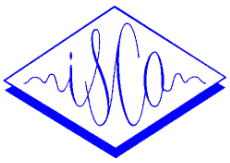

AVSP 2003 – International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4–7, 2003

Statistical Analysis of the Relationship between Audio and Video Speech Parameters for Australian English

Roland Goecke[1] *and Bruce Millar*[2]

[1]Human-Centered Interaction & Technologies, Fraunhofer IGD, Rostock, Germany
[2]Research School of Information Sciences and Engineering, Australian National University, Canberra, Australia
Roland.Goecke@ieee.org Bruce.Millar@anu.edu.au

Abstract

After decades of research, automatic speech processing has become more and more viable in recent years. Audio-video speech recognition has been shown to improve the recognition rate in noise-degraded environments. However, which audio and video speech parameters to choose for an optimal system and how they are related is still an open research issue. Here we present a number of statistical analyses that aim at increasing our understanding of such audio-video relationships. In particular, we look at the canonical correlation analysis and the coinertia analysis which investigate the relationship of linear combinations of parameters. The analyses are performed on Australian English as an example.

1. Introduction

Human-Computer Interaction (HCI) is a fast growing field these days. More and more computer systems are used in every aspect of our life. An important aspect of such technology is the way in which people interact with it. Traditional means of HCI are often cumbersome or simply impractical in many new application areas. Research in recent years has more and more focussed on ways of "tracking" the user, understanding what the person does or says, and then reacting appropriately. Although automatic speech recognition (ASR) systems have improved significantly in recent years, they still have some limitations with respect to the environment in which they can be used. Current commercially available ASR systems employ statistical models of spoken language and enable continuous speech recognition in reasonably good acoustic conditions. However, they can fail unpredictably in noisy conditions. One way of overcoming some of the limitations of audio-only ASR systems is to use the additional visual information of the act of speaking similar to what humans do when facing adverse acoustic conditions (e.g. [1, 2, 3, 4]).

Various research groups around the world have shown that the incorporation of visual speech information in an ASR system can significantly improve the recognition rate, in particular in noisy conditions (e.g. [5, 6]). However, how the audio and video speech parameters are related to one another and how these relationships can be exploited best, still remains an open research issue. Yehia et al [7] looked at the relationship between acoustic parameters, the shape of the vocal tract, and the position of facial feature points around the mouth and lower face from a speech production and animation perspective. A strong correlation (80-91%) was found between the latter two. They also found that a large part (72-85%) of the variance observed in acoustic parameters can be determined from vocal-tract and facial data together. And even the facial data alone performed well in accounting for the acoustic parameter variance. The drawbacks of their study are the small number of speakers looked at (only 2) and the use of intrusive measurement techniques (transducers inside the oral cavity).

Our goal in the work presented here is to apply statistical analyses to increase our understanding of the relationship between audio and video speech parameters. We deliberately chose explicit geometric video speech parameters, such as the height of the mouth opening, over implicit ones (region of interest) because they facilitate the interpretation of the results. As an example, we use data for Australian English (AuE) from our AVOZES data corpus (Section 2). Section 3 describes the recording setup and the content of the AVOZES data corpus in more detail. The statistical analyses are detailed in Sections 4 and 5. Section 6 presents the results of the coinertia analysis. Finally, the results are discussed in Section 7, before we finish with the conclusions (Section 8).

2. AVOZES Data Corpus

We use the data recorded in our Audio-Video Australian English Speech (AVOZES) data corpus [8]. The recordings were made with a stereo camera system to achieve more accurate 3D measurements on the face. To exploit this fact, we developed a novel, non-intrusive 3D lip tracking algorithm [9]. It does not require any artificial markers on the face, thus allowing a more natural behaviour of the speaker and greater practical application. Through the use of a face tracking system, the speakers were allowed to move their head freely within the cameras' field of view, again facilitating a more natural way of speaking in. The speakers sit in front of a stereo camera pair with an omnidirectional microphone attached 20-25cm below their mouth (Fig. 1). The face is well illuminated by a light source just below the cameras so that no shadows appear on the face. Recordings were made to digital video (DV) tape because of its ability to playback the recordings many times without a loss of quality. The recordings were made at 30Hz video frame rate and 16bit 48kHz mono audio rate in a controlled acoustic environment (almost no external noise, some air conditioning and computer noise in the background).

The AVOZES data corpus comprises 20 native speakers of AuE (10 female and 10 male speakers). It covers all phonemes and visemes in AuE except for the neutral vowel /ə/ because of its great audio variability and the neutral consonant /h/ which adds little to the analysis. In addition, the voiced fricative /ʒ/ and the diphthong /ʊə/ were also omitted because they have a low occurence in AuE. The core part of the corpus consists of 40

Figure 1: Setup for AVOZES data corpus recordings.

sequences per speaker containing consonant-vowel-consonant- (CVC-) or vowel-consonant-vowel- (VCV-) nonsense words with the phoneme of interest in the central position. The vowel context is the wide open /ɑː/ ("ar-ar"). The voiced bi-labial /b/ is used as the consonant context ("b-b"). These words are put in a carrier phrase ("You grab *WORD* beer.") to overcome articulation patterns associated with reading words from a list. Visual segmentation is facilitated through the use of bi-labial closings before and after the CVC- or VCV-word. We are aware that a bi-labial context causes strong coarticulation effects in the formants. However, these are quite predictable for /b/ and we believe that the advantages of a bi-labial context for visual segmentation outweigh the disadvantages from coarticulation.

3. Parameter Extraction and Preprocessing

From the recorded data, we extract 5 audio speech parameters and 5 video speech parameters (Table 1) which are measured at different sampling rates (100Hz for the audio parameters and 30Hz for the video parameters). The former are extracted with the help of the *esps* software package. The latter parameters are determined by applying our face tracking and lip tracking algorithms. Feature points on the inner lip contour, such as the lip corners, are automatically determined in both camera images for each frame and the 3D coordinates calculated. From these, real 3D distances can be measured, rather than measuring parameters based on 2D image coordinates. Thus, head movements by a speaker do not lead to measurement errors.

Audio Parameters	Voice source excitation F0 Formant frequencies F1, F2, F3 RMS energy
Video Parameters	Mouth width (MW) Mouth Height (MH) Protrusion upper lip (PUL) Protrusion lower lip (PLL) Relative teeth count (RTC)

Table 1: The audio and video speech parameters (variables).

The audio speech parameters are standard acoustic, energy, and frequency parameters. The video speech parameters are explicit geometric features. The mouth width is defined as the 3D distance from lip corner to lip corner. Similarly, the mouth height is defined as the 3D distance from the midpoint of the upper lip to the midpoint of the lower lip. The lip protrusion parameters are not easy to determine without using additional fixed markers in order to measure the relative distance. We want to avoid such additional markers by defining the protrusion parameters as the respective distances from the midpoint of the line from lip corner to lip corner to each lip midpoint. In addition to these metric measures, we introduce the "relative teeth count" as a measure of the visibility of teeth which we consider to be a potentially useful measure. It is defined as the average of the number of teeth pixels c in A in the left and right camera image divided by the distance d to the cameras (as measured for the left lip corner):

$$RTC = \frac{c_l^A + c_r^A}{2} / d \quad (1)$$

where A is the rectangular area spanned by the four lip feature points already determined.

The measurements of the audio and video speech parameters contain an error component. As a result, the parameter curves can show incorrect large variation although the underlying function is actually smooth. Therefore, cubic spline smoothing is performed on the parameter curves (parameters over time). The relationship between audio and video speech parameters is determined at a phoneme level. The sample points corresponding to each phoneme of interest are determined as follows. For the vowels and diphthongs, the positions of the bi-labial closures before and after the phoneme of interest are found on the time line by looking at the mouth height parameter curve. Sample points between them are taken as belonging to the vowel or diphthong. For the consonants, first the bi-labial closures of the surrounding words are identified in the mouth height parameter curve. Then the maxima of the mouth height parameter curve corresponding to the wide open /ɑː/ context are found. Any sample points on the time line between these maxima are taken as belonging to the consonant.

To facilitate the statistical analysis of each sequence (phoneme of interest), all audio and video speech parameters are resampled to 25 observation points on the time line. A linear resampling is chosen to retain the original parameter curve shapes, whereas dynamic time warping changes the shape. By resampling, inter-speaker differences with respect to the length of a phoneme are taken care of.

4. From Correlations to Canonical Correlations

4.1. Within-Set Relationships

Firstly, correlations between pairs of parameters within each parameter set are computed to evaluate if any parameters are redundant. In combination with principal component analysis (PCA), redundant parameters in each set are identified and removed from further analysis. A PCA is applied separately to the sets of audio and video speech parameters for each phoneme.

In the case of the video speech parameters, the PCA results show that the first four principal components (PCs) already explain 96-99% of the variance, so that the fifth PC is somewhat redundant. Using the correlation analyis, we can identify that the two lip protrusion parameters are highly correlated which can be expected. Upper and lower lip are typically moved simultaneously and in a similar fashion. The correlation coeffi-

cient for the two protrusion parameters shows values between 0.79 and 0.99, with particularly high values for vowels and diphthongs. It is therefore sufficient to continue the analysis with only one protrusion parameter and we choose the upper lip protrusion parameter. No other strong correlation between any parameters is found in the video speech parameter set.

For the audio speech parameters, the PCA results also suggest that some parameters are correlated and that there is thus redundancy in the data. The first four PCs cover 90-97% of the variance. However, no single pair of parameters stands out in the correlation analysis. This suggests that it is rather a case of a combination of parameters being correlated than two single parameters. It is only after the PCA, with the orthogonal PCs forming a new coordinate system, that four 'new' parameters are able to express an average of 94% of the variance. Hence, we cannot simply take one particular parameter away but must include all five parameters in the further analyses.

Secondly, before other statistical analyses are applied, the amount of data is reduced by again using PCA but this time as a statistical shape analysis technique. For this, a PCA is applied to each parameter separately for each phoneme, that is, a PCA is performed on the temporal domain. This allows us to find the main modes of variation in the shape of the parameter curves and thus the relationship between points on the time scale and the PCs. In addition, it enables the use of a compact representation of the individual parameter curves in our further analyses.

Table 2 shows the average proportion of variance explained by the top three PCs for each parameter. Any further PCs can be neglected as the amount of variance they cover is small. To analyse what variation in the shape of the parameter curves the PCs stand for, we plot the mean parameter graph for each phoneme and each parameter and add the graphs of the PCs with ± 10 standard deviations. Figure 2 shows a typical example for the first three PCs.

	Vowels / Diphthongs			Consonants		
	PC1	PC2	PC3	PC1	PC2	PC3
F0	0.83	0.11	0.04	0.86	0.07	0.04
F1	0.50	0.21	0.12	0.44	0.26	0.14
F2	0.64	0.15	0.09	0.54	0.21	0.12
F3	0.67	0.14	0.08	0.60	0.19	0.10
RMS	0.42	0.23	0.14	0.42	0.23	0.16
MW	0.83	0.12	0.03	0.85	0.10	0.03
MH	0.66	0.18	0.10	0.65	0.20	0.11
PUL	0.54	0.22	0.12	0.50	0.24	0.12
RTC	0.65	0.22	0.09	0.80	0.13	0.05

Table 2: Average proportion of variance explained by the top three PCs for each parameter.

The PCs relate to three main modes of variation:

- a vertical shift,
- a mode related to the slope of the curve, and
- a mode describing the horizontal range or shift.

For the vast majority of parameter-phoneme pairs, the first PC is related to a vertical shift of the parameter curve (Figure 2 left). In other words, the strongest variation for the individual curves of the speakers is in these cases not related to differences in the curve shape but to a mere shift which appears to be a personal characteristic of each speaker. This shift occurs for all sample points on the time scale and is almost invariant in size. In contrast, the second and third PC (Figure 2 centre and right) express variation in the curve shape. These PCs are related to the slope of the curve and the horizontal range or shift.

We are interested in the common behaviour of parameters over time 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, we are interested in the shape, not the shift separating the curves. Consequently, the PC that expresses the vertical shift is not used as input for the analyses, as a way of normalising the data. We focus on the two shape PCs that are related to the slope of the curve and the horizontal range or shift instead.

4.2. Between-Set Relationships

First, a pairwise linear correlation analysis is performed across the parameter sets. That is, the correlations between one parameter from each set is looked at. No strong correlations are found, nor are the weaker correlations consistent for all phonemes or some subgroups of them. The correlations appear to be phoneme-specific. The largest correlation values ($|r| \approx 0.5$) are found for F0 and mouth width for phoneme /ɪ/, for F2 and mouth height for phonemes /ʌ eɪ ɔɪ/, for F3 and RTC for phoneme /eɪ/, for RMS and mouth height for phoneme /p/, and for RMS and RTC also for phoneme /p/. Thus, the data does not support a hypothesis of a direct 1-1 relationship between any of the speech parameters in the two sets. However, a combination of parameters from either set may correlate well, or the parameters could be related in a non-linear manner that is not uncovered by linear statistical methods. Hence, we next look at statistical analyses that explore the relationship between combinations of parameters.

Canonical correlation analysis (CANCOR) is a statistical analysis for the exploration of relationships of linear combinations of variables. CANCOR is a generalisation of multiple correlation analysis for sets of parameters with at least two parameters in each set [10]. Similarly to PCA, a rotation of the coordinate system is performed but rather than 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 relationships within each set are disentangled, so that the relationships between the sets become clear. The variables in the new coordinate system are linear combinations of the parameters in each set and are called canonical variates (CV). They are orthogonal to each other and successive pairs of CVs are uncorrelated. Most of the covariance between parameter sets is explained by the first few CVs.

The correlation between the linear combinations is given by the canonical correlation coefficients $r_k \in [-1, +1]$ where k is an index number between 1 and the sum of the number of parameters in the sets. r_1 refers to the highest canonical correlation coefficient. Typically, only the first 2-3 r_k's are of interest as levels of correlation drop quickly.

For small samples where the number of parameters approaches the sample size, r_1 quickly tends towards 1. Canonical correlation coefficients computed in such cases can be misleading with respect to the extent of linear relationship between the linear combinations in question. In our case, the sample size is $N = 20$, the audio set has 5 parameters, and the video set 4 parameters. Ideally, we would like to take the two shape PCs identified earlier as input. However, taking 18 parameters into a CANCOR analysis of sample size $N = 20$ will almost surely lead to $r_1 \rightarrow 1$ and thus the results would be of little value due to collinearity. In consequence, we only take the PC related to

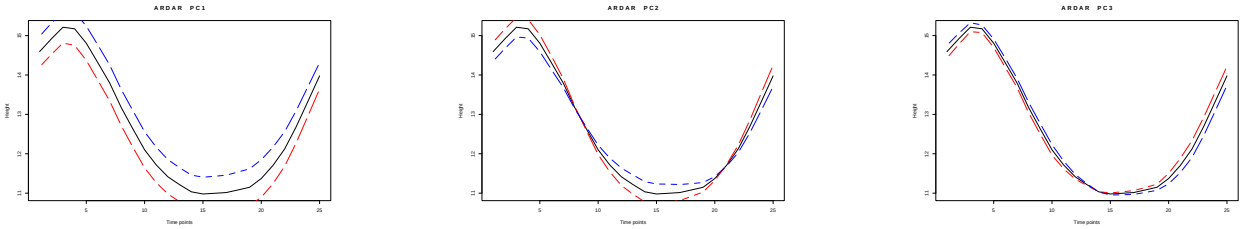

Figure 2: Typical modes of variation by the top three PCs on the example of the phoneme /d/ and the mouth height parameter. Shown are mean curve (solid line) and curves showing the effect of the PC at ± 10 standard deviations (dashed lines).

the slope of the curve as input, considering it the more important one of the two PCs. The PCs are normalised to zero mean and unit variance.

Table 3 shows for each phoneme the canonical correlation coefficient r_1. The canonical weights - the coefficients of the linear combinations (CVs) - are not shown here because they are small in magnitude (≤ 0.1) with few exceptions. Overall, no parameter can be shown to distinctively contribute more (or less) to the canonical correlation. It is the combinations of all parameters which are highly correlated, not single parameters.

Short Vowels	ɪ	ʊ	ɛ	ɒ	ʌ	æ
r_1	0.59	0.80	0.85	0.61	0.76	0.76
Long Vowels	iː	uː	ɜː	ɔː	ɑː	əː
r_1	0.85	0.76	0.83	0.80	0.81	0.87
Diphthongs	eɪ	ɔɪ	aɪ	aʊ	ɪə	əʊ
r_1	0.79	0.81	0.79	0.81	0.76	0.60
Consonants	p	b	t	d	k	g
r_1	0.84	0.68	0.87	0.83	0.83	0.64
Consonants	f	v	θ	ð	s	z
r_1	0.71	0.66	0.78	0.81	0.77	0.78
Consonants	ʃ	tʃ	dʒ	m	n	ŋ
r_1	0.70	0.69	0.65	0.75	0.82	0.77
Consonants	l	r	w	j		
r_1	0.78	0.57	0.82	0.76		

Table 3: Canonical correlation values r_1.

As Gittins [10] points out, a substantive interpretation of the pattern of canonical weights is difficult anyway. They are known for their instability. Small changes in the parameters have great effects. Factors contributing to this instability are insufficient sample size, measurement errors, and collinearity. Therefore, we look at another statistical method called coinertia analysis which does not suffer the same problems.

5. Coinertia Analysis

Coinertia analysis (COIA) is a relatively new multivariate statistical analysis for coupling two (or more) sets of parameters by looking at linear combinations of these. It was introduced for ecological studies by Dolédec and Chessel [11]. As it appears to be relatively unknown in the HCI community, we will first give some background information. In COIA, the term 'inertia' is used as a synonym for variance. The method is related to other multivariate analyses such as canonical correspondence analysis, redundancy analysis, and canonical correlation analysis [10]. COIA is a generalisation of the inter-battery analysis by Tucker [12] which in turn is the first step of partial least squares methods [13].

In many aspects, COIA is very similar to CANCOR. It also rotates the data to a new coordinate system and the new variables are linear combinations of the parameters in each parameter set. However, this time, it is not the correlation between the two sets that is maximised but the coinertia (or covariance) which can be decomposed as

$$cov(A,V) \;=\; corr(A,V) * \sqrt{var(A)} * \sqrt{var(V)}\,. \quad (2)$$

COIA finds a compromise between the correlation ($corr(A,V)$) and the variance in either set ($var(A)$, $var(V)$). It aims 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.

The advantage of COIA is its numerical stability. The number of parameters relative to the sample size does not affect the accuracy and stability of the results [11]. The method does not suffer from collinearity and the consistency between the correlation and the coefficients is very good [14]. Thus, COIA is a very well-suited multivariate method in our case. [1]

The coinertia value is a global measure of the co-structure in the two sets. If the value is high, the two parameter sets vary accordingly (or inversely). If it 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 of each set and then compare the projected variance of the separate analyses with the variance from the COIA (see the appendix of [11] for the theory). 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 taken by the coinertia vectors. It is important to compare the sum of axes, not axis by axis, because the variance projected onto the second axis depends on what is projected onto the first axis, and so on. Often it is sufficient to look at the first 2-3 axes because they already account for 90-95% of the variance. In addition, COIA computes the weights (coefficients) of the individual parameters in the linear combinations of each set, so that it becomes obvious which parameters contribute to the common structure of the two sets and which do not. As has already been pointed out, these weights are much more stable than the weights in a CANCOR analysis. Finally, a measure of overall 'relatedness' of the two domains based on the selected parameters is given by the RV coefficient [16].

[1] COIA can be computed with the ADE-4 tool [15] and is also available on the R statistical platform in the 'ade4' package.

Scores	ɪ	ʊ	ɛ	ɒ	ʌ	æ
cov	4.37	3.27	7.37	8.37	8.03	5.32
corr	0.66	0.64	0.75	0.68	0.54	0.59
Ratio Audio	0.69	0.68	0.88	0.65	0.80	0.64
Ratio Video	0.91	0.43	0.74	0.97	0.77	0.90
RV	0.23	0.14	0.37	0.23	0.17	0.25
Scores	iː	uː	ɜː	ɔː	ɑː	əː
cov	7.34	3.19	4.85	6.00	7.98	4.78
corr	0.83	0.68	0.82	0.75	0.67	0.63
Ratio Audio	0.91	0.84	0.75	0.95	0.95	0.72
Ratio Video	0.95	0.58	0.84	0.67	0.90	0.49
RV	0.50	0.26	0.33	0.34	0.32	0.17
Scores	eɪ	ɔɪ	aɪ	aʊ	ɪə	əʊ
cov	4.98	5.06	6.14	4.67	5.86	5.56
corr	0.65	0.67	0.65	0.79	0.74	0.70
Ratio Audio	0.74	0.79	0.96	0.51	0.69	0.38
Ratio Video	0.58	0.77	0.70	0.83	0.61	0.76
RV	0.22	0.33	0.24	0.31	0.26	0.18

Table 4: Coinertia scores for vowels and diphthongs.

One of COIA's biggest advantages is that it can be coupled with other statistical methods, such as correspondence analysis 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* [14] show that seen in this context, COIA is a generalisation of many multivariate methods. In our analysis it means that we can use both shape PCs as input for a COIA and are not restricted as in the case of CANCOR. As a result, the audio parameter set contains 10 parameters and the video parameter set 8 parameters made up by the two shape PCs for each parameter.

6. Results Coinertia Analysis

Tables 4 and 5 show the coinertia value, the ratio of each 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. The first three of these values exist for every coinertia axis. However, only the values for the first coinertia axis are shown which by definition is the axis onto which the largest amount of overall variance is projected and which is therefore the most important one.

The coinertia values range from 3.19 to 8.37 with a mean of 5.73 for the vowels and diphthongs, and from 3.29 to 10.38 with a mean of 5.99 for the consonants. Although the coinertia values are slightly higher for the consonants, there are no significant differences between them and the vowels and diphthongs. However, the coinertia values differ quite significantly between individual phonemes. As a rule of thumb, the higher both ratios of projected variance from separate analysis to variance from coinertia analysis are, which means the higher the amount of variance in a parameter set obtained by the coinertia axes is, the higher is the coinertia value. This follows from equation 2.

The correlation values range from 0.54 to 0.83 with a mean of 0.69 for the vowels and diphthongs, and from 0.50 to 0.82 with a mean of 0.68 for the consonants. It shows that the first coinertia vectors from each domain correlate well. Differences in the strength of the correlation exist for individual phonemes. For example, the correlation value is high for the vowels /ɛ iː ɜː ɔː/ and the consonants /p d k g v θ ð ʃ/, while it is considerably lower for the vowels /ʌ æ/ and the consonants /b tʃ ŋ l r/.

Scores	p	b	t	d	k	g
cov	9.90	5.32	10.38	4.78	5.11	5.40
corr	0.82	0.52	0.74	0.75	0.75	0.79
Ratio Audio	0.74	0.75	0.97	0.64	0.71	0.87
Ratio Video	0.95	0.94	0.90	0.76	0.70	0.69
RV	0.34	0.22	0.44	0.25	0.33	0.34
Scores	f	v	θ	ð	s	z
cov	7.16	8.42	6.17	6.53	6.12	6.22
corr	0.65	0.81	0.77	0.75	0.66	0.66
Ratio Audio	0.92	0.95	0.90	0.81	0.95	0.81
Ratio Video	0.98	0.53	0.72	0.92	0.82	0.76
RV	0.28	0.29	0.39	0.33	0.30	0.30
Scores	ʃ	tʃ	dʒ	m	n	ŋ
cov	5.73	4.66	3.59	5.46	3.86	5.74
corr	0.77	0.55	0.60	0.68	0.68	0.56
Ratio Audio	0.49	0.91	0.32	0.71	0.65	0.64
Ratio Video	0.89	0.62	0.71	0.70	0.60	0.79
RV	0.25	0.18	0.12	0.22	0.17	0.18
Scores	l	r	w	j		
cov	3.29	4.27	8.21	5.54		
corr	0.50	0.56	0.67	0.61		
Ratio Audio	0.87	0.79	0.81	0.67		
Ratio Video	0.69	0.52	0.71	0.81		
RV	0.17	0.14	0.24	0.20		

Table 5: Coinertia scores for consonants.

For the vowels and diphthongs, the amount of variance taken by the first coinertia axis ranges 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. Similarly, for the consonants, the amount of variance obtained by the first coinertia axis ranges 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. In other words, the first coinertia axis accounts for about 75% of the variance in either parameter set which confirms that (1) COIA is a suitable way to represent the data and (2) it is efficient to only look at the first axis in the analysis of the data in the AVOZES corpus.

Unlike the correlation value which belongs to a particular pair of coinertia axes, the RV coefficient takes all axes into account. For the vowels and diphthongs, it ranges from 0.14 to 0.50 with a mean of 0.27. For the consonants, the RV coefficients range from 0.12 to 0.44 with an average of 0.26. Roughly speaking, about a fifth to a third of the variance in either domain is predictable from the other domain.

Summarising the results of the computed parameter weights, we observe that all parameters contribute strongly to the linear combination for one phoneme or another, although some parameters contribute strongly significantly more times than others. For the vowels and diphthongs, the most often appearing strong parameters in the linear combinations of the first coinertia axis are the slope PC of RMS, relative teeth count, mouth height, upper lip protrusion, and F1, and the horizontal range PC of RMS (in that order). We find a similar picture for the consonants. Here, the slope PC of mouth height, RMS, F1, and relative teeth count are the parameters most often found to contribute strongly. The results show that the horizontal range or shift PC is, on a general level, not as important as the slope PC. However, on an individual level, it can be of importance.

Again, these results confirm earlier results from the CANCOR analysis that the way how parameters contribute to related linear combinations across the domains is phoneme-specific.

7. Discussion

In summary, for the data in the AVOZES data corpus, we have not found a strong correlation between single parameters of the audio parameter set and the video parameter set. Both CANCOR and COIA point to strong correlations of linear combinations of the parameters. COIA confirms largely the results from the CANCOR analysis but the analysis is much more statistically stable. Linear combinations of the parameters in each set are related well across the domains but the composition of those linear combinations is phoneme-specific. The PC related to the slope of a parameter curve contributes more to the linear combinations than the horizontal range or shift PC for most phonemes which does not surprise if we compare it with the average proportion of variance explained by the PCs (Table 2). The parameters most often contributing strongly to the linear combinations are F1, RMS, MH, and RTC. On average about 75% of the variance in each parameter set is obtained by the first coinertia vector which is sufficiently high to concentrate on that vector. In the coinertia coordinate system, the first coinertia vector from either set is correlated with an average of 66%.

The results of the RV coefficients show that about a fifth to a third of the variance in one domain is predictable from the other domain. Given that not all acoustic variation has visible consequences, this result is plausible. The amount of variance predictable from the other domain is lower than the figures reported by Yehia et al [7] which could be due to two reasons. Firstly, our video speech parameters concentrate on the lip area while the other study used a larger part of the lower face. Thus, potentially more information was available to capture the visible consequences of acoustic changes in a better way. Secondly, our study is performed on AuE. Native speakers of AuE (some but not all) are reknowned for a certain "lip laziness" and our results could support this notion to some extent. Also, AuE exists in three varieties - broad, general, and cultivated - which are combined in our study. The varieties are known to differ acoustically, mostly in the diphthongs, but they could also differ in the visual consequences of speaking. In either case, this calls for further investigation with a larger sample size and also with speakers from other dialects of English than AuE.

8. Conclusions

We applied various statistical methods to improve our understanding of the relationship between audio and video speech parameters. For the selected parameters and the data for Australian English, we found that 1-1 relationships did not exist between parameters of the two domains but linear combinations of parameters correlated well across the domains. We found that COIA is a more useful method than CANCOR to investigate the relationship of two sets of parameters in the case of small sample sizes compared to the number of parameters because it gives more stable results. The parameters most often contributing strongly to the linear combinations are F1, RMS, mouth height, and relative teeth count. Our results call for further studies with a larger sample size to investigate similarities and differences between the three varieties of AuE. In addition, curve registration as another preprocessing technique should be looked at, so that the common structure of parameter curves from different speakers can be captured in an even better way.

9. References

[1] C. Bregler and Y. Konig, ""eigenlips" for robust speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP'94*, Adelaide, Australia, 1994, vol. II, pp. 669–672.

[2] I. Matthews, T. Cootes, S. Cox, R. Harvey, and J.A. Bangham, "Lipreading using shape, shading and scale," in *Proceedings of the International Conference on Auditory-Visual Speech Processing AVSP'98*, D. Burnham, J. Robert-Ribes, and E. Vatikiotis-Bateson, Eds., Terrigal, Australia, Dec. 1998, pp. 73–78.

[3] E.D. Petajan, *Automatic Lipreading to Enhance Speech Recognition*, Ph.D. thesis, University of Illinois at Urbana-Champaign, 1984.

[4] D.G. Stork and M.E. Hennecke, Eds., *Speechreading by Humans and Machines*, vol. 150 of *NATO ASI Series*, Springer-Verlag, Berlin, Germany, 1996.

[5] A. Adjoudani and C. Benoît, "On the Integration of Auditory and Visual Parameters in an HMM-based ASR," in *Speechreading by Humans and Machines*, D.G. Stork and M.E. Hennecke, Eds., Berlin, Germany, 1996, vol. 150 of *NATO ASI Series*, pp. 461–471, Springer-Verlag.

[6] 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.

[7] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quatitative association of vocal-tract amd facial behaviour," *Speech Communication*, vol. 26, no. 1–2, pp. 23–43, 1998.

[8] 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 *Proc. 8th Australian Int. Conf. on Speech Science and Technology SST2000*, Canberra, Australia, Dec. 2000, pp. 92–97.

[9] R. Goecke, J.B. Millar, A. Zelinsky, and J. Robert-Ribes, "Automatic Extraction of Lip Feature Points," in *Proc. Australian Conf. on Robotics and Automation ACRA2000*, Melbourne, Australia, Aug. 2000, pp. 31–36.

[10] R. Gittins, *Canonical Analysis*, Springer-Verlag, Berlin, Germany, 1985.

[11] S. Dolédec and D. Chessel, "Co-inertia analysis: an alternative method for studying species-environment relationships," *Freshwater Biology*, vol. 31, pp. 277–294, 1994.

[12] L.R. Tucker, "An inter-battery method of factor analysis," *Psychometrika*, vol. 23, pp. 111–136, 1958.

[13] A. Höskuldsson, "Partial least square regression," *Journal of Chemometrics*, vol. 2, pp. 211–228, 1988.

[14] S. Dray, D. Chessel, and J. Thioulouse, "Co-inertia analysis amd the linking of ecological tables," *Ecology*, 2003, (in print).

[15] J. Thioulouse, D. Chessel, S. Dolédec, and J.-M. Olivier, "ADE-4: a multivariate analysis and graphical display software," *Statistics and Computing*, vol. 7, pp. 75–83, 1997.

[16] 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*, vol. 27, pp. 843–856, 1997.