speaking style, and personal characteristics. The first category includes issues like dialects (or accents) and first versus other languages. Usually, one would study native speakers first to characterise a particular language, but the identification of differences between native and foreign speakers is also an interesting research topic. Within a language, different dialects exist even among native speakers, and these must be considered. The second category, speaking style, determines the social and / or emotional conditioning of speaking, for example a conversational style or an excited style. It should be noted that first-time participants in a data corpus collection often feel nervous about the task ahead or are overly motivated to do the task particularly well, so that their speaking style changes from the way they would normally speak. Therefore, the familiarisation of speakers with the environment and the speech material is important. The third category deals with the 'natural' characteristics of a person, such as gender, body physique, characteristics of the vocal tract, or the amount of visible movement of the articulators.

Generally, a balanced population in a corpus is desirable. Finding a sufficient number of speakers to cover normal variation in the above categories might not always be possible, but it is suggested here to at least achieve a gender balance in groups with the same language background (native speakers, foreign speakers). As mentioned for the data collection factors, a systematic way of going through all possible combinations of these factors leads to an exponential growth of the corpus. It is, therefore, necessary to define the range of speaker-related factors, which are addressed in a corpus, in advance and to select speakers accordingly.

Speech Material Factors

These factors relate to the material that speakers are requested to speak. Such material can be letters of the alphabet, isolated words, and continuously spoken phrases. Words can be real, existing words (e.g. digits, commands) or nonsense words, designed to investigate a particular phoneme transition (e.g. ABABA). Which material

is included in a particular corpus depends on the application in mind. It is suggested here that general-purpose corpora contain some examples from all categories. It should be noted that reading lists of phones, diphones, words etc. often results in a speaking style different from what it would be, if the words were included in a phrase. The use of a carrier phrase, in which the speech material unit of interest is embedded, is therefore suggested, unless the target application in mind requires otherwise (for example, a spelling task). This problem does not occur (or only to a much smaller extent) for continuously spoken phrases.

Furthermore, the coverage of phonemes and visemes is another factor. Whenever feasible, it is suggested to cover all phonemes and visemes of a language at least once in the chosen context (phones, diphones, words etc.) for completeness. That means that except for very large corpora, not all possible diphone or triphone transitions will be covered. By using the same context, the phonemes and visemes can at least be studied in a controlled environment. Moreover, if the resources do not allow the inclusion of each phoneme, a careful selection must take place, so as to choose at least one for each viseme category, as the number of visemes is smaller than the number of phonemes (see Section 2.1.2).

4.1.2 The Proposed Framework

The ideas presented in Section 4.1.1 are extended here and a new framework for the design of AV speech data corpora is proposed. A modular approach, where each module contains certain sequences, allows for extensibility in terms of the various factors discussed in the previous subsection. For example, a data corpus could start with a small number of speakers uttering selected phoneme sequences in a noiseless audio condition. Later, more sequences can be recorded to extend the phonemic coverage, add more speakers, or repeat sequences in different noise levels. Thus, a corpus can grow over time, thereby accommodating the amount of resources it takes to create and store it, while still providing usable data from the beginning. In this context, ensuring continuity in the facilities and equipment used, as well as in terms of the speakers appearing in the corpus, is important. If recordings are made

at different points in time, the comparability of the recorded material with earlier recordings is an important issue that needs to be addressed.

As a minimum, any AV speech data corpus should contain the following three modules:

1. sampling recording setup without a speaker,

For each speaker,

2. sampling recording setup with speaker, and
3. recording of phonemes and visemes.

The module "sampling recording setup without a speaker" captures general aspects of the data collection process, such as visual background, scene illumination, and acoustic background. For every speaker, there are at least two modules. The module "sampling recording setup with speaker" shows the speaker in the scene. This can include sequences useful or necessary for the video processing, such as shots of the face from various angles. The module "recording of phonemes and visemes" contains the actual speech material sequences following the guidelines in the previous section.

Additional modules can be added easily. Some modules, for each speaker, that were considered prior to the creation of the AVOZES data corpus described in this thesis, were:

- speaker calibration,
- application sequences,
- different view angles,
- different levels of illumination, and
- different levels of acoustic noise.

The module "speaker calibration" could contain sequences which exhibit particular acoustic or visible speech patterns (for example, lip rounding). These sequences

can be used to classify speakers into different classes in the analysis stage. Longer sequences of continuous speech or command sequences would make up the module "application sequences". The other three modules comprise changes in data collection factors. From a data analysis point of view, repeating the modules "recording setup with speaker" and "coverage of phonemes and visemes" for each different condition is desirable. However, it must be noted that this may not be practically feasible, both in terms of the amount of resources and the duration of session times required, if a lot of speech material has to be covered. Speakers get tired if recording sessions become too long, so either the amount of speech material must be reduced, that is, not to cover all phonemes and visemes, or recordings must be made in different sessions, which raises questions of the comparability of the recordings because a speaker's mood or health might have changed between sessions. The longer the time span between sessions, the more pertinent these questions become.

The proposed framework enables the design of AV speech corpora in a systematic way. The modular structure gives it the flexibility required to be useful for various research themes and applications, while the minimum requirements help to achieve consistency across corpora.

4.2 The Design of the AVOZES Data Corpus

This framework was followed in the design of the *A*udio-*V*isual *OZ*stralian *E*nglish *S*peech (*AVOZES*) data corpus [Goecke 00b]. No other AV speech data corpus with stereo camera video has been published thus far.

The AVOZES data corpus has a total of six modules — one general module and five speaker-specific modules. These six modules are:

1. sampling recording setup without a speaker,

For each speaker,

2. sampling recording setup with speaker and definition of face model,
3. calibration sequences,

4. short words in carrier phrase covering phonemes and visemes,
5. application sequences - digits, and
6. application sequences - continuous speech.

These modules are described in more detail in the following sections.

4.2.1 Module 1 - Sampling Recording Setup without Speaker

This module contains only one sequence in the AVOZES data corpus. It is a 30 second sequence of the recording scene viewed by the two cameras, but without any speaker present. The sequence can be used to determine the background level of acoustic noise present in the recording studio, due to air-conditioning as well as computer and recording equipment. In addition, information about the visual background can be gained, if it is required for the segmentation of the speaker from the background in the video stream.[1] Since the sequence in this module is speaker-independent, only one recording was needed. However, if corpus recordings were made over prolonged time spans (months or years), or in intervals (for example, extending the corpus at a later stage), the sequence should be repeated once during each interval to record possible changes to the recording environment.

4.2.2 Module 2 - Sampling Recording Setup with Speaker

A sequence showing certain head movements was needed by the stereo vision face tracking system to establish the face model of each speaker. The process of building a face model required a frontal view of the face as well as views on an angle of 45° to either side (see Section 3.1.5 for details). Thus, the speaker is first shown in face frontal position for 5 seconds, then the speaker turned the head 45° to the left, kept it there for 5 seconds, then turned it 45° to the right of the frontal position and held that position for 5 seconds again. Module 2 contains one such sequence for each speaker in the AVOZES data corpus.

[1] Such information was not required in the face and lip tracking systems used in this study, but might be needed by other algorithms.

4.2.3 Module 3 - Calibration Sequences

This module comprises two sequences per speaker for the purpose of 'speaker calibration', in terms of their visible speech articulation or visual expressiveness. For (purely visual) lipreading as well as AV automatic speech recognition, the amount of visible speech articulation determines how much (additional) information can possibly be gained from the video stream. Expressive visible speech articulation offers more information than a person who does not move the visible speech articulators much (for example, a person who mumbles). Extracting lip parameters, such as mouth width or mouth height, over time enables an analysis of the visual expressiveness of a speaker, for example by analysing the maximum values reached in each cycle of lip movements. Speakers with values in the margin of the overall distribution can be excluded from the analysis or treated differently, if desired.

The two calibration sequences "ba ba ba ..." and "e o e o e o ..." recorded in the AVOZES data corpus were each repeated continuously by each speaker for about 10 seconds. Despite the artificial nature of these prompts, the first sequence can give insight into the amount of vertical lip movement, i.e. opening and closing, while the second sequence emphasises horizontal lip movement, i.e. rounding and stretching.

4.2.4 Module 4 - Short Words in a Carrier Phrase Covering Phonemes and Visemes

The sequences in this module form the core part of the AVOZES data corpus for the statistical analysis of relationships between audio and video speech parameters (see Chapters 5 and 6). There are 44 phonemes (24 consonantal and 20 vocalic phonemes) and 11 visemes (7 consonantal and 4 vocalic visemes) in AuE, according to Woodward and Barber [Woodward 60], Plant and Macrae [Plant 77], and Plant [Plant 80]. Following the ANDOSL design [Millar 94], the phonemes can be categorised into 8 classes (Tables 4.1 and 4.2).[2] Similarly for the visemes, follow-

[2] IPA refers to the *International Phonetic Association* and its alphabet. The latest version was published in 1993 and updated in 1996 [IPA 99].

ing [Plant 77] and [Plant 80], there are 11 viseme classes (Table 4.3)[3]. Plant and Macrae [Plant 77] do not label their vowel and diphthong visemes, but they are broadly:

1. front non-open vowels and front close-onset diphthongs,
2. open vowels and open-onset diphthongs,
3. back/central non-open vowels and diphthongs containing these vocalic positions, and
4. back/central open vowels and diphthongs containing these vocalic positions.

In [Plant 80], these visemes were described in terms of their mouth shape as (1) small aperture and spread lips, (2) large aperture and neutral lips, (3) small aperture and rounded lips, and (4) large aperture and rounded lips. It was also noted that the diphthong /aʊ/ appeared to be visually distinctive in a CVC-context[4] (with C=/b/), while this was not the case in the original study with a CV-context (with C=/b/). It might, therefore, be considered as an additional viseme.

The phonemes and visemes in the AVOZES data corpus were put in central position in CVC- or VCV-contexts[5] to be free of any phonological or lexical restrictions. However, wherever possible, existing English words (that follow these context restrictions) were favoured over nonsense words in order to simplify the familiarisation process of the speakers with the speech material. The vowel context for VCV-words was the wide open /ɑː/ ("ar-ar"). The voiced bilabial /b/ was used as the consonant context ("b-b") for CVC-words. The opening and closing of a bilabial viseme clearly marks the beginning and end of the vocalic nucleus, and thus facilitates the visual analysis. Using /b/ instead of /p/ lengthens each word, giving more data to analyse.

A disadvantage of the /bVb/ context is that a bilabial context causes strong coarticulation effects in the formants. However, these are quite predictable for /b/

[3] The phonemes /z/, /ʒ/, /h/, and /ŋ/ were not included in the investigation by Plant and Macrae, but are here classified into corresponding viseme classes in Table 4.3.

[4] CVC - consonant-vowel-consonant

[5] VCV - vowel-consonant-vowel

Class	Description	IPA Symbol	Example "as in ..."
Oral stops	Bilabial voiceless	p	poor
	Bilabial voiced	b	bore
	Alveolar voiceless	t	tore
	Alveolar voiced	d	door
	Velar voiceless	k	core
	Velar voiced	g	gore
Fricatives	Labio-dental voiceless	f	fan
	Labio-dental voiced	v	van
	Inter-dental voiceless	θ	thin
	Inter-dental voiced	ð	than
	Alveolar voiceless	s	sue
	Alveolar voiced	z	zoo
	Palatal voiceless	ʃ	sure
	Palatal voiced	ʒ	azure
	Glottal voiceless	h	ham
Affricates	Alveolar voiceless	tʃ	chore
	Alveolar voiced	dʒ	judge
Nasals	Bilabial closure	m	mow
	Alveolar closure	n	now
	Velar closure	ŋ	sing
Liquids and glides	Lateral	l	lull
	Rhotic	r	row
	Bilabial	w	wow
	Palatal	j	you

Table 4.1: Consonant phoneme classes in the ANDOSL database.

Class	IPA Symbol	Example "as in ..."
Short vowels	ɪ	hid
	ʊ	hood
	ɛ	head
	ə	the *(not "thee")*
	ɒ	hod
	ʌ	bud
	æ	had
Long vowels	iː	heed
	uː	who'd
	əː	there
	ɜː	heard
	ɔː	hawed
	ɑː	hard
Diphthongs	eɪ	hay
	əʊ	hoed
	ɔɪ	hoy
	aɪ	hide
	aʊ	how
	ɪə	here
	ʊə	tour

Table 4.2: Vowel phoneme classes in the ANDOSL database.

and it is believed that the advantages of a bilabial context for visual segmentation outweigh the disadvantages from coarticulation.

To overcome the typical articulation patterns associated with reading words from a list, each CVC- and VCV-word was enclosed by the carrier phrase *"You grab /WORD/ beer."* Having a bilabial opening and closing before and after the word under investigation again helps with the visual segmentation process, in particular for the VCV-words. Tables 4.4 and 4.5 show the lists of prompts and pronunciation

Viseme Description	IPA Symbols
Bilabials	p b m
Labio-dentals	f v
Inter-dentals	θ ð
Labio-velar glides	w r
Palatals	ʃ tʃ ʒ dʒ
Alveolar non-fricatives and plosives and velar plosives	l n j h g k
Alveolar fricatives and plosives	z s d t
Front non-open vowels and front close-onset diphthongs	iː ɪ ɛ ɪə
Open vowels and open-onset diphthongs	æ ɑː ɜː ʌ ə ɔː aɪ eɪ
Back/central non-open vowels and diphthongs	uː ʊ əː ɔɪ ʊə
Back/central open vowels and diphthongs	ɒ əʊ aʊ

Table 4.3: Viseme classes in Australian English.

hints, which were presented to the speakers during familiarisation and recording. Each phrase to be read out aloud by the speakers was shown at the top of the prompt message on the screen, and was followed by an example of how to pronounce the phoneme under investigation in that prompt. For an example of such a prompt message, see Figure 4.2 in Section 4.3.

4.2.5 Module 5 - Application Sequences - Digits

The sequences in this module can be used as examples of applying any results, gained from an analysis of the phonemes and visemes in the "short words" module,

Class	IPA Symbol	"You grab ... beer."	Pronunciation "as in ..."
Short vowels	ɪ	bib	ship
	ʊ	boub	should
	ɛ	beb	head
	ɒ	bob	shop
	ʌ	bub	cup
	æ	bab	had
Long vowels	iː	beeb	heed
	uː	boob	cool
	ɜː	berb	herb
	ɔː	borb	floor
	ɑː	barb	hard
	əː	bareb	bare
Diphthongs	eɪ	babe	babe
	ɔɪ	boyb	boy
	aɪ	bibe	hide
	aʊ	bowb	how
	ɪə	beerb	here
	əʊ	bobe	pope

Table 4.4: Prompts for vowels and diphthongs in the AVOZES data corpus.

to short sequences that are more application-driven. Digit recognition is a common task in automatic speech recognition, e.g. [Luettin 98, Petajan 84, Potamianos 97], and similar sequences can be found in a number of AV speech corpora, for example in DAVID [Chibelushi 96b] and Tulips1 [Movellan 95].

The AVOZES data corpus includes one sequence per digit for each speaker, spoken in order from 0 to 9. Again, each digit is enclosed by the carrier phrase *"You grab /DIGIT/ beer."* to ensure lip closure before and after the digit for ease of segmentation of the video stream.

Class	IPA Symbols	"You grab ... beer."	Pronunciation "as in ..."
Oral stops	p	arpar	par
	b	arbar	bar
	t	artar	tar
	d	ardar	dark
	k	arkar	car
	g	argar	garb
Fricatives	f	arfar	far
	v	arvar	van
	θ	arthar	thin
	ð	arthar	than
	s	arsar	sue
	z	arzar	zoo
	ʃ	arshar	sharp
Affricates	tʃ	archar	chart
	dʒ	arjar	jar
Nasals	m	armar	arm
	n	arnar	barn
	ŋ	arngar	sing
Liquids and	l	arlar	large
glides	r	ara	run
	w	arwar	wow
	j	aryar	yard

Table 4.5: Prompts for consonants in the AVOZES data corpus.

4.2.6 Module 6 - Application Sequences - Continuous Speech

This second module with application-driven sequences contains examples of continuous speech from each speaker. The three sequences are:

1. *"Joe took father's green shoe bench out."*[6]
2. *"Yesterday morning on my tour, I heard wolves here."*
3. *"Thin hair of azure colour is pointless."*

Together with the first sentence, the second and third sentences were designed in such a way that they contain all phonemes and visemes of AuE. One of the ultimate goals in automatic speech recognition is the task of continuous speech recognition in all conditions. The sequences in this module offer an initial way of applying and testing any results from an analysis of the sequences in module 4 to such a task.

4.3 Experimental Setup

To be able to perform various and repeated experiments on the same material of a speech data corpus, the sequences must be stored on a medium that allows easy repeated access, without loss of quality. For small corpora with few speakers and sequences, storage on a computer hard disk is possible. However, for large corpora with many speakers and sequences, such storage becomes quickly impossible or very expensive, despite ever-growing hard disk capacities. Video and audio compression is a way of overcoming problems with large amounts of data to some extent, but high compression is often related to loss of detail, which is clearly not desirable. Digital Video (DV) systems offer a good alternative for high-quality storage (see Appendix A for more details on the DV standard). DV tapes are inexpensive and common tape sizes can store up to 63 minutes of video and audio data. This solution was used in this project.

[6] This sentence appeared first in the corpora M2VTS and XM2VTSDB [Messer 99].

Recordings of the AVOZES data corpus were made in August 2000 and August 2001. The recordings took place in the audio laboratory of the Computer Sciences Laboratory (CSL) at the Australian National University. The same equipment was used on both occasions. The CSL audio laboratory is a soundproof room in the interior of the building, well-shielded from noise sources outside the room but with a small amount of background noise from the room's air-conditioning.

Figure 4.1 shows the recording setup. The speakers sat on a swivel chair in front of the stereo cameras, which were positioned with the help of a camera tripod. A light source was placed directly below the camera rig to illuminate the speaker's face. This light source was a normal office desk lamp with a reflective lampshade. Placing the light source below the cameras ensured a well-lit face, while blinding was reduced to a minimum. Other light source arrangements were considered, such as putting one light source on either side of the cameras — sufficiently apart so as not to blind the speaker (similar to [Petajan 84]) — or a more expensive lighting system, as used in professional photographic studios. However, these were discarded in favour of the simplicity of a single light source, which achieved the objective of removing shadows in the mouth region. In addition, there was a general illumination of the room from three ceiling lights (normal light bulbs, not fluorescent light).

An office swivel chair was used for two reasons. Firstly, the height of the seat could be adjusted easily for shorter or taller people, while leaving the camera arrangement etc. unchanged. Secondly, the process of building a face model for each speaker required sequences, in which the speaker was asked to turn the head 45° to the left and to the right of the cameras (see Section 4.2.2). The speakers were instructed to turn not just their eyes, but the whole head, so that it would point to corresponding markers on the wall. By sitting on a swivel chair, the speakers could, in fact, simply turn their whole bodies towards the markers. Keeping the vertical axis of the chair at a marked position ensured that the face was kept in the cameras' viewfields. The distance from the face to the cameras was about 600 ± 50mm, which corresponded to the distance ("depth") range that the cameras were calibrated for (see Section 3.1.3).

The speaking prompts appeared on the computer screen above the cameras.

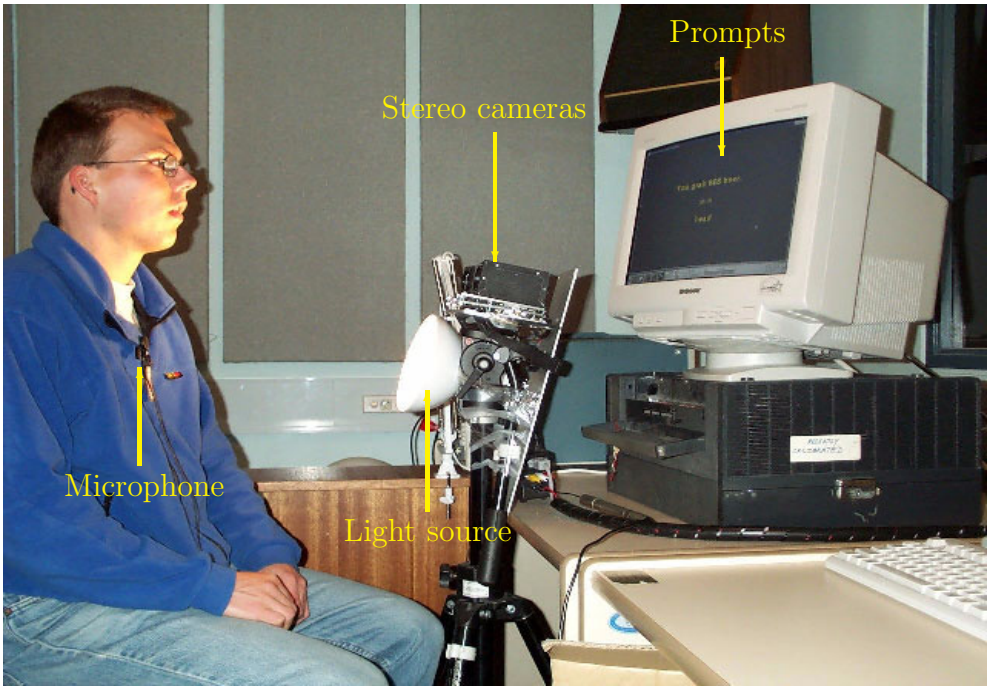


Figure 4.1: Recording setup in the CSL audio laboratory.

Figure 4.2 shows the stereo cameras in the foreground and the computer screen in the background from the viewpoint of a speaker. Prompts were advanced per mouse click by the recording assistant, when a prompt was pronounced correctly. Otherwise, the speaker was asked to repeat the phrase. The screen's background colour was swapped between a dark green and a dark blue whenever the next prompt appeared, so as to give the speaker an additional visual signal that a new prompt had appeared on the screen.

A clip-on microphone was attached to the speaker's clothes on the chest about 20cm below the mouth. The microphone was an omnidirectional Sennheiser MKE 10-3 microphone with a frequency response of 50Hz–20kHz [Sennheiser 81]. The microphone system was directly connected to the DV recorder, where the microphone's output was recorded as mono sound on DV tape with a 48kHz sampling frequency.[7] The DV recorder was a JVC HR-DVS1U miniDV/S-VHS video recorder, which also featured an IEEE-1394 DV in/out connector.

[7] In the 48kHz sampling mode, two channels are recorded for stereo audio but in case of mono audio input, both channels contain the same signal.

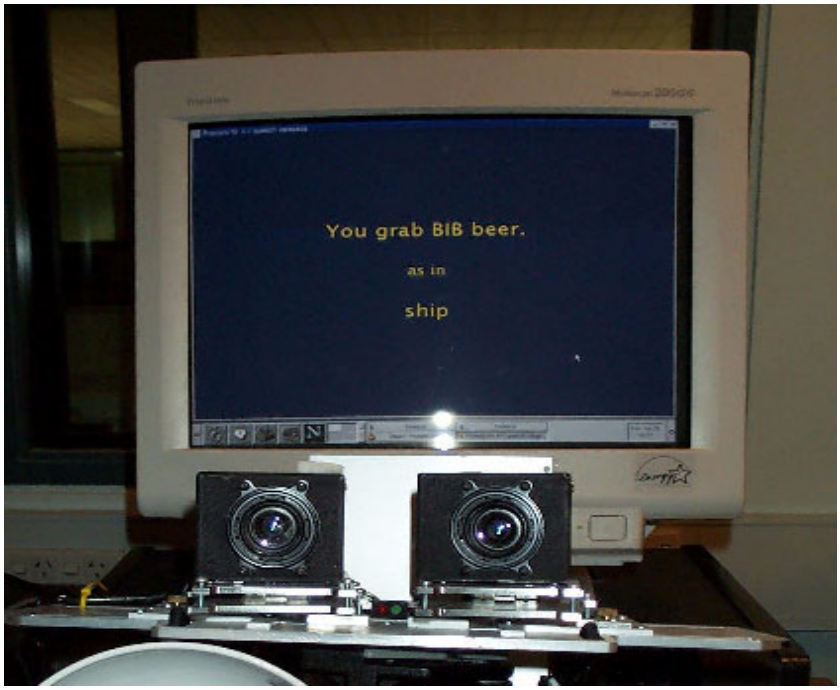


Figure 4.2: Speaker's view of the recording setup.

The output of the stereo cameras was multiplexed into one video signal, then sent to the Hitachi IP5005 video card, as described in Section 3.1. However, no face or lip tracking was performed at this time. The video signal was merely unscrambled, so that the video sequences on tape showed the output from the left camera in the top half and the output of the right camera in the bottom half of each video frame (as shown in Figure 3.5). The video signal was then sent from the video card to the DV recorder, where it was recorded as an NTSC YUV 4:1:1 signal at 30Hz frame rate (see Appendix A and the list of abbreviations in the preface for an explanation of these terms). Because of the way that the outputs from the two cameras were multiplexed, there was a 16.6ms delay between the output from each camera in any recorded video frame. While virtually undetectable by the human eye at normal video play rate (30Hz), it is a potential error source for the 3D reconstruction process, which requires the same object point to be identified in both images (and assumes that the object has the same shape in both the left and right image). However, no noticeable problems were encountered with this delay during the video analysis.

4.4 Recording

In the first set of recordings, made over the period of one week in August 2000, ten native speakers of AuE and three speakers with a different language background were recorded. Of the latter, two speakers have an English language background (United Kingdom, New Zealand), the third speaker speaks German as his first language, but has spent extended periods of time in the United Kingdom and in Australia. The second set of recordings (additional speakers) was taken over a period of two days in August 2001, using exactly the same equipment, setup, and location on both occasions. The second set contains ten native speakers of AuE and one speaker with a Chinese dialect as his first language, but who has lived in Australia for 6 years.

Beside the actual recordings, each speaker was also asked to fill in a form about personal data, so that any outstanding effects in the recorded material could be checked against these data. A similar approach was taken in the ANDOSL database [Millar 94]. It is important to collect such data in addition to the signal data, as for example professional training in singing or medical conditions of the respiratory system can have an effect on the pronunciation. Personal data collected contains:

- name, date of birth, and gender,
- level of education and current occupation,
- height and weight,
- native language of speaker, speaker's mother, and speaker's father,
- place of origin and occupation of both parents,
- extended periods outside Australia (at least 3 months) — time and place,
- singing, training in singing,
- smoking, medical conditions (e.g. asthma).

The individual information (names omitted, date of birth transformed into age) about the native speakers of AuE is presented in Appendix B. In addition, the

distance from the speaker's mouth to the microphone was also measured. The range was 150–250mm. The distance from the face of a speaker to the cameras varied between 550–650mm. Speakers were allowed to move their head freely, but were asked to keep it roughly in the same position to ensure that it was within the cameras' viewfield. The computer screen, from which the prompts were read, was another 20cm (in horizontal direction) behind the cameras (Figure 4.1). The computer and DV recorder were housed in an acoustic insulated box to reduce the amount of acoustic noise produced by the hard disk and cooling fans.

Figure 4.3 shows the faces of the native speakers of AuE. The group is gender balanced with ten female and ten male speakers. Six speakers wear glasses, three wear lip make-up, two have beards. At the time of the recordings, these speakers were between 23 and 56 years old. The speakers were approximately classified into the three speech varieties of AuE (see Section 2.2) by the recording assistant, which created groups of 6 speakers for broad AuE, 12 speakers for general AuE, and only 2 speakers for cultivated AuE. While this distribution approximately reflects the composition of the Australian population in terms of the accent varieties, it is important to point out that the individual groups are not gender balanced, and that their size is small for statistical analyses on an individual group basis. This study, therefore, concentrated on analysing the corpus as a whole. It is also worthwhile to remember that the speech varieties are not discrete entities, but rather span a continuum of accent variation, so that any classification can only be approximate. Still, the classification can be helpful for experiments and analyses that aim at identifying similarities and differences between the AuE speech varieties.

The AVOZES data corpus currently contains only frontal face ($\pm 10^{\circ}$) recordings, with no separate or simultaneous recordings from a different angle. The faces were illuminated from the front. Recordings were made for a clean audio condition. There was no particular background noise other than what has already been described in Section 4.3. However, artificial (computer-generated) noise could be added to the audio signal, if that was desired for some experiment. In that way, the control of the noise is much better, because it can be designed to suit a particular experimental situation and the AVOZES date corpus could be used for a wider

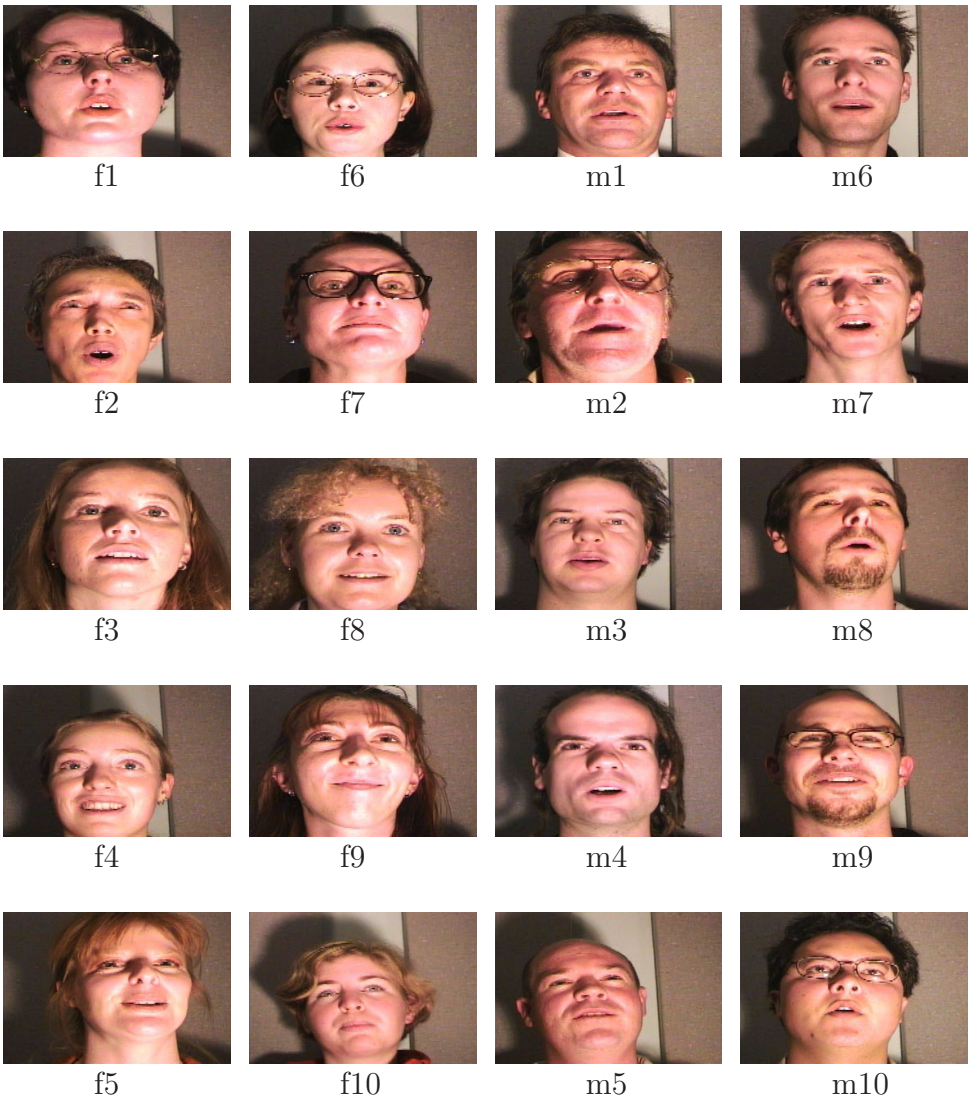


Figure 4.3: Face shots of the native speakers of Australian English in AVOZES.

range of tests. It should be noted that adding artificial noise does not take account of the Lombard effect [Junqua 93]. Recordings were made in a conversational tone.

Each speaker spent about half an hour in the recording studio. They were first familiarised with the speech material and informed about the recording procedure about to follow. Actual recordings took about five minutes per speaker. The author was present as a recording assistant, so that speakers did not have to handle any of the equipment themselves and could concentrate on the speaking task.

A total of 58 sequences were recorded per speaker (3 face model sequences and 55 speech material sequences). Two phonemes from the lists in Tables 4.1 and 4.2 were omitted (see prompt lists in Tables 4.4 and 4.5) because they have a low occurrence

in AuE. These phonemes were /ʒ/ (as in "azure") and /ʊə/ (as in "tour"). It was, therefore, quite likely that speakers would not pronounce the prompts correctly. These two phonemes were also rather difficult to achieve in the selected CVC- and VCV-contexts. Furthermore, the neutral vowel /ə/ and the neutral consonant /h/ were not recorded, because it was assumed that they add little to the statistical analysis of relationships between audio and video speech parameters due to their neutrality.[8] During the recordings it also became evident, that some speakers had difficulties in producing distinguished sounds for the voiceless and voiced inter-dental fricatives /θ/ and /ð/, as well as producing the velar closure nasal /ŋ/. The analysis of these sequences must therefore be treated with care.

4.5 Chapter Summary

In this chapter, design principles of AV speech data corpora have been discussed. Most existing AV speech data corpora were found to follow an ad-hoc approach that suited a particular research theme, but this did not allow them to be used as 'benchmark' corpora. A new framework for the design of the currently collected and future corpora has therefore been proposed in this chapter. The framework's modular approach allows for extensibility in terms of data collection factors, speaker factors, and speech material factors which should all be considered in the design of an AV speech data corpus. With the modular approach, it is easy to consider sequences that cover the phonemes and visemes of a language as the minimum common set of corpora, while project-specific sequences can be added in separate modules according to the specific requirements.

The few well-designed and comprehensive corpora presently described in the literature (cf. Section 2.3.6) are not for AuE or not publicly available. Therefore, a new corpus following the proposed framework was designed and recorded. The

[8] In hindsight, it might have been better to also record these four phonemes at the time for completeness, even if speakers had difficulties producing the correct pronunciation. However, these sequences can be added in future, due to the modular design of the data corpus.

design of the AVOZES data corpus covers all phonemes and visemes in AuE, as well as offering application-driven modules for testing. The AVOZES corpus is also the first AV speech corpus to take advantage of a stereo camera system for recording of the video data, which allows accurate 3D measurements of objects in the scene compared to monocular camera systems limited to 2D measurements.[9] AVOZES currently contains recordings from 20 native speakers of AuE plus four speakers with a different language background, but who have lived in Australia for several years at least. Recordings were made on two occasions about 12 months apart with ten speakers being recorded each time, using the audio laboratory of the Computer Sciences Laboratory at the Australian National University. The sequences of the native AuE speakers in the AVOZES data corpus were used in the statistical analysis of the relationships between audio and video speech parameters described in the following chapters.

[9] Audio and video streams were recorded on DV tape, which offers an easy way of repeated access while maintaining a consistent high quality.

Chapter 5

Analysis of Data Corpus

This chapter addresses the issues of how suitable parameters can be extracted from the audio and video modalities of the speech signal and how these parameters can be analysed in a statistical way. Section 5.1 gives a detailed overview of analysis techniques for audio speech signals to provide background information for the extraction of audio speech parameters. Methods for the extraction of such parameters are also discussed. In Section 5.2, methods for the analysis of the video modality of the speech signal are described. The extraction of geometric parameters describing the 3D shape of the mouth is based on the automatic lip tracking algorithm of Chapter 3. Another (non-geometric) parameter evaluates the visibility of teeth in the mouth opening. The use of dynamic speech parameters is discussed in Section 5.3, because they are often seen as an important part of (visual) speech information (cf. Section 2.1.6). Finally, Sections 5.4 and 5.5 describe the methods used in preprocessing and statistically analysing the audio and video speech parameters. These methods form the core of the analysis work in this study.

5.1 Audio Analysis

One area that the different approaches to audio ASR, discussed in Section 2.3.1, have in common is the need for a signal-processing front end, which converts the speech waveform to some parametric representation, before subsequent analysis and processing. This conversion also achieves a significant reduction in the information

rate of the raw speech signal, which is sampled at high frequencies (e.g. 48kHz in the AVOZES data corpus). The vocal tract system, and the articulators in particular, are under physical constraints, which prevent them from moving too quickly, and hence the raw signal contains many redundancies. By extracting a set of acoustic (or spectral) parameters (or features), the information rate can be greatly reduced. A typical sampling rate for these parameters is 100Hz (or one sample every 10ms).

Over past decades, a lot of research has been carried out involving possible parametric representations of the audio speech signal. General textbooks on automatic speech processing offer more details on these representations, than that can be mentioned here, for example, Rabiner and Juang [Rabiner 93], Rabiner and Schafer [Rabiner 78], Furui [Furui 00], and Flanagan [Flanagan 72]. Representations considered include short time energy, zero crossing rates, level crossing rates and others, but the most common one is the short time *spectral envelope* from spectral analysis. The smoothed spectral envelope is the overall spectral feature, which reflects resonance and antiresonance, as well as radiation characteristics of the vocal tract. The reasons behind the preference for a spectral representation of the speech signal over the original acoustic pressure waveform is, firstly, that speech waves can be reproduced by a sum of sinusoidal waves. Secondly, the critical features for human speech perception are mainly included in the spectral information (frequency and amplitude), with phase information playing a minor role [Furui 00]. Beside the spectral envelope, the spectral *fine structure* plays a role in determining the short-time spectrum, as it corresponds to the periodicity of the sound source, i.e. periodic patterns for voiced sounds and aperiodic patterns for unvoiced sounds.

Spectral Representations

A *spectrogram* is a common way of graphically presenting spectral information. It is a three-dimensional representation of the speech intensity in different frequency bands over time. Figure 5.1[1] shows both a wideband and a narrowband spectrogram with the corresponding speech waveform of the sentence *"You grab BAB beer."*,

[1] The graphs were created by the Entropic Signal Processing System (ESPS), which was used in the audio analysis of this study to extract audio speech parameters.

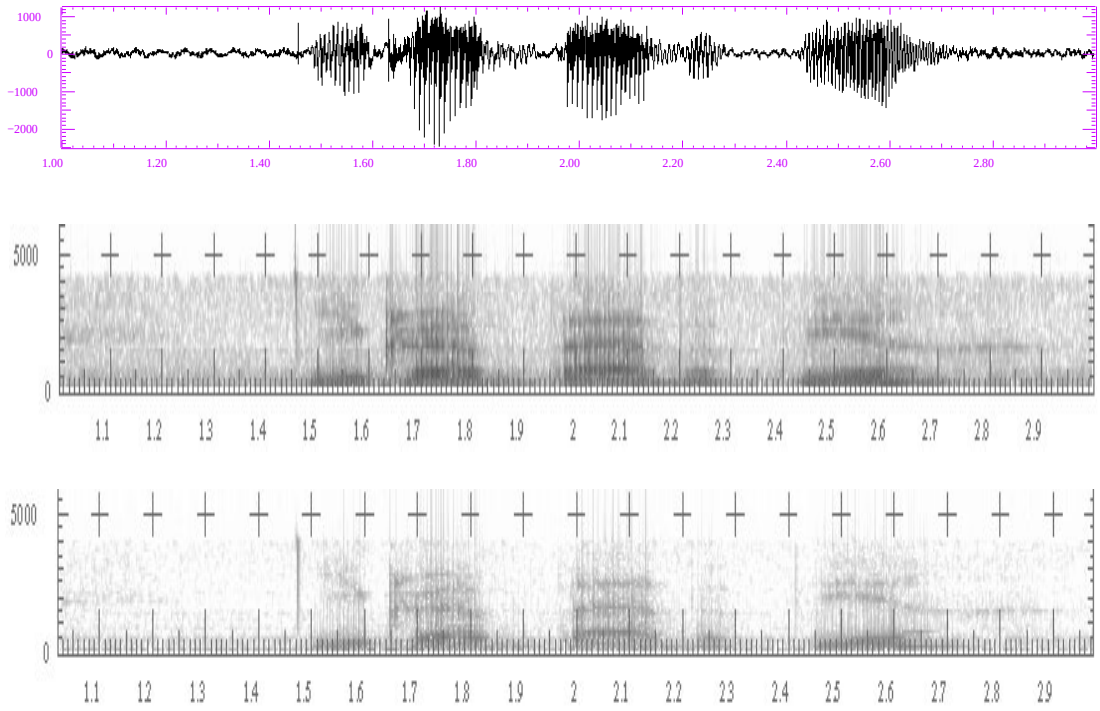


Figure 5.1: Speech waveform (top), wideband spectrogram (centre), and narrowband spectrogram (bottom) of the sentence "You grab BAB beer.", spoken by one speaker from the AVOZES data corpus. Only shown up to a frequency of 5kHz.

spoken by one of the speakers in the AVOZES data corpus. The spectral intensity is indicated by the darkness of the plot at a particular frequency. The wideband spectrogram corresponds to a spectral analysis on 15ms windows of the speech signal with a broad filter of 125Hz bandwidth. The temporal envelope is resolved, which can sometimes be seen as vertical 'lines' in spectrograms.[2] A narrowband spectrogram corresponds to a spectral analysis on 50ms windows with a narrow filter of 40Hz bandwidth. As a result, individual harmonics of the fundamental frequency are resolved and can sometimes be seen as horizontal 'lines' in spectrograms.

Also frequently used is the cepstrum representation. The cepstrum, represented by cepstral coefficients, is defined as the inverse Fourier transform of the short-time

[2] Note, the vertical lines in the voiced regions of the speech signal in Figure 5.1 are artifacts of the signal processing software (ESPS). The signal was lowpass filtered at 4kHz resulting in the visible cut-off of higher frequencies. Bold horizontal lines correspond to the formant frequencies.

logarithmic amplitude spectrum. The advantage of the cepstrum is that it allows for a separate representation of the spectral envelope and the spectral fine structure. Low order cepstral coefficients have also been found to be a more reliable feature set for automatic speech recognition than, for example, LPC coefficients [Rabiner 93].

The remainder of Section 5.1 gives an overview of the methods commonly used for the analysis of the audio speech signal. First, Section 5.1.1 gives useful background information by describing two frequently used methods — the filter bank method and the LPC method — in more detail. The LPC method forms the basis for the parameter extraction algorithms of the ESPS package that were used in this study. These parameter extraction methods are described in Sections 5.1.2 (formant extraction) and 5.1.3 (estimation of voice source excitation). Some necessary preprocessing steps are summarised in Section 5.1.4.

5.1.1 Spectral Analysis Methods

Methods for spectral analysis can be classified into two major classes: *parametric analysis* and *non-parametric analysis* (for example, see Furui [Furui 00] for a discussion). Short-time autocorrelation, short-time spectrum, cepstrum, bandpass filter bank, and zero-crossing analysis are examples of non-parametric methods. Parametric methods are, for example, analysis-by-synthesis and LPC.

Two common spectral analyses are the filter-bank method and LPC, which are discussed briefly below. The general assumption is taken that the speech signal is quasi-stationary for short periods of time.

The Filter Bank Method

The filter bank method applies a set (or 'bank') of Q bandpass filters to the raw speech signal [Rabiner 93]. The outcomes from each filter are passed through a non-linearity to shift the bandpass spectrum to the low-frequency band and a subsequent lowpass filter is applied to eliminate high-frequency noise. The lowpass-filtered signals are resampled at a lower rate, followed by a dynamic range (or amplitude) compression. These final two steps achieve a significant reduction in the data rate.

The filter bank can be *uniform* or *non-uniform*. In the first case, the bandpass filters are centred at frequency f_i with bandwidth b_i

$$f_i = \frac{f_s}{n}\, i, \qquad 1 \leq i \leq Q \tag{5.1}$$

$$b_i \geq \frac{f_s}{n} \tag{5.2}$$

where f_s is the sampling rate of the speech signal and n is the number of uniformly spaced filters required to span the frequency range. Equality in the bandwidth formula corresponds to no frequency overlap between adjacent filters, inequality means that adjacent filters overlap.

In non-uniform filters, the individual filters are spaced in frequency at different distances and with different bandwidths according to some scheme. A common scheme, instead of a linear frequency scale, is a quasi-logarithmic scale, which corresponds to human auditory perception [Stevens 37]. Some variations exist in modelling perceptual aspects (e.g. mel and bark frequency scales[3] (for example, see Davis and Mermelstein [Davis 80]). The set of Q bandpass filters can be defined by

$$b_1 = C \tag{5.3}$$

$$b_i = \alpha\, b_{i-1}, \qquad 2 \leq i \leq Q \tag{5.4}$$

$$f_i = f_1 + \sum_{j=1}^{i-1} b_j + \frac{b_i - b_1}{2} \tag{5.5}$$

where C and f_1 are the arbitrary bandwidth and centre frequency of the first filter, and α is the (logarithmic) growth factor [Rabiner 93].

The implementation of the filter bank depends mostly on the design of the individual filters. Commonly, filters are classified as *finite impulse response (FIR) filters* and *infinite impulse response (IIR) or recursive filters*. FIR filters require more computation in a straightforward implementation than IIR filters. However, FIR filters can be implemented using Fast Fourier Transforms (FFTs), which reduces the amount of computation. FIR filters have the advantage of a linear phase

[3] In mel frequency scale, or similarly in bark frequency scale, the scale is linear below a certain frequency and logarithmic above that frequency.

response, while retaining the ability to approximate ideal magnitude characteristics. Most practical digital filter bank implementations, therefore, use FIR filters in an FFT realisation. The interested reader is referred to Rabiner and Gold [Rabiner 75], and Rabiner and Schafer [Rabiner 78], for example, for a detailed discussion of filter design.

The Linear Predictive Coding Method

The technique of linear prediction of a time series was first introduced in general by Wiener [Wiener 57], and later successfully applied to speech analysis and synthesis independently by Itakura and Saito [Itakura 68], and Atal and Schroeder [Atal 68]. A detailed discussion of applying linear predictive coding (LPC) to speech can be found in Markel and Gray [Markel 76].[4]

The LPC model is based on the assumption that the speech signal with its waveform and spectrum can be represented efficiently and precisely using only a small number of parameters [Furui 00]. In LPC modelling, the vocal tract system is characterised by an all-pole model (for mathematical details, see Markel and Gray [Markel 76]). In the LPC model, a sample $s(t)$ of the speech signal, at time t, can be approximated by a linear combination of the previous p samples

$$s(t) \approx a_1\, s(t-1) + a_2\, s(t-2) + \cdots + a_p\, s(t-p) \tag{5.6}$$

where the coefficients a_1, a_2, ..., a_p are assumed to be constant over the analysis time frame. By introducing a prediction error term $e(t)$, this can be rewritten as

$$s(t) = \tilde{s}(t) + e(t) \tag{5.7}$$

It follows that

$$e(t) = s(t) - \tilde{s}(t) \tag{5.8}$$

$$= s(t) - \sum_{k=1}^{p} a_k\, s(t-k) \quad . \tag{5.9}$$

[4] Linear prediction analysis can also be extended to perceptual linear predictive (PLP) analysis to incorporate concepts from the psychophysics of hearing [Hermansky 90].

The coefficients $a_1, a_2, \ldots, a_p$ can then be determined by minimising the mean-squared prediction error over a short segment of the speech signal. For details on methods to determine the coefficients, such as the autocorrelation method or the covariance method, consult the literature, e.g. Rabiner and Schafer [Rabiner 78], and Rabiner and Juang [Rabiner 93].

LPC models of sufficiently high order provide a good model of the speech signal and, in particular, a good approximation to the vocal tract spectral envelope of the quasi-steady state voiced segments. Unvoiced segments are not as well modelled as voiced segments. LPC models also offer a good separation of sound source and vocal tract, so that the characteristics of the latter, i.e. the formants, can be determined reasonably well. As discussed earlier in Section 2.1.4, different speech sounds are produced by changing the vocal tract shape, which results in a change of the formant frequencies. LPC models have also the advantage that they are mathematically well-defined and, hence, simple and straightforward in their implementation. It has also been shown to be an effective method in ASR applications with similar recognition results to the filter bank method [White 76].

5.1.2 Formant Extraction

As discussed in Section 2.1.4, the formant frequencies are the resonance frequencies of the vocal tract. Different configurations (shape, length) of the vocal tract cause different speech sounds to be produced, resulting in different formant patterns. Thus, the formants are an important cue in the characterisation of speech sounds. Their strong relation to the position of the articulators, which change the shape and length of the vocal tract, make them a very suitable set of acoustic features for the analysis of the relationships between audio and video speech parameters. However, the automatic extraction of formant frequencies is not trivial.

Several methods of automatic formant estimation have been developed over the past decades. Usually, the first three formants are estimated, because they are considered to be the most useful ones to describe speech sounds [Fant 60]. The *analysis-by-synthesis method* [Bell 61, Olive 71] calculates a frequency spectrum

based on a speech production model, which has the fundamental frequency F_0 and the first three formants F_1, F_2, F_3 and their bandwidths as parameters. The spectrum is compared to the measured spectrum, and the parameters are modified until the difference between the two spectra is at a minimum.

The *peak-picking method* estimates the formants by locating the peaks in the (cepstrally) smoothed short-time logarithmic magnitude spectra at each time step [Schafer 70]. These peaks correspond essentially to the formants. The difficulties in the peak-picking method are the detection of spurious peaks and merged formants.

Formant estimation based on the *linear prediction analysis* was, for example, used by Markel ([Markel 72a, Markel 73], summary in [Markel 76]). Root-finding algorithms are employed to determine the zeros of the polynomial resulting from LPC analysis (e.g. Snell and Milinazzo [Snell 93]). Problematic is that the determination of formant frequencies is only successful for complex-conjugate poles and not for real poles [Welling 96]. Nasals and nasalised vowels also cause problems, because oral cavity formants are highly damped, mostly reduced due to nasal zeros, and additional nasal cavity formants are introduced [McCandless 74].

Formant estimation based on LPC analysis has been further improved by continuity constraints [McCandless 74, Seneff 76, Wagner 82], segmental phonemic information [Lee 99], pole enhancement [Duncan 86], hidden Markov models [Acero 99] and vector quantisation [Kopec 86]. Other automatic formant estimation methods are based on a digital resonator [Welling 96], linear-prediction phase spectra (as opposed to amplitude spectra) and group delay functions [Yegnanarayana 78, Murthy 91, Yegnanarayana 98], peak-picking using generalised centroids [Crowe 87], and multiple centroid analysis [Wrench 95].

The first three formants F_1, F_2, and F_3 were used as audio speech parameters in the statistical analyses in this study. Formant estimation was performed by the ESPS command `formant`. The formants were found by solving the roots of the LPC polynomial and imposing frequency continuity constraints [Entropic 93a].

Group	Methods
Waveform processing	Parallel processing Data reduction Zero-crossing count of the (possibly filtered) speech signal
Correlation processing	Autocorrelation Modified correlation Simplified inverse filter tracking Average magnitude differential function
Spectrum processing	Cepstrum Period histogram

Table 5.1: Classification of methods for F_0 estimation [Furui 00]. Details about the individual methods can be found in [Rabiner 76, Hess 83, Furui 00].

5.1.3 Estimation of Voice Source Excitation

The voice source excitation or fundamental frequency F_0 was a further audio speech parameter in this study.[5] F_0 is commonly defined as the frequency of the opening and closing of the glottis. As such, it is well-defined for voiced sounds, but not for unvoiced sounds.

The problem of reliable automatic F_0 estimation is still unsolved. Three main factors can be identified that make this task such a challenge. Firstly, vocal cord vibration is only quasi-periodic. The resulting glottal waveform varies in period, amplitude, and shape, in particular at the beginning and end of voiced segments. Secondly, separating the effects of the vocal tract from the source signal is not a trivial task. Thirdly, the dynamic range of F_0 is very large, ranging from 33–3100Hz, with the range of F_0 in conversational speech being below about 500Hz [Hess 83].

Various techniques have been proposed over the years. Furui [Furui 00] classifies the techniques into three major groups: (1) *waveform processing*, (2) *correlation*

[5] The term 'pitch' is often used synonymously in the literature, although pitch refers to the sensation of the physical property fundamental frequency [Fry 79].

processing, and (3) *spectrum processing*. Table 5.1 shows how individual methods fall into these three groups.[6] A comparative study on the performance of these methods can be found in [Rabiner 76]. The most widely used methods are the autocorrelation method [Dubnowski 76, Rabiner 77], the simplified inverse filter tracking (SIFT) method [Markel 72b], and the cepstrum method [Schafer 70].

The ESPS command `get_f0` was used to extract the F_0 parameter [Entropic 93a]. The algorithm falls into the correlation processing category and is based on applying a normalised cross-correlation function to an LPC analysis residual, plus dynamic programming as a postprocessing tool to obtain smooth contours [Secrest 83]. In addition to the F_0 value, `get_f0` also delivers the RMS energy value (see Section 2.1.4), which was used as a further audio speech parameter in the statistical analyses in this study.

5.1.4 Preprocessing

As discussed in Chapter 4, the AVOZES data corpus was recorded on DV tape. For the audio speech parameter extraction, the sequences were first processed with the Apple Final Cut Pro software package. In this process, the audio signal was separated from the video signal and stored as an individual WAV-file for each sequence.

These files were then processed with the ESPS software, by the commands mentioned in the previous two subsections, for the extraction of formant frequencies, F_0 frequency, and RMS value. Prior to the parameter extraction, the audio signal was lowpass filtered with a cut-off frequency at 4kHz to filter out unwanted high frequencies due to the studio's air-conditioning system, as well as from the internal cooling fans of the recording equipment. A lowpass filter up to 4kHz is generally sufficient to cover the first three formant frequencies [Rabiner 93]. The filter was a linear phase FIR filter based on Kaiser windowing, created by the ESPS command `win_filt` [Entropic 93b]. The passband ranged from 0 to 4000Hz, with a 20dB rejection ratio to the stopband (Figure 5.2).

[6] Rather than giving individual references to each method, the interested reader is referred to [Rabiner 76, Hess 83, Furui 00] for reviews of the methods.

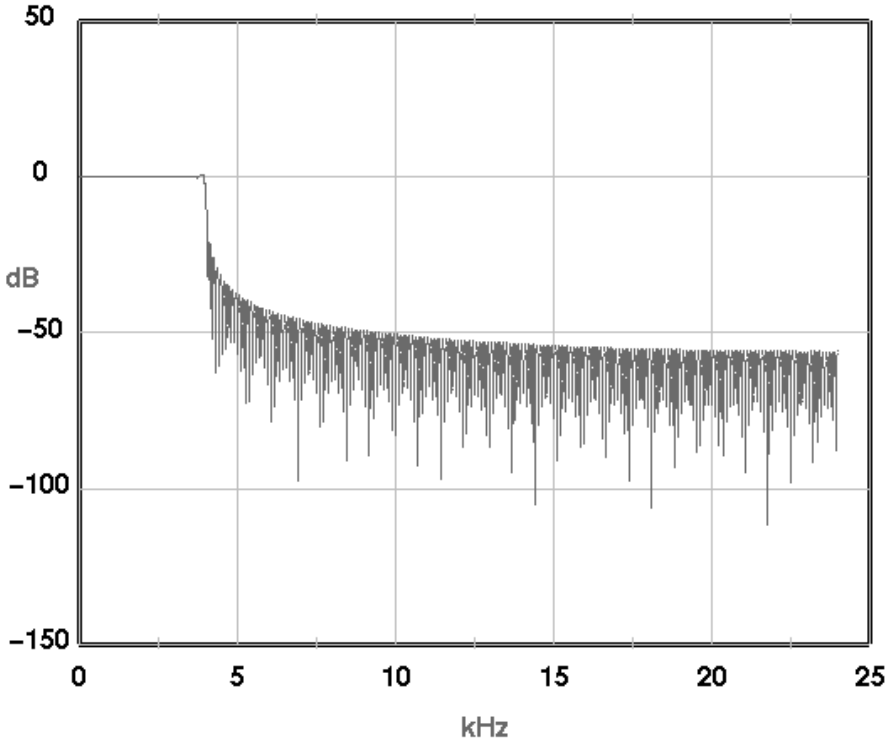


Figure 5.2: The 4kHz lowpass FIR filter used to reduce high frequency noise.

5.2 Video Analysis

In this section, the video speech parameters used in the statistical analyses in Chapter 6 are derived. Section 5.2.1 describes the geometric parameters, which are based on the lip feature points tracked by the lip tracking algorithm. A numerical parameter describing the visibility of teeth in the video frames complements the geometric parameters (Section 5.2.2).

5.2.1 Geometric Parameters

While a lot of research over the past decades has focused on finding parameters that describe the audio part of the speech signal, the issue of which parameters describe the video part of the speech signal best largely remains unclear (cf. Section 2.1.6). A geometric explicit feature extraction approach was followed in the work described in this thesis. As described in Chapter 3, a new lip tracking algorithm based on a stereo vision face tracking system was developed in the course of this study. This

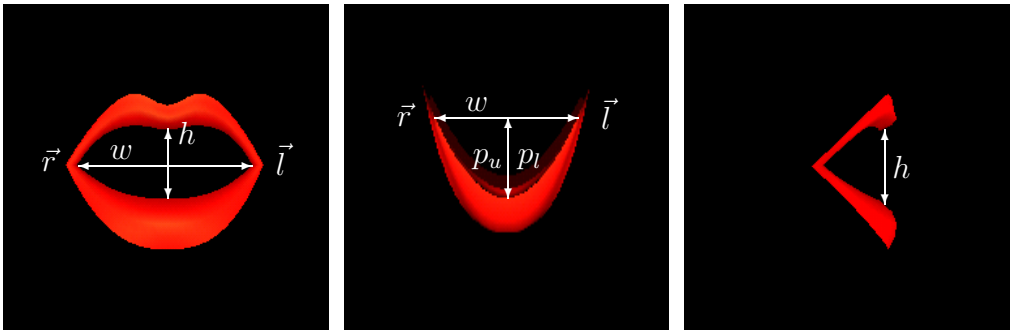
Figure 5.3: Examples of the four lip feature points being tracked automatically.

algorithm is able to determine the 3D positions of four lip feature points relative to the origin of the world coordinate system of the stereo camera system. These four lip feature points are the two lip corners as well as the midpoint of upper and lower lip located on the inner lip contour (Figure 5.3).

From these four points, geometric parameters which describe the shape of the mouth can easily be calculated. If the 3D coordinates of the speaker's left and right lip corner are denoted with $\vec{l}$ and $\vec{r}$, respectively, and the midpoint of upper and lower lip with $\vec{m}_u$ and $\vec{m}_l$, respectively, then geometric parameters can be defined by (see also Figure 5.4)

- internal width of the mouth opening $MW = w = \|\vec{l} - \vec{r}\|$
- internal height of the mouth opening $MH = h = \|\vec{m}_u - \vec{m}_l\|$
- protrusion of upper lip $PUL = p_u = \|\frac{\vec{l}+\vec{r}}{2} - \vec{m}_u\|$
- protrusion of lower lip $PLL = p_l = \|\frac{\vec{l}+\vec{r}}{2} - \vec{m}_l\|$

where $\|\cdot\|$ denotes the Euclidean norm. The width and height parameter come with a confidence value based on the confidence measures derived in Section 3.3.6. To explain the protrusion parameters in more detail, these parameters were calculated as the distance between the midpoint of the vector from the left to the right lip corner and the midpoint of upper or lower lip (Figure 5.4 (b)). This is clearly an approximation, but it was a reasonable choice because the 3D positions were taken into account. Finding a reference point for the protrusion of the lips is not a trivial task. It was decided to choose the relative protrusion of the lip midpoints compared to the contact points of upper and lower lip (or lip corners). As an alternative, a simplified way of determining the protrusion would be to use the 'depth' value (z



(a) (b) (c)

Figure 5.4: The geometric parameters describing the mouth shape as viewed from front (a), above (b), and in profile (c).

element of 3D point vectors) of the midpoints and the lip corners relative to the cameras, and to determine their difference, which gives useful values under the assumption that the speaker's face is in full frontal view.

5.2.2 Teeth Visibility Parameters

In addition to the geometric parameters, a parameter describing the visibility of teeth was also extracted, because teeth visibility depends on the mouth shape (openness and lip position) and, therefore, can be considered as a useful cue for determining the mouth shape, for example, due to the strong contrast between relatively bright teeth and the lips. The lip tracking system automatically checks for teeth in the video frames as part of the algorithm. Hence, it would be simple to derive binary parameters that describe the visibility of upper and lower teeth.

More useful for the statistical analyses in Chapter 6 than binary parameters was a numeric parameter. Such a numeric parameter was provided in the form of the novel *relative teeth count* (RTC). This parameter performs a pixelwise test for teeth on the rectangular area A spanned by the four lip feature points in each video frame. However, simply taking the number of teeth pixels counted results in a flawed parameter, because the number of pixels visible would be affected not only by the shape of the mouth, but also by the distance of the speaker's face to the cameras. To overcome this problem, the relative teeth count RTC was defined as

the average of the number of teeth pixels in A in the left and right camera image, divided by the distance d to the cameras (as measured for the left lip corner)

$$RTC \;=\; \frac{c_l^A \;+\; c_r^A}{2} \;/\; d \quad . \tag{5.10}$$

The height confidence measure was also applied to RTC, because the correct determination of the lip midpoints was the major factor in obtaining a correct RTC.

5.3 Dynamic Speech Parameters

The audio and video speech parameters discussed so far are static parameters, that is, they describe the situation at a particular point in time but not the dynamic patterns that underlie the change from one situation to the next. As discussed in Section 2.1.6, some researchers consider dynamic parameters, particularly of the video modality, at least as important as static parameters. This view was derived from studying human speech processing. As a consequence, it is important to also study dynamic speech parameters (the velocity and acceleration patterns) for speech processing by machines.

This area warrants further investigation. It is recommended to investigate the relationships of such dynamic audio and video speech parameters in future work, so that a comparison of the AV relationships for static and dynamic speech parameters may be performed.

5.4 Audio-Video Analysis — Preprocessing

Before the various statistical analyses could be performed, several preprocessing steps had to be carried out. The important issue of AV synchronisation is desribed in Section 5.4.1. Details of the already mentioned smoothing method used to lessen the effect of measurement errors can be found in Section 5.4.2. Another preprocessing step was the creation of the same number of samples for all parameters (Section 5.4.3), so as to enable certain statistical analyses of the sequences of the AVOZES data corpus, which require the same number of samples for all parameters.

5.4.1 Audio-Video Synchronisation

It is of outstanding importance for the statistical analysis of audio and video speech parameters that the two signals are in synchronisation. Both signals were recorded at the same time on DV tape and recordings were made in locked audio mode, which means that the audio samples are precisely locked to the video frames (see Appendix A). Thus, synchronisation is inherent in the use of DV equipment.

However, a delay in the order of the duration of one NTSC video frame (about 33ms, see Appendix A for details) occurred for the video signal during the recording stage. While the microphone was directly plugged into the DV recorder, the cameras' output signals travelled through the multiplexer and the video card, where the interlaced fields in the multiplexed video frames were separated again, so that one camera image is shown in the top half and the other one in the bottom half of each video frame (cf. Section 3.1.1 and Figure 3.5). The delay between the video and audio signals must be allowed for in the statistical analyses.

The information given in Appendix A also highlights that during the first second of every minute of recorded video, only 28 video frames are taken in the NTSC video format, due to the NTSC frame rate being 29.97Hz and not exactly 30Hz. In order to consistently have 30 video frames per second for the statistical analyses, resampling by piecewise linear interpolation was performed on the extracted video speech parameters when only 28 frames occurred.

5.4.2 Smoothing

Measurements of the audio and video speech parameters contain an error component (see Section 3.1.4). As a result, the parameter graphs can show incorrect large variation, although the underlying functions are actually smooth. Techniques for smoothing the data are therefore commonly applied and they were also applied here. A second purpose of smoothing is that restrictions to the smoothness can be applied, for example, to ensure smooth derivatives. This would be of importance for dynamic speech parameters (see Section 5.3). For example, smoothness up to the second derivative can be ensured by putting restrictions on the fourth derivative.

Smoothing is often done with splines using polynomial base functions, most commonly cubic polynomials [Eubank 88, de Boor 01]. As the statistical analyses in this study were performed with the R statistical software system [Ihaka 96], its cubic spline smoothing functions were used, in particular the functions `smooth.spline` in the standard distribution package 'modreg' and `smooth.basis` in Ramsay's functional data analysis package[7] 'fda' (see Section 5.5.7).

The basic idea of spline smoothing is to define a function x that fits the observed data for coordinate X, subject to a penalty placed on the roughness (or lack of smoothness) of x [Eubank 88]. The penalty term keeps function x from fitting the data precisely, while ensuring the desired amount of smoothness. A common way of determining the spline function is by least squares approximation, which also forms the basis of the smoothing criterion suggested by Ramsay [Ramsay 96]. Ramsay's smoothing criterion, which includes an additional penalty term ensuring smooth second derivatives, was also used in this study. It is defined as

$$\min \sum_i w_i \left(X_i - x(t_i)\right)^2 + \lambda \int \left(\frac{d^4x}{dt^4}\right)^2 dt \tag{5.11}$$

where i iterates through the samples, w_i are weights, t is the time parameter and λ is a (non-negative) penalty factor. The first term is the least squares approximation and without the second term — the penalty term — it would be possible to find an exact fit to the data ($\lambda = 0$). The penalty term measures the smoothness of the function x and it forces x to give up some fitting power in order to remain smooth. λ typically takes small values and a value of 0.001 was used in this study.

The weights w_i enable the determination of spline functions, which are pulled more strongly to some measurement values than to others, depending on the ratio of weights. This can be useful for smoothing the parameters with confidence measures, such as the video speech parameters mouth width MW and mouth height MH. Sample points with high confidence values are more closely fitted than those with low confidence values. For parameters without confidence measures, equal weights of $w_i = 1$ can be applied, thus fitting all sample points equally well.

[7] The package can be downloaded from Ramsay's homepage at McGill University, Montreal, Canada, URL = http://www.psych.mcgill.ca/faculty/ramsay/ramsay.html .

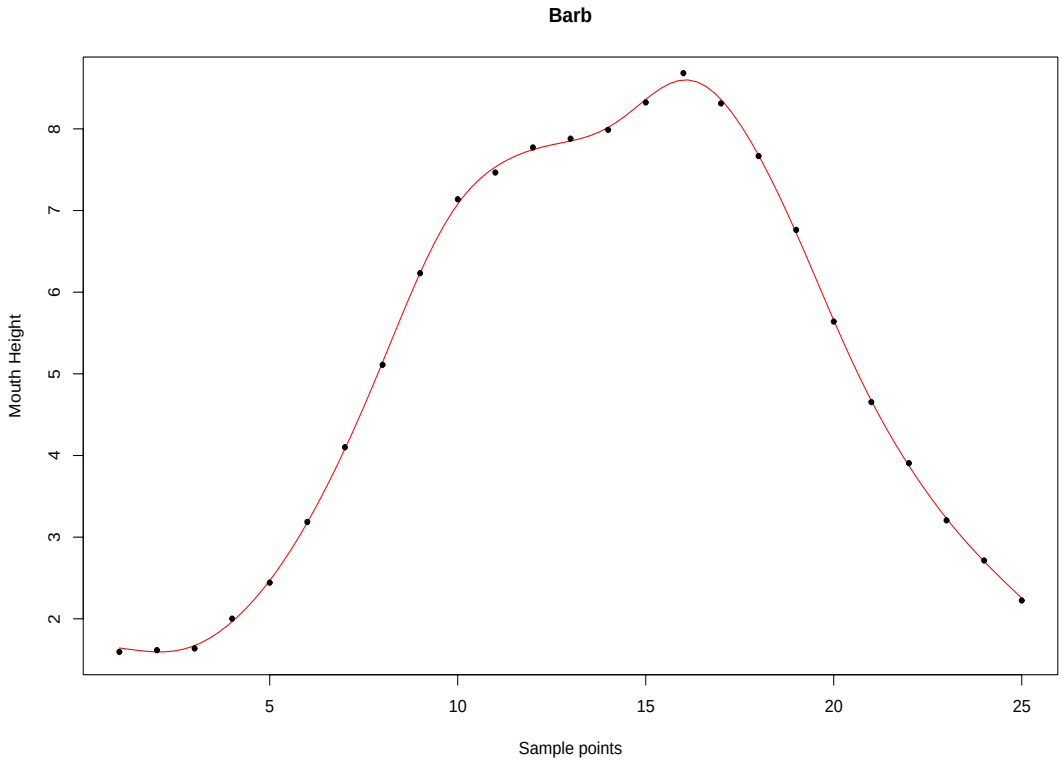


Figure 5.5: An example of a smoothed mouth height parameter curve. Black dots refer to the measurement values. The red solid line shows the smoothed curve.

Smoothing was applied to all audio and video speech parameters. Weights of $w_i = 1$ were applied to all audio speech parameters during smoothing, because confidence measures were not available. The confidence measures α_w and α_h (defined in Section 3.3.6) were applied in smoothing the video speech parameters MW and MH, respectively. The confidence measure α_h was also used in smoothing the teeth visibility parameter RTC, because it is a reasonable assumption that teeth visibility and mouth height are statistically related. The protrusion parameters PUL and PLL were smoothed with weights of $w_i = 1$.

5.4.3 Establishing the Same Number of Samples

In order to facilitate the statistical analysis, all audio and video speech parameters were resampled to the same number of sample points. Audio speech parameters were measured at a rate of 100Hz and video parameters were extracted from the data at

30Hz. For each sequence[8] under investigation, these parameters were resampled to 25 sample points on the time axis, which on average corresponded to about 0.3s of the speech signal.

Resampling in the R system can easily be done by either linear interpolation of the sample points using the command `approx`, or by evaluating the spline functions using either the `predict` command of the 'modreg' package or the `eval.fd` command of the 'fda' package (Section 5.5.7). The latter was used for the functional data analysis, while `approx` was employed for all other analyses. In either case, the values of the resampled points on the time scale formed the input and the function values at these points were returned.

Why was it necessary to resample the measured parameter values? First of all, even for the same phoneme[9], different speakers produce it with different length. These inter-speaker differences must be accounted for. The statistical analyses described in the next section require the input data vectors to be of the same length. Secondly, intra-speaker differences in the length of a phoneme can occur, when a speaker repeats a word. Since the AVOZES corpus currently does not contain

[8] Sequence here means the sample points pertaining to the central phoneme of the CVC- and VCV-words under investigation, for example, one of the vowels. In this study, the length of the central phoneme was defined as the time between the start sample point and the end sample point, both defined by analysing the MH parameter. In the CVC-words (/bVb/), the start sample point was defined by the lip closure before the central vocalic phoneme and the end sample point by the lip closure after the central vocalic phoneme. In the VCV-words (/ɑːCɑː/), the start sample point was marked by the local maximum in the MH parameter, due to the wide open /ɑː/, before the central consonantal phoneme and the end sample point similarly by the local maximum afterwards.

However, coarticulation effects (see Section 2.1.4) occurred and may have altered the results of the analyses to some extent, as compared to results for single phonemes, due consonant-to-vowel and vowel-to-consonant transition sample points being included. Coarticulation is a naturally occurring phenomenon of speech production and it is, therefore, important to study AV relationships in syllables, that contain transitions between vocalic and consonantal phonemes, as well as in isolated phonemes. The results of this study on two particular contexts are hoped to serve as a reference for future studies.

[9] Phoneme here and in the remainder of this thesis, when talking about statistical analyses, refers to the central phoneme in the CVC- and VCV-words recorded in the AVOZES data corpus.

repetitions of any sequence by a speaker, intra-speaker differences did not need to be considered in the analysis, but are mentioned here for completeness. Thirdly, different phonemes could also be produced with different lengths. That in itself is a way of distinguishing phonemes, for example, short and long vowels. However, if the values of the same parameter for different phonemes were to be compared, it is more helpful to have the same number of sample points and to remove the individual temporal information. Since phonemes were studied in CVC- and VCV-contexts, it is important to note that short central phonemes had potentially more sampling of context-related information occurring than long phonemes, which may have affected the results.

5.4.4 Outlier Analysis

Finally, an outlier analysis was performed before the statistical analyses, so that the number of outliers and their potential influence on the results could be judged and actions taken to counter the effects, where deemed necessary. Outliers are commonly defined to be measurement values outside a certain range centred at the mean value. This range is often defined by multiples of the standard deviation and was set to 3 standard deviations in the outlier analysis in this study. Three ways of handling outliers are common practice

1. outliers are deemed to have a minor or negligible influence on the results and are left unchanged,

2. outliers are substituted by the mean of the measurement values from all other observations at the same sample point, which are not outliers themselves, and

3. outliers and the entire parameter curve they belong to are completely discarded, i.e. also the values at other sample points, which are perhaps well within the accepted range, are discarded from the analysis.

The first two ways have the advantage that the sample size is not decreased. The decision of not changing any outliers can only be made after the number and magnitude of outliers has been established. For the second way, it is often argued that

introducing a sample that is not actually measured, even if it is given the mean value, has an effect on the analysis that cannot be judged easily. On the contrary, the third way has the advantage of not adding any samples that were not measured, but has the disadvantage of reducing the sample size by discarding outliers.

The results of the outlier analysis, and the discussion on which way of treating outliers was chosen for this study, can be found in Section 6.2.

5.5 Audio-Video Analysis — Statistical Analyses

In this section, the theoretical background of the statistical analyses performed in this study and details on their application for investigating the relationships between audio and video speech parameters are given. Of particular interest were multivariate analyses because they describe the relationship between two and more parameters. Sections 5.5.1 – 5.5.6 describe the analyses used here, such as linear discriminant analysis (LDA), principal component analysis (PCA), pairwise linear correlation analysis, canonical correlation analysis (CANCOR), and coinertia analysis (COIA). In Section 5.5.7, functional data analysis (FDA) is introduced, in which the data are treated as functions. Curve registration and PCA using FDA are described.

5.5.1 Multivariate Analysis (MVA) — Introduction

The investigation of the statistical relationship between various parameters (or statistical variables[10]) is a classic case of MVA (see for example [Mardia 79, Rencher 98]). MVA deals with data containing observations of p parameters measured on a set of n objects. In the work described in this thesis, the audio and video speech parameters were measured on the set of phonemes, with each element of that set containing one sample sequence (= observation) per speaker. Each observation can be written

[10] The term parameter is used synonymously for statistical variable here and in the remainder of this thesis.

as a vector

$$\mathbf{X}_{i,j}(t_i) \;=\; \begin{pmatrix} x_{1,j}(t_i) \\ x_{2,j}(t_i) \\ \vdots \\ x_{p,j}(t_i) \end{pmatrix} \tag{5.12}$$

where t_i represents the time value of the particular observation, i iterates through the observations $\mathbf{X}_{i,j}$ and j iterates through the set of n parameters.

Many techniques in MVA are extensions of univariate analysis techniques. In the following, only the techniques used in this project are described briefly. The interested reader is referred to statistical textbooks, for example Mardia *et al.* [Mardia 79], Rencher [Rencher 98], or Venables and Ripley [Venables 99] for more details.

5.5.2 MVA — Linear Discriminant Analysis

Discriminant analysis classifies an object into one of n groups based on the information from a set of p parameters. In other words, discriminant analysis aims to determine which parameters discriminate between two or more groups occurring in the observations. In the present study, the aim of the discriminant analysis was to explore the data space. Of interest was how LDA separated the phonemes into classes, which were possibly useful for the subsequent analyses of the relationships between audio and video speech parameters. The analysis was done by performing a LDA separately on the sets of vocalic and consonantal phonemes.[11] The vocalic phonemes were further split into groups of short vowels, long vowels, and diphthongs.

One of the most popular methods in discriminant analysis is *Fisher's LDA* [Mardia 79]. The idea is to find a linear combination of the original parameters

$$z \;=\; a_1 x_1 \;+\; a_2 x_2 \;+\; \cdots \;+\; a_p x_p \;=\; \mathbf{a}^T \mathbf{X} \tag{5.13}$$

that exhibits the largest ratio of between-class variance to within-class variance.

[11] Just to remind the reader, these phonemes were the central phonemes in the /bVb/- and /ɑːCɑː/-words.

z is known as Fisher's discriminant function or first canonical variate. Once z is known, an observation $\mathbf{X}$ can be allocated to one of the groups by its *discriminant score* $\mathbf{a}^T\mathbf{X}$. LDA is similar to the (multivariate) analysis of variance (ANOVA / MANOVA) in that it is based on the analysis of means and variances.

LDA was performed with the R command `lda` of the 'MASS' package. The results can be found in Section 6.3.

5.5.3 MVA — Principal Component Analysis

Generally, PCA seeks to find a few linear combinations of the original p parameters that explain as much of the total sample variance as possible. These linear combinations are called *principal components (PCs)*. The first PC is defined as the linear combination with maximal sample variance, the second PC as the orthogonal linear combination with the second largest variance, and so on. In mathematical terms, for an observation vector $\mathbf{X}$ in a sample, the linear combination

$$z \;=\; a_1x_1 \;+\; a_2x_2 \;+\; \cdots \;+\; a_px_p \;=\; \mathbf{a}^T\mathbf{X} \tag{5.14}$$

is sought with sample variance

$$s_z^2 \;=\; \mathbf{a}^T\mathbf{S}\,\mathbf{a} \tag{5.15}$$

where $\mathbf{S}$ is the sample covariance matrix and $a \;=\; (a_1, a_2, \ldots, a_p)^T$. This is equivalent to a rotation of the coordinate system.

The PCs correspond to the p eigenvalues of $\mathbf{S}$, with the first PC corresponding to the largest eigenvalue, and so on. The corresponding eigenvectors define the directions of variance and these directions are orthogonal to each other. They can be used to define a new coordinate system with the PCs as axes. If PCs exist which explain a large part of the variance, they can be used to reduce the dimensionality of the set of parameters. In this study, PCA was used first of all to check if the set of audio and video speech parameters contained redundancies.

Secondly, PCA has also gained influence in statistical shape analysis in recent years. A number of researchers have applied PCA to determine the main modes of variation of 'shapes', for example Cootes *et al.* [Cootes 95], Golland *et*

al. [Golland 99], and Golland [Golland 01]. Such shapes could be the contour lines of an object in an image but also parameter curves. The latter was used in this study, that is, PCA was performed in the temporal domain. Doing so required the application of PCA individually to the set of curves (= one observation per speaker) of each parameter for each phoneme. For each phoneme-parameter pair, PCA was thus applied across all speakers. As a result, the parts of the graphs which exhibit least and most variation were determined. The resulting PCs describe the main modes of variation and provide a compact representation of the individual parameter curves for the subsequent analyses, as usually a few PCs (say 2–3) are sufficient to extract most of the variance (say 95%).

PCA was performed by the R command `prcomp` of the package 'mva'. The results can be found in Sections 6.4 (redundancy test) and 6.5 (shape analysis).

5.5.4 MVA — Pairwise Linear Correlation Analysis

One of the most common analyses between two parameters is a pairwise linear correlation analysis. It determines the extent, measured by the correlation coefficient r, to which values of the two parameters are 'proportional' or linearly related to each other. The coefficient r ranges from -1 to $+1$, where -1 indicates that the two parameters vary exactly in opposite direction to each other and $+1$ means that they vary accordingly. A coefficient of $r = 0$ means that the two parameters vary with complete independence of each other. In practice, these extreme values of the correlation coefficient are rarely reached, so that it is customary to speak of strong correlations for $|r| \geq 0.75$ and weak correlations for $|r| \approx 0.5$.

There are various ways of computing a correlation coefficient r, the most common one being Pearson's correlation coefficient

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \cdot \sum_i (y_i - \bar{y})^2}} \tag{5.16}$$

where x_i and y_i are the samples, and $\bar{x}$ and $\bar{y}$ are the mean values [Göhler 87]. Pairwise linear correlation was performed using the R command `cor`. It was first of all applied within the sets of audio and video speech parameters together with

PCA, as described in the previous subsection, to check for redundancies in the parameter sets. The results are presented in Section 6.4. Strong correlations between two parameters within a set are an indication of redundancy. It is sufficient to consider only one of the redundant parameters in any further analyses. Secondly, pairwise linear correlation analysis was also applied between parameter pairs with one parameter each from the audio and video speech parameters. The results are shown in Section 6.6.1. A strong correlation would support a hypothesis of a linear 1–1 relationship between (some) audio and video speech parameters.

5.5.5 MVA — Canonical Correlation Analysis

While PCA considers relationships within a set of parameters, CANCOR is a statistical analysis for the exploration of relationships between two sets of parameters. Such a statistical analysis is very useful to test the hypothesis that combinations of parameters correlate better across the two modalities than single parameters. The observations of the sets of audio ($\mathbf{X_A}$) and video ($\mathbf{X_V}$) speech parameters formed the input of the CANCOR, which was performed separately for each phoneme.

CANCOR is a generalisation of multiple correlation analysis for sets of parameters with at least two parameters in each set [Hotelling 36, Gittins 85]. It allows the simultaneous analysis of both parameter sets. In general, all the information about linear relationships within and between parameter sets is summarised in the covariance matrix, which is identical to the correlation matrix for normalised (zero mean, unit variance) parameters [Gittins 85].

The correlation matrix forms the starting point for a CANCOR. Similar to PCA, a rotation of the coordinate system is performed, but instead of maximising the variance within a single set of variables as in PCA, the correlation between two sets of parameters is maximised in CANCOR. As a result, the linear relationships within each set are disentangled, so that these relationships between the sets become clear [Gittins 85]. The variables — called *canonical variates (CV)* — in the new coordinate system are linear combinations of the parameters in each set

$$\eta = \mathbf{a}^T\mathbf{X_A} \qquad \phi = \mathbf{b}^T\mathbf{X_V} \tag{5.17}$$

and found such that η and ϕ are maximally correlated [Gittins 85, Mardia 79].

Canonical correlations are computed via the eigenvalues. Let us denote the number of parameters in the two sets by p and q, respectively. The square root of the eigenvalues gives the canonical correlation coefficients r_k, with $k \in [1, p+q]$. The coefficients describe the linear relationships between the two sets and range from -1 to $+1$. It is customary to order the r_k's from highest to lowest and to report only the absolute values. There may be more than one canonical correlation relating the two sets and each such correlation represents a different, orthogonal dimension by which the two sets are related to each other. Typically, only the first 2–3 r_k's are of interest as levels of correlation decrease quickly. The eigenvalues (or r_k^2) can also be interpreted as the proportion of variation in one canonical variate predicted from its conjugate canonical variate [Gittins 85].

For small samples, where the number of parameters $p+q$ approaches the sample size N, the value of the highest canonical correlation r_1 quickly tends towards 1 (see [Gittins 85] for a detailed explanation). In such circumstances, the results can suffer from statistical instability due to collinearity. Canonical correlation coefficients computed in such cases can be misleading with respect to the extent of linear relationship between the linear combinations — the canonical variates — in question.

As the interpretation of canonical weights and correlation coefficients is often difficult, other methods of interpreting the CANCOR have been developed [Gittins 85]. These are structure correlations, variance extracted by a canonical variate, redundancy, and total redundancy. These were also analysed in the study of AV relationships described in this thesis.

Canonical correlation was calculated in R by using the command `cancor` of the 'mva' package. The results can be found in Section 6.6.2.

5.5.6 MVA — Coinertia Analysis

COIA offers a way to overcome the stability problems exhibited by CANCOR. It is a relatively new multivariate statistical analysis, introduced for ecological studies

by Dolédec and Chessel in 1994 [Dolédec 94]. Here, the term 'inertia' is used as a synonym for variability. The method is related to other multivariate analyses such as canonical correspondence analysis (CCA) [ter Braak 86], redundancy analysis (RDA) [Rao 64], and the just discussed CANCOR. COIA is a multivariate method for coupling two (or more) sets of parameters. It gives insight into the relationship between the two sets by analysing linear combinations of the parameters in each set. COIA is a generalisation of the inter-battery analysis by Tucker [Tucker 58], which in turn is the first step of partial least squares methods [Höskuldsson 88].

In many aspects, COIA is very similar to CCA and CANCOR. It also transfers the data into a new coordinate system and the new variables are linear combinations of the parameters in each set. However, in COIA, instead of the correlation between the two sets, the square covariance is maximised, which can be decomposed as

$$cov(A,V) \;=\; corr(A,V) * \sqrt{var(A)} * \sqrt{var(V)} \quad . \tag{5.18}$$

In other words, COIA finds a mathematical compromise between the correlation ($corr(A,V)$), the variance in the audio set ($var(A)$), and the variance in the video set ($var(V)$). COIA can also be seen as aiming to find orthogonal vectors — the coinertia axes — in the two sets which maximise the coinertia value. The number of axes is equivalent to the rank of the covariance matrix. It is common practice to order the axes according to the covariance value from highest to lowest.

The advantage of COIA is its numerical stability and the fact that the number of parameters relative to the sample size does not affect the accuracy and stability of the results [Dolédec 94]. The results of the method do not suffer in the presence of collinearity and the consistency between the correlation and the coefficients is very good, according to Dray *et al.* [Dray 03], which makes it a particularly well-suited multivariate analysis in this study of AV relationships.[12]

The coinertia value (covariance value) is a global measure of the co-structure in the two sets. If the value is high, the two parameter sets vary in a dependent

[12] This was discussed in great detail with Stéphane Dray from the Biométrie et Biologie évolutive - Equipe "Écologie Statistique", Université Claude Bernard Lyon 1, Villeurbanne Cedex, France.

fashion, and if the value is low, the sets vary independently. The correlation value gives a measure of the correlation between the coinertia vectors of both domains. Furthermore, one can project the variance onto the new vectors (or axes) of each set and then compare the projected variance of the separate analyses with the variance from the COIA (see the appendix of [Dolédec 94] for the theory and [Dolédec 97] for an example). The ratio of the projected variance from the separate analyses to the variance from the COIA is a measure of the amount of variance of a parameter set that is explained by each coinertia vector. It is important to compare the sum of vectors, not vector by vector, because the variance projected onto the second vector depends on what is projected onto the first vector, and so on. Often it is sufficient to analyse the first 1–3 vectors, because they typically account for 90–95% of the variance. In addition, COIA computes the weights (coefficients) of the parameters in the linear combinations of each set. The weights show which parameters contribute to the common structure of the two sets and which do not. An overall correlation value between the two sets is given by the RV coefficient (see Robert and Escoufier [Robert 76], and Heo and Gabriel [Heo 97] for more details).

COIA has the advantage that it can be coupled easily with other statistical methods, such as CCA and PCA. That is, these methods are performed on the data of the two domains separately, and then a COIA follows. In fact, Dray *et al.* [Dray 03] show that, seen in this context, COIA is a generalisation of many multivariate methods. For the analysis reported in this thesis, it means that the PCs resulting from the PCA performed to find the main modes of variation can be used as input to COIA.

COIA can be computed with the ADE-4 tool [Thioulouse 97] and is also available on the R statistical platform with the command `coinertia` in the 'ade4' package, which was used in the present study. The results are shown in Section 6.6.3.

5.5.7 Functional Data Analysis

FDA, developed by Ramsay *et al.* [Ramsay 82, Besse 86, Ramsay 96], is an alternative approach, in which the traditional MVAs are expressed in functional analytic

terms. For example, statistical analyses such as analysis of variance and principal component analysis are also available in FDA. As is the case in MVA, the data contain observations of p parameters measured on a set of n objects (see Section 5.5.1). In FDA, the data are viewed as p functions, each observed at m argument values. This view is particularly useful for time series data, such as the data extracted from the AVOZES data corpus.

FDA has the advantage that it accounts for the underlying continuity of the human speech production system. Temporal dependencies in the data are shown [Ramsay 82]. FDA also offers the possibility of studying the variation among derivatives of functions, thereby enabling the exploration of velocity and acceleration patterns in the speech parameters. Doing so requires the smoothing of parameters (see Section 5.4.2) up to the second derivatives.

Curve Registration

Ramsay *et al.* [Ramsay 96] argue that differences between curves for a particular parameter tend to be a combination of end-point variation and shape variation. End-point variation can be interpreted as a result of the individual speaker differences, e.g. the rest positions of the articulators, and coarticulation effects (see 'Coarticulation' in Section 2.1.4). As the same coarticulation context (/bVb/) was used for all CVC-words and the same context (/ɑːCɑː/) for all VCV-words, end-point variation in the data is considered to be mostly due to speaker characteristics.

On the other hand, shape variation is considered to be mostly due to differences in the parameters for the different phonemes, i.e. different vocal tract configurations based on different articulator positions and speaker characteristics. Shape variation is of more interest in the analysis of AV relationships, because it enables studying similarities and differences between phonemes.

Studying shape variation is facilitated by *curve registration*. Curve registration can be summarised as transforming the arguments — the sample points — of curves, so as to align various salient features. As Ramsay [Ramsay 03] points out, the problem has received much attention in various fields and is also known under

the term *time warping*. Warping functions $h_i(t)$ are determined which define how the arguments are transformed based on optimising some criterion [Ramsay 01]. Registration can be done by marker (landmark) registration or global registration. In the first approach, curves are aligned by identifying the timing of salient features, such as peaks or valleys. Curve registration is then achieved by computing and applying a warping function based on the marker alignment. Dynamic time warping [Sakoe 78] is one prominent example. In the second approach, the entire curve is used. This is achieved by optimising some similarity measure of the curve shapes. Ramsay [Ramsay 03] proposes to minimise the logarithm of the smallest eigenvalue of the cross-product matrix of the target curve and the curve to be registered, as it creates better registration results than, for example, a least-squares measure. See [Ramsay 01] for more details on both registration approaches.

In the work presented here, a global registration method based on FDA was used. Global curve registration works well when salient features (maxima, minima) are present, but can fail when the target curve is flat. Another important issue is that of chosing a suitable target curve to which all curves are registered. If for each parameter and for each /bVb/- or /ɑːCɑː/-word one of the curves from the set of speakers is chosen as target curve, then the question arises, which one to choose for the best results. Hence, in this work, the (pointwise) mean curve of all speakers was selected as target curve and all other curves were registered to it.

It would be simple to select the mean curve for each parameter and sequence as target curve and to register all speakers' curves to it, but doing so has the potential problem of creating different warping functions for the same speaker and the same sequence for different parameters, so that the time point t refers to different points on the curves for the same syllable. To avoid such problems and to keep the parameters time-synchronised, the following process[13] was performed:

For each syllable

1. The *RMS* parameter served as registration target. All *RMS* parame-

[13] This process is based on the theory of curve registration using FDA as detailed in [Ramsay 01].

ter curves were registered to the RMS mean curve using the R command `registerfd` in the 'fda' package.

2. The warping function $h_{i,A}(t)$ for the RMS parameter and its functional inverse[14] $h_{i,A}^{-1}(t)$ were computed, which returned the warped sampling points for each speaker's curves.

3. For each speaker, these warped sample points were applied to the other parameters in the audio and video parameter sets to perform registration.

Adhering to this process ensured that the same sample points — possibly different for each sequence — were used for all parameter curves of each sequence in each of the two modalities. The RMS parameter was chosen as registration target, because of its non-flat curve shape, which was particularly well-suited for the registration process. The process was based on the previous assumption, that the audio and video signals were in synchronisation (see Section 5.4.1 for details). Curve registration was performed on the smoothed and resampled parameter curves (see Sections 5.4.2 and 5.4.3). Curve registration results are discussed in Section 6.7.1.

It must be noted that a disadvantage of using this registration method based on FDA is that employing (cubic or higher order) spline curves has the inherent risk of creating strongly oscillating curves, which do not reflect the original curve shape well. As a counter measure, smoothing algorithms are usually applied. However, such smoothing bears the risk of deviating significantly from the original curves. Therefore, additional smoothing after curve registration was not used in this study. Curves that oscillated strongly after registration were determined in a manual process by visual inspection of the graphs, then removed, and replaced by the curves that were used as input, in order to keep the number of curves constant.

Principal Component Analysis

PCA can also be defined in functional analytic terms [Besse 86]. Similar to the statistical shape analysis described in Section 5.5.1, a PCA in the temporal domain

[14] Note that -1 is used here to refer to the inverse, not the reciprocal.

was performed on the registered parameter curves to identify the main modes of variation. As (successful) curve registration is expected to reduce the amount of phase variation, the FDA PCs should show the modes of variation due to shape variation even more clearly. This was tested by comparing the proportion of variance accounted for by each FDA PC with those previously computed. FDA PCA was performed by the R command `pca.fd` of package 'fda'. The results can be found in Section 6.7.2.

5.6 Chapter Summary

In summary, this chapter has described the methods used to explore the relationships of audio and video speech parameters based on the data in the AVOZES data corpus. This has included the extraction of audio and video speech parameters, necessary preprocessing methods, and the statistical methods for the analysis of the relationships between audio and video speech parameters.

First, the analysis of the audio signal has been described. A common way is to perform a spectral analysis. Background information on two common spectral analysis techniques — the filter bank method and the linear predictive coding method — has been given. In preparation for the analysis of AV relationships in the next chapter, details on how the voice source excitation frequency F_0, the formant frequencies F_1, F_2, F_3, and the RMS energy value can be determined have been presented. The parameter extraction was done with the ESPS package on lowpass filtered audio signals.

Secondly, methods of video analysis have been presented. Based on the automatic lip tracking algorithm described in Chapter 3, geometric parameters such as width and height of the mouth opening, as well as protrusion of upper and lower lip, were extracted. In addition, the visibility of teeth was determined and measured in a numeric fashion.

Issues of AV synchronisation have been discussed, as it is essential for a correct statistical analysis that the two signal streams are synchronised. The DV recording equipment ensures, by definition of the DV standard, that the audio and video signal

are synchronised when in locked audio mode, which was used during the recordings of the AVOZES data corpus. Due to the stereo vision system used during recording, there is a constant delay of the length of one video frame between the audio and video signal, which was accounted for by a shift on the time axis when resampling the extracted parameters. In addition, because of the definition of the NTSC video signal standard, careful resampling was required on the data of the first second of each minute of video data. Moreover, both audio and video parameter sequences were resampled to give them the same number of observation points, which is a prerequisite for many statistical analyses.

The extracted speech parameters were smoothed using cubic splines. Where possible, weights based on the confidence measures were applied, which pulled the smoothing spline function more strongly to points of high confidence. Otherwise, when confidence measures were not available, equal weights were applied. Smoothing and the statistical analyses were performed using functionality provided in the R statistical software system.

Finally, statistical analysis techniques have been detailed. MVA techniques were of particular interest, because they describe the relationship between two or more parameters, and thus were well-suited for analysis of the relationships between audio and video speech parameters. These analyses included PCA, LDA, pairwise linear correlation analysis, CANCOR, and COIA. COIA is a recently developed MVA which offers more statistical stability than other analyses. COIA has not been used in the area of AVSP before. Secondly, some of the relatively new techniques of FDA have been presented. FDA expresses traditional MVA techniques in functional analytic terms and is particularly suitable for the analysis of time series data like the parameters in the analysis of the AV relationships. It also offers ways of registering curves, which eliminates other variation sources that could hide the shape variation, which was the area of focus in this study. The results of the analyses are described in the next chapter.

Chapter 6

Results and Discussion

In this chapter, the results of the various analyses are presented and discussed. The aim was the determination of the statistical relationship between the AV speech parameters selected in the previous chapter. Firstly, the data space is explored in Sections 6.1 – 6.3. Observations made by visual inspection of the measured parameters are given in Section 6.1. This section also discusses differences in the parameter curves between male and female speakers of AuE, as well as between the three varieties of AuE. The results of the outlier analysis, described in Section 5.4.4, are discussed in Section 6.2. The treatment of outliers is also described there. The exploration of the data space is finished by a discussion of the results of applying an LDA to the data to analyse, how the parameters separate the phonemes.

The remaining sections in this chapter present the results of the statistical analyses for the determination of the AV relationships. In Section 6.4, the relationships within each parameter set are analysed. This was done by applying a PCA to the parameter sets for each phoneme, as a test for redundancies in those sets, and combining the results with the results of a linear pairwise correlation analysis for determining the redundant parameters, where possible. Then, Section 6.5 presents the results of applying PCA in the temporal domain, separately for each parameter, as a statistical shape analysis technique for determining the main modes of variation in the parameter curves for each phoneme. This is followed by Section 6.6, which presents and discusses the results of the various analyses of the AV re-

lationships across the two parameter sets. Pairwise correlation analysis, canonical correlation analysis, and coinertia analysis were performed. The latter two analyses are multivariate analyses that explore, how linear combinations of the parameters are statistically related to each other across the two modalities. Finally, an outlook is given in Section 6.7 on how FDA can aid the analysis with the help of its curve registration techniques. The results of the curve registration process are discussed, as well as the results of applying PCA in the temporal domain (as in Section 6.5) as a way of comparing the influence that curve registration had on the results.

Because of the complexity of the data, numeric results of the analyses can generally be found in the Appendices C – J, which includes data on the accompanying CD-ROM. It is also worthwhile to remind the reader that the results presented here are for the central phonemes in the vocalic (/bVb/) and consonantal (/ɑːCɑː/) contexts chosen in the AVOZES data corpus. Since coarticulation is a natural feature of (continuous) spoken language, it undoubtedly is a factor in the results, as the investigated sequences contained some samples from the contextual phonemes surrounding the central phoneme. However, it is beyond the scope of this study to investigate other vowel and consonant contexts. The reader is also reminded that some speakers in the AVOZES data corpus had problems in producing the velar closure nasal /ŋ/, as well as distinguished voiceless and voiced inter-dental fricatives /θ ð/. The results for these phonemes must therefore be treated with care.

6.1 Presentation of the Data and Some Initial Remarks

This section describes some observations made by visual inspection of the data and discusses differences in various groups of speakers. The parameter curves for all phoneme-parameter pairs are presented in Appendix C. For each parameter in the sets of audio and video speech parameters, the graphs of the parameters curves are shown side by side for all phonemes, thus enabling an easy way of visual comparison. The order of the phonemes follows that in Tables 4.4 and 4.5, that

is, short vowels, long vowels, diphthongs, and finally consonants. The parameter curves of the individual speakers are shown in green (female speakers) and black (male speakers) and the pointwise mean curve is shown in red.

As a general comment before going into detail, the individual parameter curves agreed very well for some parameters and some phonemes, but there was also a large degree of variance for some other parameters and some other phonemes. Some observations, made by visual inspection of the graphs, are discussed in more detail in the Sections 6.1.1 – 6.1.9. The sections for the formant frequencies also discuss the similarities and differences of the measurements with the findings of Lindblom and Sundberg (see Section 2.1.4 and [Lindblom 71]). Section 6.1.10 discusses visible differences between the parameter curves from female and male speakers. Finally, a comparison of the parameter curves for the three varieties of AuE (cp. Section 2.2) is presented in Section 6.1.11.

6.1.1 Voice Source Excitation Frequency F_0

In the case of the vocalic phonemes, the individual F_0 parameter curves were spread over a frequency range from about 70Hz to 280Hz. Despite that range, the curves showed a high degree of similarity in terms of their shape. The F_0 value typically increased slightly with the onset of the vocalic phoneme (following the first /b/ in the bilabial context) and decreased slightly again at the offset (going into the second /b/ of the bilabial context). In between, i.e. for the length of the vocalic phoneme, the curves were fairly flat, indicating a largely constant F_0 value (Figure 6.1 top). Since the vowels and diphthongs are voiced sounds, generated by quasi-periodic pulses of air from the glottis (see 'The Articulation of Vowels and Consonants in English' in Section 2.1.4), relatively stable F_0 values were expected.

For the consonantal phonemes, the individual F_0 curves were in the frequency range of about 70Hz to 250Hz with some outliers (higher F_0) for some female speakers. Some curves exhibited a minimum during the intervocalic consonant, others were flat. Differences between voiced and voiceless phonemes were noticed. For voiced consonantal phonemes, the curves were flatter (e.g. compare /p b/ in Fig-

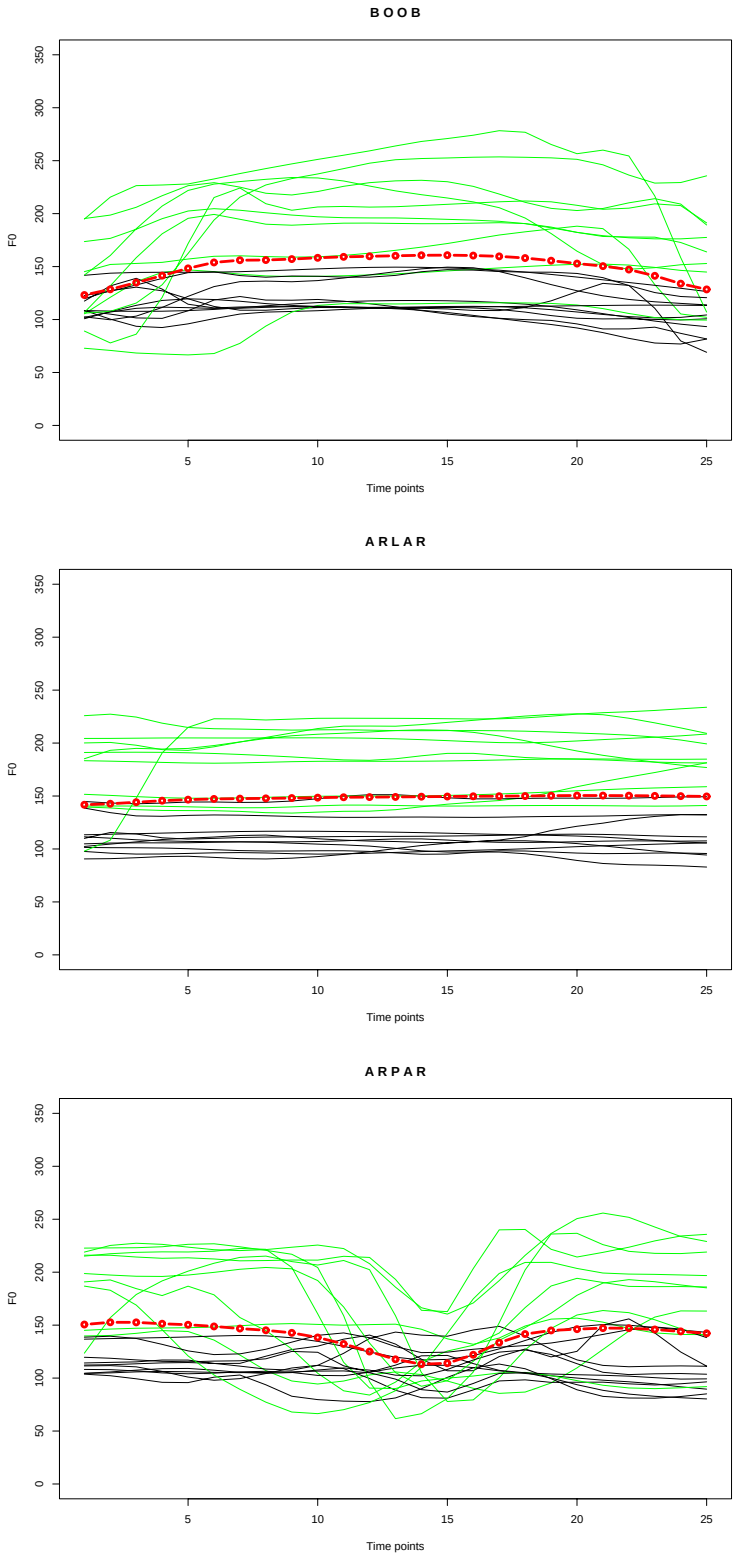


Figure 6.1: Examples of F_0 curves: /uː/ at the top, /l/ in the centre, and /p/ at the bottom.

ure C.1) and some did not show a minimum during the intervocalic consonant at all (see the liquids and glides /l r w j/, for example Figure 6.1 centre). Voiceless consonantal phonemes, however, showed a clear minimum during the intervocalic consonant (Figure 6.1 bottom).

6.1.2 Formant Frequency F_1

The F_1 values ranged from about 200Hz to 1000Hz. The individual parameter curves for the vocalic phonemes agreed fairly well for a given phoneme, perhaps with the exception of /ɑː/, where the individual curves showed more variation. Outliers were noted for /ʊ æ/. Typically, the mean curves were either fairly flat, slowly dropping off as time progressed, in the case of smaller F_1 values (up to about 400Hz), or they exhibited a mostly flat peak in the case of F_1 values above 400Hz. In the former category, the phonemes /ɪ ʊ iː uː ɪə/ were found (Figure 6.2 top). The latter category consisted of the phonemes /ɒ ʌ æ ɑː eɪ aɪ aʊ əʊ/ (Figure 6.2 centre). The phonemes /ɛ ɜː ɔː əː ɔɪ/ fell in between these two categories.

For the consonantal phonemes, the individual curves diverged more than for the vocalic phonemes. For example, the individual curves for /k/ seem to have very little in common (Figure 6.2 bottom). It was noticed that, due to the low RMS energy values for the unvoiced oral stops /p t k/, the trajectories of the voice source excitation frequency F_0 and the formant frequencies F_1, F_2, F_3 were not well-defined. A gating based on the RMS energy parameter could mark these points and exclude them from analysis. The analysis results for these phonemes must be treated with care, if no gating is performed. Outliers such as the topmost curves for /d g/ also occured (Figure C.4 in Appendix C). However, many curves seemed to differ mainly in the horizontal place (timing) and the width of the minimum (if present). The latter can be attributed to differences in the F_1 patterns for different speakers. The timing differences could be related to the originally different sampling frequencies of the audio and video speech parameters (100Hz v. 30Hz). If the start and end points of the selected sample intervals, which depended on the onset and offset found in the MH parameter sampled at 30Hz, were determined incorrectly by only

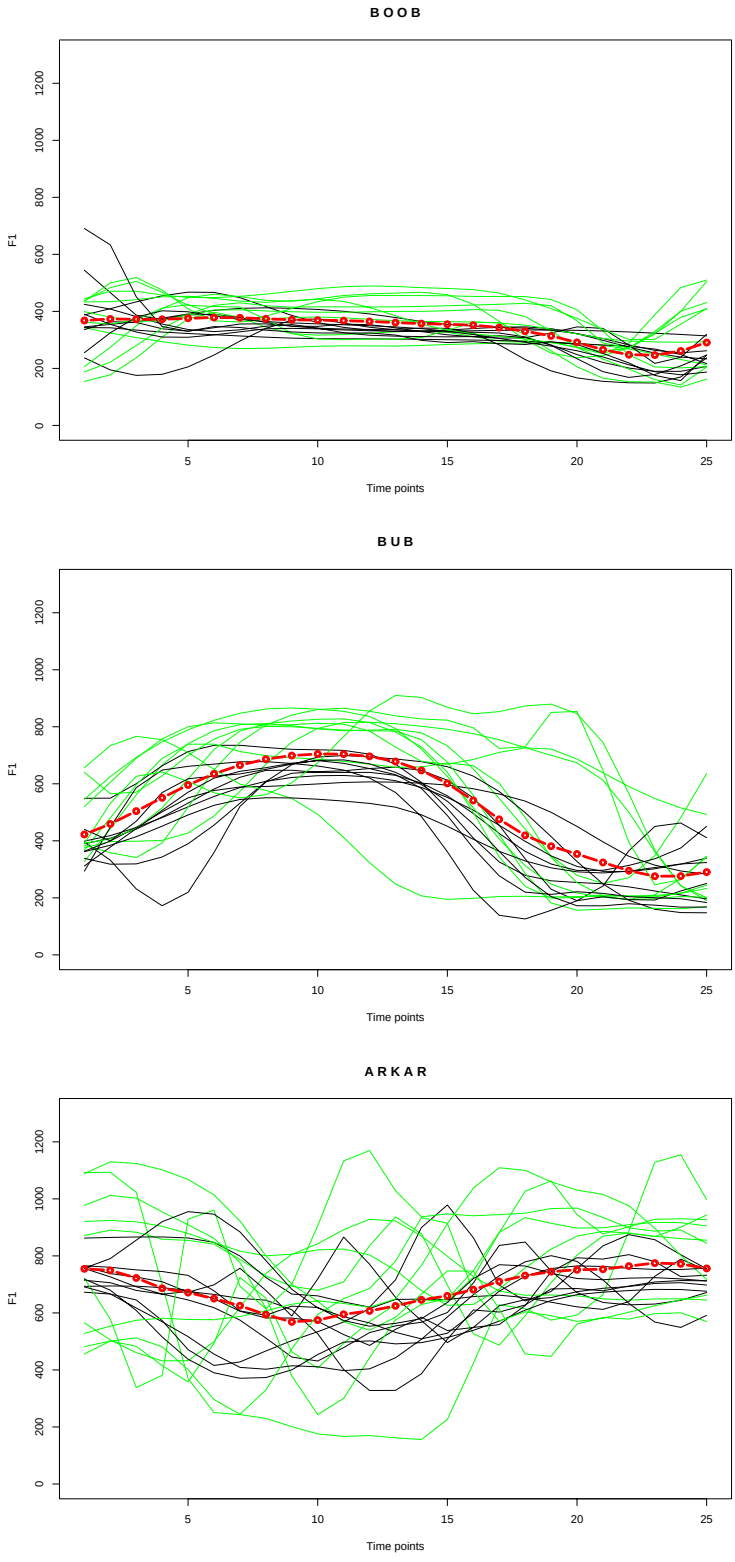


Figure 6.2: Examples of F_1 curves: /uː/ at the top, /ʌ/ in the centre, and /k/ at the bottom.

one video frame, it is possible that the data showed a horizontal displacement of about three resampled sample points (see Section 5.4.3 for a description of the resampling process). This suggests that curve registration, in which the internal AV synchronisation is maintained, could alievate the situation (see Section 6.7).

Despite this divergence, it can be seen that the curves typically have a minimum during the intervocalic consonant (Figure 6.2 bottom) with the higher surrounding F_1 values being a product of the vocalic /ɑːCɑː/-context. Looking at the mean curves, the width of the valley around the minimum clearly depends on the amount of horizontal displacement in the individual curves, which means that one has to be careful in drawing conclusions from this. However, from individual parameter curves it can be seen that for some phonemes (e.g. /m l j/ in Figure C.3) the 'valley' is broader than for other phonemes (e.g. /p b/ in Figure C.4), which appeared to be related to the length of the consonant. Almost all of the mean curve minima were at an F_1 value in the range of 400–600Hz, while individual parameter curves showed minima as low as 200Hz. Again, the horizontal displacement of individual curves played a role in this issue, due to averaging out individual minima.

Referring back to Section 2.1.4 on speech production and acoustical consequences of articulatory movements, the results agree well with the findings of Lindblom and Sundberg [Lindblom 71]. Frontal tongue position and frontal airflow constriction (tongue or lips) resulted in a decrease of F_1. For example, compare the F_1 curves of /ɪ ʌ/ (Figure C.3) with the former showing much lower F_1 values than the latter. Similarly, consonants that require more lip movement or a more frontal lip position than others, showed lower F_1 values. The measurements also confirmed that jaw opening leads to higher F_1 values in vowels and diphthongs.

6.1.3 Formant Frequency F_2

The F_2 frequency values ranged from about 500Hz to almost 3000Hz. The individual curve shape for any phoneme agreed very well for vocalic phonemes. Low outliers occured for /ɪ ɛ æ ɜː ɔɪ aɪ əʊ/. Two kinds of mean curves were found: (1) curves with a maximum during the interconsonantal vowel and (2) curves with a

minimum during that time. In the first category, the phonemes /ɪ ɛ iː əː/ were found (Figure 6.3 top). The second category consisted of the phonemes /ʊ ɒ ʌ ɔː/ (Figure 6.3 centre). The vowels /æ uː ɜː ɑː/ fell in between these two categories and showed mostly flat curves. Horizontal displacement between individual curves did not appear to occur for F_2, with the exception of /ɔɪ/.

It can clearly be seen in the graphs of Figure C.5 in Appendix C that in the case of the diphthongs, the vowels on which the diphthongs are based were matched in the F_2 curve. A prime example is the phoneme /ɔɪ/. First, the lower F_2 values of /ɔ/ can be seen, followed by the higher F_2 values of /ɪ/.

Generally, there was more variance in the individual F_2 parameter curves for the consonantal phonemes. Typically, a maximum during the intervocalic consonant can be seen (Figure 6.3 bottom). Differences existed between individual curves in the amplitude and when the maximum occurred (shift on the time scale). It also appeared that voiceless consonants led to larger divergence among the individual curves than voiced consonants due to a lack of energy (for example, compare the curves for /s z/ in Figure C.6 in Appendix C). *RMS* gating could be used to mark sample points with little energy and to exclude them from analysis.

The phoneme /w/ presented an exception with the F_2 curve having a distinct minimum rather than a maximum during the intervocalic consonant. In addition, the F_2 parameter curves of the phonemes /l r/ were more or less flat with no clear maximum or minimum. Flat individual parameter curves were also found for some speakers at other phonemes (e.g. /m n/), but no consistent trend could be identified.

Again, the observations here agreed with the findings of Lindblom and Sundberg [Lindblom 71]. A tongue constriction closer to the velum and closer to the front of the oral cavity led to a significant increase in the values of F_2. The effect was stronger for spread lips. Examples are the F_2 curves for the vowels /iː ɔː/ (Figure C.5). As Lindblom and Sundberg point out, the effect of tongue position and shape is primarily on F_2. Neutral and back tongue constrictions resulted in lower F_2 frequencies than frontal constrictions. This was also evident in the consonants where, for example, /ʃ tʃ/ had higher F_2 values than /l r/. Lip rounding as in the vowel /ɔː/ and the consonant /w/ resulted in a significant decrease in F_2.

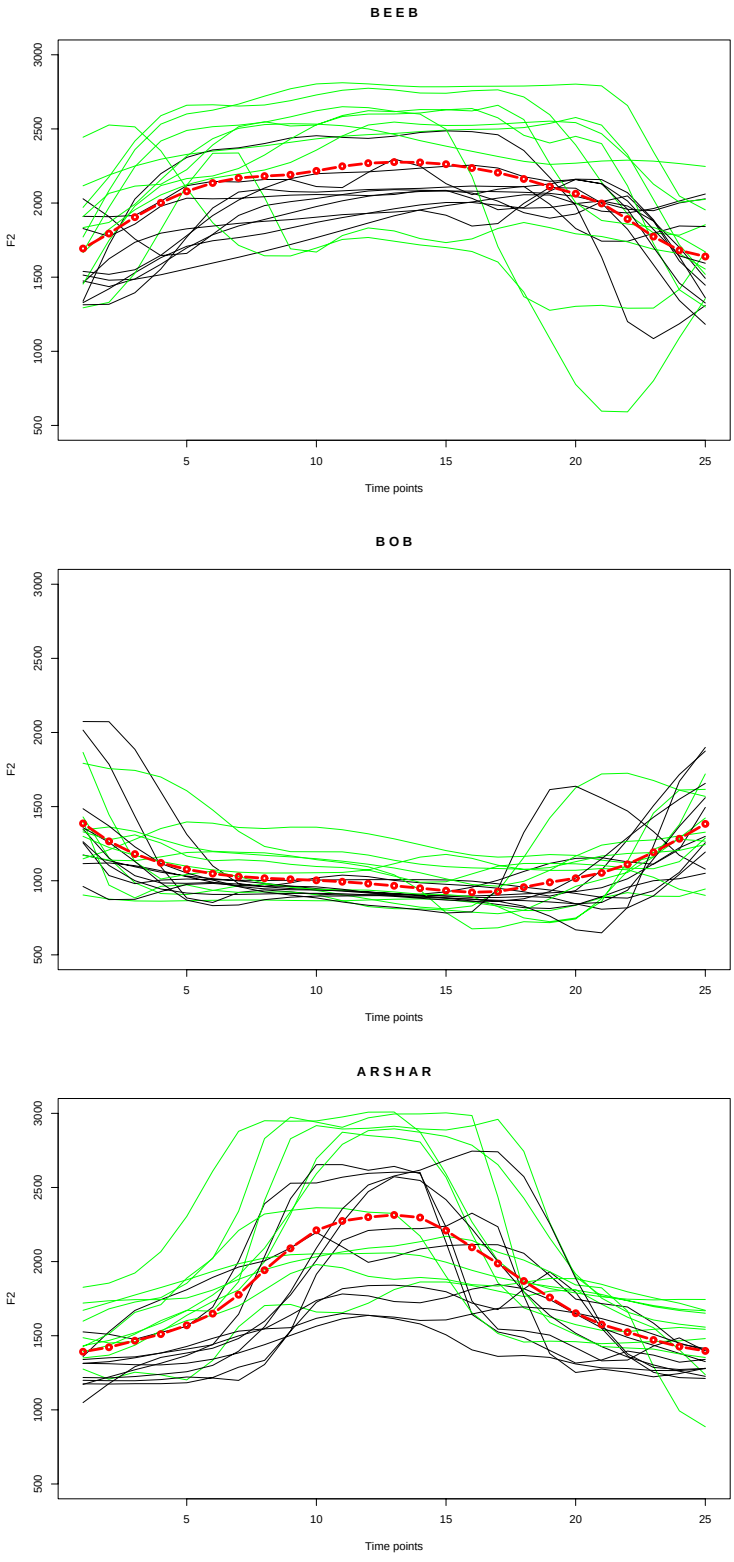


Figure 6.3: Examples of F_2 curves: /iː/ at the top, /ɒ/ in the centre, and /ʃ/ at the bottom.

6.1.4 Formant Frequency F_3

The F_3 values lay in a range from about 1500Hz to 3500Hz. For the vocalic phonemes, the individual F_3 parameter curves of the speakers agreed well for a given phoneme. They were typically fairly flat in shape (Figure 6.4 top). Little difference in terms of the frequency values was found in the mean curves for different vocalic phonemes, apart from a slight increase for the high and front vowels /ɪ iː/ (Figure C.7). In this case, the increase was more prominent for the corresponding long vowels (and the diphthongs /eɪ ɪə/), because coarticulation constraints due to the /bVb/-context would limit the reaching of the articulator targets more for short vowels than for long vocalic phonemes.

The decrease in F_3 at the onset of the interconsonantal vowel (following the first /b/ in the /bVb/-context) and the increase again at its offset (going into the second /b/ of the /bVb/-context) appeared to be an artifact of the formant frequency determination process, which is unstable for sample points with low energy. A gating based on the RMS energy parameter could mark these points. For voiceless phonemes and outside of the /bVb/ words, F_3 was often found 'hovering' around in the range of 2500–3000Hz without following a definite 'track'.

For the consonantal phonemes, again more variance was found between individual F_3 parameter curves (Figures C.7 and C.8). While there was good agreement for some consonants (e.g. /r w/), there were a lot of different curve shapes for other consonants (e.g. /k m/). For most phonemes, the F_3 parameter curves had a maximum during the intervocalic consonant (Figure 6.4 centre). This maximum was the result of an increase in F_3 by about 200–500Hz. The bilabial stops /p b/, the alveolar closure nasal /n/, the lateral liquid /l/, and the bilabial glide /w/ showed no distinct maximum. Their F_3 curves were rather flat in shape. Another striking exception was the rhotic consonant /r/, which presented a distinct minimum during the consonant (Figure 6.4 bottom). There, the decrease in frequency was about 500Hz on average.

Lindblom and Sundberg [Lindblom 71] mentioned a sharp decrease in F_3 for frontal tongue positions and spread lips and a slow rise in F_3 values for neutral

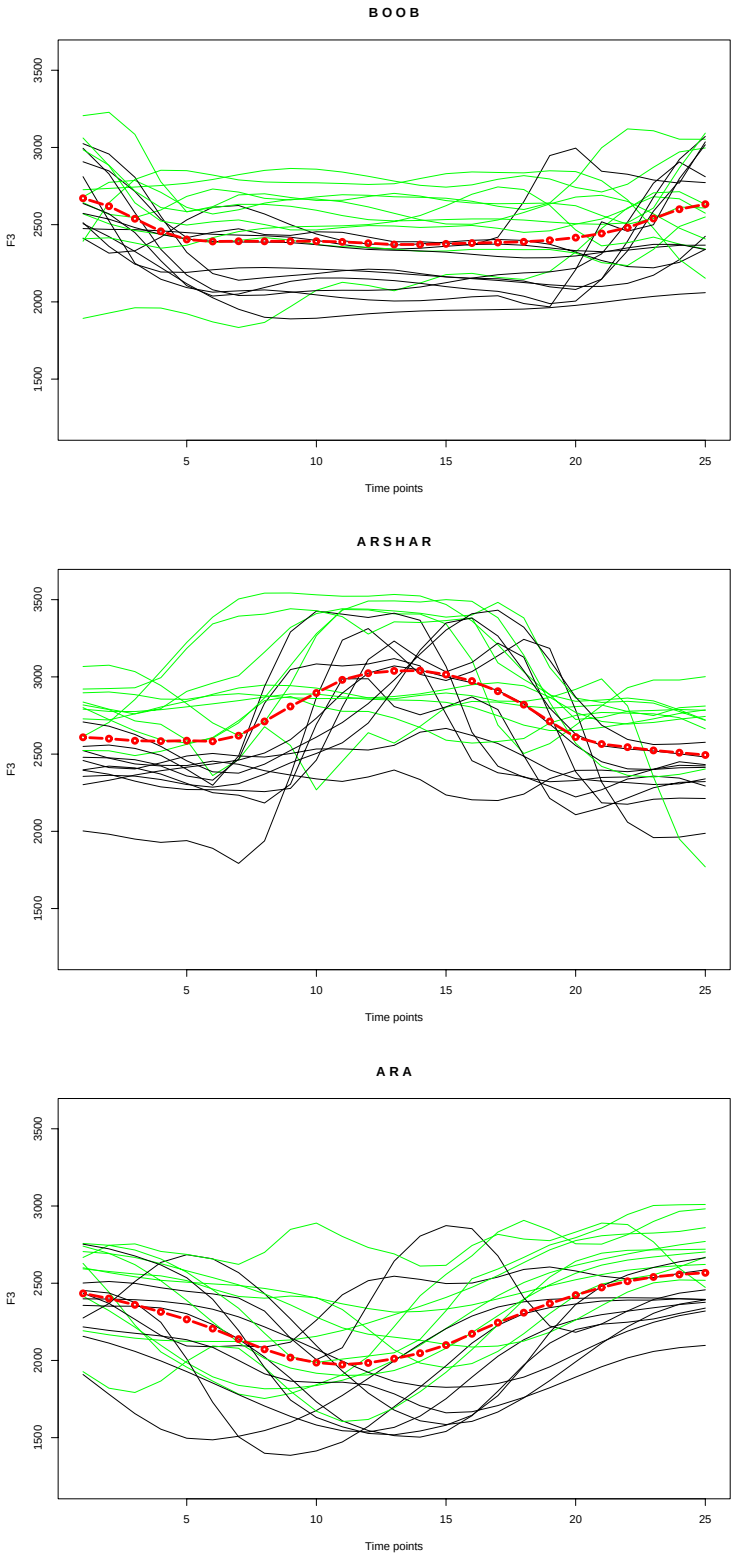


Figure 6.4: Examples of F_3 curves: /uː/ at the top, /ʃ/ in the centre, and /r/ at the bottom.

and back tongue positions. These changes in F_3 were not evident in the data from the AVOZES data corpus, which could be due to the tendency of generally speaking with spread lips in (particularly broad and general) AuE. Lindblom and Sundberg also described a significant decrease in F_3 for lip rounding. This was most prominently visible for the rhotic consonant /r/, but also for the long vowel /uː/ when, for example, compared to the long vowel /iː/.

6.1.5 RMS energy

The individual RMS parameter curves for a given phoneme agreed very well for both vocalic and consonantal phonemes. In the case of the vocalic phonemes (Figure C.9), the voicing of the vocalic sound resulted in a quick increase in RMS value at the onset and a reasonably quick decrease at the offset, although this decrease was somewhat slower than the increase (Figure 6.5 top). Diphthongs either showed a small second peak for the second vowel (/aʊ ɪə əʊ/, Figure 6.5 centre), or they showed a slow, almost constant rate decrease after the first maximum (/eɪ ɔɪ aɪ/).

For the consonantal phonemes, a minimum close to 0 in the RMS parameter curve during the intervocalic consonant was found for oral stops, fricatives, and affricates (Figure 6.5 bottom). Nasals, liquids, and glides did not present such a distinct minimum. Their RMS parameter values were relatively constant in a low energy range during the intervocalic consonant, so that the parameter curves were flatter in shape (Figure C.10 in Appendix C).

Generally, the RMS parameter curves showed a reasonable amount of horizontal displacement. A curve registration process, which maintains AV synchronisation, could reduce the effects this has on any statistics (see Section 6.7).

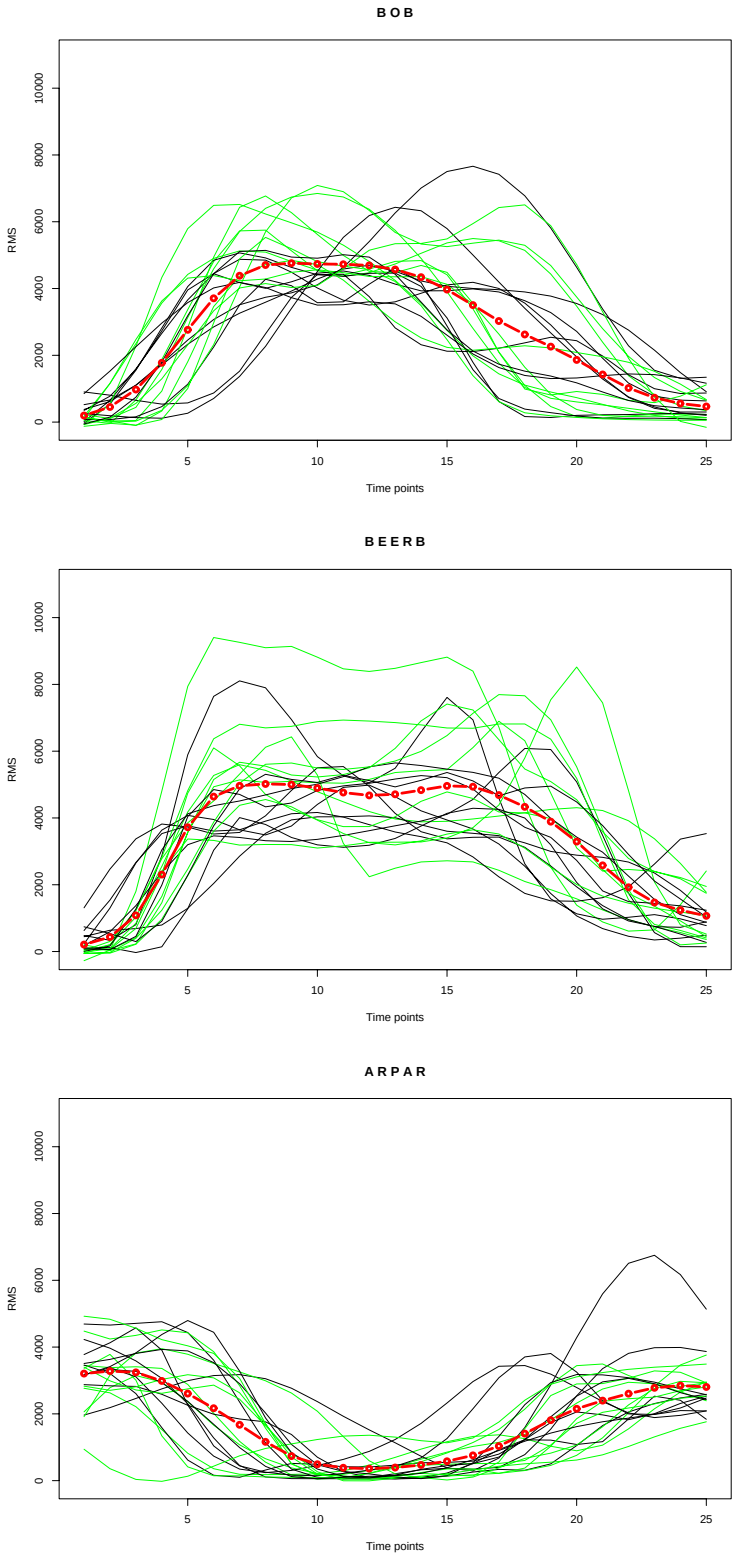


Figure 6.5: Examples of *RMS* curves: /ɒ/ at the top, /ɪə/ in the centre, and /p/ at the bottom.

6.1.6 Mouth Width

Examining the individual parameter curves for mouth width, it was noticed that the curves of the various speakers exhibited a similar shape for a given phoneme, but were spread over a range from about 30mm to 60mm. This spread can be explained with the width of the mouth in neutral state being a characteristic that varies from person to person.[1] As a consequence, statistics that analyse the shape of the curves rather than the absolute values were better suited, which led to a PCA being applied first to the parameter curves as a statistical shape analysis, as described in Section 5.5.3. Furthermore, outliers were found for some vocalic and consonantal phonemes (see Section 6.2 for a discussion). For example, see the curves for /ɜː ʃ/.

For the vocalic phonemes, the shape of the curves was generally fairly flat. However, it can be seen that some speakers decreased the width of the mouth at the onset of the interconsonantal vowel or diphthong and then increased the width again in the second half of it. This effect can be seen in /ɛ ʌ æ iː ɑː ɔː eɪ aɪ/ (Figures 6.6 top and C.11). It is, therefore, hypothesised that this effect occurs typically in front and low central vowels and diphthongs containing these.

In the case of the consonantal phonemes, most of the curves were again fairly flat (Figure 6.6 centre). Nevertheless, a few exceptions were noticed. For the bilabials /p b m w/, the labio-dentals /f v/, and the rhotic /r/, a decrease in mouth width was observed first, followed by an increase again, in the second half of the intervocalic consonant, i.e. at the transition from the consonant to the vocalic context used (Figure 6.6 bottom).

[1] Such a personal characteristic, of course, also affects the formant patterns. It is possible that narrow-mouthed people have a different pattern of AV relationships than wide-mouthed people, because of the effect of a smaller mouth opening on the vocal tract system, but such differences did not become evident in this study. If such different patterns existed, they would have consequences on the benefit of including video speech parameters in ASR systems. Further investigations of this issue are needed in future work.

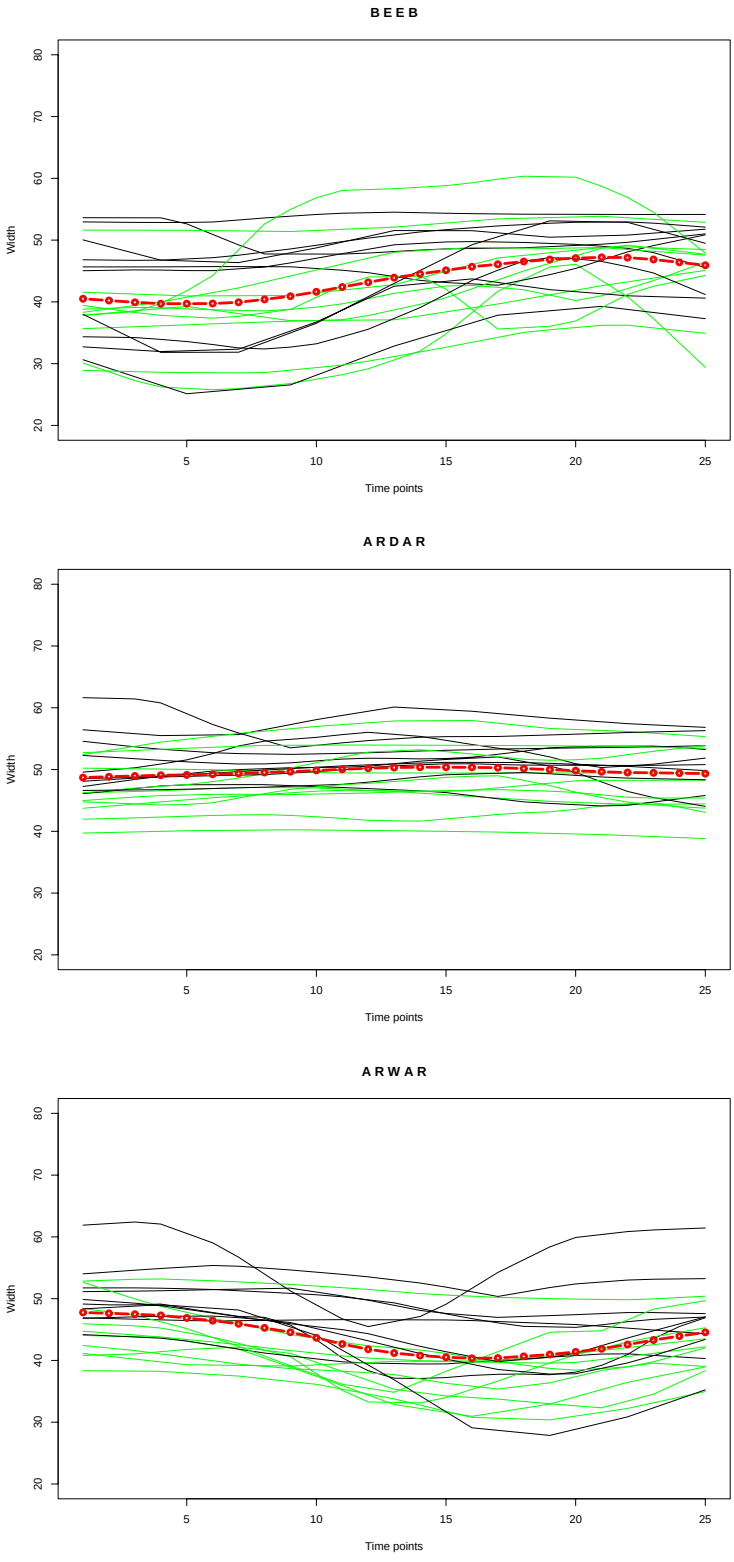


Figure 6.6: Examples of *MW* curves: /iː/ at the top, /d/ in the centre, and /w/ at the bottom.

6.1.7 Mouth Height

The individual parameter curves for mouth height agreed fairly well between speakers for the vocalic phonemes (Figure C.13). Some outliers were present due to tracking failures (see 'Outlier Analysis' in Section 6.2). The largest MH values were found for the vowels /æ iː ɑː əː/ and diphthongs /eɪ aɪ aʊ/ with more than 10mm height of the mouth opening on average at the maximum (Figure 6.7 top). MH values were, thus, larger for front vowels than for back vowels (and diphthongs containing these). The comparitively strong maximum for /iː/ is perhaps surprising. Generally, the parameter curves rose slowly from the total lip closure forced by the bilabial context towards the maximum value, which was reached after about two-thirds of the vowel length. After that, the curves fell quickly towards the second lip closure of the bilabial context /bVb/.

For the consonantal phonemes, the individual parameter curves of the speakers agreed well in their shape, but a vertical displacement could be seen, which was related to how far the mouth was opened for the vocalic context /ɑːCɑː/ (Figure C.14). Again, the shape of the curves was more important for the statistical analysis than the absolute parameter values, as the common characteristics of the individual curves were the area of focus for the analysis of the AV relationships. In the /ɑːCɑː/-context, the consonants led to a decrease in mouth height. The strength of the decrease depended on the consonant. It was strongest for the bilabials /p b m w/ and the labio-dentals /f v/, and smallest for the velar stops /k g/ and the lateral liquid /l/ (Figure 6.7 centre and right).

6.1.8 Protrusion of Upper and Lower Lip

The protrusion parameters PUL and PLL are discussed together here, because of the high degree of correlation found (see Figure 6.8 top and centre for an example and Section 6.4.1 for numeric results). Such similarity was not suprising. Purely from observing a person's lips while speaking, one would expect both lips to move back and forth in a similar fashion and simultaneously. The measurements on the AVOZES data corpus confirmed this expectation.

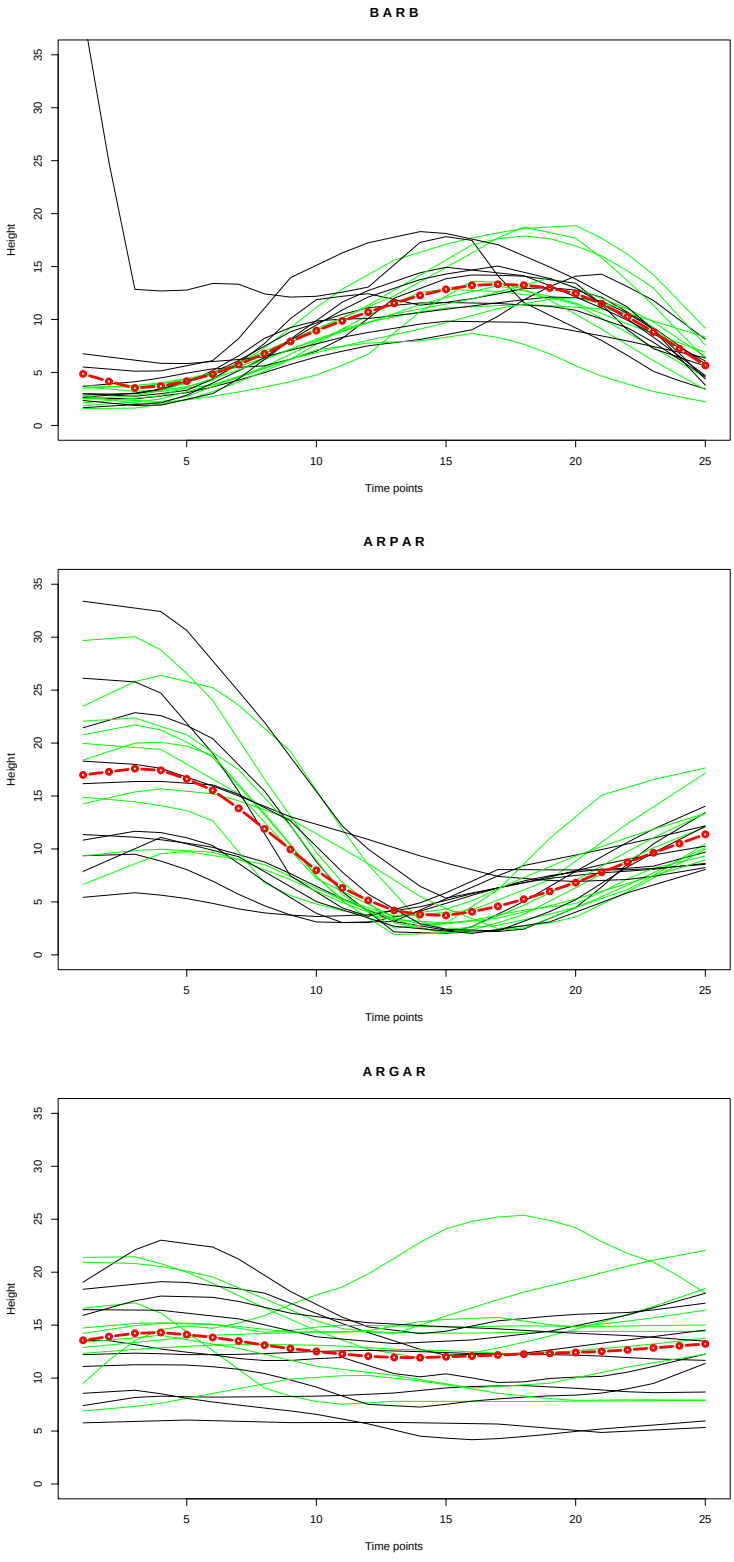


Figure 6.7: Examples of MH curves: /ɑː/ at the top, /p/ in the centre, and /g/ at the bottom.

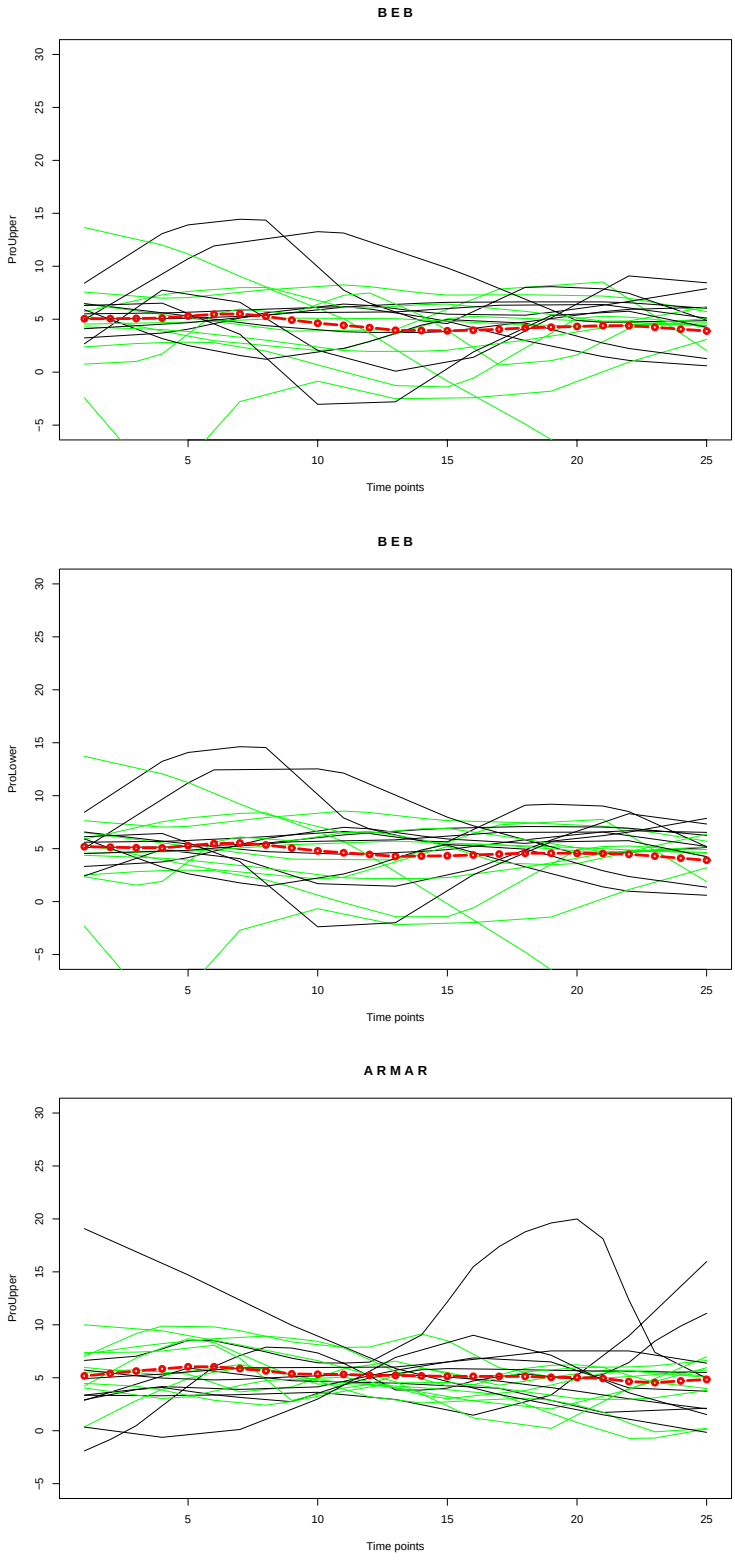


Figure 6.8: Examples of protrusion curves: The top and centre graphs show the similarity between PUL and PLL parameters on the example of /ɛ/. PUL curves of /m/ at the bottom.

Examining the individual curves of the protrusion parameters, a large amount of variance between curves was noticed (Figures C.15 – C.18). This resulted in mean curves being mostly flat as the differences in curve shapes occurred at all sample points and often cancelled each other out. A clear trend was not visible. Outliers were found which resulted from tracking failures (see Figure 6.7 bottom for an example and Section 6.2 for an analysis of outliers). Any results on the protrusion parameters in the statistical analyses must, therefore, be treated with care.

6.1.9 Relative Teeth Count

The individual parameter curves of the relative teeth count for a given phoneme showed a large degree of similarity for both vocalic and consonantal phonemes (Figures C.19 and C.20). The shape of the curves were similar, but the curves differed in the amplitude. This seemed strongly related to personal characteristics of each speaker, that is, how much the lips covered the teeth in a neutral lip position.

In the case of the vocalic phonemes, the *RTC* parameter was zero during the bilabial closures in the /bVb/-context (Figure 6.9 top). During the vocalic phoneme, it reached a maximum after about two-thirds of the length. The maximum was stronger for front and low central vowels and diphthongs containing these (/ɪ æ iː ɑː ɔː eɪ aɪ aʊ ɪə/). Thus, the *RTC* parameter, or a similar teeth visibility parameter, could be used to distinguish front and low central vowels from back and high central vowels (and diphthongs containing these).

For the consonantal phonemes, two classes of curves were found. First, there were curves with a distinct minimum during the intervocalic consonant (Figure 6.9 centre). This class was formed by the bilabials /p b m w/ and to a lesser extent by the labio-dental fricatives /f v/ and the rhotic /r/. Secondly, all other phonemes had curves which were more or less flat in shape (Figure 6.9 bottom). This second class exhibited a large degree of vertical displacement between the curves of individual speakers, which was due to personal differences in the visibility of teeth in neutral lip position.

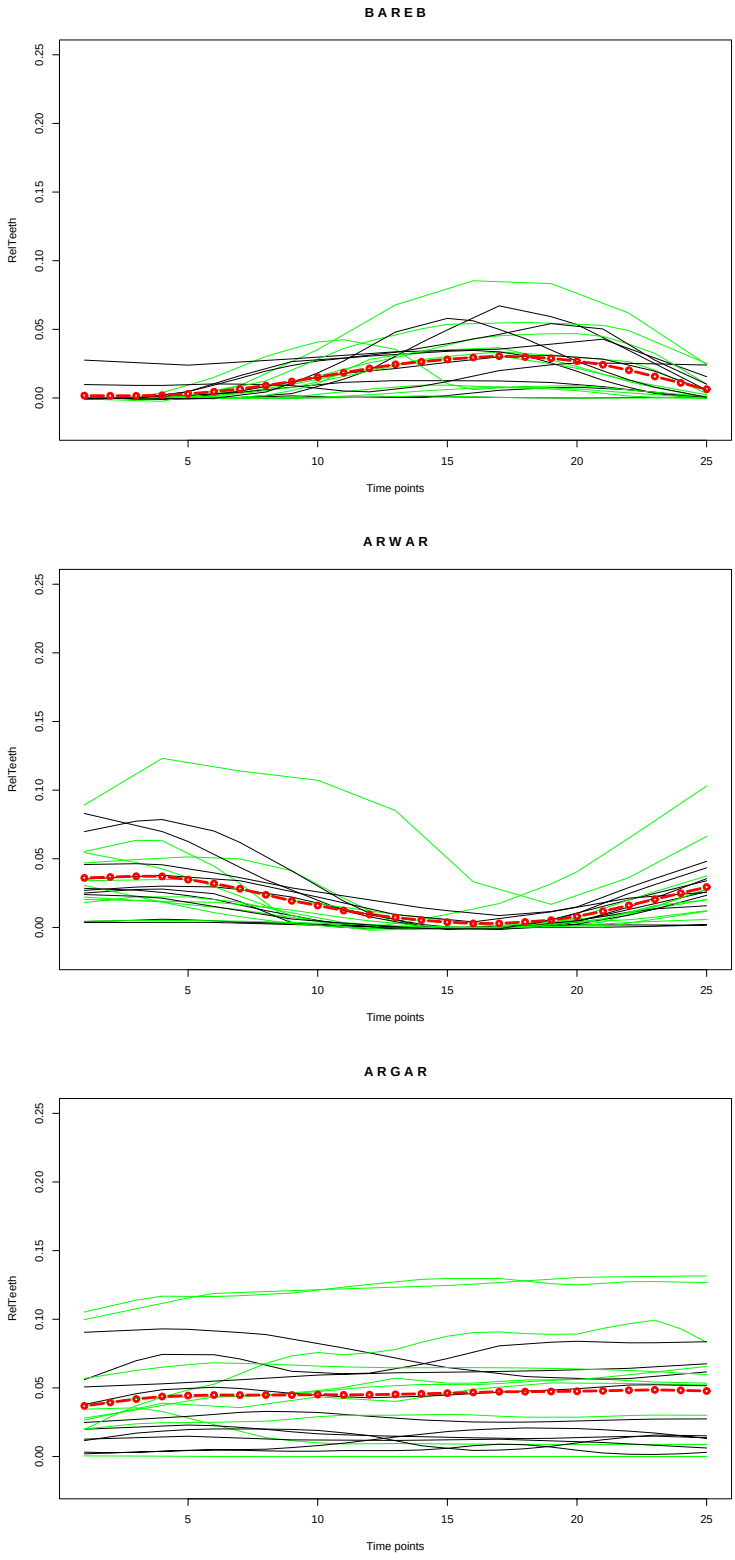


Figure 6.9: Examples of *RTC* curves: /əː/ at the top, /w/ in the centre, and /g/ at the bottom.

6.1.10 Gender Issues

In this section, visible differences in parameter values between the groups of male and female speakers in the AVOZES data corpus are discussed. Such differences were mostly found in the audio speech parameters. Differences occurred in the F_0 parameter, where female speakers generally showed higher F_0 values than male speakers. This was consistent with expectations based on different vocal tract geometries (see Section 2.1.4). Male speakers hardly reached F_0 values beyond 150Hz, while female speakers reached up to 250Hz. There was some overlap in the F_0 frequency range between the two groups for frequencies between 100–150Hz.

Similar observations were made for the formant frequencies F_1, F_2, and F_3, although there was more overlap in the frequency range of men and women for these than was the case for F_0. Again, these observations were in agreement with expectations based on the (usually) longer vocal tracts of men, which results in a lowering of all formant frequencies (see Section 2.1.4). For both F_1 and F_2, the overlap was largest for back and central-back vowels /ʊ ɒ uː ɔː/. For all other vocalic phonemes, female speakers had slightly higher F_1 and F_2 values than male speakers. In the case of the consonantal phonemes and F_1, female and male speakers shared roughly the same frequency range with the exception of the bilabial stops /p b/, where the F_1 values were higher for women than for men. In the case of the consonants and F_2, the tendency was that female speakers had higher F_2 values than male speakers, except for the velar stops /k ɡ/, the labio-dental fricatives /f v/, the inter-dental fricatives /θ ð/, the voiceless alveolar fricative /s/, the alveolar affricates /tʃ dʒ/, and the velar closure nasal /ŋ/, where there was more overlap in the frequency range between the two groups of speakers.

The observation for F_3 was that female speakers tended to have higher F_3 values than male speakers by about 500Hz on average in the AVOZES data corpus. This was the case for all vocalic phonemes, but could also be seen in the consonants /p b t d g v z n l r w j/. However, it should be noted that there was generally some overlap in the frequency range between the two groups of speakers for F_3.

No noticeable differences between the groups of female and male speakers was

found for the RMS, MH, PUL, PLL, and RTC parameters. For the MW parameter, the tendency was towards female speakers having smaller parameter values than the male speakers. A plausible explanation is that, on average in this sample, men's mouths were wider than those of women in a neutral lip position, and that this prevailed during speech production as well.

6.1.11 A Comparison of Varieties of Australian English

Another interesting aspect is the comparison of the varieties of AuE. In Section 2.2, the differences have been presented, as they are described in the literature (e.g. [Harrington 97]). The main finding was that differences mainly exist in the diphthongs and to a lesser extent in the vowels. Effects due to variety differences were mostly found in F_1 and F_2. Video speech parameters were not investigated in those studies. In general, the varieties span a continuum of accent variation rather than being defined by distinct boundaries. Furthermore, a speaker may very well pronounce some phonemes in a way characteristic of one variety, while pronouncing other phonemes in a manner typical of one of the other varieties, so that any grouping of speakers by varieties will always be a difficult task and an auxiliary means at best.

The differences reported by Harrington *et al.* [Harrington 97] cannot be confirmed unambiguously by the data of the AVOZES corpus.[2] No significant differences were apparent between the parameter curves for speakers from different varieties. At a first glance it might appear, as if the speakers of broad AuE exhibited lower F_0, F_1, F_2, and F_3 values on average for some phonemes — for example, /aɪ/ — but it is important to recall the composition of the variety groups in the AVOZES data corpus (Appendix B). Although the corpus is gender balanced on an overall level, the groups are not. The group of broad AuE speakers consisted largely of male speakers, while on the other hand the group of cultivated AuE speakers was made up by female speakers only. Therefore, it cannot be verified exactly, if the

[2] The interested reader can find the graphical representations of the parameter curves in the PDF file `curvesVarieties.pdf` on the accompanying CD-ROM.

lower formant frequency values were a result of the accent variation in broad AuE or simply due to differences between female and male speakers, as discussed before.

Finally, splitting the sample group into three groups according to the variety of AuE mostly spoken by the speaker, created groups of 6 speakers for broad AuE, 12 speakers for general AuE, and only 2 for cultivated AuE. While this reflected the typical composition of the Australian population in terms of the accent varieties well, any statistical analyses applied to the groups, in particular the smaller ones, faced stability issues in the results. This calls for further investigation in the future with a larger sample size. In this study, the focus was on analysing the group of speakers as a whole, rather than individual groups.

6.2 Outlier Analysis

In this section, the results of the outlier analysis, described in Section 5.4.4, are discussed. Outliers were defined as sample points with parameter values outside a range of three standard deviations from the pointwise mean. Tables C.1 and C.2 in Appendix C summarise the results by showing for each phoneme and parameter the total percentage of outliers and the number of speakers with outliers.

Overall, it can be judged that outliers were not occurring at a high rate and that their influence on the results of the statistical analyses was small. The highest total percentage of outliers for any phoneme-parameter pair was 5%. The average outlier occurrence rate was 0.8%. The parameters F_0, F_3, and MW had the fewest outliers overall, while MH, PUL, and PLL had the most. Outliers typically occurred for only some sample points of only one or two speakers in each phoneme-parameter pair. These patterns were not consistent, that is, outliers occurred for different speakers across the phoneme-parameter pairs, not always for the same speakers.

Generally, the outliers for the audio speech parameters were so small in number, that no further treatment (see Section 5.4.4) was deemed necessary, as their influence on the results would have been minor. More outliers were found for the video speech parameters, except for the MW parameter, which had a very low outlier occurrence rate, so that it is worthwhile to investigate the reasons more closely.

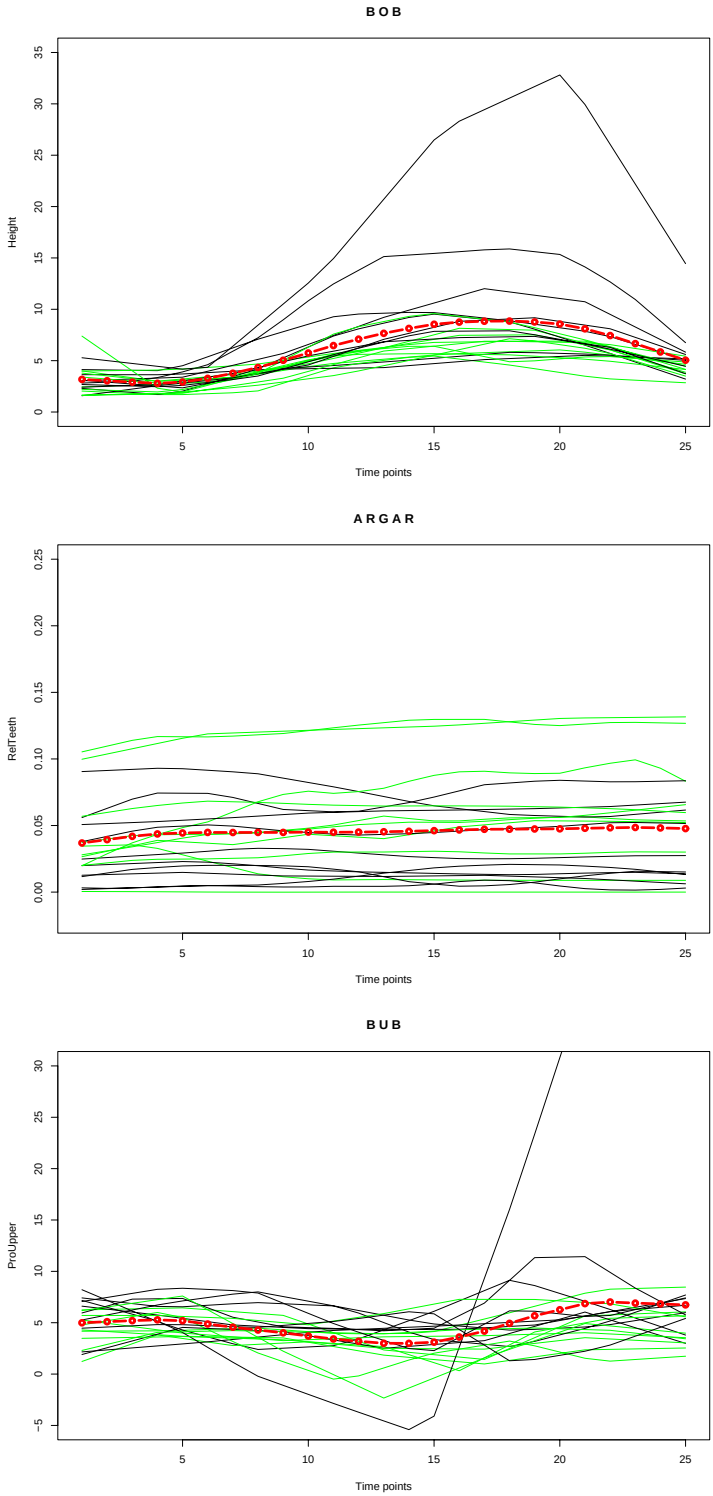


Figure 6.10: Examples of outliers: At the top, parameter MH for phoneme /əː/ and parameter RTC in the centre for phoneme /g/ are examples for outliers likely to be related to personal characteristic. At the bottom, an example of a tracking failure for parameter PUL and phoneme /ʌ/ can be seen.

Generally, the magnitude of the outliers appeared to be larger for the video speech parameters than for the audio speech parameters. Two reasons can be given. Firstly, some of the outliers once again appeared to be related to personal characteristics and the way of speaking. For example, compare the different curves in Figure 6.10 top and centre. The curves show a very similar behaviour, but are shifted on the vertical axis and differ in amplitude. Apart from the LDA (Sections 5.5.2 and 6.3), where the resampled parameter values were used, the differences between the curves were handled by first performing a statistical shape analysis (Sections 5.5.3 and 6.5) and then only using the desired components, which left the sample size unchanged and hence supported the stability of some of the analyses. In the LDA, outliers and their corresponding curves were completely removed to improve the accuracy of the classification.

Secondly, some outliers must be attributed to lip tracking failures. As has been discussed in Section 3.4, lip tracking failures can occur, although at a low rate, due to incorrectly located lip feature points. In particular, the protrusion parameters PUL and PLL suffered from incorrectly located lip midpoints. The main reasons were the difficulty of determining corresponding points on the internal lip contour and the fact that small changes in the determined position in the 2D camera images could lead to large changes in the reconstructed 3D coordinates (see also Figure 3.14 left). For example, in the graphs for the phoneme /ʌ/ in Figure 6.10 bottom, the curve of one male speaker (black curve) can be seen to suddenly rise sharply after sample point 15. A change of more than 35mm in lip protrusion within a few milliseconds did not happen in reality, in particular given that lip protrusion usually changed within a range of 10mm around the 5mm mark.[3]

The tracking failures happened for different speakers and different phonemes. Where there were only few outliers, the curve smoothing (Section 5.4.2) process 'eliminated' them well by giving them low weights based on the low confidence measures. The confidence measures generally detected short tracking failures well. However, when a larger number of tracking failures occurred, the smoothing process

[3] Note, the protrusion of lip midpoints was measured with reference to the lip corners. See Sections 3.3 and 5.2 for details.

did not eliminate the outliers, because the number of well-tracked sample points was too small. Based on the rare occurrence of such drastic outliers, they were judged to only have a minor effect on the results and were left unchanged for all analyses except LDA, where the outliers and corresponding curves were removed completely for improved accuracy.

6.3 Linear Discriminant Analysis

After the detailed overview of the observations made on the measured data in the previous sections, this section describes the statistical analyses performed to further explore the data space, before the results of the statistical analyses for the investigation of the AV relationships are reported and discussed in Sections 6.4 – 6.7. An overview of differences between parameter curves for different phonemes, both in shape and length, is given in Section 6.3.1. Then, details on the LDA and reclassification of phonemes after cross-validation are presented in Section 6.3.2. Finally, the results are discussed in Section 6.3.3.

6.3.1 Introductory Comments

Simply from visually comparing the parameter curves for the RMS energy parameter (Figures C.9 and C.10 in Appendix C), much higher values can be observed for vocalic phonemes than for consonantal ones. Given the /bVb/- and /ɑːCɑː/-contexts used in the AVOZES data corpus, the RMS parameter curves for the vowels and diphthongs showed clearly one or two maxima, while the curves exhibited a minimum or a flat shape for the consonants. Thus, the RMS value is a way of distinguishing vocalic phonemes from consonantal phonemes.

Another way of differentiating some groups of phonemes is to compare their length (before resampling the measurements to 25 sample points). Tables 6.1 and 6.2 show the mean length and the length's standard deviation of each phoneme. As a reminder (see Section 5.4.3), the length of the vocalic phonemes in this study was defined by the bilabial context, that is, the time that passed from the first

Short Vowels	ɪ	ʊ	ɛ	ɒ	ʌ	æ
Mean	230	267	268	267	268	291
s.d.	53	53	40	37	48	49
Long Vowels	iː	uː	ɜː	ɔː	ɑː	əː
Mean	307	300	328	340	357	353
s.d.	53	58	61	57	58	58
Diphthongs	eɪ	ɔɪ	aɪ	aʊ	ɪə	əʊ
Mean	343	357	373	370	350	355
s.d.	41	60	67	53	60	91

Table 6.1: Average length of vocalic phonemes: Shown are the mean value and the standard deviation for each phoneme (in milliseconds).

bilabial closure to the second one, as observed in the MH parameter. The length of a consonant was not as easy to define as for a vowel or diphthong. For example, is an oral stop simply defined by a particular point in time or can it be assigned a length value? In this study, the length of the intervocalic consonants was based on the /ɑːCɑː/-context, that is, the time that passed from the maximum observed in the MH parameter before the consonant to the maximum after the consonant.

As can be seen from Table 6.1, the length of short vowels was clearly shorter than that of long vowels and diphthongs. The high, front[4] vowel /ɪ/ was the shortest with an average length of 230ms. The short vowels /ʊ ɛ ɒ ʌ/ were on average almost 40ms longer than /ɪ/, while /æ/ was the longest of the short vowels with about 290ms on average. For the long vowels, it was noticed that the high vowels /iː uː/ (≈300–310ms) had shorter durations than the lower vowels /ɜː ɔː ɑː əː/ (≈330–360ms). Comparing short and long vowels with the same tongue position, they could clearly be distinguished by their length. The average difference in duration was 33–89ms. The duration of the long vowels /ɔː ɑː əː/ was similar to that of the diphthongs, so that they could not be simply distinguished by their length. The diphthong /eɪ/ showed the shortest length on average, while /aɪ aʊ/ showed the longest. It can be hypothesised that the length of a diphthong is greater, the longer

[4] With respect to tongue position. See vowel quadrilateral in Figure 2.1.

Consonants	p	b	t	d	k	g	f	v	θ	ð	s
Mean	347	365	382	333	387	362	375	375	358	360	391
s.d.	42	104	66	100	72	97	76	87	84	92	84
Consonants	z	ʃ	tʃ	dʒ	m	n	ŋ	l	r	w	j
Mean	353	388	378	375	337	355	370	315	313	342	368
s.d.	91	89	104	79	67	109	84	59	80	79	79

Table 6.2: Average length of consonantal phonemes: Shown are the mean value and the standard deviation for each phoneme (in milliseconds).

the way for the articulators, in particular the tongue, from their position in the first vowel target to their position in the second vowel target is. For example, the tongue has to travel a longer distance from a low position to a high position for /aɪ aʊ/ than for /eɪ/, where both vowel targets require a mid to high, front tongue position.

The length of consonantal phonemes, shown in Table 6.2, was similar to that of long vowels and diphthongs (≈310–390ms), possibly due to using the long vowel context /ɑːCɑː/ and the inclusion of some of the vocalic sample points surrounding the consonantal sample points. However, the larger standard deviations point to larger differences between the speakers. The liquid /l/ and the rhotic glide /r/ exhibited the shortest lengths on average and the fricatives /s ʃ/ the longest lengths. Consonants, or groups thereof, could not be differentiated based on the lengths.

Finally, it was also investigated, whether any difference between female and male speakers could be found with respect to the length. For the short vowels, no significant difference was apparent. The average lengths differed only in the order of 10ms or less. For the long vowels, male speakers tended to exhibit a longer vowel length than female speakers by 10–40ms on average. A similar picture was found for the diphthongs, where male speakers exhibited a longer vowel length than female speakers by 20–80ms. In contrast, female and male speakers produced the consonantal phonemes with a similar length. The differences were usually in the order of 10ms or less. The largest differences were found for /f z ŋ/, where the pronunciation by the male speakers was about 30ms longer on average than

by the female speakers. The length differences clearly highlighted the need for establishing the same number of samples for all individual parameter curves to be able to compare them in the chosen statistical analyses, as has been discussed in Sections 5.4.3 and 6.2, and was done for the analyses in Sections 6.4 – 6.7.

In summary, the RMS parameter and the phoneme lengths provided some measures to separate some of the phonemes into classes which could be useful in the subsequent time-normalised analyses. Sections 6.3.2 and 6.3.3 explore the parameter space further by describing the LDA performed and its results.

6.3.2 Analysis

The theoretical background of the LDA has been described in Section 5.5.2.[5] First of all, the sample points used for the LDA needed to be determined. Not all the sample points of a parameter curve related to a phoneme could be used, due to the computational complexity involved. It was therefore decided to use the following method. For the vowels, first, the sample points with maximum RMS and maximum MH values, respectively, were determined. The audio speech parameter values used in the LDA were those corresponding to the maximum RMS sample point of the mean curve and, similarly, the video speech parameter values used were those corresponding to the maximum MH sample point of the mean curve. The reasoning behind this method was the assumption that the two sample points with the highest energy value and the largest vertical mouth opening, respectively, were the most defining points in the parameter curves. For the consonants, the sample points used in the LDA were determined in a similar manner, except that the minimum RMS and MH sample points were sought. As diphthongs could not be characterised in the same way as vowels, because of their two vowel targets, the positions were manually set based on a visual analysis of the RMS and MH parameter curves. The chosen sample points are shown in Tables H.1 – H.3 in Appendix H. As can be seen from the tables, the sample positions for the audio

[5] The LDA were performed using the R statistical software package as well as the SPSS statistical software because of the additional reclassification functionality provided there.

and video parameters could differ. No temporal correspondence was enforced.

One central issue in the LDA is the handling of outliers, which differ from the group mean value by more than a certain amount (see Sections 5.4.4 and 6.2). This amount was set to 3 standard deviations in this analysis. Of the various ways of handling outliers described in Section 5.4.4, option (3) was chosen, that is, outliers and corresponding curves were removed completely for improved accuracy.

All calculations were performed with normalised values. The normalisation was done by subtracting the mean from the sample values and then dividing the result by the standard deviation. Next, a stepwise LDA was performed where at each step the parameter, that maximised the Mahalanobis distance between the two closest phonemes, was selected for the discriminant function. Similarly, when the discriminant function contained at least three parameters, a parameter could be removed again and replaced by another parameter, if that maximised the discriminant score. The maximally allowed significance for adding a parameter was $p_{in} \leq 0.05$ and the minimally required significance for the removal of a parameter was $p_{out} \leq 0.10$. The calculations were stopped, when the maximum discriminant score was found or the selection of another parameter led to an increase in the discriminant score that was smaller than the pre-set threshold. For n parameters, up to $n-1$ discriminant functions were formed to separate the phonemes from one another. Fewer discriminant functions sufficed at times, depending on how the phonemes were situated in the parameter space.

In this way, the parameters chosen for this study could be analysed in terms of their relative functionality in contributing to the distinctiveness of phonemes. Parameters selected first for LDA's discriminant functions contributed more to this distinctiveness.

6.3.3 Results and Discussion

The results of the LDA are shown in Appendix H. An LDA was performed separately for the vowels (Table H.4), the diphthongs (Table H.10), and the consonants (Table H.12). The vowels were also further split into groups of short vowels (Table

H.6) and long vowels (Table H.8). The summary on the left-hand side of the tables provides the following information (described in more detail in the next paragraphs): the parameters selected for the discriminant functions, the χ^2 value, its significance value p, and the overall accuracy of the reclassification of all phonemes in one group (vowels, diphthongs, consonants) after cross-validation. The parameters are shown in the order that they were selected for the discriminant functions, i.e. the first parameter listed was the one that helped most to separate the phonemes and the other parameters are given in order of their decreasing contribution in discriminating the subsets of phonemes.

The χ^2 *value* is a measure of how well the parameters selected by LDA for its discriminant functions can discriminate the phonemes (or subsets thereof). The null hypothesis is that the measured parameter values of the phonemes do not allow the phonemes to be discriminated. Hence, a large χ^2 value and a small significance value p signify that at least *one* phoneme can be separated well from the others, which was the case in all LDAs performed in this study. However, such results do not mean that *all* phonemes can be separated well from each other by the computed discriminant functions. A complete separation through step by step elimination (and recomputation of the discriminant functions) may, however, be possible [Mardia 79].

More informative are the overall accuracy of the reclassification of all phonemes after cross-validation, as well as the sensitivity and predictivity values for each phoneme. The *accuracy* of the discriminant functions was tested by reclassification after cross-validation for each group of phonemes following the leave-one-out method [Lachenbruch 68]. Cross-validation is a simulated prediction. One individual[6] at a time was removed from the set, the discriminant functions were recomputed, and then the discriminant score was computed for the left out individual. Each individual was thus classified (reclassification) by the discriminant functions derived

[6] This terminology is borrowed from the life sciences, where discriminant analysis is often performed on groups of 'individuals'. In this study, individual generally refers to the parameter curves for each speaker consisting of the sample values. In the LDA described in this section, individual means the chosen sample points (Section 6.3.2).

from all other individuals. Cross-validation was only performed for the individuals in the analysis, i.e. the individuals corresponding to outliers were not considered in the cross-validation. The results are shown on the right-hand side of the above mentioned tables. The *sensitivity* value describes the percentage of individuals for each phoneme that were correctly classified. *Predictivity* refers to the percentage of individuals classified as belonging to a phoneme that were really belonging to it. The confusion matrices for the various groups of phonemes are presented in Table H.5 for all vowels, Table H.7 for the short vowels, Table H.9 for the long vowels, Table H.11 for the diphthongs, and Table H.13 for the consonants.

Vocalic Phonemes

The overall accuracy of the reclassification was only 56.7%, if short and long vowels were considered together, but this improved to 73.3% and 80.2%, respectively, if they were considered separately. As can be seen in the confusion matrix in Table H.5, a large number of misclassifications occurred between pairs of short and long vowels that corresponded to the same articulatory position. This source of error was removed when separating short and long vowels (Tables H.7 and H.9), for example, by their length, as discussed in Section 6.3.1. In the analysis of all vowels, the non-open short vowels /ɪ ʊ ɛ/ and the non-open long vowel /əː/ stood out with low correct classification counts. These three short vowels not only showed a high confusion with their longer counterparts, but also with other vowels and among each other. This was also demonstrated by the confusion matrix for the short vowels only (Table H.7). /əː/ could not be distinguished well from /ɛ/ and /iː/. Open vowels were distinguished better than non-open vowels. It can be hypothesised that long vowels were more correctly classified than short vowels, because the influence of coarticulation was smaller. There was more time to reach the vowel target in long vowels than in short vowels, which influenced the audio and video speech parameters and, thus, may have resulted in lower correct classification scores for short vowels.

The highest overall accuracy was achieved for the diphthongs with 94.7%. Here, misclassifications were rare and sensitivity and predictivity values were often 100%.

This suggests that the classification based on two sample points — one for each vowel target — worked well. /aɪ/ exhibited the most misclassifications with two of its individuals being classified as /ɔɪ/ and one as /əʊ/, possibly due to less internal movements of the articulators, which led to the confusion.

For the vowels, the formant frequencies F_1, F_2, and F_3 were the parameters first selected for the discriminant functions and they were, therefore, the most important parameters in the LDA (Table 6.3). They were followed by the F_0 parameter and the video speech parameters MH and MW. In the analysis of all vowels, MH was first added to the functions, but later removed again, as a combination of other parameters led to a higher discriminant score. In this case, the RTC parameter was added. For the diphthongs, the picture was similar. Here, all parameters but the MW and RMS parameters were used (Table 6.3). However, of the two sample points — one for each vowel target — selected for the diphthongs in this LDA, the video speech parameters seemed to be added mostly with their second sample point. Only the lip protrusion parameter PUL was also added with its first (diphthong) sample point. However, as one of the last parameters to be added, it added relatively little to the discriminant function. In summary, vocalic phonemes were first of all discriminated by the audio speech parameters, except for the RMS parameter, which did not seem to play a significant role here. The video speech parameters played a minor role for the discrimination of vowels. Of them, the MH parameter was the most informative one.

Consonantal Phonemes

The overall accuracy for discriminating consonantal phonemes was low at 44.4%. Here, the bilabials /p b m/, the voiceless velar stop /k/, and the liquids and glides /l r w j/ were discriminated best. They were well separated from other consonants by the computed discriminant functions. At the other extreme, the voiced alveolar stop /d/, the voiced inter-dental fricative /ð/, the alveolar affricates /tʃ dʒ/, and the alveolar and velar nasals /n ŋ/ showed very low sensitivity and predictivity scores. In particular, the alveolar phonemes had many individuals misclassified as

Phoneme class	Parameters in discriminant functions
All vowels	F_1, F_3, F_2, RTC, MW, F_0
Short vowels	F_2, F_1, F_3, F_0, MH
Long vowels	F_3, F_1, F_2, F_0, MH, MW
Diphthongs	F_2^1, F_1^2, F_0^2, F_2^2, F_1^1, MH^2, RTC^2, PUL^2, F_3^1, F_0^1, PUL^1, F_3^2
All consonants	MH, F_2, RMS, RTC, F_1, PUL, MW, F_3

Table 6.3: Summary of parameters selected by LDA for its discriminating functions. Parameters are listed in the order they were selected. The superscript in the parameters for the diphthongs refers to the vowel target position in the diphthong.

other alveolar phonemes, which was possibly due to the comparatively little visible speech articulation in these phonemes. Similarly, /ð/ can be produced not only in an inter-dental manner but also in an alveolar way, so that a confusion with alveolar phonemes was not surprising. This suggests that the parameters selected by LDA were not well-suited to discriminate the alveolar phonemes.

For the consonantal phonemes, the MH parameter was the first and thus most important parameter in the discriminant functions (Table 6.3). It was followed by F_2, RMS, F_1, PUL, MW, and F_3 (in the order they were added to the discriminant functions). Thus, unlike for the vocalic phonemes, the video speech parameters played a significant role in the discrimination of the consonantal phonemes. I believe, such differences have not been reported in the AVSP literature before. Comparisons with AV speech data from other languages are required to determine, if the differences are a particular AuE phenomenon. It could point to the so-called 'lip laziness' of AuE speakers, which may be stronger for vocalic phonemes than for consonantal phonemes. Given the strong role of the MH parameter in the discriminant functions, it did not surprise that the bilabial consonants were distinguished well from the other consonants.

Discussion

Generally, it can be observed in both the sensitivity and predictivity values as well as the confusion matrices that, based on the computed discriminant functions, some phonemes were discriminated well against the others, while other phonemes had a high confusion count. There are two possible explanations for the poor results for some phonemes. Firstly, the (static) parameters used in the analysis may not distinguish well between some phonemes or may require a step by step elimination to separate the phonemes. In addition, the discriminative power of dynamic speech parameters based on the parameters used here should be investigated. Secondly, the LDAs were based on one sample value per speaker per phoneme only (see Section 6.3.2) except for the diphthongs, where there were two sample values, one for each vowel target. Interestingly, the classification results were best for the diphthongs. This leads to the hypothesis that more sample points could also improve the discrimination of phonemes in the groups of vowels and consonants through a contour analysis. A test of this hypothesis is suggested for future work. Applying curve registration before the LDA is also suggested to improve the alignment of individual parameter curves (see Section 5.5.7). Furthermore, speaker normalisation by vocal tract normalisation (Cohen *et al.* [Cohen 95]) or formant frequency warping (Lee and Rose [Lee 98]) could improve the performance of the LDA by normalising the audio speech parameters. However, the issue remains how speakers can be normalised in the video modality. To the best of my knowledge, no studies on speaker normalisation on video speech parameters have been published so far.

6.4 Within-Set Correlation

After the exploration of the data space for the audio and video speech parameters, this section starts the presentation and discussion of the results of the statistical analyses to determine the AV relationships. As described in Section 5.5.3, PCA can be used to check for redundancies in the parameter sets because of its dimensionality reduction property. For each phoneme, a PCA was applied separately to the sets of

audio and video speech parameters, each containing five parameters (cp. Sections 5.1, 5.2, and 6.1). With the help of the PCA and a linear correlation analysis (see Section 5.5.4), redundant parameters were searched for in each set, which could be eliminated from the further analysis. Results are summarised in the remainder of this section, while individual results can be found in Appendix D.

6.4.1 Video Parameter Set

The case was very clear-cut for the video parameters. For all phonemes, the first four PCs explained at least 96% of the variance (Tables D.1 and D.2). For more than half of the phonemes, the first four PCs already explained 99% of the variance in the data. This was particularly the case for the vocalic phonemes, but also occurred for many consonantal phonemes. In other words, the parameters contained a considerable amount of redundancy.

Identifying a redundant parameter (or redundant parameters in general) is generally not a simple straightforward task, because the PCA transforms the coordinate system of the data into a new one according to the orthogonal PCs. A PC can be the new representation of only one parameter in the old space, but it can also be the new representation of a combination of parameters, in which case the identification of the redundant information is difficult.

However, in the case of the video parameter set, it was hypothesised that the redundant information was most likely to be found in the two lip protrusion parameters. Upper and lower lip are typically moved simultaneously and in a similar fashion, so that a high correlation between the two parameters was expected. This view was confirmed by the results of a pairwise linear correlation analysis (Tables D.3 and D.4). For vocalic phonemes, the correlation coefficient r of the parameters PUL and PLL ranged from 0.91 to 0.99. The correlation coefficient r was slightly smaller for consonantal phonemes, ranging from 0.79 to 0.95, but still confirmed a strong correlation. The hypothesis was further supported by leaving one of the protrusion parameters out, for example the lower lip protrusion parameter PLL, and then repeating the PCA. The results showed very little difference to the previous

ones. The cumulative proportion of the variance explained by the first three PCs changed by less than 0.05, with the first PC showing a change in the proportion of variance of <0.1, and second and third PCs showing even smaller changes. The video speech parameter set, therefore, can be reduced to four parameters for the remaining analyses by eliminating one of the redundant lip protrusion parameters. Without loss of generality, the PLL parameter was eliminated in this study.

No other strong correlation ($r \geq 0.75$) between any other parameters was found. Weaker correlations were found for some other pairs of parameters. These correlations were more apparent in the consonantal phonemes than in the vocalic ones, but in either case occurred only for some but not all phonemes. In particular, the correlation between MH and RTC, as well as between MW and MH stood out.

The former was found with correlation values between 0.43 and 0.55 in the bilabial stops /p b/, the labio-dentals /f v/, the voiced alveolar /d/, and the bilabial glide /w/. Also, even weaker, it was found in the rhotic /r/, the bilabial closure /m/, the velar closure /ŋ/, and the voiceless inter-dental fricative /θ/. All these phonemes have in common that the lips close completely or almost completely, so that it was expected to see the RTC parameter decrease as the MH parameter decreases and vice versa. The data supported this expectation. Overall, the correlation was clearly lower for vocalic phonemes than for consonantal ones. On a relative comparison, the correlation values for low vowels (cp. Figure 2.1 in Section 2.1.3), and diphthtongs containing these vocalic positions, appeared to be higher than those for high vowels and diphthongs.

The latter parameter pair, MW and MH, was found with a stronger correlation ($0.42 \leq r \leq 0.60$) for the bilabial glide /w/, the short vowel /æ/, the long vowels /ɑː iː/, and the diphthong /aʊ/. A weaker correlation ($0.31 \leq r \leq 0.40$) was experienced for the bilabial stops /p b/, the long vowel /ɜː/, and the diphthongs /eɪ ɪə/. This correlation was also found with negative correlation values in some phonemes. The slightly stronger correlation ($-0.43 \leq r \leq -0.48$) was demonstrated in the bilabial closure nasal /n/, the voiceless alveolar stop /t/, and the voiced alveolar fricative /z/. Indications of weaker correlations ($-0.30 \leq r \leq -0.33$) were seen in the voiceless velar stop /k/, the velar closure nasal /ŋ/, and the voiceless

alveolar fricative /s/.

What all these phonemes have in common is their place of articulation, which is front or front-central. The bilabial glide /w/ is the prime example for the positive correlation values. Both MW and MH decrease and increase simlutaneously in the process of rounding the lips to articulate this phoneme. The negative correlations on the other hand were found for phonemes, that do not require a fully closed mouth, but rather have the lips apart at a small vertical distance and also result in an increase in mouth width. Again, these results matched the expectations.

6.4.2 Audio Parameter Set

The situation was not as clear-cut for the set of audio speech parameters, as it was for the video speech parameters. Examining the results of the PCA with five parameters (Tables D.5 and D.6), it can be seen that the first four PCs cover 90–97% of the variance, which suggests that some parameters were correlated and that there was thus redundancy in the data. However, as is shown in Tables D.7 and D.8, no single pair of parameters stood out as in the case of the video speech parameters. This suggests that it was rather a case of more than one parameter being correlated with one or more than one other parameter. In that case, it was only after the PCA, with the orthogonal PCs forming a new coordinate system, that four 'new' parameters were able to express an average of 94% of the variance. Hence, it was not possible to eliminate one particular parameter. All five parameters were included in the further analyses.

However, a few general points can be made from the results of the correlation analysis. Table 6.4 summarises these by showing the phonemes for each pair of parameters in the set of audio speech parameters, where the absolute value of the correlation coefficient $r \geq 0.5$. No phonemes had correlation values $r \geq 0.5$ for the parameter pairs F_0 - F_1, F_0 - F_2, and F_0 - F_3. This is in line with a study by Kosiel [Kosiel 73] on Polish vowels, which also found no correlation between the voice source excitation frequency F_0 and the formant frequencies F_1 - F_4. In the present study, the highest correlation values (for almost all phonemes) were found

Parameter Pair	+/−	Vowel & Diphthong	Consonant
F_0 - RMS	pos.	uː	v m r w j
F_1 - F_2	pos.		w
	neg.		d z ʃ dʒ j
F_1 - F_3	pos.		r
	neg.		j
F_1 - RMS	pos.	ɪ ʊ ɒ ʌ æ əː ɔː ɑː ɜː eɪ ɔɪ aɪ aʊ əʊ	b d g v ð z dʒ
F_2 - F_3	pos.	ɪ iː əː eɪ ɪə	t f s ʃ tʃ dʒ j
F_2 - RMS	pos.	ɛ iː əː	
	neg.	ʊ ɔː	t k g f θ ð s z ʃ tʃ dʒ
F_3 - RMS	neg.		s ʃ tʃ dʒ

Table 6.4: Phonemes where the values of pairwise parameter correlation within the audio set fulfilled $|r| \geq 0.5$. Empty fields mean there were no phonemes with $|r| \geq 0.5$ for that parameter pair.

between F_1 and RMS or between F_2 and RMS. The values reached up to $r = 0.78$. Exceptions were

- the long mid-high to high, central to back vowel /uː/ ($r = 0.58$) and the bilabial closure nasal /m/ ($r = 0.56$) where F_0 - RMS was strongest,
- the bilabial glide /w/ ($r = 0.80$) and palatal glide /j/ ($r = -0.85$) where F_1 - F_2 was strongest (Note: highest correlation of all phonemes),
- the rhotic /r/ ($r = 0.68$) where F_1 - F_3 was strongest,
- the vowels /ɛ/ ($r = 0.60$) and /ɔː/ ($r = -0.75$), the voiceless alveolar and velar stops /t/ ($r = -0.65$) and /k/ ($r = -0.63$), and the voiceless fricatives /f/ ($r = -0.61$), /θ/ ($r = -0.51$), /s/ ($r = -0.68$), and /ʃ/ ($r = -0.73$) where F_2 - RMS was strongest, and
- the high front vowels and diphthongs /iː/ ($r = 0.62$), /eɪ/ ($r = 0.62$), /ɪə/ ($r = 0.71$), the palatal voiceless fricative /ʃ/ ($r = 0.73$), and the affricates /tʃ/

($r = 0.75$) and /dʒ/ ($r = 0.68$) where F_2 - F_3 was strongest.

In general, strong correlations of formant frequencies with the RMS parameter were expected based on considering the source-filter model, the shape of the excitation spectrum, the lip openness, and the placement of the vowels in the F_1 - F_2 plane. The data confirmed these expectations.

While correlations between F_0 and RMS as well as F_1 and RMS were positive for all phonemes, correlations between F_2 and RMS were negative for some phonemes. For the vocalic phonemes, these were /ʊ ɒ ɔː ɔɪ aɪ/. Medium strong positive correlations ($r \geq 0.5$) were found for the front and central, mid-high to high vowels and diphthongs /ɛ iː əː ɪə/. For the consonantal phonemes, medium strong negative correlations with $r \leq -0.5$ occurred for the alveolar and velar stops /t k g/, the fricatives /f θ ð s z ʃ/, and the affricates /tʃ dʒ/. It must also be kept in mind that the /bVb/ and /ɑːCɑː/ contexts used in this study were both voiced contexts, which influenced the central phoneme to some extent due to coarticulation. Nevertheless, the fact remains that correlation was stronger for voiced phonemes than for unvoiced phonemes.

For the consonantal phonemes, it was also noticed that the correlation between F_1 and RMS was always stronger for the voiced phoneme in a pair of voiceless and voiced phonemes like /p b/, for example. For the vocalic phonemes, no relationship between tongue position and correlation value was evident for this pair of parameters. However, correlation between F_1 and RMS was generally fairly strong for the vocalic phonemes ($r \geq 0.5$), which are voiced phonemes, so that it can be reasoned that voicing generally leads to a strong correlation between F_1 and RMS.

Also, correlations between F_2 and F_3 were medium strong to strong ($r \geq 0.40$) for some phonemes. These were the front and central vowels and diphthongs /ɪ æ iː əː eɪ ɪə/, as well as consonants produced by a constriction of the airflow towards the front of the oral cavity (typically without completely blocking the airflow as in the bilabials), such as the alveolar stops /t d/, the fricatives /f θ ð s z ʃ/, the alveolar affricates /tʃ dʒ/[7], and the palatal glide /j/.

[7] /ʃ tʃ dʒ/ had the highest correlation overall between F_1 and F_2 with $0.68 \leq r \leq 0.75$.

Finally, a medium strong negative correlation ($r \leq -0.5$) between F_3 and RMS was found for the fricatives /s ʃ/ and affricates /tʃ dʒ/. It was noticed that when F_3 - RMS were correlated strongly, so were F_2 - RMS, but this relationship did not hold in the opposite direction (e.g. see the velar stops /k g/).

6.4.3 Summary Within-Set Correlation

The results of a PCA and linear pairwise correlation analysis separately on each parameter set revealed that redundancy was present in each parameter set. For the video speech parameter set, the redundancy could be identified in the lip protrusion parameters PUL and PLL, which were very strongly correlated. It was therefore decided, to eliminate one lip protrusion parameter from the subsequent analyses. Some other correlation trends were found between parameters MH and RTC as well as between MW and MH, but these correlations were more phoneme-specific. For the audio speech parameters, the PCA results also indicated redundancy between the parameters, but the linear pairwise correlation analysis revealed no parameter pair that was strongly correlated across all phonemes, as for the video speech parameters. Some phoneme-specific strong correlations were found, mostly between F_1 and RMS, and between F_2 and RMS. The results suggest that the redundancy in the audio speech parameters was spread over the parameters. Hence, no parameter was eliminated from the subsequent analyses. It shall be pointed out again, that the samples of the central phonemes in the /bVb/- and /ɑːCɑː/-contexts contained some sample points with sample values influenced by the context phonemes due to coarticulation. Therefore, it is possible that some of the correlation found (and similarly, some of the correlation not found) in the results of the pairwise correlation analysis was precluded by the use of the contexts. However, by using the same vocalic and consonantal contexts for all vocalic and all consonantal phonemes, respectively, the results are comparable between phonemes. It is suggested for future work, to investigate the correlation for other contexts, such as /iːCiː/, /uːCuː/, /ɔːCɔː/ etc. Together with the current results on these phonemes, it can then be analysed how much influence the context has on the results.

PC	F_0	F_1	F_2	F_3	RMS	MW	MH	PUL	RTC
1	0.83	0.50	0.64	0.67	0.42	0.83	0.66	0.54	0.65
2	0.11	0.21	0.15	0.14	0.23	0.12	0.18	0.22	0.22
3	0.04	0.12	0.09	0.08	0.14	0.03	0.10	0.12	0.09
1	0.86	0.44	0.54	0.60	0.42	0.85	0.65	0.50	0.80
2	0.07	0.26	0.21	0.19	0.23	0.10	0.20	0.24	0.13
3	0.04	0.14	0.12	0.10	0.16	0.03	0.11	0.12	0.05

Table 6.5: Average proportion of variance (rounded to 2 decimal places) explained by the top three PCs for each parameter. Top: Vocalic phonemes. Bottom: Consonantal phonemes.

6.5 Shape Analysis of Parameter Curves

In the previous section, the results of applying PCA to each set of parameters for each phoneme to check for redundancies in the sets have been discussed. However, as described in Section 5.5.3, PCA has also gained importance as a statistical shape analysis technique in recent years. For this, a PCA was applied to each parameter separately for each phoneme. In other words, PCA was performed on the temporal domain. The main modes of variation in the shape of the parameter curves were determined and thus the relationship between sample points and PCs was revealed. In addition, it allowed for a compact representation of the individual parameter curves in the further analyses.

The numeric results for each phoneme can be found in the tables E.1 – E.40 in Appendix E, which show the cumulative proportion of the variance explained by the first six PCs. This is followed by a visualisation — in the form of star charts — of the amount of variation explained by the first and second PCs (Figures E.1 – E.4). The visualisation helps to find similarities and differences between the phonemes (see below). The shape of a star chart offers a quick way of identifying parameters with unusual proportions of variance. A summary of the average individual proportion of variance expressed by the top three PCs is provided in Table 6.5.

Phoneme	ɪ	ʊ	ɛ	ɒ	ʌ	æ	iː	uː	ɜː	ɔː	ɑː	əː	eɪ	ɔɪ	aɪ	aʊ	ɪə	əʊ
F_0	2	2	2	2	2	2	2	2	2	2	2	1	2	2	2	1	2	3
F_1	4	3	5	5	4	4	4	4	4	4	4	4	5	4	5	4	5	4
F_2	3	3	3	5	4	3	4	3	2	4	3	4	4	4	4	4	3	3
F_3	4	3	3	4	4	4	4	3	3	4	2	4	4	5	3	3	4	4
RMS	3	4	4	5	4	4	4	4	4	5	5	5	5	5	5	5	5	4
MW	1	1	2	1	2	2	2	1	2	2	3	2	2	2	2	2	2	1
MH	1	1	2	2	2	3	3	2	2	1	3	3	4	3	3	4	3	3
PUL	3	3	3	2	3	4	4	2	5	2	4	4	4	4	5	5	4	3
RTC	2	1	2	3	2	3	3	2	3	2	3	3	4	3	3	3	3	3

Table 6.6: Number of principal components needed to explain $\geq 90\%$ of the temporal variance: Vocalic phonemes.

The results shown in Table 6.5 are similar for both vocalic and consonantal phonemes. For the F_0 and MW parameters, the first PC already explained on average about 85% of the variance. The second PC for these parameters covered about 10% of the variance. The influence of the third and further PCs was very small. A second group of parameters was formed by those, where the first PC explained on average between 60–80% of the variance. These parameters were F_2 and F_3 of the audio speech parameter set, and MH and RTC of the video speech parameter set. Here, the second PC expressed about 15–20% of the variance and the third PC about 5–10%. Finally, only about 40–55% of the variance was explained on average by the first PC for the parameters F_1, RMS, and PUL[8]. The second PC covered about 20–25% of the variance and the third PC about 10–15%.

However, it must be noted that the results differed considerably for certain phonemes, which would perhaps not be expected from the average results. Tables 6.6 and 6.7 and Figures E.1 – E.4 show the results in more detail. The number of PCs required for each parameter and each phoneme to express $\geq$90% of the variance

[8] Note that a strong correlation between protrusion of upper and lower lip was shown in the previous section. One protrusion parameter can be substituted by the other.

Phoneme	p	b	t	d	k	g	f	v	θ	ð	s
F_0	3	1	3	1	2	2	2	1	2	2	2
F_1	4	4	4	4	5	4	5	5	5	5	4
F_2	4	4	4	3	4	3	5	5	5	4	5
F_3	3	2	4	3	4	3	5	4	3	4	4
RMS	4	5	5	4	5	5	5	5	4	5	5
MW	2	2	2	1	1	2	2	2	1	1	3
MH	3	3	3	3	2	2	3	3	3	3	3
PUL	4	4	4	4	4	4	1	3	4	3	4
RTC	3	3	2	1	1	1	3	3	3	2	2
Phoneme	z	ʃ	tʃ	dʒ	m	n	ŋ	l	r	w	j
F_0	1	4	3	2	1	1	1	1	1	1	2
F_1	4	5	5	4	3	4	4	4	3	4	3
F_2	3	4	5	4	3	3	4	2	3	3	3
F_3	4	4	4	4	2	2	4	2	3	4	3
RMS	5	5	5	5	3	3	4	4	3	4	4
MW	1	2	2	2	1	2	1	1	2	2	1
MH	3	3	3	2	3	3	2	3	3	3	3
PUL	5	4	4	5	4	3	3	3	4	3	4
RTC	3	2	1	2	3	1	1	1	2	2	1

Table 6.7: Number of principal components needed to explain $\geq 90\%$ of the temporal variance: Consonantal phonemes.

are presented. As can be seen in the tables, only two PCs were typically needed for the F_0 and MW parameters, for many phonemes even just one. Examining the star charts, the exceptions in F_0 occurred for the phonemes /əʊ p t ʃ tʃ/ and in F_1 for /ɑː p ʃ n/. Significantly more PCs were needed for the formant frequency F_1 with an average of four PCs to reach 90% variance. The much smaller (more centrally located) F_1 results can also clearly be seen in the star charts, with the exception of /ɪ iː/. The MH parameter was also fairly consistent in the number of

PCs required for 90% variance across all phonemes. Here, three PCs were needed on average. The star charts also show the exceptions with a smaller first PC but a second PC larger than that for other phonemes. These exceptions were the vowels /aː əː/, the diphthongs /eɪ ɔɪ aɪ aʊ əʊ/, and the consonants /f v m w/. For all other parameters, the number of PCs required ranged from 1 to 5 with no obvious systematic pattern.

Modes of Shape Variation

An interesting aspect was now to analyse what variation in the shape of the parameter curves the PCs stood for. This was done by computing the pointwise mean parameter curve, as well as the pointwise standard deviation (based on the distribution of the individual parameter curves from the 20 speakers) for each phoneme-parameter pair. Then, for each phoneme-parameter pair and each PC, the mean curve and two curves with ± 10 standard deviations were drawn in a graph. Figure 6.11 shows a typical example for the first three PCs.[9] There were three main modes of variation

- a vertical shift,
- a mode related to the slope of the curves, and
- a mode describing the horizontal range or a horizontal shift.

In a striking way, the first PC was in 88% of all phoneme-parameter pairs related to a vertical shift of the parameter curve (Figure 6.11 left). In other words, the strongest variation for the individual curves of the speakers was in these cases not related to differences in the curve shape (e.g. the slope of a curve) but to a mere shift (or offset), which appeared to be a personal characteristic of each speaker. This shift occurred for all sample points and was almost invariant in size. In contrast, the second and third PC expressed variation in the curve shape. These PCs were

[9] This kind of visualisation leads to a very large number of graphs which could not possibly be included in this thesis. However, the interested reader can find them on the accompanying CD-ROM in the directory 'PCExplanation'.

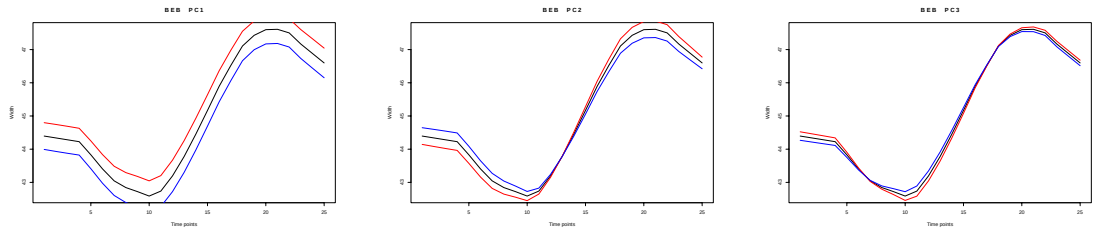


Figure 6.11: Typical modes of variation by the top three PCs on the example of the phoneme /ɛ/ and the *MW* parameter. Shown are the mean curve (black) and curves showing the effect of the PC at ±10 standard deviations (red and blue).

related to the slope of the curve and the horizontal range or shift. For some sample points, these PCs had no effect, while their effect was considerable at other sample points (Figure 6.11 centre and right). Both modes of variation occurred in either or both the second and third PC, i.e. in some cases, the slope was found in the second PC and the horizontal range in the third PC. In other cases, it was the opposite way, or the second and third PCs expressed a mix of the two modes of variation.

In the other 12% of the phoneme-parameter pairs, the first PC did not express a vertical shift. Rather, it expressed one of the other modes of shape variation, such as variation in the slope or in the horizontal range. The vertical shift still existed, but was of a lesser degree. It occurred as only the second or third PC. The parameters, where this behaviour was most common, were the *RMS* parameter (25 out of 40 phonemes) and the *PUL* parameter (12 phonemes). Table 6.8 presents a list of the affected phonemes. It can be seen that the F_0 and *MW* parameters had no exception to the vertical shift being the strongest PC. Referring back to the observations in Section 6.1, this was no surprise, because the individual parameter curves exhibited a similar curve shape but at a different vertical place. The F_0 and *MW* parameters were also the ones with the highest proportion of variance explained in the first PC (Table 6.5).

Consequences

Finding these three main modes of shape variation had consequences for the further analyses in terms of what data was used as input. The focus of interest in this study

Parameter	Phonemes
F_0	none
F_1	v z w j
F_2	θ ŋ
F_3	s
RMS	ɪ ʊ ɛ ɒ ʌ ɔː əː aɪ ɪə p b t d k g f v θ ð s z ʃ tʃ dʒ w
MW	none
MH	m
PUL	ɒ ʌ ɜː ɑː eɪ iː p t v z ŋ w
RTC	ʌ ɔː

Table 6.8: List of phonemes for each parameter, where the vertical shift was not represented by the first PC (but by the second or third PC).

was on the common characteristics of the individual parameter curves for a certain phoneme, i.e. what are the similarities for all speakers. For example, having two curves of similar shape, but with a vertical shift between them, the focus is on the similar shape, not the vertical shift which is related to a particular speaker. Consequently, the PC, which expressed the vertical shift, was not used as input for the further analyses. Instead, the analyses concentrated on the two PCs that were related to the slope of the curve and the horizontal range or shift. Other PCs were too small in their proportion of variance to have a considerable influence on the results and were thus neglected.

With the vertical shift being such an important factor in the variation of many parameter curves, curve registration (see Section 5.5.7) is again a process worth considering. Curve registration would minimise the influence of a vertical shift between curves on the modes of variation found in a PCA by minimising differences between curves on the y-axis. It can be considered as a form of speaker normalisation.[10] However, curve registration is a complex task in its own right and the time and effort required to perform it on a large scale is beyond the scope of the work

[10] Another option would be a zero mean normalisation of the parameter curves.

presented here. For the further analyses, the PC related to the vertical shift was eliminated and the analyses performed with the other two main modes of variation. Nevertheless, some experiments with a curve registration process were performed and the results of these experiments are described in Section 6.7.

6.6 Between-Set (Audio-Video) Correlation

After examining the within-set correlation for the sets of audio and video speech parameters in Section 6.4, the focus is now on the between-set or AV relationship, which is the central theme of this thesis. The theoretical background of the statistical analyses has been given in Sections 5.5.4 – 5.5.6. First, the results of a pairwise linear correlation analysis across the two sets are analysed in Section 6.6.1. This is followed by the presentation and discussion of the results of the canonical correlation analysis in Section 6.6.2. Issues with collinearity and statistical stability in that analysis led to the application of the coinertia analysis, which produces stable results. These results are presented and discussed in Section 6.6.3. Finally, a summary of the results of the between-set relationship is given in Section 6.6.4.

As before, it is important to keep in mind that all results reported here came from the analysis of the central phonemes in the /bVb/- and /ɑːCɑː/-contexts. The terms 'vocalic phonemes' and 'consonantal phonemes', respectively, refer to the central phonemes in these CVC- and VCV-words. Because of the analysis of phonemes in such contexts, the effect of coarticulation on the results must be kept in mind.

6.6.1 Pairwise Correlation

Similar to the pairwise correlation analysis within each parameter set (discussed in Section 6.4), pairwise correlations across the two sets were performed. The numeric results can be found in Tables F.1 – F.5 in Appendix F. For this analysis, the smoothed and resampled parameter values were used (cp. Section 5.4).

Generally, the pairwise correlations between audio and video speech parameters

were small in value. No parameter pair showed a very strong correlation ($|r| \geq 0.75$) across all phonemes for the speakers recorded in the AVOZES data corpus. Thus, the data did not support a hypothesis of a direct 1–1 relationship between any of the speech parameters in the two sets. However, this does not mean that the parameters were unrelated, as the possibility existed of a combination of audio speech parameters correlating well with a combination of video speech parameters, or a non-linear relationship between parameters, that was not uncovered by the statistical analyses presented here. Combinations of parameters were investigated using CANCOR and COIA (see below Sections 6.6.2 and 6.6.3). Also, it should be noted that some weak and medium strong pairwise correlations and correlation trends were found in the results, which are described in the next paragraphs.

For the vocalic phonemes, the following parameter pairs and phonemes were found with $|r| \geq 0.40$:

- F_1 - MW: /eɪ/ ($r = -0.41$),
- F_2 - MH: /ʌ/ ($r = -0.48$), /eɪ/ ($r = 0.46$), /ɔɪ/ ($r = 0.51$),
- F_2 - RTC: /ʌ/ ($r = 0.43$), /eɪ/ ($r = 0.41$), and
- F_3 - RTC: /eɪ/ ($r = -0.51$).

No particular pattern was evident. Furthermore, it was noticed that the correlation values for the parameter pair F_0 - MW were always negative and for RMS - MH always positive, although $|r| = 0.38$ at most. The former trend was related to the observation, that the F_0 frequency rose slightly at the onset of vowel or diphthong and then decreased again at the end, while the mouth width often decreased slightly at the onset and then rose again slowly. Negative correlation values were thus expected. In the case of the second trend, both RMS and MH rose from small values to a maximum during the vocalic phoneme and then returned to small values again due to the bilabial context. Positive correlation values were therefore expected. Both trends were related to the uniform bilabial context /bVb/ used in the AVOZES corpus.

For the consonantal phonemes, the parameter pairs and phonemes with $|r| \geq 0.40$ were:

- F_0 - MW: /b/ ($r = -0.40$), /t/ ($r = -0.45$), /θ/ ($r = -0.41$), /m/ ($r = -0.46$), /r/ ($r = -0.51$),
- F_0 - RTC: /t/ ($r = 0.40$), /d/ ($r = 0.42$), /g/ ($r = 0.42$), /tʃ/ ($r = 0.47$), /ŋ/ ($r = 0.48$),
- RMS - MH: /p/ ($r = 0.54$), and
- RMS - RTC: /p/ ($r = 0.51$).

The reader is reminded that the problem of a reliable automatic F_0 estimation for all phonemes is not yet solved (see Section 5.1.3). While the voice source excitation frequency is well-defined for voiced phonemes, it is not for unvoiced phonemes. In addition to coarticulation effects, the algorithm used in the ESPS command `get_f0` includes some smoothing, which has the potential of bridging over short unvoiced segments such as, for example, the intervocalic /t/. The results for the unvoiced phonemes must, therefore, be treated with some care.

In addition, similar to the case of the vocalic phonemes, a trend of weak to medium strong negative correlation for all phonemes (exception /ʃ/) for the parameter pair of F_0 and MW (maximum $|r| = 0.51$) and a trend of a weak to medium strong positive correlation for many phonemes for the parameter pair of RMS and MH (maximum $|r| = 0.54$) were noticed. Again, it is important to keep in mind that the results were specific for the /ɑːCɑː/-context. In the first parameter pair, examining again the parameter curves in Appendix C, both parameters F_0 and MW had mostly flat curves, so that a weak correlation was found. The F_0 values showed more variation during the intervocalic consonant for the voiceless consonants, than for the voiced consonants. As a result, the correlation values for the parameter pair of F_0 and MW were smaller for the voiceless consonants, with the exception of the alveolar voiceless stop /t/ and voiceless inter-dental fricative /θ/. For the second correlation trend, pairwise correlation values were higher for phonemes, where the RMS values as well as the MH values first decreased towards a minimum and

then increased again during the intervocalic consonant. No correlation existed for phonemes like /l n/, where RMS and MH parameter curves showed more variation and less common behaviour. Very low correlation values were also found for the bilabial closure /m/ and the bilabial glide /w/. Here, the MH parameter went to a minimum during lip closure but the RMS parameter did not show such a minimum.

In summary, pairwise correlations between parameters of the audio and video sets were generally weak. No strong correlation was found for any pair across all or a large number of phonemes. However, some trends of weak to medium strong correlation were found for many phonemes for the pairs of F_0 and MW as well as RMS and MH. It seems that these parameters were related in some way, but this way was not a linear pairwise relationship. Hence, the results of statistical analyses that explore the relationship between combinations of parameters were analysed and are described in the following subsections.

6.6.2 Canonical Correlation Analysis

CANCOR is a statistical analysis for the exploration of relationships of linear combinations of variables (see Section 5.5.5). Since a linear (pairwise) relationship of audio and video speech parameters was not found for the data in the AVOZES corpus and the parameters measured, the investigation was continued with a check for relationships between combinations of parameters. In summary, CANCOR performs a rotation of the coordinate system, such that the correlation between the linear combinations is maximised. Most of the covariance and the highest correlation values between parameter sets are found in the first few canonical variates, so that it often suffices to analyse the first 1–3 canonical variates. Successive pairs of canonical variates are uncorrelated.

For small samples, where the number of parameters approaches the sample size, CANCOR can suffer from statistical instability, as has been described in Section 5.5.5. In this study, the sample size was $N = 20$, the audio set had $p = 5$ parameters, and the video set $q = 4$ parameters. Ideally, both the PCs — identified in Section 6.5 — related to the slope of the parameter curves and the horizontal range or

shift, respectively, would have formed the input for CANCOR. However, taking 18 parameters ($(p + q = 9) * 2$) into a CANCOR of sample size $N = 20$ was almost surely leading to $r_1 \rightarrow 1$ and thus the results would have been of little value due to collinearity. As a consequence, only the PC related to the slope of the curve was taken into account, considering it the more important one compared to the horizontal range PC. The PCs were normalised to zero mean and unit variance as a prerequisite for CANCOR.

Canonical Weights and Correlation Coefficients

The results of the CANCOR can be found in Section F.2 in Appendix F. Tables F.6 – F.8 show for each phoneme the computed canonical weights for each parameter and the canonical correlation coefficient r_1 of the highest canonical correlation. The *canonical weights* are coefficients, which when applied to the measured parameters, result in the canonical variates, i.e. the variables in the new coordinate system. In principle, the magnitude and sign of canonical weights can be used to indicate the importance and effect of the parameters. Scaling effects were removed by the applied normalisation. As Gittins [Gittins 85] points out, canonical weights relate to the unique part of variables, rather than the common parts. In other words, the weights indicate the variables contribute something distinct to the canonical variate. However, this distinct 'something' is nevertheless related to the linear combination of the other parameter set. It should also be noted that the sign of the canonical weights is of little consequence, as a reversal of all the signs does not affect the analysis. It is the pattern of the signs that is of interest. For easier comparison, it was arbitrarily chosen to set a positive sign for F_0 and to change the signs of all other weights accordingly, where necessary.

Some words of caution, a substantive interpretation of the pattern of weights is generally difficult [Gittins 85]. In the same book, Gittins further remarks that the weights depend on the selection of parameters for the analysis and on the sample-specific variation. According to Gittins, canonical weights are known for their instability and for small changes in the parameters to have great effects. Factors

contributing to this instability are insufficient sample size, measurement errors, and collinearity of the variables. Nevertheless, an analysis of the canonical weights can give some valuable insight into the relationships between the parameters that constitute the linear combinations.

For the vocalic phonemes, the first canonical correlation coefficient r_1 was highest for /əː/ with a value of 0.87 and smallest for /ɪ/ ($r_1 = 0.59$), /əʊ/ ($r_1 = 0.60$), and /ɒ/ ($r_1 = 0.61$). The latter three phonemes were also the only ones for which $r_1 < 0.75$. The average first canonical correlation coefficient was $\bar{r}_1 = 0.77$. This points towards a strong correlation between the first canonical variates, which supports the hypothesis that combinations of parameters from each modality are highly related, not single parameters.

Most canonical weights for the interconsonantal vocalic phonemes were small in magnitude — of the order of 0.10 and less — with few exceptions. The largest weights were found for the MW weight of vowel /ʊ/ at 0.22 and for the MH weight of the diphthong /əʊ/ at 0.21. Most canonical weights, however, were very similar in magnitude and a sign pattern was not evident. No parameter(s) could be singled out, which contributed notably more than others to the relationship for all vocalic phonemes or for any obvious subset of them. Again, it was the linear combination of the unique parts of the parameters that correlated well across the two sets. Based on the CANCOR of the data in the AVOZES data corpus, it can be said that all nine parameters had something to contribute in relating the audio to the video speech parameters and vice versa.

For the consonantal phonemes, the highest r_1 was found for /t/ with a value of 0.87 and the lowest r_1 for /r/ at 0.57. Other consonantal phonemes for which $r_1 < 0.75$ were /b g f v ʃ tʃ dʒ/. The average first canonical correlation coefficient was $\bar{r}_1 = 0.75$, which was very similar to the value of 0.77 for the vocalic phonemes. As in the case of the vocalic phonemes, a generally strong correlation between the first canonical variates was found. It can therefore be hypothesised, that linear combinations of the parameters in the two sets were similarly well related for all phonemes.

The canonical weights for the intervocalic consonantal phonemes were also mostly

small in magnitude — of the order of 0.10 and less — with some exceptions. The largest weight was found for the PUL parameter of fricative /f/ at -0.48. Other canonical weights, that were larger than the vast majority of weights, were the F_0 weight for the fricative /v/ (0.39) and the nasal /n/ (0.31), the MH weight for the fricative /f/ (0.22) and the glide /j/ (0.24), and the RTC weight for the oral stop /k/ (0.20), the fricatives /z/ (-0.26), and the liquid /l/ (-0.23). Most other canonical weights were very similar in magnitude. Overall, no parameter could be shown to distinctively contribute more (or less) to the canonical correlation than other parameters. Similar to the vocalic phonemes, it was the combination of all the parameters in one set that resulted in a high correlation with a combination of all the parameters in the other set, not single parameters. With respect to the signs of the weights, it appeared that F_0 on the one hand and F_1 and F_2 on the other hand often had opposite signs, but a systematic pattern for all consonantal phonemes — or for certain groups of them — was not evident.

Condition Number

The tendency of r_1 towards unity, when the sample size is similar to the number of parameters, has already been commented on. In this study, an analysis with 9 parameters and a sample size of $N = 20$ was performed. If the measurement error variances are known or can be estimated reliably, canonical correlations and variates can be corrected for attenuation due to measurement error [Gittins 85]. Unfortunately, the errors are unknown for the AVOZES data and cannot be reliably estimated, as the 'ground-truth' is unknown. Collinearity in the parameters often leads to ill-determined canonical weights. The *condition number* $\kappa(\mathbf{X})$, where $\mathbf{X}$ is a data matrix of full rank, however, is a sensitive indicator of collinearity [Gittins 85]. It is defined as the ratio of the largest to the smallest singular value of $\mathbf{X}$:

$$\kappa(\mathbf{X}) \;=\; d_{max} \,/\, d_{min}. \tag{6.1}$$

For ease of interpretation, the reciprocal condition number $1/\kappa(\mathbf{X})$ is often analysed and it was also used in this work (Tables F.6 – F.8). A value close to 0 signifies collinearity and hence instability, while a value tending to 1 strongly supports the

accuracy of a statistical analysis. It is not clearly documented in the literature, what reciprocal condition number can be considered as a clear sign of collinearity. In this study, it was judged that values below 0.20 pointed to some element of collinearity in the data, while $1/\kappa(\mathbf{X}) < 0.05$ was a strong sign for collinearity.

The reciprocal condition numbers for the data matrices used in the CANCOR are also shown in Tables F.6 – F.8. For the AVOZES data matrices, the reciprocal condition numbers lay mostly between 0.20 and 0.50. No number was below 0.05, but some numbers below 0.20 were found. For the audio matrices, this occurred for the phonemes /eɪ m n v w j z/. For the video matrices, the phonemes /d k/ were found with condition numbers below 0.20. Hence, there is support, that the data matrices used in the CANCOR, were mostly free of collinearity between the vectors making up the matrices. In some cases, the condition numbers pointed to some collinearity, which could have affected the CANCOR. It is possible to eliminate collinearity by an orthogonal transformation (e.g. a PCA on the data matrix and then using the resulting PCs), but this has the disadvantage that the meaning and relationships of the resulting variables are not clear and, thus, an interpretation of the results is difficult. No such transformation was performed in this study.

Because of the difficult interpretation of canonical weights and correlation coefficients, other methods of interpreting the CANCOR have been developed [Gittins 85]. These are structure correlations, variance extracted by a canonical variate, redundancy, and total redundancy, which are described and analysed on the next pages.

Structure Correlations

One method of interpreting canonical variates is to analyse the correlation between the vectors forming the data matrix — the parameters — and the canonical variates. These correlations are also known as *structure correlations*. Compared to canonical weights, they have the advantage of smaller standard errors and greater stability, in particular for small and medium sized samples, such as in this study, or in measurements, where parameters of either or both sets are intercorrelated [Gittins 85].

The absolute value and the sign of structure correlations yields information about which parameters contribute most to a particular canonical variate and the direction of their effect. *Intraset structure correlations* examine the correlation between the parameters and the canonical variates of the same domain, while *interset structure correlations* analyse the correlation across the domains.

Intraset Structure Correlations. These structure correlations express the contribution of the parameters of a set to the canonical variates of that set. The results of the intraset structure correlation are shown in Tables F.9 – F.22 in Section F.2.2 in Appendix F. There were two sets of intraset correlations corresponding to the two parameter sets. Shown in the tables are the square values of the correlation coefficients, which give the proportion of variance of a parameter, which was explained by a canonical variate of the same set. Columns add up to 1.00 (or 100% if taken as percentage values). It can be noticed that while for some phonemes a parameter was largely explained by a single canonical variate, and hence strongly correlated to it, for other phonemes, parameters were explained by more than one canonical variate. As an example, the proportions of variance explained by the five audio canonical variates for parameter F_0 in vowels /ɪ/ and /ʊ/ can be compared (Table F.9). For /ɪ/, CV 4 explained most of the variance. For /ʊ/, CVs 4 and 5 explained about equal amounts of variance.

Similarly, a canonical variate may explain the largest parts of variance of more than one parameter, while other canonical variates do not explain large proportions of variance of any parameter. As an example, CV 1 and 2 for the audio parameters and the phoneme /v/ can be compared in Table F.13.

Furthermore, it was observed that the canonical variate of the strongest canonical correlation, CV 1, was related most strongly to different parameters for different phonemes. Similar differences existed for the other canonical variates. If these patterns can be shown to be stable, this would offer a way to distinguish phonemes by computing the interset structure correlations. To do so, a larger sample size than the current 20 speakers in the AVOZES data corpus is required. As has already been pointed out, CANCOR can lead to unstable results for small sample sizes

compared to the size of the parameter sets.

Interset Structure Correlations. Of more interest in the analysis of relationships between audio and video speech parameters than the intraset correlations were the interset structure correlations. They characterise the relationships between the parameters of one set and the canonical variates of the other set. The results of the interset correlation can be found in Tables F.23 – F.36 in Section F.2.3 in Appendix F. There were two sets of interset correlations: audio speech parameters with video canonical variates and video speech parameters with audio canonical variates.[11] Rather than the correlation coefficients, the square of the coefficients is presented in the tables, because it specifies the proportion of variance of a parameter, which was predictable by a canonical variate of the second parameter set.

First of all, it was observed that the proportions of variance of a parameter explained by a canonical variate of the other set were generally low, i.e. under 10% of the variance of that parameter. However, for all phonemes, at least one interset structure correlation with a higher proportion of variation explained was found and hence a higher correlation between the parameter and the canonical variate. The maximum for the audio parameter - video canonical variate correlation was reached for the oral stop /k/ between parameter RMS and CV 1 at 58%. Similarly, for the video parameter - audio canonical variate correlations, the maximum was reached for the long vowel /iː/ between parameter MW and CV 1 at 70%. It was noticed that the highest proportion of variance for any parameter was almost always found in the first two canonical variates for any given phoneme, which is not automatically the case in interset structure correlations. This further supports the hypothesis, that linear combinations of parameters are well-correlated across the two modalities and that the parameters, identified as contributing strongly to the canonical variates, are the ones which account most for the variance in the other domain.

An in-depth analysis of these strong correlations was performed using the following procedure. First, it was identified which parameter and canonical variate

[11] To be precise, the interset correlations between the selected PC of the speech parameters of one set and the canonical variates of the other parameter set were analysed.

showed a correlation with $r \geq 0.30$ in the interset structure correlations. Secondly, it was checked in the intraset structure correlations, which parameters, from the same set as the canonical variates, were most accounted for by those canonical variates. As a result, the parameters in the two sets, which were related through the linear combinations, were identified. Overall, the correlations corresponded to different parameter - canonical variate pairs for different phonemes. Correlations were found for all possible parameter - canonical variate pairs, but certain pairs occurred more often for some phonemes than for others. In the case of audio speech parameters and video canonical variates, the strong correlations most often corresponded to

1. F_1 - PUL,
2. F_2 - MW and F_3 - PUL,
3. F_0 - RTC, F_2 - RTC, RMS - MW, and RMS - RTC,
4. F_1 - MW, F_1 - MH, F_2 - MH, and F_2 - PUL.

In the case of the video parameters and audio canonical variates, the following pairs were found most often corresponding to the strong correlations

1. MW - RMS, MH - RMS, PUL - F_1, RTC - F_2, and RTC - RMS,
2. MW - F_2 and RTC - F_1,
3. MW - F_1, MH - F_0, and RTC - F_0,
4. MH - F_1, MH - F_3, PUL - F_3, and PUL - RMS.

In summary, the interset structure correlations based on a CANCOR of the PCs related to the slope of the parameter curves — usually the second PC (see Section 6.5) — showed that there was little explanatory power between the canonical variates of one set and the parameters of the other set for most phonemes. However, strong correlations existed between some parameters and canonical variates, and thus between parameters of the two sets, for some phonemes. For these phonemes, linear combinations of parameters from either parameter set can be used to predict

the data from the other modality. These results showed that the AV relationships were phoneme-specific. A larger sample size than the 20 speakers currently recorded in the AVOZES data corpus would be required (a) to give the CANCOR more statistical stability and (b) to include other PCs from the PCA such as the horizontal range or shift PCs, so that an improved analysis can be performed.

The next two subsections discuss further information, that can be derived from the CANCOR, further interpreting the results of the CANCOR, in particular for those phonemes, where structure correlations did not yield much information.

Variance Extracted by a Canonical Variate

The *variance extracted by a canonical variate* is useful for determining which canonical variate has a stronger influence than others for certain measurements (the parameter values measured for the phonemes in this study). It is defined as the proportion of the total variance of a measurement domain — the audio and the video modality in this study — which is accounted for by a canonical variate of that domain [Gittins 85]. It represents the amount of variance common to both the measurement domain and the particular canonical variate. The variance extracted is calculated as the sum of squared intraset structure correlations divided by the number of parameters. Thus, there were two sets of variances extracted: one for the audio set and one for the video set. The sum of variances extracted always equals 1.

The variances extracted by canonical variates of the same domain are shown in Tables F.37 and F.38 in Section F.2.4 in Appendix F. The proportions of total variance extracted by a canonical variate ranged from 0.06 to 0.44 for the audio domain and from 0.09 to 0.49 for the video domain (over all phonemes). For most phonemes, however, the difference between the highest and lowest variance extracted was smaller, with the majority of values lying in the vicinity of the average value. This showed that no particular canonical variate accounted for a larger proportion of the total variance for all phonemes than the others. Nevertheless, some canonical variates had a stronger influence for a particular phoneme than others.

Based on the results of the analysis of the data in the AVOZES data corpus, there appeared to be no systematic pattern in which of the canonical variates accounted for a larger proportion of the total variance in its domain for a particular phoneme or subset of phonemes. The internal structure, defining which canonical variate accounted for more of the total variance than another canonical variate, was clearly phoneme-specific. Further experiments with a larger sample size are necessary to check, if the structures found in this study can be generalised for all speakers of AuE. It is possible, that such dependencies are also speaker or group specific, in which case finer grained analyses, than the one performed here, are required. If all these fail, a statistical learning process based on an analysis of a large number of speakers may provide some solution.

Redundancy and Total Redundancy

The interset analog of the variance extracted is called *redundancy* (or *explained variance*). It is of particular interest in the analysis of AV relationship because redundancy is the proportion of the total variance of a measurement domain predictable from a canonical variate of the other domain [Gittins 85]. It is calculated as the mean of the sum of squared interset structure correlations. Again, there were two sets of results: one for the total variance of the audio set predictable from the video canonical variates and one for the total variance of the video set predictable from the audio canonical variates. As redundancy is defined across two domains, the sum of redundancies equals 1 only if one domain can be predicted completely by the measurements in the other domain.

The results can be found in Tables F.39 and F.40 in Section F.2.5 in Appendix F. For almost every phoneme, the first two canonical variates of either domain accounted for the largest amount of variance of the other domain that was predictable from the canonical variates. The individual proportions of variance for these canonical variates ranged from about 0.05 to 0.30. For the other two canonical variates, the variance explained across the domains was generally below 0.05. A measure of how well the two domains or parameter sets were predictable from one another, is

given by the total redundancy value referred to below.

Inspecting the first two canonical variates of each domain more closely, revealed which parameters they corresponded to. This was done again by analysing the intraset structure correlations. There were considerable differences between the phonemes. Such differences were expected because of the different behaviour of the parameters for different phonemes. Over all phonemes, the correspondence of parameters to canonical variates averaged out, with the exception of the F_3 parameter, which showed a lesser degree of correspondence to the first two canonical variates than the other audio parameters. The other audio parameters were on equal footing in their occurrence. The first two video canonical variates corresponded most often to the MW parameter and the PUL parameter, followed by the MH parameter and last the RTC parameter.

Due to the large differences between phonemes, it is not possible to make a generalised statement about which parameter combinations of one domain accounted for a large proportion of the variance in the other domain. For example, for the short vowel /ɪ/, the 17% of total audio domain variance, that were predictable from the first two video canonical variates, came from the RTC parameter and the MW parameter, and the 11% of total video domain variance predictable from the first two audio canonical variates stemmed from the F_1, F_2, and RMS parameters. In contrast, for the long vowel /iː/, the 42% of total audio domain variance in the first two video canonical variates came from the MW parameter and the PUL parameter, while the 31% of total video domain variance predictable from the first two audio canonical variates stemmed from the F_1, F_2, F_3, and RMS parameters. In summary, the correspondences were very phoneme-specific. The individual results can be found in the above mentioned tables. It remains to be tested in future work, if these correspondences hold for a larger sample of speakers of AuE.

Total redundancy is a measure of the variance in one parameter set accounted for by the parameters in the other set [Gittins 85]. It is defined as the sum of the redundancy values and, hence, there were again two sets of results. Where canonical correlation coefficients expressed the relationship between linear combinations of parameters in each set, total redundancy is a direct expression of the interrelat-

edness of the measurement domains themselves. In other words, total redundany is a measure of how much information in the audio speech parameters can be expressed by information in the video speech parameter set and vice versa. It is a global measure of parametric interaction between the two modalities.

The results are also shown in Tables F.39 and F.40. Overall, the two total redundancy values for each phoneme were generally of similar magnitude, with the exception of the alveolar voiced affricate /dʒ/. The total redundancy of the two parameter sets given the video canonical variates ranged from 0.14 to 0.47 for the vocalic phonemes, and from 0.07 to 0.39 for the consonantal phonemes. Similarly, the total redundancy given the audio canonical variates ranged from 0.11 to 0.39 for the vocalic phonemes, and from 0.10 to 0.31 for the consonantal phonemes.

In conclusion, based on the CANCOR on the PCs related to the slope of the parameter curves, it was found that on average about a fifth to a third of the variance in the audio speech parameters was predictable from the video speech parameters and vice versa. This amount is expected to be higher, if more PCs could be included in the analysis, which requires a larger sample size than is currently available in the AVOZES data corpus. Nevertheless, the analysis of variances extracted and explained by a canonical variate showed, that some of the variance in either parameter set can be predicted from the other parameter set, as was expected from the theory of speech production (cp. Section 2.1.4). However, it is also clear from vocal tract theory, that not all of the variance in the audio domain can be predicted from video speech parameters, because not all changes in vocal tract geometry result in visible changes on the lips or on the face in general. Similarly, not all of the variance in the video domain can be expressed by the audio speech parameters, because there are multiple articulations (at least for some phonemes) which produce perceptionally the same acoustic result.

Comparing these results with results on the perceptual benefits of AV speech over audio-only speech in a study on French phonemes by Benoît *et al.* (cp. Section 2.1.6 and [Benoît 96]), the results from both studies agreed well in that lip information alone is able to restore about one third of the information from the audio modality. Such similar results for two different languages and different stimuli could

point to a more general relationship between visible speech information carried by the lips and acoustic speech information.

Summary Canonical Correlation Analysis

The CANCOR as a statistical analysis to relate linear combinations of parameters has been explored. Earlier analyses of the data in the AVOZES corpus had shown, that the parameters of the audio modality and the video modality were not related in a 1–1 relationship, but that some form of relationship existed between the parameters and was also expected from vocal tract theory. The resulting canonical correlation coefficients showed, that there was a significant amount of correlation between linear combinations of parameters of each domain (roughly 60–85% correlation). It has been discussed that the interpretation of canonical weights in a CANCOR is not a straightforward task and that these weights suffer from statistical instability in small to medium sample sizes (20 speakers in this study) compared to the number of parameters (9 in the presented analysis) due to collinearity. As a result, only one kind of PC could be included in the CANCOR. It was decided to use the PC related to the slope of the parameter curves, because it was the PC corresponding to the largest amount of the variation in the parameter curves, after the influence of a vertical shift in the parameter curves (PC 1) had been removed as a means of speaker normalisation. Using the condition number of a matrix as a tool to test for collinearity, some amount of collinearity was found in the data, but, with the exception of a few phonemes, it appeared to be at an acceptable level.

Since the interpretation of canonical correlation coefficients and canonical weights is generally difficult, other ways of interpreting the results of the CANCOR were used, such as structure correlations, variance extracted by a canonical variate, redundancy, and total redundancy. All these analyses pointed to phoneme-specific relationships between the audio speech parameters and the video speech parameters (or the chosen PCs thereof). A general, fixed relationship between combinations of some of the parameters was not evident. Rather, the combinations of parameters, that showed a high correlation across the two modalities, varied from phoneme to

phoneme. It would be necessary to investigate a larger sample to answer the question whether the relationships found in the analysis are stable. There was also no evidence for the exclusion of any parameter from the two sets, because all parameters appeared in linear combinations with strong correlations across the modalities for one phoneme or another. On average, about a fifth to a third of the variance in either modality was predictable from the parameters of the other modality.

The CANCOR has been helpful in supporting the hypothesis that the relationship between modalities lies in the combination of parameters, not in single parameter relationships. The problems with statistical instability that a CANCOR can suffer from have also been discussed. Therefore, other statistical analyses that explore the relationship between sets of parameters were investigated. One such analysis is the coinertia analysis, whose results are discussed next.

6.6.3 Coinertia Analysis

The relatively new multivariate statistical analysis of COIA has been described in detail in Section 5.5.6. It had so far not been used in the area of speech processing. COIA is a multivariate method for coupling two (or more) sets of parameters. It gives insight into the relationship between the two sets by analysing linear combinations of the parameters in each set like CANCOR. COIA has the advantage of numerical stability and independence from the sample size (see Section 5.5.6).

One of the shortcomings of CANCOR was that only one of the three main modes of variation, described in Section 6.5, could be used as input due to the sample size and CANCOR's instability problems. In contrast, COIA allowed the inclusion of all the PCs related to the two main modes of variation of interest — the slope of the parameter curves and the horizontal shift or range — in the analysis. The main mode of variation related a vertical shift of the parameter curves was left out as before, because it appeared to be related personal characteristics, not the common characteristics in the parameter curves. As a result, the audio parameter[12] set

[12] The term 'parameter' refers here to the PCs related to the slope of the parameter curves and the horizontal range or shift. In turn, these PCs are linked to the original parameter sets (see Section 6.5).

contained 10 parameters and the video parameter set 8 parameters for the COIA.

Results of Coinertia Analysis

The results are shown in the tables G.1 – G.7 in Appendix G. The tables G.1 and G.2 present the coinertia scores, which are covariance (or coinertia) value, the correlation value, the ratio of projected variance from the separate analysis of each parameter set to the variance from the coinertia analysis for both audio and video parameter set, and the RV coefficient as a measure of overall 'relatedness' of the two domains based on the selected parameters. The first three of these values existed for every coinertia axis (or vector). However, only the values for the first coinertia axis are shown, which was the axis onto which the largest amount of overall variance was projected and which was therefore the most important one. The tables G.3 – G.7 show the parameter weights, which were the coefficients of the input parameters in the linear combinations formed in the analysis, for each phoneme.

Covariance. The covariance value is a global measure of the co-structure in the input parameter sets. If the value is high, the sets are related strongly and vary in a dependent fashion. If the value is low, they vary independently. The covariance values ranged from 3.19 to 8.37 with a mean of 5.73 for the vocalic phonemes, and from 3.29 to 10.38 with a mean of 5.99 for the consonantal phonemes. In other words, although the covariance values were slightly higher for the consonantal phonemes, there were no significant differences in the covariance values between vocalic and consonantal phonemes. However, the covariance values differed considerably between individual phonemes. They were smallest for the high, central-to-back vowels /ʊ uː/, the alveolar closure nasal /n/, and the lateral liquid /l/. It appeared that the covariance value was larger for phonemes, which typically exhibit a larger degree of visible speech articulation, although the bilabials /b m/ showed only average covariance values. As a rule of thumb, the higher both ratios of projected variance from separate analysis to variance from coinertia analysis were, which means the higher the amount of variance in a parameter set obtained by the coinertia axes was, the higher was the covariance value. This follows from equation 5.18.

Correlation. The correlation values ranged from 0.54 to 0.83 with a mean of 0.69 for the vocalic phonemes, and from 0.50 to 0.82 with a mean of 0.68 for the consonantal phonemes. They showed that the first (and dominant) coinertia vectors were strongly correlated across the domains. Differences in the strength of the correlation existed for individual phonemes. For example, the correlation value was high (≥ 0.75) for the short front vowel /ɛ/, the front-to-central long vowels /iː ɜː/, the long back vowel /ɔː/, the diphthong /aʊ/, the oral stops /p d k g/, and the fricatives /v θ ð ʃ/. All these phonemes have a comparitively strong visible speech articulation in common, be it lip rounding, lip spreading, lip closure, or teeth visibility. The correlation value was smaller (< 0.60) for the mid-low vowels /ʌ æ/, the bilabial voiced oral stop /b/, the alveolar voiceless fricative /tʃ/, the velar closure nasal /ŋ/, the liquid /l/, and the rhotic glide /r/. These phonemes showed less visible speech articulation, with the exception of the bilabial /b/, whose low correlation value cannot readily be explained. In summary, the correlation between the linear combinations of input parameters, that maximised the covariance between the input parameter sets, was stronger for phonemes with more visible speech articulation. This result resonates with the expectations.

Ratio of Projected Variance. For the vocalic phonemes, the amount of variance explained by the first coinertia vector ranged from 0.38 to 0.96 with a mean of 0.75 for the audio parameter set, and from 0.43 to 0.97 with a mean of 0.74 for the video parameter set. In other words, the first coinertia vector accounted for about 75% of the variance in either set. Similarly, for the consonantal phonemes, the amount of variance obtained by the first coinertia vector ranged from 0.32 to 0.97 with an average of 0.77 for the audio parameter set, and from 0.52 to 0.98 with a mean of 0.76 for the video parameter set. Again, the first coinertia vector accounted for about 75% of the variance in either parameter set.

RV Coefficient. The RV coefficient is an overall measure of how well the two domains are related given the parameters in the two sets. It is in that sense similar to the total redundancy measure in CANCOR. It takes all coinertia axes into account.

For the vocalic phonemes, the RV coefficients ranged from 0.14 to 0.50 with a mean of 0.27 (see Table G.1 for individual results). For the consonantal phonemes, the RV coefficients ranged from 0.12 to 0.44 with an average of 0.26 (see Table G.2 for individual results). These results were broadly equivalent to earlier results using the total redundancy measure in CANCOR, where about a fifth to a third of the variance in either domain was predictable from the other domain.

Parameter Weights. Analysing the (normalised) parameter weights was useful for identifying the input parameters, which contributed most to the linear combination. In particular, it was possible to analyse the influence of the two chosen PCs — forming the input parameters of the COIA — on the linear combinations and thereby the importance of the original parameters in the AV relationships. This novel technique for speech processing therefore allowed a much more detailed analysis of the relationships than CANCOR. The weights are the coefficients of the linear combinations of the parameters in each set, that maximise the covariance between the parameter sets under the constraint of orthogonality. For each phoneme, the coefficients for the two chosen PCs of each parameter are presented and marked with an 'S' for the PC related to the slope of the parameter curve and an 'H' for the PC related to the horizontal shift or range (Tables G.3 – G.7). Larger coefficients mean that these parameters contributed more to the linear combination than coefficients close to 0.

It was found that all parameters contributed strongly to the linear combination for one phoneme or another, but some parameters contributed strongly notably more times than others. Again, these results resonated with earlier results from CANCOR. The pattern of parameters contributing to linear combinations related across the domains was phoneme-specific (see Tables G.3 – G.7 for individual results). Table 6.9 summarises the results by showing for each input parameter a list of phonemes, where that parameter contributed strongly to the linear combinations. A relationship between the patterns of parameters contributing strongly to one of these linear combinations and articulatory positions was not readily apparent. For example, even for short and long vowels produced with the tongue in a

Parameter	PC	Vocalic Phonemes	Consonantal Phonemes
F_0	S	uː	t j
	H	–	–
F_1	S	ɪ ʊ ɛ ʌ æ ɜː ɑː əʊ	b t v ð s ŋ l r w j
	H	æ uː əʊ	d dʒ m j
F_2	S	ɪ ɒ iː ɜː ɔː aʊ	p ð s m n w
	H	ɔɪ aʊ	f ʃ tʃ dʒ ŋ
F_3	S	ɒ ɜː ɔɪ	θ r
	H	iː ɔɪ	ŋ
RMS	S	ʊ ɛ ɒ iː ɔː ɑː əː eɪ aɪ ɪə	p b t d k g f v θ ð s z tʃ dʒ n l w
	H	ɪ ʊ ɛ ʌ əː ɪə əʊ	d ð dʒ
MW	S	ʌ uː ɜː ɔː ɔɪ	p b θ ʃ
	H	–	–
MH	S	ʊ ʌ æ iː ɑː aʊ ɪə əʊ	b t d k g f v ð s z tʃ dʒ m n l w
	H	eɪ əʊ	v s l j
PUL	S	ʊ ɛ ɒ ʌ iː uː ɔː əː ɔɪ	p t d ð ŋ r j
	H	ʊ ɑː	g s z r w j
RTC	S	ɪ ʊ ɛ ʌ æ ɜː ɔː ɑː ɔɪ aɪ aʊ ɪə	b t k f v ð s dʒ w
	H	əʊ	θ

Table 6.9: Phonemes for each parameter, where the parameter weights were ≥ 0.40.

similar position (e.g. /ɪ/ and /iː/), differing coefficients were observed. Of course, the effects of coarticulation were stronger on a short vowel than on a long vowel in the /bVb/-context used in the AVOZES corpus. More experiments with a different sample are needed to test, if the individual results for each phoneme found here can be generalised for all speakers of AuE or certain groups thereof.

For the vocalic phonemes, the most often appearing strong parameters in the linear combinations of the first and most important coinertia vector were the slope PC of RMS, RTC, MH, PUL, and F_1, and the horizontal range PC of RMS (in that order). A similar picture was found for the consonantal phonemes. Here, the

slope PC of MH, RMS, F_1, and RTC were the parameters most often found to contribute strongly. Overall, these four parameter were also the ones contributing strongly most often to the linear combinations corresponding to the first coinertia vector. The mean absolute weight values in Table G.7 reveal that the weights associated with the slope PC of RTC, RMS, PUL, MH, F_1, and F_2 (in that order), and the horizontal range PC of RMS were strongest for the vocalic phonemes.

Similarly, for the consonantal phonemes, the weights associated with the slope PC of MH, RMS, F_1, RTC, PUL, and F_2 were found to be strongest. Both ways of analysing the contribution of parameters showed that the horizontal range or shift PC was, on a general level, not as important as the slope PC. On an individual phoneme level, however, it was of importance for some phonemes (for example, see the coefficients of /tʃ dʒ/ for F_1 and F_2 in Table G.6). It should also be pointed out that coarticulation could have had an effect on these findings, because the /bVb/- and /ɑːCɑː/-words, respectively, each cover the same phonemic space apart from the central phoneme. The measured parameter values at the start and end of the intervocalic or interconsonantal phonemes, as defined in Section 5.4.3, would show some mix of vocalic and consonantal information and this mix could have influenced the results to some extent. However, coarticulation is a naturally occurring part of spoken language and it is, therefore, important to study AV relationships in such environments. It is suggested for future work to repeat the analyses performed in this study for other vocalic-consonantal and consonantal-vocalic phoneme transitions and to compare the results to find out, it the results reported here are specific to the chosen contexts.

Summary Coinertia Analysis

In summary, COIA enabled a more detailed analysis of the AV relationships than CANCOR. Where results of the two analyses are methodologically comparable, the COIA results resonated largely with the results from CANCOR. COIA's advantage is its increased stability over CANCOR, as reported by others (see Dolédec and Chessel [Dolédec 94] and Dray *et al.* [Dray 03]), particularly for analyses where

the sample size and the number of parameters are similar. Linear combinations of the parameters in each set, that maximised the covariance between the parameter sets, were strongly related across the domains. The composition of these linear combinations was phoneme-specific with no generalisation based on, for example, similar articulatory positions being apparent. As has been discussed before for the CANCOR results, an analysis with a different and possibly extended sample would be required to test the stability of the results found here. The PC related to the slope of the parameter curves contributed more to these linear combinations than the horizontal range or shift PC for most phonemes, which was expected based on a comparison of the average proportion of variance explained by the PCs (Table 6.5). On average about 75% of the variance in each parameter set was obtained by the first coinertia vector, which was judged to be sufficiently high to concentrate on that vector. In the coinertia coordinate system, the first coinertia vectors from either set were correlated with an average of 66% across all phonemes. The results of the RV coefficients showed that about a fifth to a third of the variance in one domain was predictable from the other domain.

6.6.4 Summary Between-Set (Audio-Video) Analysis

Summarising Section 6.6, the results of various analyses for characterising the between-set or AV relationships of the parameters in the audio and video speech parameter sets have been presented and discussed. Pairwise correlations were performed first which confirmed expectations. No strong correlations were found for any AV parameter pair for any phonemes. Some weak correlations were noticed, mostly for the parameter pairs of F_0 and MW, and F_2 and MH. As a result, multivariate statistical methods that explore the relationship of linear combinations of parameters were investigated.

First, canonical correlation analysis was applied. The canonical correlation coefficients suggested that linear combinations of the input parameters — in this instance, the PC of each parameter related to the main mode of variation 'slope of the parameter curves' — from either modality explained about 75% of the variance

in linear combinations of the other modality. This supports the hypothesis that indeed combinations of parameters are related across the two domains, not individual parameters. Investigating which parameters were more important than others in the linear combinations has proven to be difficult, because CANCOR results can suffer from numerical instability and in the presence of collinearity, and hence varying canonical weights — the coefficients of the parameters in the linear combinations — when the number of input parameters approaches the sample size. As a consequence, only the PC related to the slope of the parameter curve was selected as input into the analysis, when preferably both the PC related to the slope of the parameter curves as well as the PC related to the horizontal range or shift would have formed the input. With the help of the condition number, it was established that the input data used did not show strong signs of collinearity, but some level of it was present for some phonemes. This means that the results of the CANCOR for these phonemes have to be treated with care.

Due to the difficulties in interpreting canonical weights, other ways of interpreting the results of CANCOR have been used. This has involved structure correlations, which are the correlations between the input data and the canonical variates resulting from the analysis. Both intraset and interset structure correlations showed that correlations between input parameters and canonical variates, and thereby between parameters across the domains, were phoneme-specific. Further experiments with a larger sample size are necessary to test, if the relationships found can be generalised for speakers of AuE. It is possible, that such dependencies are also speaker or group specific, in which case analyses targeting these groups are required. Alternatively, a statistical learning process based on an analysis of a large number of speakers may also provide a solution. The phoneme-specific behaviour was further supported by the results of the variance extracted by a canonical variate, the redundancy, and the total redundancy. The correspondence of canonical variates to parameters with a higher than average correlation value differed from phoneme to phoneme. All parameters played a strong role for one phoneme or another. The audio speech parameters F_0, F_1, F_2, and RMS were most often contributing strongly to the canonical variates associated with the audio parameter set. For the video

parameter set, the parameters MW, MH, and PUL were most often contributing strongly to the canonical variates. The total redundancy values showed that about a fifth to a third of the variance in one modality was predictable from the other modality using the selected parameters.

Secondly, a coinertia analysis was applied to the data, because of its numerical stability even for small sample sizes. Here, both the slope PC and the horizontal range PC of each parameter could be included in the analysis. The results confirmed the CANCOR findings. The composition of strongly related linear combinations across the modalities was phoneme-specific. The first coinertia vectors were shown to correlate well across the domains and obtained about 75% of the variance in the parameters. The RV coefficients supported the view, that about a fifth to a third of the variance in the either modality could be predicted from the other modality with the selected parameters. Through the coinertia weights it was found, that the PC related to the slope of the parameter curves played a more important role in the linear combinations than the PC related to the horizontal range or shift. The weights showed again, that the composition of the linear combinations was phoneme-specific and that the other parameters occurred with strong weights as well for some phonemes. The most important parameters were F_1, F_2, and RMS of the audio speech parameter set, and MH, RTC, and PUL of the video speech parameter set. Based on these findings, it has been hypothesised that these parameters were the ones, that were most related — through linear combinations — across the two modalities.

6.7 Curve Registration

It has already been mentioned that some of the parameter curves were not aligned well, despite the synchronisation through the bilabial closure. This means that salient curve features like maxima and minima could be misaligned by a few sample points. This may have caused problems in the statistical analyses, as sample points were compared that did not represent comparable features of the curves. As has been discussed in Section 5.5.7, two different kinds of variation can be distinguished:

end-point variation and shape variation, of which the latter was of more interest in the analysis of AV relationships.

In this study, a global registration method based on FDA was used (see Section 5.5.7 for details). Curve registration was performed on the smoothed and resampled parameter curves (see Sections 5.4.2 and 5.4.3). The RMS parameter curves were first registered, separately for each phoneme, with the pointwise mean curve — with outliers removed — serving as registration target. By choosing the mean curve over any speaker's curve as registration target, issues like which speaker to choose or what to do in the case of chosen parameter curves not being well-suited for the registration process, are avoided. A manual selection of a curve as registration target each time based on some criterion of goodness would have been possible in principle, but appeared to be less appropriate than a fixed choice like the mean curve, which also had the advantage of being done in an automated process. Each phoneme's RMS warping function was then also applied to the other parameters in the audio and video speech parameter sets. The results of this registration process are summarised and discussed in Section 6.7.1. Section 6.7.2 then presents the results of a PCA applied to the registered parameter curves in the temporal domain, similar to the statistical shape analysis described in Sections 5.5.3 and 6.5. The resulting main modes of variation are compared to the results of the shape analysis without curve registration.

6.7.1 Discussion Results of Curve Registration

The results are shown in Appendix I. They can be compared to the original parameter curves in Appendix C. Visual comparison of the unregistered and registered parameter curves showed that, overall, the registration process was successful. Examining the RMS parameter curves first, it can be seen that the curves were better aligned than before, i.e. the horizontal shifts were smaller after registration (the syllable containing /b/ is a prominent example). The parameter curves for the other parameters also appeared to be better aligned to each other, as extrema shared the same timing. Generally, curves with salient features such as maxima and min-

ima were registered better than flat curves, due to the nature of the registration algorithm. Without a dominant feature, curves were not registered well.

Despite the overall goodness of the registration process, some parameter curves for some phonemes were not registered well. For example, the F_1 curves for /f/ diverged largely, which also had an effect on the mean curve. Here, the registration process did not resolve the divergence in the unregistered curves. Technically, it would be possible to register some of these curves better by using their own mean curve as target curve for the registration process, rather than using the warping information from the registration process of the RMS parameter. However, that would lead to a loss of comparable timing information and was, therefore, not appropriate in this investigation.

It was noticed that some curves suffered from oscillation, due to spline curves being used as part of the FDA. Curves that oscillated by more than an order of magnitude from the mean curve were removed manually and replaced with their unregistered counterparts, so as not to decrease the sample size. The remaining oscillations were considered to only have a neglible influence on the results of the statistical shape analysis, which are described in the next subsection.

6.7.2 PCA on Registered Curves

Similar to the shape analysis in Section 6.5, a PCA in the temporal domain was performed on the registered parameter curves. The results were used to test, if the curve registration lead to an improved analysis. The amount of phase variation was expected to be reduced for well-registered curves. In particular, this means that the proportion of variance in the curve shape analysis related to the third main mode (see Section 6.5), describing the horizontal range or shift of the curves, was expected to be smaller than for unregistered curves. In contrast, the proportion of variance for the other two main modes of variation — related to a vertical shift (amplitude difference) and to the slope of the curves — was expected to remain about the same or even increase. For badly registered curves, including curves where oscillation occurred, and for curves that were already well-aligned before the

PC	F_0	F_1	F_2	F_3	RMS	MW	MH	PUL	RTC
1	0.88	0.53	0.68	0.73	0.67	0.84	0.90	0.64	0.87
2	0.08	0.21	0.16	0.11	0.14	0.12	0.07	0.19	0.09
3	0.03	0.12	0.08	0.07	0.07	0.03	0.02	0.09	0.03
1	0.87	0.50	0.58	0.65	0.54	0.86	0.85	0.59	0.86
2	0.07	0.23	0.21	0.16	0.21	0.09	0.08	0.19	0.09
3	0.04	0.13	0.10	0.09	0.10	0.03	0.03	0.10	0.03

Table 6.10: Average proportion of variance (rounded to 2 decimal places) explained by the top three PCs for each parameter. Top: Vocalic phonemes. Bottom: Consonantal phonemes.

registration process, only small changes were expected.

The numeric results for each syllable can be found in the tables in Appendix J, which show the cumulative proportion of variance explained by the first six PCs. These results can be compared to the results of the PCA on the unregistered curves (Appendix E). For the vast majority of syllables and parameters, the proportion of variance expressed by the first PC increased as a result of the registration process. The amount of variance explained by the second and third PCs remained at about the same level or was reduced slightly. These findings are also shown in the results in Table 6.10, which presents the average individual proportion of variance in the first three PCs. The largest change compared to the results for the unregistered curves (Table 6.5 in Section 6.5) was found in the parameters RMS, MH, and RTC. Here, the average proportion of variance in the first PC increased the most.

Tables 6.11 and 6.12 show the number of PCs required to express at least 90% of the variance. As was the case with the unregistered curves, results differed considerably for different phonemes. However, comparing the results with Tables 6.6 and 6.7, it can be seen that the amount of 90% was reached by fewer PCs than in the case of unregistered curves. Parameters F_0, MW, MH, and RTC required only one or two PCs for almost all phonemes. More PCs were typically needed for all other parameters to explain 90% of the variance.

Phoneme	ɪ	ʊ	ɛ	ɒ	ʌ	æ	iː	uː	ɜː	ɔː	ɑː	əː	eɪ	ɔɪ	aɪ	aʊ	ɪə	əʊ
F_0	2	2	2	2	1	2	2	1	2	1	1	1	2	2	2	2	1	3
F_1	3	3	5	5	4	4	5	4	4	5	4	4	4	5	4	4	4	4
F_2	3	2	2	4	4	3	3	2	2	4	3	3	3	4	4	4	3	3
F_3	4	4	2	4	3	3	4	3	2	3	2	3	4	4	3	3	3	3
RMS	2	2	3	4	5	3	3	3	3	4	5	5	4	5	5	4	4	5
MW	1	1	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	1
MH	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	1	2	2
PUL	3	3	3	1	2	3	4	2	4	2	4	3	4	4	4	5	3	3
RTC	2	1	2	1	1	1	2	1	2	1	2	2	2	2	2	2	1	2

Table 6.11: Number of principal components needed to explain $\geq 90\%$ of the temporal variance in the registered parameter curves: Vocalic phonemes.

It was again possible to explore the mode of variation, that each PC stood for, by computing the pointwise mean parameter curve as well as the pointwise standard deviation (based on the distribution of the individual parameter curves from the 20 speakers) for each phoneme-parameter pair. Then, for each phoneme-parameter pair and each PC, the mean curve and two curves with $\pm x$ standard deviations were drawn in a graph, where x was an appropriate scaling factor for the parameter.[13] The main modes of variation found for the registered curves were basically the same as for the unregistered curves. Even more often than before, the first PC was related to a vertical shift between the parameter curves. Due to the minimised phase variation, a larger amount of the remaining variation was now a result of such amplitude differences, which was the goal of the registration process. It confirmed the previous finding, that the first PC was not of much relevance in the description of similarities of curve shapes, as its vertical shift appeared to be more a personal

[13] Again, the number of graphs created by this kind of visualisation is too large to be included in this thesis. However, the interested reader can find the figures on the accompanying CD-ROM in the directory 'PCExplanationRegistered'.

Phoneme	p	b	t	d	k	g	f	v	θ	ð	s
F_0	3	1	3	1	2	1	2	1	2	2	2
F_1	4	4	4	3	5	4	4	5	4	5	4
F_2	4	4	4	3	4	3	5	4	4	4	5
F_3	2	2	4	3	5	3	4	3	3	3	4
RMS	3	5	5	4	5	4	4	5	5	5	4
MW	2	2	2	1	1	2	2	2	1	2	3
MH	1	2	2	2	2	1	2	2	1	2	2
PUL	3	3	4	4	5	5	1	3	4	2	4
RTC	2	1	2	1	1	1	2	2	2	1	2
Phoneme	z	ʃ	tʃ	dʒ	m	n	ŋ	l	r	w	j
F_0	1	4	3	2	1	1	1	1	1	1	2
F_1	4	5	4	4	3	4	4	4	4	3	4
F_2	2	3	5	4	3	2	4	2	3	3	3
F_3	3	4	5	4	3	2	4	2	4	3	3
RMS	5	6	6	5	3	3	4	3	4	4	3
MW	1	1	2	2	1	2	1	1	2	2	1
MH	2	2	2	2	2	3	2	2	2	3	3
PUL	5	3	4	4	4	2	1	3	4	2	4
RTC	2	2	1	2	2	1	1	1	2	2	1

Table 6.12: Number of principal components needed to explain $\geq 90\%$ of the temporal variance in the registered parameter curves: Consonantal phonemes.

characteristic of each speaker.[14]

In contrast, the second and third PC expressed variation in the curve shape. Again, these PCs were related to the slope of the curves and the horizontal range. The horizontal shift, which had been overlaid on top of the horizontal range, was minimised in the registration process and had therefore no noticeable influence

[14] This PC could, however, be of interest in a speaker identification / verification task.

anymore. The horizontal range was related to the timing of changes in the shape of a parameter curve. Both modes of variation occurred in the second or third PC, depending on the variation found in a particular phoneme-parameter pair.

These results confirmed the previous findings about the main modes of variation occurring in the parameter curves. It can be deduced that the results of the statistical analyses draw an accurate picture of the phoneme-specific relations between audio and video parameters in the investigated /bVb/- and /ɑːCɑː/-syllables. In summary, curve registration was an adequate method to minimise the influence of phase variation on the results. It should generally be applied, where it is necessary or desired to statistically compare curves with different timing characteristics and where phase variation can be regarded as measurement noise.

6.8 Chapter Summary

The information contained in the recordings of the /bVb/- and /ɑːCɑː/-syllables in the AVOZES data corpus has been comprehensively analysed in this chapter. The data space was first explored thoroughly by visual inspection and a linear discriminant analysis, before various statistical methods were applied to the data to determine the relationships between the various audio and video speech parameters. Pairwise correlation analyses between any two parameters was applied to both the audio and the video parameter set separately, as well as across the two modalities, to look for 1–1 relationships between parameters. The only consistent correlation found for all syllables was between the protrusion parameters for upper and lower lip. Due to the high redundancy between these two parameters, it was sufficient to only analyse one of the two protrusion parameters. The protrusion of the upper lip mid-point was chosen here. The PCA applied to the audio speech parameter set also showed the presence of redundancies, but no single pair of parameters was correlating strongly, so that all parameters were included in the further analyses.

Next, a shape analysis of the parameter curves was performed by applying PCA in the temporal domain to each parameter separately for each syllable. This way, the main modes of variation were identified. These were related to a vertical shirt

between the curves, to the slope of the curves, and to the horizontal range or shift. While the first main mode of variation appeared to be related to some personal characteristics of each speaker and thus represented differences between the speakers, the latter two described variations to the shape of the curve and were, therefore, more suitable for expressing similarities in the parameter curves of different speakers. The further analyses focussed on these two modes of variation, as the focus of this study was on analysing the common characteristics in the parameter curves from different speakers and to compare them for the various phonemes.

Canonical correlation analysis was performed to explore the relationships of linear combinations of parameters, given that no linear relationship between single audio and video speech parameters had been found. The CANCOR results showed a significant amount of correlation between linear combinations of parameters of each domain (roughly 60–85% correlation), which supported the hypothesis of combinations of audio and video speech parameters being related across modalities. Unfortunately, CANCOR results can suffer from statistical instability in the calculated canonical weights, when the sample size (20 speakers in this study) is relatively small compared to the number of parameters (9 if only one PC is used, 18 for two PCs) due to collinearity in the data. To manage the stability problem, only the PC related to the slope of the curves could be included in the analysis. Various ways of interpreting the CANCOR results were explored, which all supported the hypothesis that relationships between the audio and video speech parameters are phoneme-specific.

The problems of statistical instability for a small sample size were overcome by the application of a multivariate method called coinertia analysis. It also analyses the relationship between linear combinations of parameters in different sets, but maximises the covariance rather than the correlation. This ensures that the number of parameters relative to the sample size does not affect the accuracy and stability of the results. COIA's results resonated with the results from CANCOR. That is, linear combinations of the parameters in each set were related strongly across the domains. The composition of these linear combinations was phoneme-specific. These results were in agreement with the expectations, but did not reveal any

subsets of the phonemic space. Of the two modes of variation considered in the analysis, the one related to the slope of a curve contributed more to the linear combinations than the one related to the horizontal range or shift. The COIA and CANCOR results showed that on average about a fifth to a third of the variance in one domain was predictable from the other domain, for the audio and video speech parameters used in this study. A clear negative result of the analyses was that no subsets of the phonemic space were found, in which common AV relationships could be assumed for all speakers. It was hoped at the beginning of this study that such subsets could be identified. The statistically small number of speakers precluded the analysis of potentially similar subsets.

Finally, curve registration was examined as a tool for improving the accuracy of the analysis. Using functional data analysis, a global registration algorithm was applied to the curves to reduce the amount of phase variation between curves, so that the common shape characteristics became clearer. In order to ensure synchronisation between the parameters, registration was first performed on the RMS parameter and the warping functions found there served as reference for the other parameters in each parameter set. Registration was generally successful, although a risk of oscillating curves due to the use of spline curves in FDA was present. The results of a PCA performed on the registered curves confirmed the previous findings on the main modes of shape variation. It can, therefore, be deduced that the results of the various statistical analyses accurately described the phoneme-specific relationships of audio and video speech parameters.

Chapter 7

Conclusions

7.1 Summary

The aim of this study has been to investigate the relationship between audio and video speech parameters on the example of AuE to enhance the scientific understanding of the interplay of the auditory and the visual side of the speech signal. Previous studies had shown that adding visual speech information can improve the recognition process. However, the interaction of the two modalities remains an ongoing area of research.

The outcomes of this project include both new methodology from an engineering point of view, as well as an in-depth analysis of AV relationships in AuE. Before the statistical analyses could be performed, several prerequisites needed to be created, which in themselves formed an important part of this multi-disciplinary study. They may also serve as a resource for future investigations. First of all, a way to measure video speech parameters was required. Based on a stereo vision face tracking system, a non-intrusive, real-time lip tracking algorithm was developed, which does not require any artificial markers or made-up lips. By using stereo vision in a calibrated camera system, the 3D coordinates of object points could be recovered. Combined with the lack of artificial markers on the face or head, this technology facilitated speakers to act normally and move freely within the limits set by the camera system, thus simplifying the familiarisation process for speakers and facilitating natural

speech. The lip tracking algorithm is based on the analysis of colour information, as well as the use of *a priori* knowledge about the structure of the mouth area. It tracks the lip corners and lip midpoints in an iterative multi-stage process. To the best of my knowledge, this is the first lip tracking algorithm using stereo vision for AVSP.

Furthermore, an AV speech data corpus for AuE was needed. A comprehensive analysis of the corpora described in the literature revealed, that most AV speech corpora were developed with a particular application in mind. They were often found to have only a limited coverage of the phonemes of a language and, therefore, were not well-suited for the purpose of this study. The few well-designed and comprehensive corpora were not for AuE and often not publicly available, so that the design and implementation of a new AV speech corpus for AuE became an essential prerequisite for any subsequent analyses of the AV relationships.

Both for the design of the AV speech corpus for AuE, as well as to provide a resource for the design of future AV speech corpora, a modular framework was proposed. It consists of essential and non-essential modules. AV speech sequences, which cover the phonemes and visemes of a language, form the minimum common set for corpora following this framework, while other modules can contain project specific sequences. If a common essential module was used, AV speech corpora could be made comparable, so that the benchmarking of algorithms becomes easier than it is at the moment. The modular approach of the proposed framework ensures the extensibility of corpora. Following this framework, the Audio-Visual Australian English Speech (AVOZES) data corpus was designed and implemented to cover the phonemes and visemes of AuE. To the best of my knowledge, it is the first AV speech data corpus that used a stereo camera system for the recording of the video data. Such stereo vision recordings enable 3D measurements of facial feature points independent of the head pose, whereas the 2D measurements of conventional monocular camera systems are pose-dependent.

Of the various parts of the AVOZES corpus, the sequences covering the phonemes and visemes of AuE uttered by 20 native speakers were used for the statistical analyses of the relationships between audio and video speech parame-

ters. Voice source excitation frequency F_0, formant frequencies F_1, F_2, and F_3, as well as RMS energy were chosen as audio speech parameters, because of their more direct relation to the articulators and the vocal tract compared to other parameters like, for example, LP coefficients. Mouth width MW, mouth height MH, protrusion of upper lip midpoint PUL, protrusion of lower lip midpoint PLL, and the novel measure of teeth visibility 'relative teeth count' RTC formed the set of video speech parameters. Since these parameters were directly related to the visible articulators, a relationship between the chosen audio and video speech parameters was expected based on articulatory theory. Of course, the visible speech articulators form only one part of the articulatory system and, thus, not all the information contained in the audio speech parameters was going to be matched by information in the video speech parameters. In addition, different articulator positions can result in perceptually similar acoustic information, so that it was expected that not all the information in one modality would be related to the information in the other modality.

An extensive exploration of the data space was performed before the statistical analyses of the AV relationships. This included observations made by visual inspection of the parameter curves, an outlier analysis, and a linear discriminant analysis (LDA). Outliers — defined in this study as being sample values with a difference of more than three standard deviations from the mean value — were not occurring at a high rate. The average outlier rate was 0.8%, with the highest rate for any phoneme-parameter pair being 5%. Outliers occurred for different phoneme-parameter pairs and different speakers. Most outliers were eliminated by the cubic spline smoothing performed in the preprocessing step. Any remaining outliers were judged having a negligible influence. LDA was performed as another data exploration tool. It helped gaining a better understanding of the placement of the parameter values in the parameter space for the various phonemes. The results showed that some phonemes were discriminated well against others, while other phonemes had a high confusion count. A separation by stepwise elimination could provide a solution for separating those latter phonemes. The phoneme groups resulting from LDA were later considered in the multivariate analyses (MVA), when looking for subsets of

phonemes with similar AV relationships across all speakers. As expected, the formant frequencies F_1, F_2, and F_3 were the parameters best separating the vocalic phonemes. For the consonantal phonemes, the video speech parameter played a stronger role, in particular the MH parameter. Such specific analytical differences have not been reported in the literature for AuE (or, to the best of my knowledge, for any other language). Comparisons with AV speech data from other languages are needed to determine, if such detail is specific to AuE.

Given the different articulatory positions for the production of different phonemes, the hypothesis was that there exist phoneme-specific relationships between the parameters. In other words, different phonemes were expected to result in different statistical relationships of the parameters, with the expectation that similar articulatory positions would result in similar statistical relationships. To test this hypothesis for AuE, several statistical analyses, which in a broader sense were related to 'correlation', were performed on the data. The utterances took the form of /bVb/- and /ɑːCɑː/-words with the phoneme under investigation being in the central position. Any language in general, and AuE in this particular instance, offer far more combinations of phonemes than could be tested in such an analysis, so the reader is reminded that the results have been presented for these particular contexts. These vocalic and consonantal contexts were chosen, because they facilitated the visual segmentation of the data in the data extraction phase. Coarticulation was, of course, also present in the chosen contexts. It is an inherent phenomenon of speech production, which occurs when the articulators move from one phoneme's target position to the next phoneme's target position, before the first target position has been fully reached. Coarticulation was expected to be strong particularly for the /bVb/-context, but it was judged that the advantages of a simplified and more accurate visual segmentation of the video data outweighed the disadvantages due to strong coarticulation.

Various statistical analyses were performed on the time-synchronised data for each utterance. Of all the sample points of an utterance, the ones belonging to the central phoneme in the /bVb/- and /ɑːCɑː/-syllables were selected for the analyses, so that data for each phoneme was available (with some samples affected

by coarticulation). The simultaneous recording of both the audio and video streams on DV tape ensured time synchronisation.[1] The following analyses were performed on the extracted data of each sequence:

- correlation analyses between single parameters both within each modality as well as across modalities,
- principal component analyses (PCA) separately on all parameters in each modality as a check for redundancies,
- PCAs separately on each parameter as a statistical shape analysis technique,
- canonical correlation analyses for the exploration of relationships of linear combinations of parameters across modalities, and
- coinertia analyses, which also explore the relationships of linear combinations of parameters across the two modalities.

The results of this study are summarised and discussed in the next section.

7.2 Results and Discussion

The key results of this multi-disciplinary study are

- a stereo vision, real-time lip tracking algorithm that does not require any artificial markers on the face,
- a proposed framework for the design of AV speech data corpora,
- an AV speech data corpus for AuE with stereo-vision video data,
- an extensive analysis of the relationships between audio and video speech parameters in AuE, including the first time application of coinertia analysis in speech processing.

The outcomes, thus, include both new engineering methods and new experimental results in the area of AVSP.

[1] As pointed out in Section 5.4.1, some adjustment was necessary, due to a constant delay between the two signals as a result of the stereo camera system used.

7.2.1 Stereo Vision Lip Tracking

The lip tracking algorithm developed in the course of this study (Chapter 3) has shown that accurate real-time lip tracking can be achieved using normal PC hardware and software. It was also shown that the use of artificial markers or made-up lips is not required for accurate lip tracking. It is believed that such a non-intrusive system has clear advantages over intrusive systems (optical markers, LEDs, IREDs etc., cp. for example, Yehia *et al.* [Yehia 97, Yehia 98]), because the familiarisation process for the speakers is much easier and it is, therefore, more likely that the recordings, and the results based on them, reflect the conditions in normal spoken language. Stereo vision has the advantage that 3D coordinates of face points can be recovered, so that measurements of distances reflect the real distances and not only 2D image distances. In addition, the measurements are independent of the head pose towards the camera (within obvious limits), because they are in 3D.

7.2.2 A Framework for AV Speech Data Corpora

In the course of this project, it was found that many existing AV speech corpora do not allow for an easy comparison of the results, because the design process was strongly application-driven. As thus, an incomplete coverage of the phonemes and visemes of a language were not uncommon. A new framework for future AV speech data corpora has been proposed here (see Section 4.1), which aims at improving this situation. A modular approach allows for extensibility. It is proposed that corpora following this framework contain as a minimum a so-called module (a selection of sequences) covering the phonemic and visemic space of the particular language, thereby enabling the comparison of any results derived from the analysis of such a common module in different corpora. Other required modules are a sampling of the recording setup, both with and without speakers. Further modules can be added, which can contain specific sequences desired for a particular application.

7.2.3 The AVOZES Data Corpus for Australian English

The proposed design was followed in the creation of the AVOZES data corpus (see Section 4.2). It was necessary to create a new corpus, because no AV speech data corpus for AuE was available. AVOZES is the first AV speech data corpus that used stereo vision equipment to record the video data, thus allowing for more accurate measurements of video speech parameters. AVOZES contains utterances from 20 native speakers of AuE and 4 non-native speakers. In this study, the utterances from the native speakers were analysed. The sequences include a coverage of the phonemes and visemes of AuE, as well as some examples of continuous speech and application-driven sequences. It is planned to make the AVOZES data corpus available to the research community.

In the utterances covering the phonemic space of AuE, the same VCV- and CVC-contexts were used for all consonantal and vocalic phonemes, respectively. All results are, therefore, not only phoneme-specific, but actually 'phoneme-in-context'-specific. The natural occurrence of coarticulation has affected the results. It is, therefore, suggested for future work to investigate the AV relationships for other contexts. The modular setup of AVOZES allows for such an extension of the data corpus.

7.2.4 Analysis of AV Relationships

The experimental and analytical focus of this study was on the exploration of the relationships between audio and video speech parameters in AuE (see Chapters 5 and 6). No such extensive exploration had been done before for AuE. The individual statistical analyses have already been mentioned in the summary at the beginning of this chapter. Some results and their implications are discussed here.

Pairwise Correlation and Principal Component Analyses

PCAs and linear pairwise correlation analyses were applied to the whole set of parameters in each modality to check for redundancies, using the dimensionality reduction property of PCA. For the set of video speech parameters, a redundancy

in the original parameter set was found in the two lip protrusion parameters. Lips are typically moved simultaneously and in a similar fashion, so that the results did not surprise. The two lip protrusion parameters were strongly correlated ($r = 0.91$ – 0.99 for /bVb/-words and $r = 0.79$ – 0.95 for /ɑːCɑː/-words) for all phonemes. The video speech parameter set was therefore reduced to four parameters by eliminating the lower lip protrusion parameter from further analyses. Weaker correlations ($r \leq 0.55$) were found for some other parameter pairs for some, but not all phonemes. These weaker correlations occurred most often for the parameter pairs of $MH - RTC$ and $MW - MH$. The former parameter pair was found for consonantal phonemes involving considerable lip movement, while the latter was prominent for phonemes with a front or front-central place of articulation.

For the set of audio speech parameters, the PCA results also suggested a considerable amount of redundancy in the parameter set. However, the pairwise linear correlation analysis did not show any parameter pair to stand out as being very strongly correlated. This suggested that most of the redundancy lay in a combination of parameters, which meant that no audio speech parameter could be excluded from further analysis. Nevertheless, some correlations with $|r| \geq 0.5$ were found, in particular for F_1 - RMS, F_2 - RMS, and F_2 - F_3, showing that these parameters contained some redundant information. RMS is related to mouth openness and thus the vertical position of the jaw, and thereby also related to F_1 and F_2 (cp. Section 2.1.4). The relationships between formant frequencies have been studied by Badin *et al.* [Badin 90]) and similar results were found. The lack of strong correlations between the voice source excitation frequency F_0 and the formant frequencies F_1, F_2, and F_3 is in agreement with other studies (e.g. [Kosiel 73] on Polish vowels). No redundancy was therefore expected between these parameters.

Similar to the correlation analysis within each modality's parameter set, a pairwise linear correlation analysis was performed across the two modalities. Generally, the pairwise correlations were small in value. Some trends of weak to medium strong correlations ($|r| = 0.3$ – 0.5) were found for many phonemes for the parameter pairs $F_0 - MW$ and $RMS - MH$. Overall however, the results did not support a hypothesis of a direct linear relationship between any of the speech parameters

across the two sets. Instead, the hypothesis, that combinations of parameters were related across the modalities, was favoured. Such a hypthesis is consistent with expectations based on articulatory theory (see, for example, Fant [Fant 60]), where a direct mapping between any single acoustic parameter and any single parameter describing an articulatory position can be considered as unlikely. Rather, the relationship between these parameters, inherent in the speech production process, was expected to be found in combinations of parameters.

Statistical Shape Analysis

To gain a better understanding of the parameter behaviour during the production of the central phoneme in the /bVb/- and /ɑːCɑː/-syllables, a statistical shape analysis was performed by applying a PCA separately to each parameter (see Section 6.5). The resulting principal components described the main modes of variation and provided a compact representation of the individual parameter curves for the subsequent analyses.

The first three PCs explained about 85–98% of the variance, so that any higher PCs were considered to have a negligible impact on the curve shape. The first PC was overwhelmingly often related to a vertical shift of the parameter curves. Thus, the largest amount of variation (40–80%) in the curves was not related to the shape of the curves at all, but to a mere shift, which can be attributed to a speaker's personal characteristics. While the extent of the vertical shift might be of interest in a speaker verification task, it was of no interest in the analyses of multi-parameter statistical relationships. The PC related to the vertical shift was, therefore, omitted from these analyses. The second and third PC expressed the modes of variation in the slope of the curves as well as the horizontal range or shift of the curves, respectively. These two PCs described the common characteristics of the parameter curves, which was the object of interest. Therefore, they formed the input for the canonical correlation (CANCOR) and coinertia (COIA) analyses.

Excluding such a large amount in variation from the analyses raises, of course, the issue of how the experiments could be improved, so as to avoid such a step. One

way would be to normalise the data to zero mean, thereby, for example, reducing differences between male and female speakers in the formant frequencies or in the *MW* parameter. Another way would be to investigate dynamic parameters based on derivatives of the parameters used in this study, which would also reduce the effects of a vertical shift between parameter curves.

Multivariate Analyses between Parameter Sets

In the first statistical analysis for the exploration of relationships of linear combinations of parameters, CANCOR (see Section 6.6.2), methodological issues with its numerical stability in the presence of collinearity meant that only one PC could be taken into account in the analysis and it was decided to take the PC related to the slope of the curves because of its greater importance to the curve shape. Ideally, both previously identified curve shape PCs — curve slope and horizontal range / shift — would have formed the input for CANCOR.

On average, the first canonical correlation coefficient was about 0.75, which supported the hypothesis, that combinations of parameters were related across the modalities. Overall, no single parameter or parameters were found to contribute notably more (or less) to the canonical correlation than the other parameters. Various measures for the interpretation of the CANCOR results were analysed and they all pointed to phoneme-specific relationships between the audio and video speech parameters. That is, the combinations of parameters, which showed a high correlation, varied from phoneme to phoneme. A relationship between the patterns of parameter combinations and articulatory positions was not readily apparent for any obvious subset of phonemes, based on similar articulator positions or the results of the LDA (see Section 6.3). Further experiments with a different or larger sample size are necessary to check, if the parameter combinations found for each phoneme are stable and can be generalised for all speakers, or groups of speakers, of AuE, as well as for different phoneme contexts. On average, about 20–35% of the variance in either modality was predictable from the other modality.

To overcome the stability problems of the CANCOR analysis, COIA was applied

to the data and analysed (see Section 6.6.3). COIA is another multivariate statistical analysis describing the relationship between sets of parameters. It was first used for ecological studies and to the best of my knowledge, it has not been used in speech processing (or more generally signal processing) before. COIA's advantage is its numerical stability and its property that the number of parameters relative to the sample size does not affect the accuracy of the results. Thus, both the PC related to the slope of the parameter curves and the PC related to the horizontal range or shift could be used as input of the COIA.

The COIA results agreed with the previous CANCOR results. They confirmed that about 20–35% of the variance in either modality was predictable from the other modality. The coinertia vectors — the resulting linear combinations of input parameters — correlated well across the modalities, with the first coinertia vector having an average correlation coefficient of about 0.50–0.80 and explaining about 75% of the variance in the parameters. Again, the composition of the linear combinations was found to be phoneme-specific (see Table 6.9 for a summary), which matched the expectations. All parameters contributed strongly to the linear combination for one phoneme or another, although some parameters contributed strongly considerably more times than others. No patterns between the phonemes and the parameters, which contributed strongly to the linear combinations, were found for any obvious subset of the phoneme space. Analysing the coinertia weights, it was found that the PC related to the slope of the curves played a more important role in the linear combinations than the PC related to the horizontal range or shift. F_1, F_2, and RMS of the audio speech parameters and MH, RTC, and PUL of the video speech parameters were identified as the most important parameters. Viewed on an overall level, it was hypothesised that these were the parameters which were most related — through linear combinations — across the two modalities. Again, an analysis with a larger or different sample is needed to test, if the phoneme's individual relationships are stable for a larger number of speakers and different phoneme contexts, and if they thus can be generalised for all speakers, or certain groups of speakers, of AuE. If such a test fails, a statistical learning process based on an analysis of a large number of speakers may provide some solution.

Comparison with Other Studies

In summary, some aspects of AuE matched the expectations based on articulatory theory. Similar AV relationships, compared to those reported in the literature for other languages, were shown for AuE (cp. Section 2.4). For the investigated audio and video speech parameters, statistical relationships were found for combinations of parameters, rather than for single parameters. About a fifth to a third of the variance in either modality could be recovered from the other modality using these parameter combinations. This finding is of importance to practical applications, like AV ASR, as it gives an upper bound of how much information can be gained from the additional visual speech information. It agrees with a study by Benoît *et al.* [Benoît 96], which showed that about 30% of audio information lost in noisy audio conditions can be regained from the lips alone.

Studies by Yehia *et al.* [Yehia 97, Yehia 98] on two speakers of American English and Japanese, respectively, had found that about 70% of the variance in their acoustic and facial geometric parameters — different from the ones in this study — could be accounted for by the other modality (see Section 2.4). Three possible explanations can be given for this higher amount of variance. Firstly, the use of infrared LEDs for the measurement of the video speech parameters resulted in an artificially higher accuracy, which led to improved estimation results. While the accuracy of the LED-Optotrak system is higher than that of the purely vision-based, marker-free lip tracking system used in this study, in my opinion, the difference is not large enough to be the only or dominating force in the noted differences in the amount of variance accounted for. However, the larger number of measured facial feature points and the higher sampling frequency can well be expected to increase the amount of facial movement captured compared to the lip-based feature points in this study. Secondly, the parameters chosen in the study by Yehia *et al.* may have captured other information, that was not available in the lip-based parameters used here. In particular, Yehia *et al.* mention the strong influence of a feature point on the cheeks below the eyes. It seemed to capture valuable information about skin movement as part of the speech articulation. Obviously, such information was

not available in our study. Thirdly, AuE may differ from American English and Japanese in the strength of AV relationships. If this was indeed the case, it would correlate with the colloquial notion that (some) speakers of AuE move their lips to a lesser extent than speakers of other varieties of English or of other languages.

Similar results to those by Yehia *et al.* were reported by Jiang *et al.* [Jiang 02], who studied the correlation between facial movements, tongue movements, and speech acoustics in American English. Jiang *et al.* used audio speech parameters similar to those measured by Yehia *et al.*, but the video speech parameters were relative distances between optically tracked facial feature points, as in the present study, although feature points on the entire lower face were used. On average across the four speakers, about 69% of the information in the video speech parameters was accounted for by the audio speech parameters and 47% of the acoustic information could be recovered from the video speech parameters. Both figures are higher than the ones in the present study, which can be attributed to the higher number of video speech parameters, including non-lip feature points, the use of marker-based optical tracking, and differences in the audio speech parameters.

A study by Barker and Berthommier [Barker 99] on AV relationships in French reported that audio speech parameters accounted for about 75% of the variance in the video speech parameters. In contrast, video speech parameters accounted only for about 55% of the variance in the audio speech parameters. Barker and Berthommier attributed this to the video speech parameters being only based on the lips and the jaw, not on the entire lower face as by Yehia *et al.* It is, therefore, plausible that the relatively small video speech parameter set in the present study only accounted for about 20–35% of the variance in other modality.

The Nature of the AV Relationships

The AV relationships found in this study of AuE support the view that the linear combinations of parameters, which are strongly correlated across the two modalities, are phoneme-specific, i.e. their composition differs from phoneme to phoneme and possibly also for a given phoneme in different contexts ('phoneme-in-context'-

specific). Given the different articulatory positions in the production of speech sounds, such phoneme-specificity was expected. Attempts at identifying patterns between phonemes with a similar vocal tract configuration and the structure of the linear combinations were not successful for any obvious subset of phonemes.

Implications for Automatic Speech Recognition

If the phoneme-specific AV relationships identified in the analyses in this study can be shown to be consistent for a larger number of speakers, or certain groups of speakers of AuE (for example, groups based on the three AuE varieties), then these relationships can be utilised in an AV ASR system. Such a system would analyse the audio and video speech parameters and identify phoneme candidates by their AV relationships. AV ASR systems have the advantage over audio-only ASR systems that they can recover part of the speech information from the video modality, when there is noise on the measurements of the audio modality.

Adequacy of the Data Corpus

Despite the data from 20 speakers in the AVOZES data corpus being more than what is available in many other AV speech corpora, many statistical analyses require even more data, before they reach stability, which is needed to build a good model of the AV relationships. However, the AVOZES data corpus was sufficient to explore the trends and the general behaviour of audio and video speech parameters for AuE.

It would be particularly interesting to sample more speakers from the three AuE varieties, so that the statistical analyses can be applied to the varieties separately and the resulting AV relationships compared for the different varieties. On a second dimension, it is suggested to extend the AVOZES data corpus to include sequences covering other parts of the phonetic space using different phonemic contexts.

The Speaker Dimension

Differences between female and male speakers were not noticeable by visual inspection for the video speech parameters, but some differences were found in the audio

speech parameters F_0, F_1, F_2, and F_3. These were attributed to the (usually) longer vocal tracts of men. It can be hypothesised that such differences in the parameters would also affect the results of the analyses to some extent. However, it was judged that groups of 10 speakers each would be too small for the multivariate statistical analyses (CANCOR, COIA) used in this study, in particular having in mind that the groups of male and female speakers could be further subdivided into gender-based groups of speakers of the three varieties of AuE. More speakers and more samples through repetition of utterances and inclusion of different phonemic contexts would be required for an investigation of the AV relationships for such groups of speakers, which is suggested to be undertaken in future work. Such work would further the understanding, whether the phoneme-specific AV relationships are speaker-independent, or also contain an extrinsic speaker-specific component.

7.3 Future Work

In this final section, an overview of opportunities for future work based on the results of this study is given.

Using Better Prediction in Lip Tracking

The lip tracking algorithm could be further improved to reduce the number of tracking failures and, thus, to improve the accuracy. In the current algorithm, a feature point position was determined from the tracking result in the current frame and the previously found position in the last video frame. These two positions were combined using the confidence measure as a weight to scale the contribution from these two frames. If the confidence measure for the position estimation in the current frame was high, the position of a feature point was more based on this frame, and vice versa. This algorithm could be improved by using more of the previous frames to determine the feature point position in the current frame. Such an algorithm could use predictive filtering, e.g. Kalman filtering, to predict the position in the current frame based on the previous frames and to combine the prediction with the current measurement to achieve more stable feature point

tracking, which reduces the occurrence of outliers and thus their impact on the analyses. In the present study, the impact of outliers was reduced by applying a *post-hoc* spline smoothing algorithm with certain smoothness constraints, but avoiding outliers due to tracking failure altogether would be a better way.

Using a More Complex Lip Model

A second way to improve the lip tracking algorithm would be the use of a more complex lip model. The currently used model was based on only four lip feature points, namely the lip corners as well as the mid-point of upper and lower lip. More complex models (e.g. using active shape models or active appearance models) have been used by others and are described in the literature (see Section 'Deformable 2D Models' in 2.3.1). However, with the increased complexity of such models, the computational requirements also increase, so that real-time lip tracking is often not (yet) possible. One aim of this study was to show that real-time, natural lip tracking can be achieved and useful parameters for the AV analysis be extracted. With future advances in computational power, the deployment of more complex lip models will become feasible in real-time applications. Currently, a trade-off between accuracy and computational complexity is still required.

Tracking More Facial Feature Points

A way to increase the information contained in the video speech parameters would be to track more facial feature points. This study concentrated on lip feature points because they carry the majority of visible speech information (see Section 2.1.6). Benoît *et al.* [Benoît 96] showed that displaying the lips alone restored about one third of the missing information, when the audio signal was noise degraded, but that displaying the whole face improved the intelligibility even further. Hence, tracking facial feature points on other parts of the lower face can be expected to capture more information about the movements of the visible speech articulators (cp. Yehia *et al.* [Yehia 98], Jiang *et al.* [Jiang 02]). The problem with currently available methods is that they require the use of artificial markers (coloured dots, LEDs, etc.) because

natural salient features cannot be expected to occur on every part of the face. The use of artificial markers also raises the question of the applicability in real-world scenarios and the naturalness of the resulting speech. At a minimum, such markers require some amount of familiarisation for the speaker.

Using Other Audio and Video Speech Parameters

The analysis of other audio and video speech parameters than the ones chosen here is suggested. For the video speech parameters, it was decided to use explicit, geometric parameters because of their relatedness to speech articulators. Other explicit and implicit parameters should be investigated. Similarly, other audio speech parameters are possible as well (extracted versus measured parameters, e.g. LP coefficients, line spectrum pairs, etc. instead of formant frequencies). The speech parameters in this study were chosen for their relatedness to the vocal tract geometry and hence the articulators, so as to facilitate the interpretation of the results. This choice helped to identify and illustrate AV relationships in the less complex statistical analyses but was not required in the more complex analyses of relationships between linear combinations. Besides analysing other parameters, the investigation of dynamic speech parameters like the first and second derivatives derived from the current sets of parameters is suggested. Some studies comparing static and dynamic parameters suggested (see Section 2.1.6) that the dynamic behaviour of parameters is equally or even more important than the static parameters, so by examining the velocity and acceleration patterns, more insight can be gained.

Investigating Non-Linear AV Relationships

This present study and previous studies have focussed on linear relationships between audio and video speech parameters. While the analyses were able to explain some of the AV relationships and account for some of the common information in both modalities, non-linear relationships should also be investigated. With the advances in computational power, non-linear statistical analyses become feasible in practical terms. Barker and Berthommier [Barker 99] studied linear and non-linear

models for AV relationships in French. They demonstrated that non-linear models were able to represent the AV relationships better than linear models, although the linear models provided a good first approximation. It would be interesting to compare the results when such non-linear models are applied to AuE data, such as the AVOZES corpus. Would the AV relationships be similar for French and AuE? Are there common AV relationships across languages? Or is there a strong language-specific component in the AV relationships?

Analysing Other Sequences

Finally, the additional sequences in the AVOZES data corpus would offer further material to be analysed. In particular, the continuous speech sequences could be used to apply and test found AV relationships for the different phonemes in an environment closer to real-world scenarios than the /bVb/- and /ɑːCɑː/-words. These sequences also offer a way to test differences in coarticulation, because they contain the same phonemes as in the other sequences but in different phonemic contexts. It has also been mentioned before in this thesis, that the recording of new sequences is suggested, in which (part of) the phonetic space is investigated via other phonemic contexts. For example, one could investigate the bilabial stop /p/ in (all) different short and long vowel VCV-contexts, such as /iːCiː/, /uːCuː/, /ɔːCɔː/, etc. In that way, it would be possible to find out, what the AV relationships specific to /p/ are and how the relationships are affected by the coarticulation due to the context. Furthermore, for the study of AV relationships in AuE, more data from more speakers would be needed to increase the stability of the results in some analyses and to test if the results found here can be generalised to all speakers, or certain groups of speakers, of AuE. More data would also be necessary to perform detailed analyses like CANCOR and COIA for the different AuE varieties (broad, general, and cultivated AuE). The production of such comprehensive AV speech corpora is a time-consuming process but undoubtedly important for thorough investigations and of benefit to the AV speech community as a whole.

Appendix A

Digital Video Format

A short overview of the Digital Video (DV) format is given here for the interested reader. Detailed information can be found in the international standards document IEC 61834 [IEC 01].[1] The original DV format (or Digital Video Cassette (DVC)) standards document is the so-called "Blue Book" [Blue Book 94]. The DV format should not be confused with standards for DVD (Digital Video Disc or Digital Versatile Disc) or DVB (Digital Video Broadcasting), which are different.

DV is an international standard for a consumer digital video format created by a consortium of ten companies. The companies originally involved in creating the standard were Matsushita Electric Industrial Corp (Panasonic), Sony Corp, Victor Corporation of Japan (JVC), Philips Electronics N.V., Sanyo Electric Co. Ltd., Hitachi Ltd., Sharp Corporation, Thompson Multimedia, Mitsubishi Electric Corporation, and Toshiba Corporation. Since then others have joined; there are now over 60 companies in the DV consortium.

The sampled video is compressed using a Discrete Cosine Transform (DCT), the same sort of compression used in motion-JPEG and MPEG. However, DV's DCT allows for more local optimisation (of quantising tables) within the video frame than do JPEG compressors, thus allowing for higher quality at the nominal 5:1 compression factor than a JPEG frame would show. DV uses intraframe compression; each

[1] The information presented here is largely taken from the websites www.dvformat.com and www.adamwilt.com .

compressed frame depends entirely on itself, and not on any data from preceding or following frames. However, it also uses adaptive interfield compression. If the compressor detects little difference between the two interlaced fields — the odd and even fields — of a frame, it will compress them together, freeing up some of the 'bit budget' to allow for higher overall quality. In theory, this means that static areas of images will be more accurately represented than areas with a lot of motion. In practice, this can sometimes be observed as a slight degree of 'blockiness' in the immediate vicinity of moving objects.

There are different colour sampling models for digital video depending on the original input (NTSC, PAL, etc.). The colour sampling used in the work presented in this thesis was NTSC YUV 4:1:1. The first number refers to the sampling rate of the luminance (Y), the other two numbers refer to that of the colour difference signals (U and V) relative to the first one. In a 4:1:1 system, the colour difference signals are sampled every fourth luminance sample. Other common sampling structures are 4:2:2 (D-1, D-5, etc.) and 4:2:0 (PAL).

In NTSC DV format, the resolution is 720×480 pixels. DV sampling is mostly said to be at exactly 30Hz frame rate (or 60Hz field rate)[2] but it is actually at a frame rate of 29.97Hz. To keep in synchronisation with the NTSC TV frame rate of exactly 30 frames per second, the video sequence of one second per minute — usually the first — DV contains only 28 frames! It is important to take this into account when analysing DV data, for example by interpolating the 28 frames to 30 frames.

Audio sampling in the DV standard is PCM (Pulse Code Modulation) at 48kHz with 16bit (2 channels), at 32kHz with 12bit (4 channels), or at 44.1kHz with 16bit (2 channels, same as audio CD (CD-DA) sampling). PCM is a way to digitise analogue signals by sampling the signal at constant time intervals (e.g. [Gibson 93]). The amplitude at each sampling point is rounded off to the nearest of several specific, predetermined levels in a process called quantisation. The error between the exact sample value and the assigned level is called quantisation error. The number

[2] Note, NTSC is an interlaced video standard.

of levels is defined by the number of bits assigned to represent each sample value, e.g. 16bit gives $2^{16} = 65,536$ levels. The 48kHz and 32kHz sampling rates of the DV standard can be used in locked mode, the 44.1kHz sampling rate only in unlocked mode. In locked mode, the audio sample clock is precisely locked to the video sample clock such that there is exactly the same number of audio samples recorded per video frame (or multiples of one video frame). To ensure synchronisation between the audio and video signals, locked mode is generally preferable.

Finally, the DV format is very well suited for transfer via an IEEE-1394 'FireWire' link to other equipment. For example, a computer with an IEEE-1394 compliant I/O-card can be linked to a DV recorder, so that there is a fully digital transfer of the data stored on DV tape to the computer for processing, analysis etc., thereby eliminating potential quality losses due to digital-to-analogue and analogue-to-digital conversion.

Appendix B

Speaker Data

This appendix contains background information on the speakers in the AVOZES data corpus. The tables on the following pages summarise some background information on the speakers in the AVOZES data corpus. The speakers were asked to fill in a questionnaire at the time of recording and the answers are presented here. The kind of information collected is detailed below (Table B.1). All speakers are considered to be native speakers of Australian English.

Speaker	Identifier, f1–f10 for female speakers, m1–m10 for male speakers
AuE variety	Variety of AuE: broad, general, or cultivated
Age	At the time of recording (in years)
Height	In cm
Weight	In kg
Level of education	Secondary, Tertiary, etc
Time abroad	Significant time spent abroad by the speaker (where, how long for, and at what age)
Singing / Training	Does the speaker sing? Has the speaker received training?
Smoking	Is or was the speaker a smoker?
Medical conditions	Related to respiratory system or otherwise potentially affecting the speech production
Country of origin	Of the speaker's parents
Native language	Of the speaker's parents
Occupation	Of the speaker's parents

Table B.1: Description of column headers in the tables on the following pages.

Speaker	AuE Variety	Age	Height (in cm)	Weight (in kg)	Level of education	Time abroad (where, how long)	Singing / Training	Smoking	Medical conditions
f1	Cultivated	23	163	62	Tertiary	–	Yes / No	No	–
f2	General	47	168	58	Tertiary	Hungary 2yr age 0–2 Canada 4yr age 22–26	No / No	No	–
f3	General	23	169	72	Tertiary	–	No / No	No	–
f4	Broad	22	170	54	Tertiary	–	No / No	No	Fibromyalgia since age 19
f5	General	32	163	55	Tertiary	Malaysia 2yr age 21–23	Yes / Yes	No	–
f6	Cultivated	28	164	56	Tertiary	–	No / No	No	–
f7	General	38	172	63	Tertiary	USA 8yr age 29–37	No / No	Yes	–
f8	General	29	175	70	Tertiary	–	No / No	Yes	–
f9	General	24	164	50	Tertiary	–	No / No	No	–
f10	General	24	172	60	Tertiary	–	Yes / Yes	No	–

Table B.2: Speaker data for female speakers. For a description of the column headers, see the beginning of this section. The native language of all speakers is English.

Speaker	AuE Variety	Age	Height (in cm)	Weight (in kg)	Level of education	Time abroad (where, how long)	Singing / Training	Smoking	Medical conditions
m1	General	40	178	90	Tertiary	Europe 6mo age 32 Japan 3yr age 33–35	No / No	No	–
m2	Broad	56	175	92	Tertiary	USA 1yr age 27	No / No	No	–
m3	Broad	27	188	95	Tertiary	USA 6mo age 8 UK 6mo age 8	No / No	No	Perforated ear drum, age 12
m4	Broad	26	175	62	Tertiary	–	No / No	No	Could not breathe through nose until age 13
m5	Broad	33	174	87	Tertiary	–	No / No	No	Asthma
m6	General	26	198	91	Tertiary	Scotland 1yr age 1 England 6mo age 13 USA 5mo age 15	No / No	No	–
m7	General	28	178	57	Tertiary	–	No / No	No	Mild asthma
m8	Broad	27	188	95	Tertiary	Argentina 8mo age 25	No / No	No	–
m9	General	27	190	75	Tertiary	NZ 15yr age 0–14 Japan 5mo age 25	No / No	No	–
m10	General	28	175	70	Tertiary	Switzerland 3mo age 17	Yes / No	No	–

Table B.3: Speaker data for male speakers. For a description of the column headers, see the beginning of this section. The native language of all speakers is English.

Speaker	Mother's			Father's		
	Country of Origin	Native Language	Occupation Occupation	Country of Origin	Native Language	Occupation
f1	Australia	English	Secretary	Australia	English	Retired
f2	Hungary	Hungarian	Education research officer	Hungary	Hungarian	Industrial chemist
f3	Australia	English	Consultant in education	Australia	English	Teacher
f4	Australia	German	Primary teacher	Australia	English	Furniture maker
f5	Australia	English	Housewife	Australia	English	Electrician
f6	Australia	English	Historian	Australia	English	Air Force Pilot
f7	Australia	English	Retired	Australia	English	Retired
f8	Australia	English	Hairdresser	Australia	English	Financier
f9	Australia	English	Caterer	Egypt	Greek	Architect
f10	Australia	English	Psychologist	Australia	English	Geologist

Table B.4: Family background for female speakers. For a description of the column headers, see the beginning of this section. The native language of all speakers is English.

Speaker	Mother's			Father's		
	Country of Origin	Native Language	Occupation	Country of Origin	Native Language	Occupation
m1	China	Russian	Housewife	Russia	Russian	Teacher
m2	Australia	English	Housewife	Australia	English	Plumber
m3	Australia	English	Librarian	New Zealand	English	Academic
m4	Australia	English	Physiotherapist	Australia	English	Storeman
m5	Australia	English	Retired	Australia	English	Retired
m6	Australia	English	Administrator	Australia	English	Mathematician
m7	Australia	English	Clinical nurse consultant	Australia	English	Courier
m8	Australia	English	Librarian	Australia	English	Engineer
m9	New Zealand	English	Accountant	New Zealand	English	Painter
m10	Australia	English	Manager of accounts	Australia	English	Medical doctor

Table B.5: Family background for male speakers. For a description of the column headers, see the beginning of this section. The native language of all speakers is English.

Appendix C

Parameter Curves

This appendix shows the parameter curves after smoothing and resampling to 25 points each on the time line as an aid to the discussion in Section 6.1. Shown in black are the individual parameter curves of the 10 male speakers, while the green curves denote the individual parameter curves of the 10 female speakers. The red curve shows the mean (pointwise average) of all individual parameter curves. The way the plots are ordered is first by parameter, then by phoneme. The order of the parameters is audio speech parameters first (voice source excitation frequency F_0, formant frequencies F_1, F_2, and F_3, root mean squared energy RMS), then the video speech parameters (mouth width MW, mouth height MH, protrusion of upper lip PUL, protrusion of lower lip PLL, relative teeth count RTC). There are two pages of graphs per parameter. The first page exhibits the vocalic phonemes as well as the consonantal phonemes /p b t/, while the other consonants are shown on the second page. The scale on the axes is always the same for all plots belonging to one particular parameter in order to facilitate the visual comparison of the curves. Drastic outliers in the individual parameter curves were the result of tracking failures.

Due to the very large number of graphs, the pages of this appendix can be found on the accompanying CD-ROM in the file `appendixC.pdf`.

Appendix D

Results Redundancy Analysis

This appendix contains the tables (D.1 – D.8) with the results of the redundancy (or within-set correlation) analysis for all phonemes, as described in Sections 5.5.3 and 5.5.4, and discussed in Section 6.4. First, the results for the set of video speech parameter (mouth width MW, mouth height MH, protrusion of upper lip PUL, protrusion of lower lip PLL, and relative teeth count RTC) are presented in Section D.1. The PCA on the set of video speech parameters showed that the first four principal components accounted for at least 96% of the variance. There was clearly a strong correlation between the two lip protrusion parameters, so that it was sufficient to include only one of the two in the further analyses.

In Section D.2, the results for the set of audio speech parameters (voice source excitation frequency F_0, formant frequencies F_1, F_2, and F_3, root mean squared energy RMS) are given. The PCA suggested that four of the five parameters were also sufficient to explain about 94% on average of the variance but the case was not as straightforward as for the video parameter set in terms of which parameter that was. Medium to strong correlations were found but for varying pairs of parameters, so that a generalisation appears impossible. The results suggest that it was rather a case of more than one parameter being correlated with one or more than one other. In that case, it was only after the PCA, with the orthogonal principal components forming a new coordinate system, that four 'new' parameters were able to express an average of 94% of the variance. Hence, all five parameters were included in the

further analyses.

Due to the large number of tables, the pages of this appendix can be found on the accompanying CD-ROM in the file `appendixD.pdf`.

Appendix E

Results PCA

This appendix contains the tables (E.1 – E.40) with the results of the PCA in the temporal domain for all central phonemes in the /bVb/- and /ɑːCɑː/-words in the AVOZES data corpus, as described in Section 5.5.3 and discussed in Section 6.5. In the tables below, phonemes are identified by their IPA symbol as well as the prompts used in the carrier phrases spoken by the speakers (compare Tables 4.4 and 4.5). The tables show the cumulative proportion of the variance explained by the first six principal components. The top half of the tables show the audio speech parameters (voice source excitation frequency F_0, formant frequencies F_1, F_2, and F_3, root mean squared energy RMS), while the bottom half displays the video speech parameters (mouth width MW, mouth height MH, protrusion of upper lip PUL, protrusion of lower lip PLL, relative teeth count RTC). First, the results for the vocalic phonemes are presented in Section E.1, then for the consonantal phonemes in Section E.2.

After the numeric results in the tables, a visualisation of the proportion of variance expressed by the first and second principal components of each parameter for each phoneme is given in the form of 'star charts' in the Figures E.1 – E.4. The first figure contains the star charts for the vocalic phonemes plus those of the consonantal phonemes /p b/. The second figure displays the star charts of the remaining consonantal phonemes. Phonemes are again identified by both their corresponding IPA symbol and the prompts used in the AVOZES data corpus.

Due to the large number of tables and figures, the pages of this appendix can be found on the accompanying CD-ROM in the file `appendixE.pdf`.

Appendix F

Results Correlation Analysis

In this appendix, the numeric results of the correlation analysis are presented. First, the results of the between-set correlation analysis, described in Section 5.5.4 and discussed in Section 6.6.1, are shown in Section F.1. A linear correlation analysis similar to the within-set analysis presented in Appendix D was performed. The correlation values of all pairs of audio and video speech parameters for each phoneme are given in the Tables F.1 – F.5. Tables F.1 and F2 show the results for the vocalic phonemes in the /bVb/-words, followed by the results for the consonantal phonemes in the /ɑːCɑː/-words in Tables F.3 – F.5.

Secondly, Section F.2 contains the numeric results of the canonical correlation analysis of the audio and video speech parameter sets, as described in Section 5.5.5 and discussed in Section 6.6.2. Section F.2.1 presents for each central phoneme in the /bVb/- and /ɑːCɑː/-words and each parameter the computed canonical weights, the canonical correlation value r_1 of the highest canonical correlation, and the reciprocal condition number $1/\kappa$. The intraset structure correlations are shown in Section F.2.2, followed by the interset structure correlations in Section F.2.3. The variance extracted by a canonical variate is summarised in the tables in Section F.2.4. Finally, Section F.2.5 contains the redundancy and total redundancy values.

The input parameters for the analyses were the five audio speech parameters (voice source excitation frequency F_0, formant frequencies F_1, F_2, and F_3, root mean squared energy RMS) and the four video speech parameters (mouth width MW,

mouth height MH, protrusion of upper lip PUL, relative teeth count RTC).

Due to the large number of tables, the pages of this appendix can be found on the accompanying CD-ROM in the file `appendixF.pdf`.

Appendix G

Results Coinertia Analysis

In this appendix, the results of the coinertia analysis (COIA) are presented (see Section 6.6.3 for a discussion). COIA is a multivariate statistical method similar to methods such as canonical correspondence analysis or canonical correlation analysis. However, instead of maximising the correlation, it maximises the covariance between two parameter sets (see Section 5.5.6 for details). Unlike other methods, COIA results do not suffer from instability and in the presence of collinearity, even if the number of parameters approaches the sample size. Thus, it was possible to incorporate both the principal component related to the slope of the parameter curves as well as the principal component related to the horizontal range or shift in the COIA analysis.

First, Section G.1 presents the 'coinertia scores'. These are the covariance (or coinertia) value, the correlation value, the ratios of the projected variance from separate analysis of each parameter set to the variance from the coinertia analysis for both audio and video parameter set, and the RV coefficient as a measure of overall 'relatedness' of the two domains given the selected parameters. The first three of these four scores are measures that exist for every coinertia vectors. However, shown are only the values for the first coinertia vector, which explains the largest amount of variance and is therefore the most important one.

Secondly, Section G.2 presents the parameter weights resulting from COIA. These weights are the coefficients of the parameters in the linear combinations of

the sets that are related. Parameters with larger weights have a stronger influence than others.

Again, the audio speech parameters and the principal components related to them are denoted by F_0 for the voice source excitation frequency, F_1, F_2, and F_3 for the first three formant frequencies, and RMS for the root mean squared energy. The video speech parameters and the principal components related to them are denoted by MW for the mouth width, MH for the mouth height, PUL for the protrusion of the upper lip, and RTC for the relative teeth count.

Due to the large number of tables, the pages of this appendix can be found on the accompanying CD-ROM in the file `appendixG.pdf`.

Appendix H

Results Linear Discriminant Analysis

This appendix presents the results of the linear discriminant analysis (LDA), as described in Section 5.5.2 and discussed in Section 6.3. Discriminant analysis classifies objects into groups based on information from a set of parameters. LDA finds a linear combination of selected original parameters that exhibits the largest ratio of between-class variance to within-class variance.

The sample points used in the LDA are given in Tables H.1 – H.3. Section 6.3.2 describes the selection process. LDAs were performed separately for the (short and long) vowels (Tables H.4, H.6, H.8), the diphthongs (Table H.10), and the consonants (Table H.12). As is the case for all analyses mentioned in this study, these phonemes refer to the central phonemes in the /bVb/- and /ɑːCɑː/-words. The summary on the left-hand side of the tables details the parameters selected for the discriminant functions, the χ^2 value, the significance value p, and the overall accuracy of the reclassification of all phonemes in one group (vowels, diphthongs, consonants) after cross-validation. The parameters are shown in the order that they were selected for the discriminant functions, i.e. the parameter listed first was the one that helped most to separate the phonemes and the other parameters are given in order of their decreasing contribution.

The accuracy of the discriminant functions was tested by reclassification after

cross-validation for each group of phonemes following the leave-one-out method [Lachenbruch 68]. The results are shown on the right-hand side of the above mentioned tables. Sensitivity describes the percentage of individuals for each phoneme that were correctly classified. Predictivity refers to the percentage of individuals classified as belonging to a phoneme that were really belonging to it. Table H.5 for all vowels, Table H.7 for the short vowels, Table H.9 for the long vowels, Table H.11 for the diphthongs, and Table H.13 for the consonants present the confusion matrices after reclassification by cross-validation. The rows refer to the actual phonemes, the columns to the classification results. Cross-validation was only performed for the individuals in the analysis, i.e. the individuals corresponding to outliers were not considered in the cross-validation, which led to some of the rows in the confusion matrices not adding up to a total of 20.

Due to the large number of tables, the pages of this appendix can be found on the accompanying CD-ROM in the file `appendixH.pdf`.

Appendix I

Registered Parameter Curves

The following pages show the parameter curves after curve registration, as described in Section 5.5.7 and discussed in Section 6.7. Shown in black are the individual parameter curves of the 10 male speakers, while the green curves denote the individual parameter curves of the 10 female speakers. The red curve shows the mean (pointwise average) of all individual parameter curves. The way the plots are ordered is the same as for the unregistered parameter curves in Appendix C, that is, first by parameter, then by phoneme. The order of the parameters is audio speech parameters first (voice source excitation frequency F_0, formant frequencies F_1, F_2, and F_3, root mean squared energy RMS), then the video speech parameters (mouth width MW, mouth height MH, protrusion of upper lip PUL, relative teeth count RTC). There are two pages of graphs per parameter. The first page exhibits the vocalic phonemes as well as the consonantal phonemes /p b t/, while the graphs for the remaining consonantal phonemes are shown on the second page.

The scale on the axes is always the same for all plots belonging to one particular parameter in order to facilitate the visual comparison of the curves. Drastic outliers in the individual parameter curves were the result of tracking failures in the measurement of the original data or due to oscillating curves as a result of the registration process.

Due to the very large number of graphs, the pages of this appendix can be found on the accompanying CD-ROM in the file `appendixI.pdf`.

Appendix J

Results of PCA on Registered Parameter Curves

This appendix contains the tables (J.1 – J.40) with the results of the PCA, as described in Section 5.5.7 and discussed in Section 6.7.2, applied to the registered parameter curves in the temporal domain for all /bVb/- and /ɑːCɑː/-words in the AVOZES data corpus. That is, a registration process (see Section 5.5.7 for details and Section 6.7 and Appendix I for results) was applied to the smoothed and resampled parameter curves before a PCA was performed. Parameters investigated were the voice source excitation frequency F_0, the formant frequencies F_1, F_2, and F_3, RMS energy, mouth width MW, mouth height MH, protrusion of upper lip PUL, and relative teeth count RTC. In the tables below, syllables are identified by the central phoneme (IPA symbol) in the /bVb/- and /ɑːCɑː/-words as well as by the prompts used in the carrier phrases spoken by the speakers (compare Tables 4.4 and 4.5). The tables show the cumulative proportion of the variance explained by the first six principal components. First, the results for the vocalic phonemes are presented, then for the consonantal phonemes.

Due to the large number of tables, the pages of this appendix can be found on the accompanying CD-ROM in the file `appendixJ.pdf`.

Bibliography

[Acero 99] A. Acero. *Formant Analysis and Synthesis Using Hidden Markov Models.* In Proceedings of the 6th European Conference on Speech Communication and Technology EUROSPEECH'99, Volume 3, pages 1047–1050, Budapest, Hungary, September 1999. European Speech Communication Association ESCA.

[Adjoudani 93] A. Adjoudani. *Élaboration d'un modèle de lèvres 3D pour animation en temps réel.* Mémoire de D.E.A. Signal Image Parole, INPG, Grenoble, France, 1993.

[Adjoudani 96] A. Adjoudani and C. Benoît. *On the Integration of Auditory and Visual Parameters in an HMM-based ASR.* In D.G. Stork and M.E. Hennecke, editors, Speechreading by Humans and Machines, Volume 150 of *NATO ASI Series*, pages 461–471, Berlin, Germany, 1996. Springer-Verlag.

[Ainsworth 76] W.A. Ainsworth. *Mechanisms of Speech Recognition*, Volume 85 of *International Series in Natural Philosophy.* Pergamon Press, Oxford, United Kingdom, 1976.

[Atal 68] B.S. Atal and M.R. Schroeder. *Predictive Coding of Speech Signals.* In Y. Kohasi, editor, Reports of the 6th International Conference on Acoustics, pages C–5–4, Tokyo, Japan, 1968.

[Badin 90] P. Badin, P. Perrier, L.-J. Boë, and C. Abry. *Vocalic nomograms: Acoustic and articulatory consideration upon formant convergences.* Journal of the Acoustical Society of America, 87(3):1290–1300, March 1990.

[Bailly 98] G. Bailly, P. Badin, and A. Vilain. *Synergy between Jaw and Lips/Tongue Movements: Consequences in Articulatory Modelling.* In R.H. Mannell and J. Robert-Ribes, editors, Proceedings of the 5th International Conference on Spoken Language Processing ICSLP'98, Volume 5, pages 1859–1862, Sydney, Australia, December 1998. Australian Speech Science and Technology Association (ASSTA).

[Barker 99] J.P. Barker and F. Berthommier. *Estimation of Speech Acoustics from Visual Speech Features: A Comparison of Linear and Non-Linear Models.* In

D.W. Massaro, editor, Proceedings of the International Conference on Auditory-Visual Speech Processing AVSP'99, pages 112–117, Santa Cruz (CA), USA, August 1999.

[Basu 98] S. Basu, N. Oliver, and A. Pentland. *3D lip shapes from video: A combined physical-statistical model.* Speech Communication, 26(1–2):131–148, October 1998.

[Bell 61] C.G. Bell, H. Fujisaki, J.M. Heinz, K.N. Stevens, and A.S. House. *Reduction of Speech Spectra by Analysis-by-Synthesis Techniques.* Journal of the Acoustical Society of America, 33(12):1725–1736, December 1961.

[Benoît 96] C. Benoît, T. Guiard-Marigny, B. Le Goff, and A. Adjoudani. *Which Components of the Face Do Humans and Machines Best Speechread?* In D.G. Stork and M.E. Hennecke, editors, Speechreading by Humans and Machines, Volume 150 of *NATO ASI Series*, pages 315–328, Berlin, Germany, 1996. Springer-Verlag.

[Bernard 81] J.R.L. Bernard. *Australian Pronunciation.* In The Macquarie Dictionary, pages 18–27, North Ryde, Australia, 1981. Macquarie Library.

[Bernstein 03] L.E. Bernstein, S. Takayanagi, and E.T. Auer, Jr. *Enhanced Auditory Detection with AV Speech: Perceptual Evidence for Speech and Non-Speech Mechanisms.* In J.L. Schwartz, F. Berthommier, M.A. Cathiard, and D. Sodoyer, editors, Proceedings of the ISCA Tutorial and Research Workshop on Audio Visual Speech Processing AVSP2003, pages 13–17, St Jorioz, France, September 2003. ICP, INP Grenoble, Université Stendhal.

[Besse 86] P. Besse and J.O. Ramsay. *Principal Component Analysis of Sampled Functions.* Psychometrika, 51(2):285–311, June 1986.

[Beyerlein 98] P. Beyerlein. *Discriminative Model Combination.* In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP1998, Volume 1, pages 481–484, Seattle (WA), USA, May 1998. IEEE.

[Blair 93] D. Blair. *Australian English and Australian national identity.* In G. Schulz, editor, The languages of Australia, pages 62–70, Canberra, Australia, 1993. Australian Academy of the Humanities.

[Blue Book 94] Blue Book. *Specifications of Consumer-Use Digital VCRs using 6.3mm magnetic tape.* HD Digital VCR Conference, December 1994.

[Boë 94] L.-J. Boë, J.-L. Schwartz, and N. Vallée. *The prediction of vowel systems: Perceptual contrast and stability.* In E. Keller, editor, Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges, pages 185–214, Chichester, United Kingdom, 1994. John Wiley & Sons.

[Bothe 96] H.-H. Bothe. *Relations of Audio and Visual Speech Signals in a Physical Feature Space: Implications for the Hearing Impaired.* In D.G. Stork and M.E. Hennecke, editors, Speechreading by Humans and Machines, Volume 150 of *NATO ASI Series*, pages 445–460, Berlin, Germany, 1996. Springer-Verlag.

[Bregler 94a] C. Bregler and Y. Konig. *"Eigenlips" for Robust Speech Recognition.* In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP'94, Volume II, pages 669–672, Adelaide, Australia, 1994.

[Bregler 94b] C. Bregler and S.M. Omohundro. *Surface learning with applications to lipreading.* In J.D. Cowan, G. Tesauro, and J. Alspector, editors, Advances in Neural Information Processing Systems, Volume 6, pages 43–50, San Francisco (CA), USA, 1994. Morgan Kaufmann.

[Brunelli 93] R. Brunelli and T. Poggio. *Face Recognition: Features versus Templates.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 15(10):1042–1052, October 1993.

[Burnham 96] D. Burnham and B. Dodd. *Auditory-Visual Speech Perception as a Direct Process: The McGurk Effect in Infants and Across Languages.* In D.G. Stork and M.E. Hennecke, editors, Speechreading by Humans and Machines, Volume 150 of *NATO ASI Series*, pages 103–114, Berlin, Germany, 1996. Springer-Verlag.

[Burnham 02] D. Burnham and K. Sekiyama. *Investigating Auditory-Visual Speech Perception Development using the Ontogenetic and Differential Language Methods.* In E. Vatikiotis-Bateson, P. Perrier, and G. Bailly, editors, Advances in Auditory-Visual Speech Processing, Cambridge (MA), USA, 2002. MIT Press. in press.

[Campbell 96] R. Campbell. *Seeing Brains Reading Speech: A Review and Speculations.* In D.G. Stork and M.E. Hennecke, editors, Speechreading by Humans and Machines, Volume 150 of *NATO ASI Series*, pages 115–133, Berlin, Germany, 1996. Springer-Verlag.

[Cathiard 96] M.-A. Cathiard, M.-T. Lallouache, and C. Abry. *Does Movement on the Lips Mean Movement in the Mind ?* In D.G. Stork and M.E. Hennecke, editors, Speechreading by Humans and Machines, Volume 150 of *NATO ASI Series*, pages 211–219, Berlin, Germany, 1996. Springer-Verlag.

[Chan 01] M.T. Chan. *HMM-Based Audio-Visual Speech Recognition Integrating Geometric- and Appearance-Based Visual Features.* In J.-L. Dugelay and K. Rose, editors, Proceedings of the 2001 IEEE Fourth Workshop on Multimedia Signal Processing, pages 9–14, Cannes, France, October 2001. IEEE.

[Chen 01] T. Chen. *Audiovisual Speech Processing.* IEEE Signal Processing Magazine, 18(1):9–21, January 2001.

[Chibelushi 96a] C.C. Chibelushi, F. Deravi, and J.S. Mason. *Survey of Audio Visual Speech Databases.* Technical report, Department of Electrical and Electronic Engineering, University of Wales, Swansea, UK, 1996.

[Chibelushi 96b] C.C. Chibelushi, S. Gandon, J.S. Mason, F. Deravi, and D. Johnston. *Design Issues for a Digital Integrated Audio-Visual Database.* In IEE Colloquium on Integrated Audio-Visual Processing for Recognition, Synthesis and Communication, pages 7/1–7/7, London, UK, Digest Reference Number 1996/213, November 1996.

[Christiansen 99] M.H. Christiansen and N. Chater. *Connectionist Natural Language Processing: The State of the Art.* Cognitive Science, 23(4):417–437, October-December 1999.

[Clark 89] J. Clark. *Regional Dialects in Australian English phonology.* In D. Blair and P. Collins, editors, Australian English: the language of a new society, pages 205–213, St Lucia, Australia, 1989. University of Queensland Press.

[Clark 95] J. Clark and C. Yallop. *An Introduction to Phonetics and Phonology.* Blackwell, Oxford, United Kingdom, Cambridge, USA, 2nd edition, 1995.

[Cochrane 89] G.R. Cochrane. *Origins and development of the Australian accent.* In D. Blair and P. Collins, editors, Australian English: the language of a new society, pages 176–186, St Lucia, Australia, 1989. University of Queensland Press.

[Cohen 95] J. Cohen, T. Kamm, and A.G. Andreou. *Vocal tract normalization in speech recognition: Compensating for systematic speaker variability.* In 129th Meeting of the Acoustical Society of America, *Journal of the Acoustical Society of America*, Volume 97, pages 3246–3247, Washington (DC), USA, May 1995.

[Colin 98] C. Colin, M. Radeau, and P. Deltenre. *Intermodal Interactions in Speech: A French Study.* In D. Burnham, J. Robert-Ribes, and E. Vatikiotis-Bateson, editors, Proceedings of the International Conference on Auditory-Visual Speech Processing AVSP'98, pages 55–60, Terrigal, Australia, December 1998.

[Colmenarez 96] A.J. Colmenarez and T.S. Huang. *Maximum Likelihood Face Detection.* In Proceedings of the Second International Conference on Automatic Face and Gesture Recognition FG'96, pages 307–311, Killington (VT), USA, October 1996. IEEE.

[Conrey 03] B.L. Conrey and D.B. Pisoni. *Audiovisual Asynchrony Detection for Speech and Nonspeech Signals.* In J.L. Schwartz, F. Berthommier, M.A. Cathiard, and

D. Sodoyer, editors, Proceedings of the ISCA Tutorial and Research Workshop on Audio Visual Speech Processing AVSP2003, pages 25–30, St Jorioz, France, September 2003. ICP, INP Grenoble, Université Stendhal.

[Cootes 95] T. Cootes, C. Taylor, D. Cooper, and J. Graham. *Active Shape Models - their training and applications.* Computer Vision and Image Understanding, 61(1):38–59, January 1995.

[Cootes 96] T. Cootes and C. Taylor. *Tracking and Learning Graphs and Pose on Image Sequences of Faces.* In Proceedings of the Second International Conference on Automatic Face and Gesture Recognition FG'96, pages 204–209, Killington (VT), USA, October 1996. IEEE.

[Cosi 96] P. Cosi and E.M. Caldognetto. *Lips and Jaw Movements for Vowels and Consonants: Spatio-Temporal Characteristics and Bimodal Recognition Applications.* In D.G. Stork and M.E. Hennecke, editors, Speechreading by Humans and Machines, Volume 150 of *NATO ASI Series*, pages 291–313, Berlin, Germany, 1996. Springer-Verlag.

[Crowe 87] A. Crowe and M.A. Jack. *Globally Optimising Formant Tracker Using Generalised Centroids.* Electronics Letters, 23(19):1019–1020, September 1987.

[Dalton 96] B. Dalton, R. Kaucic, and A. Blake. *Automatic Speechreading using Dynamic Contours.* In D.G. Stork and M.E. Hennecke, editors, Speechreading by Humans and Machines, Volume 150 of *NATO ASI Series*, pages 373–382, Berlin, Germany, 1996. Springer-Verlag.

[Davis 80] S.B. Davis and P. Mermelstein. *Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences.* IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4):357–366, August 1980.

[de Boor 01] C. de Boor. *A Practical Guide to Splines*, Volume 27 of *Applied Mathematical Sciences.* Springer-Verlag, Berlin, Germany, revised edition, 2001.

[Denes 93] P.B. Denes and E.N. Pinson. *The Speech Chain: The Physics and Biology of Spoken Language.* W.H. Freeman and Company, New York (NY), USA, 2nd edition, 1993.

[Dolédec 94] S. Dolédec and D. Chessel. *Co-inertia analysis: an alternative method for studying species-environment relationships.* Freshwater Biology, 31:277–294, 1994.

[Dolédec 97] S. Dolédec and D. Chessel. *Co-structure between two principal components analyses.* Topic documentation 4.1, Université Lyon 1, France, July 1997.

[Dray 03] S. Dray, D. Chessel, and J. Thioulouse. *Co-inertia analysis amd the linking of ecological tables*. Ecology, 2003. (in press).

[Dubnowski 76] J.J. Dubnowski, R.W. Schafer, and L.R. Rabiner. *Real-time Digital Hardware Pitch Detector*. IEEE Transactions on Acoustics, Speech, and Signal Processing, 24(1):2–8, February 1976.

[Duncan 86] G. Duncan and M.A. Jack. *Improved Algorithm Based on Pole Enhancement for Estimation of the Vocal Tract Frequency Response*. Electronics Letters, 22(23):1213–1214, November 1986.

[Dunn 50] H.K. Dunn. *The calculation of vowel resonances, and an electrical vocal tract*. Journal of the Acoustical Society of America, 22:151–166, 1950.

[Elman 97] J.L. Elman and J.L. McClelland. *Exploiting lawful variability in the speech wave*. In J.S. Perkell and D.H. Klatt, editors, Invariance and variability in speech processes, pages 360–380. Lawrence Erlbaum Associates, Hillsdale (NJ), USA, 1997.

[Entropic 93a] Entropic. *ESPS Programs A-L*. Entropic Research Laboratory, Inc., 1993. Version 5.0.

[Entropic 93b] Entropic. *ESPS Programs M-Z*. Entropic Research Laboratory, Inc., 1993. Version 5.0.

[Eubank 88] R.L. Eubank. *Spline Smoothing and Nonparametric Regression*, Volume 90 of *Statistics: Textbooks and Monographs*. Marcel Dekker, Inc., New York (NY), USA, 1988.

[Eveno 01] N. Eveno, A. Caplier, and P.-Y. Coulon. *A New Color Transformation for Lips Segmentation*. In J.-L. Dugelay and K. Rose, editors, Proceedings of the 2001 IEEE Fourth Workshop on Multimedia Signal Processing, pages 3–8, Cannes, France, October 2001. IEEE.

[Fant 60] G. Fant. *Acoustic Theory of Speech Production*. Mouton, The Hague, The Netherlands, 1960.

[Faugeras 86] O.D. Faugeras and G. Toscani. *The calibration problem for stereo*. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR'86, pages 15–20, Miami Beach, USA, June 1986. IEEE.

[Fisher 68] C.G. Fisher. *Confusions among visually perceived consonants*. Journal of Speech and Hearing Research, 11:796–804, 1968.

[Flanagan 72] J.L. Flanagan. *Speech Analysis, Synthessis and Perception*. Springer-Verlag, Berlin, Germany, 2nd edition, 1972.

[Foley 96] J.D. Foley, A. van Dam, S.K. Feiner, and J.F. Hughes. *Computer Graphics - Principles and Practice.* Addison-Wesley, Reading (MA), USA, 1996.

[Fry 76] D.B. Fry. *Acoustic Phonetics.* Cambridge University Press, Cambridge, United Kingdom, 1976.

[Fry 79] D.B. Fry. *The Physics of Speech.* Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge, United Kingdom, 1979.

[Furui 89] S. Furui. *Digital Speech Processing, Synthesis, and Recognition.* Electrical engineering and electronics. Marcel Dekker, New York (NY), USA, 1989.

[Furui 00] S. Furui. *Digital Speech Processing, Synthesis, and Recognition*, Volume 7 of *Signal Processing and Communications.* Marcel Dekker, New York (NY), USA, 2nd edition, 2000.

[Fuster-Duran 96] A. Fuster-Duran. *Perception of Conflicting Audio-Visual Speech: an Examination Across Spanish and German.* In D.G. Stork and M.E. Hennecke, editors, Speechreading by Humans and Machines, Volume 150 of *NATO ASI Series*, pages 135–143, Berlin, Germany, 1996. Springer-Verlag.

[Ganapathy 84] S. Ganapathy. *Decomposition of Transformation Matrices for Robot Vision.* In Proceedings of the IEEE International Conference on Robotics and Automation, pages 130–139, Atlanta (GA), USA, March 1984.

[Gay 81] T. Gay, B.E.F. Lindblom, and J. Lubker. *Productions of bite-block vowels: Acoustical equivalence by selective compensation.* Journal of the Acoustical Society of America, 69(3):802–810, March 1981.

[Gibson 93] J.D. Gibson. *Principles of digital and analog communications.* Macmillan, New York (NY), USA, 2nd edition, 1993.

[Girin 01] L. Girin, A. Allard, and J.-L. Schwartz. *Speech Signals Separation: A New Approach Exploiting the Coherence of Audio and Visual Speech.* In J.-L. Dugelay and K. Rose, editors, Proceedings of the 2001 IEEE Fourth Workshop on Multimedia Signal Processing, pages 631–636, Cannes, France, October 2001. IEEE.

[Gittins 85] R. Gittins. *Canonical Analysis.* Springer-Verlag, Berlin, Germany, 1985.

[Glotin 01] H. Glotin, D. Vergyri, C. Neti, G. Potamianos, and J. Luettin. *Weighting Schemes for Audio-Visual in Speech Recognition.* In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP2001, Salt Lake City (UT), USA, May 2001. IEEE. On CD-ROM.

[Goecke 00a] R. Goecke, J.B. Millar, A. Zelinsky, and J. Robert-Ribes. *Automatic Extraction of Lip Feature Points.* In Proceedings of the Australian Conference on Robotics and Automation ACRA2000, pages 31–36, Melbourne, Australia, August 2000.

[Goecke 00b] R. Goecke, Q.N. Tran, J.B. Millar, A. Zelinsky, and J. Robert-Ribes. *Validation of an Automatic Lip-Tracking Algorithm and Design of a Database for Audio-Video Speech Processing.* In Proceedings of the 8th Australian International Conference on Speech Science and Technology SST2000, pages 92–97, Canberra, Australia, December 2000. Australian Speech Science and Technology Association (ASSTA).

[Göhler 87] W. Göhler. *Höhere Mathematik.* Deutscher Verlag für Grundstoffindustrie, Leipzig, Germany, 10th edition, 1987.

[Goldschen 96] A.J. Goldschen, O.N. Garcia, and E.D. Petajan. *Rationale for Phoneme-Viseme Mapping and Feature Selection in Visual Speech Recognition.* In D.G. Stork and M.E. Hennecke, editors, Speechreading by Humans and Machines, Volume 150 of *NATO ASI Series*, pages 505–515, Berlin, Germany, 1996. Springer-Verlag.

[Golland 99] P. Golland, W.E.L. Grimson, and R. Kikinis. *Statistical Shape Analysis Using Fixed Topology Skeletons: Corpus Callosum Study.* In Proceedings of the 16th International Conference on Information Processing and Medical Imaging IPMI'99, LNCS 1613, pages 382–387, Visegrád, Hungary, June 1999. Springer-Verlag, Berlin, Germany.

[Golland 01] P. Golland. *Statistical Shape Analysis of Anatomical Structures.* PhD thesis, Massachusetts Institute of Technology, Cambridge (MA), USA, August 2001.

[Graf 96] H.P. Graf, E. Cosatto, D. Gibbon, M. Kocheisen, and E. Petajan. *Multi-Modal System for Locating Heads and Faces.* In Proceedings of the Second International Conference on Automatic Face and Gesture Recognition FG'96, pages 88–93, Killington (VT), USA, October 1996. IEEE.

[Grant 00] K.W. Grant and P.F. Seitz. *The use of visible speech cues for improving auditory detection of spoken sentences.* Journal of the Acoustical Society of America, 108(3):1197–1208, 2000.

[Grant 01a] K.W. Grant. *The effect of speechreading on masked detection thresholds for filtered speech.* Journal of the Acoustical Society of America, 109(5):2272–2275, 2001.

[Grant 01b] K.W. Grant and S. Greenberg. *Speech Intelligibility Derived from Asynchronous Processing of Auditory-Visual Information.* In Proceedings of the In-

ternational Conference on Auditory-Visual Speech Processing AVSP2001, pages 132–137, Aalborg, Denmark, September 2001.

[Grant 03] K.W. Grant, V. van Wassenhove, and David Poeppel. *Discrimination of Auditory-Visual Synchrony.* In J.L. Schwartz, F. Berthommier, M.A. Cathiard, and D. Sodoyer, editors, Proceedings of the ISCA Tutorial and Research Workshop on Audio Visual Speech Processing AVSP2003, pages 31–35, St Jorioz, France, September 2003. ICP, INP Grenoble, Université Stendhal.

[Green 94] K.P. Green. *The influence of an inverted face on the McGurk effect.* Journal of the Acoustical Society of America, 95:3014, 1994.

[Green 96] K.P. Green. *The Use of Auditory and Visual Information in Phonetic Perception.* In D.G. Stork and M.E. Hennecke, editors, Speechreading by Humans and Machines, Volume 150 of *NATO ASI Series*, pages 55–77, Berlin, Germany, 1996. Springer-Verlag.

[Guiard-Marigny 94] T. Guiard-Marigny, A. Adjoudani, and C. Benoît. *A 3-D model of the lips for visual speech synthesis.* In Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis, pages 49–52, New Paltz (NY), USA, September 1994.

[Guiard-Marigny 96] T. Guiard-Marigny, N. Tsingos, A. Adjoudani, C. Benoît, and M.-P. Gascuel. *3D Models of the Lips for Realistic Speech Animation.* In Proceedings of Computer Graphic 96, Geneva, Switzerland, 1996.

[Guiard-Marigny 97] T. Guiard-Marigny, A. Adjoudani, and C. Benoît. *3D models of the lips and jaw for visual speech synthesis.* In J.P.H. van Santen, R.W. Sproat, J.P. Olive, and J. Hirschberg, editors, Progress in Speech Synthesis. Springer-Verlag, Berlin, Germany, 1997.

[Harrington 97] J. Harrington, F. Cox, and Z. Evans. *An Acoustic Phonetic Study of Broad, General, and Cultivated Australian English Vowels.* Australian Journal of Linguistics, 17:155–184, 1997.

[Harrington 99] J. Harrington and S. Cassidy. *Techniques in Speech Acoustics.* Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999.

[Harshman 77] R. Harshman, P. Ladefoged, and L. Goldstein. *Factor analysis of tongue shapes.* Journal of the Acoustical Society of America, 62(3):693–707, September 1977.

[Heinz 61] J.M. Heinz and K.N. Stevens. *On the Properties of Voiceless Fricative Consonants.* Journal of the Acoustical Society of America, 33(5):589–596, May 1961.

[Heinzmann 97] J. Heinzmann and A. Zelinsky. *Robust Real-Time Face Tracking and Gesture Recognition.* In Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence IJCAI-97, Volume 2, pages 1525–1530, Nagoya, Japan, August 1997.

[Heinzmann 98] J. Heinzmann, Y. Matsumoto, J. Kieffer, and A. Zelinsky. *Smart Interfaces + Safe Mechanisms = Human Friendly Robots.* In Proceedings of International Workshop on Humanoid and Human-Friendly Robots, Tsukuba, Japan, October 1998.

[Heinzmann 99] J. Heinzmann and A. Zelinsky. *Building Human-Friendly Robot Systems.* In Proceedings of the International Symposium of Robotics Research ISRR'99, Salt Lake City (UT), USA, October 1999.

[Hennecke 96] M.E. Hennecke, D.G. Stork, and K.V. Prasad. *Visionary Speech: Looking Ahead to Practical Speechreading Systems.* In D.G. Stork and M.E. Hennecke, editors, Speechreading by Humans and Machines, Volume 150 of *NATO ASI Series*, pages 331–350, Berlin, Germany, 1996. Springer-Verlag.

[Heo 97] M. Heo and K.R. Gabriel. *A permutation test of association between configurations by means of the RV coefficient.* Communications in Statistics - Simulation and Computation, 27:843–856, 1997.

[Hermansky 90] H. Hermansky. *Perceptual linear predictive (PLP) analysis of speech.* Journal of the Acoustical Society of America, 87(4):1738–1752, April 1990.

[Hess 83] W. Hess. *Pitch Determination of Speech Signals*, Volume 3 of *Springer Series in Information Sciences.* Springer-Verlag, Berlin, Germany, 1983.

[Holden 00a] E.J. Holden, G. Loy, and R. Owens. *Accommodating for 3D Head Movement in Visual Lipreading.* In Proceedings of the IASTED International Conference on Signal and Image Processing SIP2000, pages 166–171, Las Vegas (NV), USA, November 2000.

[Holden 00b] E.J. Holden and R. Owens. *Visual Speech Recognition using Cepstral Images.* In Proceedings of the IASTED International Conference on Signal and Image Processing, pages 331–336, Las Vegas (NV), USA, 2000.

[Holden 02a] E. Holden and R. Owens. *Lip tracking using pattern matching snakes.* In Proceedings of the 5th Asian Conference on Computer Vision, Volume 1, pages 273–278, Melbourne, Australia, 2002.

[Holden 02b] E.J. Holden and R. Owens. *Automatic Facial Point Detection.* In Proceedings of the 5th Asian Conference on Computer Vision, Volume 2, pages 731–736, Melbourne, Australia, 2002.

[Hoole 87] P. Hoole. *Bite-Block Speech in the Absence of Oral Sensibility.* In Proceedings of the 11th International Congress on Phonetic Sciences, Volume 4, pages 16–19, Tallinn, Estonia, 1987.

[Horn 81] B.K.P. Horn and B.G. Schunck. *Determining Optical Flow.* Artificial Intelligence, 17:185–203, 1981.

[Höskuldsson 88] A. Höskuldsson. *Partial least square regression.* Journal of Chemometrics, 2:211–228, 1988.

[Hotelling 36] H. Hotelling. *Relations between two sets of variates.* Biometrika, 28:321–377, 1936.

[IEC 01] IEC. *61834.* International Electrotechnical Commission, consolidated edition, 2001.

[Ihaka 96] R. Ihaka and R. Gentleman. *R: A Language for Data Analysis and Graphics.* Journal of Computational and Graphical Statistics, 5(3):299–314, 1996.

[Ince 92] A. Nejat Ince, editor. *Digital Speech Processing: Speech Coding, Synthesis and Recognition.* Kluwer Academic Publishers, Boston (MA), USA, 1992.

[IPA 99] IPA. *Handbook of the International Phonetic Association.* Cambridge University Press, Cambridge, United Kingdom, 1999.

[Itakura 68] F. Itakura and S. Saito. *Analysis Synthesis Telephony Based on the Maximum Likelihood Method.* In Y. Kohasi, editor, Reports of the 6th International Conference on Acoustics, pages C–5–5, C17–20, Tokyo, Japan, 1968.

[Iyengar 01a] G. Iyengar and C. Neti. *Detection of Faces under Shadows and Lighting Variations.* In J.-L. Dugelay and K. Rose, editors, Proceedings of the 2001 IEEE Fourth Workshop on Multimedia Signal Processing, pages 15–20, Cannes, France, October 2001. IEEE.

[Iyengar 01b] G. Iyengar, G. Potamianos, C. Neti, T. Faruquie, and A. Verma. *Robust Detection of Visual ROI for Automatic Speechreading.* In J.-L. Dugelay and K. Rose, editors, Proceedings of the 2001 IEEE Fourth Workshop on Multimedia Signal Processing, pages 79–84, Cannes, France, October 2001. IEEE.

[Jiang 02] J. Jiang, A. Alwan, L.E. Bernstein, P. Keating, and E. Auer. *One the Correlation between Facial Movements, Tongue Movements and Speech Acoustics.* Special Issue on Joint Audio-Visual Speech Processing of EURASIP Journal on Applied Signal Processing, 2002(11):1174–1188, November 2002.

[Jones 17] D. Jones. *An English Pronouncing Dictionary.* J.M. Dent & Sons Limited, London, United Kingdom, 1917.

[Jones 03] J.A. Jones and K.G. Munhall. *Learning to produce speech with an altered vocal tract: The role of auditory feedback*. Journal of the Acoustical Society of America, 113(1):532–543, January 2003.

[Joos 48] M. Joos. *Acoustic Phonetics*. Supplement to Language: Journal of the Linguistic Society of America, 24(2):5–137, April-June 1948.

[Juang 00] B.-H. Juang and S. Furui. *Automatic Recognition and Understanding of Spoken Language - A First Step Toward Natural Human-Machine Communication*. Proceedings of the IEEE, 88(8):1142–1165, August 2000.

[Junqua 93] J.C. Junqua. *The Lombard reflex and its role on human listeners and automatic speech recognizers*. Journal of the Acoustical Society of America, 93(1):637–642, 1993.

[Kanade 73] T. Kanade. *Picture Processing System by Computer Complex and Recognition of Human Faces*. PhD thesis, Kyoto University, Japan, November 1973.

[Kass 88] M. Kass, A. Witkin, and D. Terzopoulos. *Snakes: Active Contour Models*. International Journal on Computer Vision, 1(4):321–331, 1988.

[Kim 01] J. Kim and C. Davis. *Visible Speech Cues and Auditory Detection of Spoken Sentences: an Effect of Degree of Correlation Between Acoustic and Visual Properties*. In Proceedings of the International Conference on Auditory-Visual Speech Processing AVSP2001, pages 127–131, Aalborg, Denmark, September 2001.

[Kim 03] J. Kim and C. Davis. *Testing the cuing hypothesis for the AV speech detection advantage*. In J.L. Schwartz, F. Berthommier, M.A. Cathiard, and D. Sodoyer, editors, Proceedings of the ISCA Tutorial and Research Workshop on Audio Visual Speech Processing AVSP2003, pages 9–12, St Jorioz, France, September 2003. ICP, INP Grenoble, Université Stendhal.

[King 00] S.A. King, R.E. Parent, and B.L. Olsafsky. *An anatomically-based 3D parametric lip model to support facial animation and synchronized speech*. In Proceedings of Deform 2000, pages 7–19, Geneva, Switzerland, November 2000.

[Kjeldsen 96] R. Kjeldsen and J. Kender. *Finding Skin in Color Images*. In Proceedings of the Second International Conference on Automatic Face and Gesture Recognition FG'96, pages 312–317, Killington (VT), USA, October 1996. IEEE.

[Klatt 79] D.H. Klatt. *Speech perception: A model of acoustic phonetic analysis and lexical access*. Journal of Phonetics, 7:279–312, 1979.

[Klatt 81] D.H. Klatt. *Lexical representation for speech production and perception*. In T. Myers, J. Laver, and J. Anderson, editors, The cognitive respresentation of speech, pages 11–31. North-Holland, Amsterdam, The Netherlands, 1981.

[Klatt 89] D.H. Klatt. *Review of selected models of speech perception.* In W.D. Marslen-Wilson, editor, Lexical representation and process, pages 169–226. MIT Press, Cambridge (MA), USA, 1989.

[Kohlrausch 00] A. Kohlrausch. *Perceptual consequences of asynchrony in audio-visual stimuli.* In IPO Annual Progress Report, Volume 35, pages 140–149. IPO, Center for User-System Interaction, Technische Universiteit Eindhoven, Einhoven, The Netherlands, 2000.

[Kopec 86] G.E. Kopec. *Formant Tracking Using Hidden Markov Models and Vector Quantization.* IEEE Transactions on Acoustics, Speech, and Signal Processing, 34(4):709–729, August 1986.

[Kosiel 73] U. Kosiel. *Correlations between Fundamental Frequency and Formant Frequencies in Polish Vowels.* Speech Analysis and Synthesis, 3:117–120, 1973.

[Kricos 96] P.B. Kricos. *Differences in Visual Intelligibility Across Talkers.* In D.G. Stork and M.E. Hennecke, editors, Speechreading by Humans and Machines, Volume 150 of *NATO ASI Series*, pages 43–53, Berlin, Germany, 1996. Springer-Verlag.

[Krumm 00] J. Krumm, S. Harris, B. Meyers, M. Hale B. Brumitt, and S. Shafer. *Multi-Camera Multi-Person Tracking for EasyLiving.* In S. Maybank and T. Tan, editors, Proceedings of the Third IEEE International Workshop on Visual Surveillance VS'2000, pages 3–10, Dublin, Ireland, July 2000. IEEE Computer Society.

[Kuhl 92] P.K. Kuhl, K.A. Williams, F. Lacerda, K.N. Stevens, and B. Lindholm. *Linguistic experience alters phonetic perception in infants by 6 months of age.* Science, 255:606–608, 1992.

[Kuhl 94] P.K. Kuhl, M. Tsuzaki, Y. Tohkura, and A.N. Meltzoff. *Human processing of auditory-visual information in speech perception: Potential for multimodal human-machine interfaces.* In Proceedings of the International Conference on Spoken Language Processing ICSLP'94, pages 539–542, Yokohama, Japan, 1994.

[Lachenbruch 68] P. Lachenbruch and R.M. Nickey. *Estimation of error rates in discriminant analysis.* Technometrics, 10:1–11, 1968.

[Ladefoged 78] P. Ladefoged, R. Harshman, L. Goldstein, and L. Rice. *Generating vocal tract shapes from formant frequencies.* Journal of the Acoustical Society of America, 64(4):1027–1035, October 1978.

[Ladefoged 79] P. Ladefoged and R. Harshman. *Formant frequencies and movements of the tongue.* In B. Lindblom and S. Ohman, editors, Frontiers of Speech Communication Research, pages 25–34, New York (NY), USA, 1979. Academic Press.

[Lee 98] L. Lee and R. Rose. *A Frequency Warping Approach to Speaker Normalization.* IEEE Transactions on Speech and Audio Processing, 6(1):49–60, January 1998.

[Lee 99] M. Lee, J. van Santen, B. Möbius, and J. Olive. *Formant Tracking Using Segmental Phonemic Information.* In Proceedings of the 6th European Conference on Speech Communication and Technology EUROSPEECH'99, Volume 6, pages 2789–2792, Budapest, Hungary, September 1999. European Speech Communication Association ESCA.

[Liberman 57] A.M. Liberman, K.S. Harris, H.S. Hoffmann, and B.C. Griffith. *The discrimination of speech sounds within and across phoneme boundaries.* Journal of Experimental Psychology, 54:358–368, 1957.

[Liberman 67] A.M. Liberman, F.S. Cooper, D.S. Shankweiler, and M. Studdert-Kennedy. *Perception of the Speech Code.* Psychological Review, 74:431–461, 1967.

[Liberman 76] A.M. Liberman, P.C. Delattre, F.S. Cooper, and L.J. Gerstman. *The Role of Consonant-Vowel Transitions in the Perception of the Stop and Nasal Consonants.* In D.B. Fry, editor, Acoustic Phonetics, chapter 21, pages 315–347. Cambridge University Press, Cambridge, United Kingdom, 1976. Appeared first in *Psychological Monographs*, 68(8).

[Liberman 85] A.M. Liberman and I.G Mattingly. *The motor theory of speech revised.* Cognition, 21:1–36, 1985.

[Lindblom 71] B.E.F. Lindblom and J.E.F. Sundberg. *Acoustical Consequences of Lip, Tongue, Jaw, and Larynx Movement.* Journal of the Acoustical Society of America, 50(4):1166–1179, 1971.

[Longuet-Higgins 81] H.C. Longuet-Higgins. *A computer algorithm for reconstructing a scene from two projections.* Nature, 293:133–135, September 1981.

[Loy 00a] G. Loy, R. Goecke, S. Rougeaux, and A. Zelinsky. *Stereo 3D Lip Tracking.* In Proceedings of the Sixth International Conference on Control, Automation, Robotics and Computer Vision ICARCV2000, Singapore, December 2000. On CD-ROM.

[Loy 00b] G. Loy, E. Holden, and R. Owens. *A 3D Head Tracker for an Automatic Lipreading System.* In Proceedings of the Australian Conference on Robotics and Automation ACRA2000, pages 37–42, Melbourne, Australia, August 2000.

[Luettin 96] J. Luettin, N.A. Thacker, and S.W. Beet. *Active Shape Models for Visual Speech Feature Extraction.* In D.G. Stork and M.E. Hennecke, editors, Speechreading by Humans and Machines, Volume 150 of *NATO ASI Series*, pages 383–390, Berlin, Germany, 1996. Springer-Verlag.

[Luettin 98] J. Luettin and S. Dupont. *Continuous Audio-Visual Speech Recognition*. In Proceedings of the Fifth European Conference on Computer Vision ECCV'98, Volume II of *Lecture Notes in Computer Science LNCS 1407*, pages 657–673, Freiburg, Germany, June 1998. Springer-Verlag, Berlin, Germany.

[Luettin 01] J. Luettin, G. Potamianos, and C. Neti. *Asynchronous Stream Modeling for Large Vocabulary Audio-Visual Speech Recognition*. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP2001, Salt Lake City (UT), USA, May 2001. IEEE. On CD-ROM.

[Luong 93] Q.-T. Luong, R. Deriche, O.D. Faugeras, and T. Papadopoulo. *On Determining the Fundamental matrix: Analysis of Different Methods and Experimental Results*. Technical Report 1894, Unité de Recherche INRIA-Sophia Antipolis, Institut National de Recherche en Informatique et en Automatique, Sophia-Antipolis, France, April 1993.

[Luong 96] Q.-T. Luong and O.D. Faugeras. *The Fundamental matrix: theory, algorithms, and stability analysis*. International Journal of Computer Vision, 17(1):43–76, 1996.

[MacDonald 78] J.W. MacDonald and H. McGurk. *Visual influences on speech perception processes*. Perception & Psychophysics, 24(3):253–257, 1978.

[Maeda 79] S. Maeda. *An articulatory model of the tongue based on a statistical analysis*. Journal of the Acoustical Society of America, 65:22, 1979.

[Maeda 82] S. Maeda. *A Digital Simulation Method of the Vocal-Tract System*. Speech Communication, 1(3–4):199–229, December 1982.

[Maeda 88] S. Maeda. *Improved articulatory models*. Journal of the Acoustical Society of America Supplement 1, 84:146, October 1988.

[Mardia 79] K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis*. Probability and mathematical statistics. Academic Press, London, UK, 1979.

[Markel 72a] J.D. Markel. *Digital Inverse Filtering - A New Tool for Formant Trajectory Estimation*. IEEE Transactions on Audio and Electroacoustics, 20(2):129–137, June 1972.

[Markel 72b] J.D. Markel. *The SIFT Algorithm for Fundamental Frequency Estimation*. IEEE Transactions on Audio and Electroacoustics, 20(5):367–377, December 1972.

[Markel 73] J.D. Markel. *Application of a Digital Inverse Filter for Automatic Formant and F_0 Analysis*. IEEE Transactions on Audio and Electroacoustics, 21(3):154–160, June 1973.

[Markel 76] J.D. Markel and A.H. Gray. *Linear Prediction of Speech*, Volume 12 of *Communications and Cybernetics*. Springer-Verlag, Berlin, Germany, 1976.

[Mase 91] K. Mase and A. Pentland. *Automatic Lipreading by Optical-Flow Analysis*. Systems and Computer in Japan, 22(6):67–76, 1991.

[Massaro 83] D.W. Massaro and M.M. Cohen. *Evaluation and Integration of Visual and Auditory Information in Speech Perception*. Journal of Experimental Psychology: Human Perception and Performance, 9(5):753–771, 1983.

[Massaro 87] D.W. Massaro. *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Lawrence Erlbaum Associates, Hillsdale (NJ), USA, 1987.

[Massaro 92] D.W. Massaro. *Broadening the Domain of the Fuzzy Logical Model of Perception*. In H.L. Pick, Jr., P. van den Broek, and D.C. Knill, editors, Cognition: Conceptual and Methodological Issues, pages 51–84, Washington (DC), USA, 1992. American Psychological Association.

[Massaro 96] D.W. Massaro. *Bimodal Speech Perception: A Progress Report*. In D.G. Stork and M.E. Hennecke, editors, Speechreading by Humans and Machines, Volume 150 of *NATO ASI Series*, pages 79–101, Berlin, Germany, 1996. Springer-Verlag.

[Massaro 98] D.W. Massaro and D.G. Stork. *Speech Recognition and Sensory Integration*. American Scientist, 86(3):236–244, 1998.

[Matsumoto 97] Y. Matsumoto, T. Shibata, K. Sakai, M. Inaba, and H. Inoue. *Real-Time Color Stereo Vision System for a Mobile Robot based on Field Multiplexing*. In Proceedings of the IEEE International Conference on Robotics and Automation ICRA'97, pages 1934–1939, Albuquerque (NM), USA, April 1997. IEEE.

[Matsumoto 99] Y. Matsumoto, J. Heinzmann, and A. Zelinsky. *The Essential Components of Human-Friendly Robot Systems*. In Proceedings of the International Conference on Field and Service Robotics FSR'99, pages 43–51, Pittsburgh (PA), USA, August 1999.

[Matsumoto 00] Y. Matsumoto and A. Zelinsky. *An Algorithm for Real-Time Stereo Vision Implementation of Head Pose and Gaze Direction Measurement*. In Proceedings of the Fourth IEEE International Conference on Face and Gesture Recognition FG'2000, pages 499–505, Grenoble, France, March 2000. IEEE.

[Matthews 98] I. Matthews, T.F. Cootes, S. Cox, R. Harvey, and J.A. Bangham. *Lipreading using Shape, Shading and Scale*. In D. Burnham, J. Robert-Ribes, and E. Vatikiotis-Bateson, editors, Proceedings of the International Conference on Auditory-Visual Speech Processing AVSP'98, pages 73–78, Terrigal, Australia, December 1998.

[Matthews 02] I. Matthews, T.F. Cootes, J.A. Bangham, S. Cox, and R. Harvey. *Extraction of Visual Features for Lipreading.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(2):198–213, February 2002.

[Maurer 96] T. Maurer and C. von der Malsburg. *Tracking and Learning Graphs and Pose on Image Sequences of Faces.* In Proceedings of the Second International Conference on Automatic Face and Gesture Recognition FG'96, pages 176–181, Killington (VT), USA, October 1996. IEEE.

[McCandless 74] S.S. McCandless. *An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra.* IEEE Transactions on Acoustics, Speech, and Signal Processing, 22(2):135–141, April 1974.

[McClelland 86] J.L. McClelland and J.L. Elman. *The TRACE model of speech perception.* Cognitive Psychology, 18:1–86, 1986.

[McGurk 76] H. McGurk and J. MacDonald. *Hearing lips and seeing voices.* Nature, 264:746–748, December 1976.

[McKenna 96] S. McKenna and S. Gong. *Tracking Faces.* In Proceedings of the Second International Conference on Automatic Face and Gesture Recognition FG'96, pages 271–276, Killington (VT), USA, October 1996. IEEE.

[Medler 76] D.A. Medler. *A Brief History of Connectionism.* Neural Computing Surveys, 1:61–101, 1976.

[Meier 96] U. Meier, W. Huerst, and P. Duchnowski. *Adaptive Bimodal Sensor Fusion For Automatic Speechreading.* In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP'96, pages 833–836, Atlanta (GA), USA, May 1996.

[Meier 00] U. Meier, R. Stiefelhagen, J. Yang, and A. Waibel. *Towards Unrestricted Lipreading.* International Journal of Pattern Recognition and Artificial Intelligence, 14(5):571–585, 2000.

[Messer 98] K. Messer, J. Matas, and J. Kittler. *Acquisition of a large database for biometric identity verification.* In Proceedings of BIOSIGNAL 98, pages 70–72, Brno, Czech Republic, June 1998.

[Messer 99] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. *XM2VTSDB: The Extended M2VTS Database.* In Proceedings of the Second International Conference on Audio and Video-based Biometric Person Authentication AVBPA'99, pages 72–77, Washington (DC), USA, March 1999.

[Millar 94] J.B. Millar, J.P. Vonwiller, J.M. Harrington, and P.J. Dermody. *The Australian National Database Of Spoken Language.* In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing ICASSP'94, Volume 1, pages 97–100, Adelaide, Australia, 1994.

[Millar 97] J.B. Millar, J.M. Harrington, and J.P. Vonwiller. *Spoken Language Resources for Australian Speech Technology.* Journal of Electrical and Electronic Engineers (Australia), 17(1):13–23, 1997.

[Millar 99] J.B. Millar and D.R.L. Davies. *A Reassessment of Temporal Information in Speech Processing.* In Proceedings of the Workshop on Innovative Speech Processing, Stratford-on-Avon, UK, August 1999.

[Mitchell 46] A.G. Mitchell. *The Pronunciation of English in Australia.* Angus and Robertson, Sydney, Australia, 1946.

[Mitchell 65] A.G. Mitchell and A. Delbridge. *The Pronunciation of English in Australia.* Angus and Robertson, Sydney, Australia, revised edition, 1965.

[Movellan 95] J.R. Movellan. *Visual speech recognition with stochastic networks.* In G. Tesauro, D. Touretzky, and T. Leen, editors, Advances in Neural Information Processing Systems, Volume 7, pages 851–858, Cambridge (MA), USA, 1995. MIT Press.

[Movellan 96] J.R. Movellan and G. Chadderdon. *Channel Separability in the Audio-Visual Integration of Speech: A Bayesian Approach.* In D.G. Stork and M.E. Hennecke, editors, Speechreading by Humans and Machines, Volume 150 of *NATO ASI Series*, pages 473–487, Berlin, Germany, 1996. Springer-Verlag.

[Murthy 91] H.A. Murthy and B. Yegnanarayana. *Formant extraction from group delay function.* Speech Communication, 10(3):209–221, August 1991.

[Nakadai 01] K. Nakadai, K. Hidai, H.G. Okuno, and H. Kitano. *Real-Time Multiple Speaker Tracking by Multi-Modal Integration for Mobile Robots.* In P. Dalsgaard, B. Lindberg, H. Benner, and Z.-H. Tan, editors, Proceedings of the 7th European Conference on Speech Communication and Technology, Aalborg, Denmark, September 2001. ISCA. On CD-ROM.

[Neti 00] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou. *Audio-Visual Speech Recognition.* Workshop report, CSLP / Johns Hopkins University, Baltimore, USA, 2000.

[Neti 01] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, and D. Vergyri. *Large-vocabulary audio-visual speech recognition: A summary of the Johns Hopkins Summer 2000 Workshop.* In Proceedings of the 2001 IEEE Fourth

Workshop on Multimedia Signal Processing MMSP-01, pages 619–624, Cannes, France, 2001.

[Newman 99a] R. Newman. *Head Pose and Gaze Point Estimation System Version 2.0.* Technical report, Robotic Systems Laboratory, Research School of Information Sciences and Engineering, Australian National University, Canberra, Australia, 1999.

[Newman 99b] R. Newman and A. Zelinsky. *Error Analysis of Head Pose and Gaze Direction from Stereo Vision.* In Proceedings of the Australian Conference on Robotics and Automation ACRA'99, pages 114–118, Brisbane, Australia, March 1999.

[Newman 00] R. Newman, Y. Matsumoto, S. Rougeaux, and A. Zelinsky. *Real-Time Stereo Tracking for Head Pose and Gaze Estimation.* In Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition FG'2000, pages 122–128, Grenoble, France, March 2000.

[Ney 03] H. Ney. *Maschinelle Sprachverarbeitung. Der statistische Ansatz in der Spracherkennung und Sprachübersetzung.* Informatik Spektrum, pages 94–102, May 2003.

[Nordstrand 03] M. Nordstrand, B. Granström G. Svanfeldt, and D. House. *Measurements of Articulatory Variation and Communicative Signals in Expressive Speech.* In J.L. Schwartz, F. Berthommier, M.A. Cathiard, and D. Sodoyer, editors, Proceedings of the ISCA Tutorial and Research Workshop on Auditory-Visual Speech Processing AVSP2003, pages 233–237, St Jorioz, France, 2003. ICP, INP Grenoble, Université Stendhal.

[Oden 78] G.C. Oden and D.W. Massaro. *Integration of Featural Information in Speech Perception.* Psychological Review, 85(3):172–191, 1978.

[Öhman 98] T. Öhman. *An audio-visual speech database and automatic measurements of visual speech.* Quarterly Status and Progress Report TMH-QPSR 1-2/1998, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden, 1998.

[Olive 71] J.P. Olive. *Automatic Formant Tracking by a Newton-Raphson Technique.* Journal of the Acoustical Society of America, 50(2):661–670, 1971.

[Padmanabhan 02] M. Padmanabhan and M. Picheny. *Large-Vocabulary Speech Recognition Algorithms.* Computer, 35(4):42–50, April 2002. IEEE Computer Society.

[Patterson 02] E.K. Patterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy. *CUAVE: A New Audio-Visual Database for Multimodal Human-Computer Interface Research.* In Proceedings of the IEEE International Conference on Acoustics, Speech,

and Signal Processing ICASSP2002, Volume 2, pages 2017–2020, Orlando (FL), USA, May 2002. IEEE.

[Petajan 84] E.D. Petajan. *Automatic Lipreading to Enhance Speech Recognition*. PhD thesis, University of Illinois at Urbana-Champaign, 1984.

[Petajan 96] E.D. Petajan and H.P.Graf. *Robust Face Feature Analysis for Automatic Speachreading (sic) and Character Animation*. In D.G. Stork and M.E. Hennecke, editors, Speechreading by Humans and Machines, Volume 150 of *NATO ASI Series*, pages 425–436, Berlin, Germany, 1996. Springer-Verlag.

[Plant 77] G.L. Plant and J.J. Macrae. *Visual Perception of Australian Consonants, Vowels and Diphthongs*. Australian Teacher of the Deaf, 18:46–50, July 1977.

[Plant 80] G.L. Plant. *Visual identification of Australian vowels and diphthongs*. Australian Journal of Audiology, 2(2):83–91, 1980.

[Potamianos 97] G. Potamianos, E. Cosatto, H.P. Graf, and D.B. Roe. *Speaker Independent Audio-Visual Database for Bimodal ASR*. In C. Benoît and R. Campbell, editors, Proceedings of the ESCA Workshop on Audio-Visual Speech Processing AVSP'97, pages 65–68, Rhodes, Greece, September 1997. ESCA.

[Potamianos 00] G. Potamianos, A. Verma, C. Neti, G. Iyengar, and S. Basu. *A Cascade Image Transform for Speaker Independent Automatic Speechreading*. In Proceedings of the 2000 IEEE International Conference on Multimedia and Expo, Volume 2, pages 1097–1100, New York (NY), USA, August 2000. IEEE.

[Potamianos 01] G. Potamianos, J. Luettin, and C. Neti. *Hierarchical Discriminant Features for Audio-Visual LVCSR*. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP2001, Salt Lake City (UT), USA, May 2001. IEEE. On CD-ROM.

[Pratt 78] W.K. Pratt. *Digital Image Processing*. John Wiley & Sons, New York (NY), USA, 1978.

[Rabiner 75] L.R. Rabiner and B. Gold. *Theory and Application of Digital Signal Processing*. Prentice-Hall Signal Processing Series. Prentice-Hall, Englewood Cliffs (NJ), USA, 1975.

[Rabiner 76] L.R. Rabiner, M.J. Cheng, A.E. Rosenberg, and C.A. McGonegal. *A Comparative Performance Study of Several Pitch Detection Algorithms*. IEEE Transactions on Acoustics, Speech, and Signal Processing, 24(5):399–418, October 1976.

[Rabiner 77] L.R. Rabiner. *On the Use of Autocorrelation Analysis for Pitch Detection*. IEEE Transactions on Acoustics, Speech, and Signal Processing, 25(1):24–33, February 1977.

[Rabiner 78] L.R. Rabiner and R.W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall Signal Processing Series. Prentice-Hall, Englewood Cliffs (NJ), USA, 1978.

[Rabiner 93] L.R. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall Signal Processing Series. Prentice-Hall, Englewood Cliffs (NJ), USA, 1993.

[Ramsay 82] J.O. Ramsay. *When the Data Are Functions*. Psychometrika, 47(4):379–396, December 1982.

[Ramsay 96] J.O. Ramsay, K.G. Munhall, V.L. Gracco, and D.J. Ostry. *Functional data analyses of lip motion*. Journal of the Acoustical Society of America, 99(6):3718–3727, June 1996.

[Ramsay 01] J.O. Ramsay. *A Guide to Curve Registration*. McGill University, Montreal, Canada, January 2001.

[Ramsay 03] J.O. Ramsay. *Matlab, R and S-PLUS Functions for Functional Data Analysis*. McGill University, Montreal, Canada, April 2003.

[Rao 64] C.R. Rao. *The use and interpretation of principal component analysis in applied research*. Sankhya A, 26:329–359, 1964.

[Reinders 96] M.J.T. Reinders, R.W.C. Koch, and J.J. Gerbrands. *Locating Facial Features in Image Sequences Using Neural Networks*. In Proceedings of the Second International Conference on Automatic Face and Gesture Recognition FG'96, pages 230–235, Killington (VT), USA, October 1996. IEEE.

[Rencher 98] A.C. Rencher. *Multivariate Statistical Inference and Applications*. Wiley series in probability and statistics. John Wiley & Sons, New York (NY), USA, 1998.

[Repp 84] B.H. Repp. *Categorical perception: Issues, methods, findings*. In N.J. Lass, editor, Speech and language: Advances in basic research and practice, pages 243–335. Academic Press, San Diego (CA), USA, 1984.

[Revéret 98] L. Revéret and C. Benoît. *A new 3D Lip Model for Analysis and Synthesis of Lip Motion*. In D. Burnham, J. Robert-Ribes, and E. Vatikiotis-Bateson, editors, Proceedings of the International Conference on Auditory-Visual Speech Processing AVSP'98, pages 207–212, Terrigal, Australia, December 1998.

[Robert-Ribes 94] J. Robert-Ribes, J.-L. Schwartz, and P. Escudier. *Audio-Visual Recognition of Speech Units: A Tentative Functional Model Compatible with Psychological Data*. In Proceedings of the Australian International Conference on

Speech Science and Technology SST-94, Volume 2, pages 448–453, Perth, Australia, December 1994. Australian Speech Science and Technology Association (ASSTA).

[Robert-Ribes 95a] J. Robert-Ribes. *Modèles d'intégration audiovisuelle de signaux linguistiques: de la perception humaine à la reconnaissance automatique des voyelles*. PhD thesis, Institut National Polytechnique de Grenoble, France, 1995.

[Robert-Ribes 95b] J. Robert-Ribes, J.-L. Schwartz, and P. Escudier. *A Comparison of Models for Fusion of the Auditory and Visual Sensors in Speech Perception*. Artificial Intelligence Review, 9(4-5):323–346, 1995.

[Robert-Ribes 96] J. Robert-Ribes, M. Piquemal, J.-L. Schwartz, and P. Escudier. *Exploiting Sensor Fusion Architectures and Stimuli Complementarity in AV Speech Recognition*. In D.G. Stork and M.E. Hennecke, editors, Speechreading by Humans and Machines, Volume 150 of *NATO ASI Series*, pages 193–210, Berlin, Germany, 1996. Springer-Verlag.

[Robert 76] P. Robert and Y. Escoufier. *A Unifying Tool for Linear Multivariate Statistical Methods: The RV-Coefficient*. Applied Statistics, 25(3):257–265, 1976.

[Rogozan 97] A. Rogozan, P. Deléglise, and M. Alissali. *Adaptive Determination of Audio and Visual Weights for Automatic Speech Recognition*. In C. Benoît and R. Campbell, editors, Proceedings of the ESCA Workshop on Audio-Visual Speech Processing AVSP'97, pages 61–64, Rhodes, Greece, September 1997.

[Rosenblum 96] L.D. Rosenblum and H.M. Salda na. *An audiovisual test of kinematic primitives for visual speech perception*. Journal of Experimental Psychology: Human Perception and Performance, 22(2):318–331, 1996.

[Russ 95] J.C. Russ. *The Image Processing Handbook*. CRC Press, Boca Raton (FL), USA, 2nd edition, 1995.

[Sakoe 78] H. Sakoe and S. Chiba. *Dynamic Programming Algorithm Optimization for Spoken Word Recognition*. IEEE Transactions on Acoustics, Speech, and Signal Processing, 26(1):43–49, 1978.

[Sams 97] M. Sams, V. Surakka, P. Helin, and R. Kättö. *Audiovisual fusion in Finnish syllables and words*. In C. Benoît and R. Campbell, editors, Proceedings of the ESCA Workshop on Audio-Visual Speech Processing AVSP'97, pages 101–104, Rhodes, Greece, September 1997.

[Scanlon 01] P. Scanlon and R. Reilly. *Feature Analysis for Automatic Speechreading*. In J.-L. Dugelay and K. Rose, editors, Proceedings of the 2001 IEEE Fourth Workshop on Multimedia Signal Processing, pages 625–630, Cannes, France, October 2001. IEEE.

[Schafer 70] R.W. Schafer and L.R. Rabiner. *System for Automatic Formant Analysis of Voiced Speech.* Journal of the Acoustical Society of America, 47(2):634–648, 1970.

[Schwartz 02] J.-L. Schwartz, F. Berthommier, and C. Savariaux. *Audio-Visual Scene Analysis Evidence for a "Very-Eearly" Integration Process in Audio-Visual Speech Perception.* In Proceedings of the 7th International Conference on Spoken Language Processing ICSLP2002, Volume 3, pages 1937–1940, Denver (CO), USA, September 2002.

[Schwartz 03] J.-L. Schwartz, F. Berthommier, and C. Savariaux. *Auditory Syllabic Identification Enhanced by Non-Informative Visible Speech.* In J.L. Schwartz, F. Berthommier, M.A. Cathiard, and D. Sodoyer, editors, Proceedings of the ISCA Tutorial and Research Workshop on Audio Visual Speech Processing AVSP2003, pages 19–24, St Jorioz, France, September 2003. ICP, INP Grenoble, Université Stendhal.

[Secrest 83] B.G. Secrest and G.R. Doddington. *An Integrated Pitch Tracking Algorithm for Speech Systems.* In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP'83, pages 1352–1355, Boston (MA), USA, May 1983. IEEE.

[Sekiyama 93] K. Sekiyama. *Inter-language differences in the influence of visual cues in speech perception.* Journal of Phonetics, 21:427–444, April 1993.

[Sekiyama 98] K. Sekiyama. *Face or Voice? Determinant of Compellingness to the McGurk Effect.* In D. Burnham, J. Robert-Ribes, and E. Vatikiotis-Bateson, editors, Proceedings of the International Conference on Auditory-Visual Speech Processing AVSP'98, pages 33–36, Terrigal, Australia, December 1998.

[Seneff 76] S. Seneff. *Modifications to Formant Tracking Algorithm of April 1974.* IEEE Transactions on Acoustics, Speech, and Signal Processing, 24(2):192–193, April 1976.

[Senior 99] A.W. Senior. *Face and Feature Finding for a Face Recognition System.* In Proceedings of the Second International Conference on Audio and Video-based Biometric Person Authentication AVBPA99, pages 154–159, Washington (DC), USA, March 1999.

[Sennheiser 81] Sennheiser. *User's Guide: Handgrip/Powering Module K3N / K3U and Microphone Heads MKE 10-3 / ME 20 / ME 40 / ME 80 / ME 88.* Sennheiser Electronic KG, Wedemark, Germany, December 1981.

[Smeele 96] P.M.T. Smeele. *Psychology of Human Speechreading.* In D.G. Stork and M.E. Hennecke, editors, Speechreading by Humans and Machines, Volume 150 of *NATO ASI Series*, pages 3–15, Berlin, Germany, 1996. Springer-Verlag.

[Snell 93] R.C. Snell and F. Milinazzo. *Formant Location From LPC Analysis Data*. IEEE Transactions on Speech and Audio Processing, 1(2):129–134, April 1993.

[Sobottka 96] K. Sobottka and I. Pitas. *Segmentation and Tracking of Faces in Color Images*. In Proceedings of the Second International Conference on Automatic Face and Gesture Recognition FG'96, pages 236–241, Killington (VT), USA, October 1996. IEEE.

[Sodoyer 03] D. Sodoyer, L. Girin, C. Jutten, and J.-L. Schwartz. *Further experiments on audio-visual speech source separation*. In J.L. Schwartz, F. Berthommier, M.A. Cathiard, and D. Sodoyer, editors, Proceedings of the ISCA Tutorial and Research Workshop on Audio Visual Speech Processing AVSP2003, pages 145–150, St Jorioz, France, September 2003. ICP, INP Grenoble, Université Stendhal.

[Stein 93] B.E. Stein and M.A. Meredith. *The Merging of the Senses*. MIT Press, Cambridge (MA), USA, 1993.

[Stevens 37] S.S. Stevens, J. Volkmann, and E.B. Newmann. *A scale for the measurement of a psychological magnitude: Pitch*. Journal of the Acoustical Society of America, 8(1):185–190, January 1937.

[Stevens 55] K.N. Stevens and A.S. House. *Development of a Quantitative Description of Vowel Articulation*. Journal of the Acoustical Society of America, 27(3):484–493, May 1955.

[Stevens 56] K.N. Stevens and A.S. House. *Studies of Formant Transitions Using a Vocal Tract Analog*. Journal of the Acoustical Society of America, 28(4):578–585, July 1956.

[Stevens 67] K.N. Stevens and M. Halle. *Remarks on analysis by synthesis and distinctive features*. In W. Wathen-Dunn, editor, Models for the perception of speech and visual form, pages 88–102. MIT Press, Cambridge (MA), USA, 1967.

[Stevens 72] K.N. Stevens. *The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data*. In E.E. David, Jr. and P.B. Denes, editors, Human Communication: A Unified View, Volume 15 of *Inter-University Electronics Series*, chapter 3, pages 51–66. McGraw-Hill Book Company, New York (NY), USA, 1972.

[Stevens 89] K.N. Stevens. *On the quantal nature of speech*. Journal of Phonetics, 17:3–45, 1989.

[Summerfield 87] Q. Summerfield. *Some preliminaries to a comprehensive account of audio-visual speech perception*. In B. Dodd and R. Campbell, editors, Hearing by Eye, pages 3–51. Lawrence Erlbaum Associates, Hillsdale (NJ), USA, 1987.

[ter Braak 86] C.J.F. ter Braak. *Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis*. Ecology, 67:1167–1179, 1986.

[Thioulouse 97] J. Thioulouse, D. Chessel, S. Dolédec, and J.-M. Olivier. *ADE-4: a multivariate analysis and graphical display software*. Statistics and Computing, 7:75–83, 1997.

[Tran 00] Q.N. Tran. *Show Me Your Lips*. Technical report, Computer Sciences Laboratory, Research School of Information Sciences and Engineering, Australian National University, Canberra, Australia, 2000.

[Trucco 98] E. Trucco and A. Verri. *Introductory Techniques for 3-D Computer Vision*. Prentice Hall, Upper Saddle River (NJ), USA, 1998.

[Tsai 86] R.Y. Tsai. *An Efficient and Accurate Camera Calibration Technique for 3D Machine Vision*. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR'86, pages 364–374, Miami Beach (FL), USA, June 1986. IEEE.

[Tucker 58] L.R. Tucker. *An inter-battery method of factor analysis*. Psychometrika, 23:111–136, 1958.

[van den Berg 58] J. van den Berg. *Myoelastic-aerodynamic theory of voice production*. Journal of Speech and Hearing Research, 1:227–244, 1958.

[Vatikiotis-Bateson 95] E. Vatikiotis-Bateson and D.J. Ostry. *An analysis of the dimensionality of jaw motion in speech*. Journal of Phonetics, 23:101–117, 1995.

[Vatikiotis-Bateson 96] E. Vatikiotis-Bateson, K.G. Munhall, M. Hirayama, Y.V. Lee, and D. Terzopoulos. *The Dynamics of Audiovisual Behaviour in Speech*. In D.G. Stork and M.E. Hennecke, editors, Speechreading by Humans and Machines, Volume 150 of *NATO ASI Series*, pages 221–232, Berlin, Germany, 1996. Springer-Verlag.

[Venables 99] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S-PLUS*. Statistics and Computing. Springer-Verlag, New York (NY), USA, 3rd edition, 1999.

[Vogt 96] M. Vogt. *Fast Matching of a Dynamic Lip Model to Color Video Sequences Under Regular Illumination Conditions*. In D.G. Stork and M.E. Hennecke, editors, Speechreading by Humans and Machines, Volume 150 of *NATO ASI Series*, pages 399–407, Berlin, Germany, 1996. Springer-Verlag.

[Vroomen 92] J.M.H. Vroomen. *Hearing Voices and Seeing Lips: Investigations in the Psychology of Lipreading*. PhD thesis, Tilburg University, The Netherlands, 1992.

[Wagner 82] M. Wagner. *Formant Extraction Algorithm in Error.* IEEE Transactions on Acoustics, Speech, and Signal Processing, 30(3):520, June 1982.

[Wark 98] T. Wark and S. Sridharan. *A Syntactic Approach to Automatic Lip Feature Extraction for Speaker Identification.* In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP'98, Volume 6, pages 3693–3696, Seattle (WA), USA, May 1998.

[Wark 01] T. Wark and S. Sridharan. *Adaptive Fusion of Speech and Lip Information for Robust Speaker Identification.* Digital Signal Processing, 11:169–186, 2001.

[Wegmann 96] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin. *Speaker normalization on conversational telephone speech.* In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP'96, Volume 1, pages 339–343, Atlanta (GA), USA, May 1996.

[Welling 96] L. Welling and H. Ney. *A Model for Efficient Formant Estimation.* In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP'96, pages 797–800, Atlanta (GA), USA, May 1996. IEEE.

[White 76] G.M. White and R.B. Nelly. *Speech Recognition Experiments with Linear Predication, Bandpass Filtering and Dynamic Programming.* IEEE Transactions on Acoustics, Speech, and Signal Processing, 24(2):183–188, April 1976.

[Wiener 57] N. Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications.* John Wiley & Sons, New York (NY), USA, reprint of 1949 edition, 1957.

[Wojdel 01a] J.C. Wojdel and L.J.M. Rothkrantz. *Robust Video Processing for Lipreading Applications.* In Proceedings of the 6th Annual Scientific Conference on Web Technology, New Media, Communications and Telematics Theory, Methods, Tools and Applications EUROMEDIA2001, Valencia, Spain, April 2001.

[Wojdel 01b] J.C. Wojdel and L.J.M. Rothkrantz. *Using Aerial and Geometric Features in Automatic Lip-reading.* In P. Dalsgaard, B. Lindberg, and H. Benner, editors, Proceedings of the 7th European Conference on Speech Communication and Technology EUROSPEECH2001, pages 2463–2466, Aalborg, Denmark, September 2001. ISCA.

[Wood 79] S. Wood. *A radiographic analysis of constriction locations for vowels.* Journal of Phonetics, 7:25–43, 1979.

[Woodward 60] M.F. Woodward and C.G. Barber. *Phoneme Perception in Lipreading.* Journal of Speech Hearing Research, 3(3):212–222, September 1960.

[Wrench 95] A.A. Wrench. *Analysis of Fricatives Using Multiple Centres of Gravity*. In Proceedings of the 13th International Congress of Phonetic Sciences ICPhS95, Volume 4, pages 460–463, Stockholm, Sweden, August 1995.

[Xu 96] G. Xu and Z. Zhang. *Epipolar Geometry in Stereo, Motion, and Object Recognition: A Unified Approach*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[Yakel 95] D.A. Yakel, D. Rosenblum, Lawrence, K.P. Green, R.A. Vasquez, and C. Bosley. *The effect of face and lip inversion on audiovisual speech integration*. Journal of the Acoustical Society of America, 97:3286, 1995.

[Yang 96] J. Yang and A. Waibel. *A Real-Time Face Tracker*. In Proceedings of the Third IEEE Workshop on Applications of Computer Vision WACV'96, pages 142–147, Sarasota (FL), USA, 1996.

[Yang 98] J. Yang, R. Stiefelhagen, U. Meier, and A. Waibel. *Real-time Face and Facial Feature Tracking and Applications*. In D. Burnham, J. Robert-Ribes, and E. Vatikiotis-Bateson, editors, Proceedings of the International Conference on Auditory-Visual Speech Processing AVSP'98, pages 79–84, Terrigal, Australia, December 1998.

[Yegnanarayana 78] B. Yegnanarayana. *Formant extraction from linear-prediction phase spectra*. Journal of the Acoustical Society of America, 63(5):1638–1640, May 1978.

[Yegnanarayana 98] B. Yegnanarayana. *Extraction of Vocal-Tract System Characteristics from Speech Signals*. IEEE Transactions on Speech and Audio Processing, 6(4):313–327, July 1998.

[Yehia 97] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson. *Quantitative Association of Orofacial and Vocal-Tract Shapes*. In C. Benoît and R. Campbell, editors, Proceedings of the ESCA Workshop on Audio-Visual Speech Processing AVSP'97, pages 41–44, Rhodes, Greece, September 1997. ESCA.

[Yehia 98] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson. *Quantitative association of vocal-tract amd facial behaviour*. Speech Communication, 26(1–2):23–43, 1998.

[Yuille 92] A.L. Yuille, P.W. Hallinan, and D.S. Cohen. *Feature Extraction from Faces Using Deformable Templates*. International Journal of Computer Vision, 8(2):99–111, 1992.

[Zelinsky 99] A. Zelinsky, Y. Matsumoto, J. Heinzmann, and R. Newman. *Towards Human Friendly Robots: Vision-based Interfaces and Safe Mechanisms*. In P. Corke and J. Trevelyan, editors, Proceedings of the Sixth International Symposium

on Experimental Robotics ISER'99, Lecture Notes in Control and Information Sciences LNCIS 250, pages 487–498, Sydney, Australia, March 1999. Springer-Verlag, London, United Kingdom.