CURRENT TRENDS IN JOINT AUDIO-VIDEO SIGNAL PROCESSING: A REVIEW

Roland Goecke[1,2]

[1]Autonomous System and Sensing Technology, National ICT Australia, Canberra, Australia
[2]Department of Information Engineering, Australian National University, Canberra, Australia
Email: roland.goecke@nicta.com.au

ABSTRACT

Multimodal signal processing has gained a lot of significance in recent years due to advances in computer technology as well as more sophisticated sensors being available. One example is the joint processing of audio and video signals in a variety of applications. This paper serves as a broad introduction to the special session on "Audio-Video Signal Processing and its Applications". The paper reviews current trends and developments in joint audio-video (AV) signal processing and gives an overview of current issues in theory and application in this area. We focus on speech processing, person authentication, and affective sensing as examples. An overview of available AV data corpora is given.

1. INTRODUCTION

The future of human-computer interaction will be quite different from conventional keyboards, mice, and screens. While these have and continue to have their place in certain applications (e.g. office desk environment), the increasingly widespread use of computer systems in everyday circumstances, such as the household or the car, requires a shift towards more human-like means of interaction such as gesture recognition and automatic speech recognition (ASR). Reliability in such systems cannot be achieved by single sensors in single modalities. Rather, multimodal signal processing is a key element to robustness in a large variety of environments, similar to humans employing their different senses in a coordinated way. Human perception is multi-sensory.

In the following, a selection of recent developments and current issues in joint AV signal processing are presented. Section 2 focuses on the area of audio-video speech processing (AVSP). Section 3 reviews current trends in AV person authentication. An overview of existing AV data corpora is given in Section 4. The application of AV signal processing to the new field of affective sensing is described in Section 5. Finally, a summary is provided in Section 6.

National ICT Australia is funded by the Australian Government's Department of Communications, Information Technology, and the Arts and the Australian Research Council through Backing Australia's Ability and the ICT Research Centre of Excellence programs.

2. AUDIO-VIDEO SPEECH PROCESSING

Human perception of spoken language is no exception to the multi-sensory, multimodal perception of the environment. To the naive observer, speech perception is a unimodal process, purely based on the audio modality. However, humans make also use of visual speech information, provided by the facial movements during speech production. Such information contributes not only in noise-degraded conditions or when the listener is hearing-impaired, but also in clear audio conditions (e.g. McGurk effect [1]).

The addition of visual speech information has been shown to improve the recognition rate of audio-only ASR systems, particularly in conditions degraded by acoustic noise [2, 3]. Auditory speech sounds can be categorised into basic units called *phonemes*. Different phonemes serve to distinguish the meaning of one word from another. By analogy, a *viseme* is a member of the set of visually distinguishable articulations. Generally, more phonemes exist than visemes, for example, in Australian English 44 phonemes exist but only 11 visemes [4]. However, even without a 1-1 phoneme-to-viseme mapping, the visual speech information is useful because acoustically ambiguous phonemes fall into different viseme categories and vice versa.

While much research has been done on audio feature extraction, open research issues to the visual front end still remain [3], including robust face and region of interest (ROI) localisation across individual facial appearance variation, different head poses, and illumination changes. Face localisation methods can be classified into two classes: traditional image processing methods [5] (e.g. skin colour segmentation, template matching, edge detection, and thresholding) and statistical approaches [6] (e.g. artificial neural networks). The ROI is typically the lower half of the face.

Once a ROI has been localised, visual speech features must be extracted, which capture visible evidence of speech articulation. Features fall into one of three categories: geometric (shape-based, explicit) features, appearance-based (implicit) features, and combinations of the previous two categories. In geometric features, image processing techniques are used to extract the position of mouth features [7],

such as mouth height, width, area, and visibility of teeth, or to track the lip contours with statistical shape models (e.g. active shape models [8]). A common problem is the reliance on 2D image data, which makes it difficult to adjust for head pose variation, and the reliance on artificial facial markers. [9] proposes a non-invasive, real-time 3D lip tracking algorithm using stereo video to overcome these problems.

Appearance features are based on the assumption that all pixels in the ROI encode visible speech information. Feature vectors are formed directly from the pixel values. Techniques such as PCA or LDA are often employed to reduce the dimensionality of the feature vector [10]. Appearance features avoid the difficulties of explicitly finding facial feature points. Their main disadvantage, apart from the potential high dimensionality, is the inability to cope with a freely moving head, because the systems can only be trained for one specific angle of the face towards the camera (with some tolerance for small head rotations).

Features from these two categories can be combined into joint geometric and appearance feature vectors for the recognition step [3, 11] or form part of a joint statistical model as in active appearance models [3, 12].

2.1. Audio-Video Fusion

Another open research issue concerns the integration of audio and visual speech information. The aim is to develop a joint AV recogniser which outperforms single-modality classifiers. The AV classifier needs to be able to handle input signal streams with varying levels of confidence in the measurements, so that more emphasis can be placed on the input signal that is trusted more at a particular time. For example, in strong acoustic background noise, the AV recogniser should rely more on the visual speech features, and vice versa in the presence of visual noise.

AV fusion methods can be classified into two classes: *feature fusion* and *decision fusion*. The former train a speech recogniser on concatenated A+V feature vectors, using the same techniques as in single modality recognisers [2, 7], also [Lewis, this session]. In decision fusion, separate recognisers are trained for the audio and visual speech features. Their outputs are linearly combined into a joint AV recognition score using the individual likelihoods of the two recognisers [13, 14]. Through appropriate weights it is possible to take the reliability of the extracted features into account. A taxonomy of signal fusion models in human speech perception is studied in [15]. A motor theory model, in which the signals are recoded into speech motor space, explains the experimental evidence best. However, a hybrid model of feature and decision fusion might also explain the data. Recently, another hybrid approach has been proposed [3], in which a discriminant feature extraction [14] is taken as a stream in a multistream-based decision fusion. This hybrid fusion performs better than either feature or decision fusion.

3. AUDIO-VIDEO PERSON AUTHENTICATION

Auditory and visual speech features can also be used in person authentication (or speaker recognition) [16]. An audio-only system is prone to accept a replayed pre-recorded sample of speech as coming from the real person and video-only systems could be attacked by showing a photograph of the real person's face. Using joint AV features is less vulnerable to attacks than single modality authentication.

Each of the modalities can be used separately to give a verification result. In the audio modality, text-independent and text-prompted speaker verification systems have been proposed [16]. The former build a model of the person's whole range of speech sounds, so that the identity can be identified irrespective of the text spoken. Text-prompted systems train models of specific passwords or passphrases spoken by the true person. During verification, the claimant is prompted to utter these words or phrases and their characteristics in the audio speech features are compared with the ones in the model. In the video modality, much research has been done on face recognition [17]. Similar to AVSP, both geometric and appearance features can be used to identify a face. An often applied method is *eigenfaces* [18], which performs PCA on the pixel data and thus falls into the category of appearance features.

Fusion is also an issue in AV person authentication [16], also [Dean, this session]. The AVSP fusion methods can also be applied here. Another issue is the 'liveness' verification of the signal input [Chetty, this issue], as current AV person authentication systems mostly verify a static image of a person's face. Using stereo cameras it is possible to build a 3D head model of the person and to check the correct 3D position of facial feature points [9]. Alternatively, the 'liveness' of the input can be tested by checking the synchrony of the acoustic and visual speech input [19].

4. AUDIO-VIDEO DATA CORPORA

Data corpora form an important tool in AV signal processing. However, partly because of the field still being young, partly because of the time and resources it takes to record an AV data corpus, the number of existing AV data corpora is small compared to the number of audio-only speech data corpora, which have been collected for a long time.

Table 1 lists some existing AV data corpora, their sizes, characteristics, and intended purpose. Included are AVOZES [20], BANCA [21], CUAVE [22], DAVID [23], the proprietary IBM LVCSR AV corpus [14], M2VTS [24], VidTIMIT [25], and XM2VTS [26]. Often these corpora are also used for different purposes than the original intention. Since creating AV data corpora is resource-intensive, it is important that efforts are made at the time of recording to maximise the reusability of corpora [27].

Corpus	Language	Subjects	Video format	Audio kHz / bits	Sessions	Intended purpose	Environment
AVOZES	En	10f, 10m	NTSC DV	48 / 16	1	AV ASR	Studio
BANCA	En/Fr/Sp/It	$4\times$(26f, 26m)	PAL DV	32 / 16, 12	12	Authentication	Different backgr.
CUAVE	En	17f, 19m + 20 pairs	NTSC DV, MPEG-2	44 / 16	1	AV ASR	Subjects moving, pairs of subjects
DAVID	En	SC1: 2f, 5m SC2: 61f, 62m SC3: 2f, 3m SC4: 61f, 62m			1 5 1 5	Face segmentation AV ASR, Authent. Compr., synthesis AV ASR, Authent.	Variable visual scenarios, some speakers with highlighted lips
IBM AV	En	290	NTSC, MPEG-2	22 / 16	1	AV ASR	Studio
M2VTS	Fr	37	Hi8, CIF	48 / 16	5	Authent., AV ASR	Const. conditions
VidTIMIT	En	19f, 24m	PAL DV, JPEG	32 / 16	3	AV ASR, Authent.	Noisy office
XM2VTS	Fr	295	PAL DV	32 / 16	4	Authent., AV ASR	Const. conditions

Table 1: An overview of some existing AV data corpora and their sizes, characteristics, and intended purpose.

5. APPLICATION TO AFFECTIVE SENSING

Affective computing aims to develop computer systems capable of sensing a person's affective state and taking that state into account in their own actions [28]. Application areas include patient monitoring in health care applications, operator monitoring in safety-critical environments (e.g. air-traffic control, driver assistance systems), and educational software. As single modality sensors can often give ambiguous results, affective sensing uses multimodal sensors such as video cameras, microphones, and physiological sensors to infer a person's affective state.

Audio and video signals play an important role in affective sensing. A prosody analysis of the audio signal can deliver information about the affective state [29]. Prosody investigations typically analyse F_0, utterance intensity, high-frequency energy, and speech rate. Similarly, facial expressions can also help to infer the affective state [30]. Some of the techniques used in visual speech feature extraction, e.g. [12], can also be used for facial expression tracking. The positional information of facial feature points (e.g. eyebrows, lips, and cheeks) then forms the input for the affective state classifier. Recent advances and lower prices in infrared (IR) imaging technology also offer a completely new look at facial video data for affective sensing. Video IR technology now allows quantitative thermal imaging at an accuracy of 0.1K and can thus monitor changes to the facial bloodflow, which are at least partly a result of changes in affective state.

Similar to AVSP, the input signals from the various sensors must be integrated in a robust and meaningful way that takes the different reliabilities of the sensors into account. Research in this area is still in its infancy, but similar fusion approaches as for AVSP seem a natural choice.

6. SUMMARY

A broad overview of developments and current issues in AV signal processing as an example of multimodal signal processing has been given. The examples of speech processing, person authentication, and affective sensing have been presented. Other areas, not discussed here, are the translation of sign language to and from spoken language [Holden, this session], music perception, and joint AV gestures.

7. REFERENCES

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, Dec. 1976.

[2] T. Chen, "Audiovisual Speech Processing," *IEEE Signal Proc. Mag.*, vol. 18, no. 1, pp. 9–21, Jan. 2001.

[3] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent Advances in the Automatic Recognition of Audiovisual Speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, Sept. 2003.

[4] G.L. Plant and J.J. Macrae, "Visual Perception of Australian Consonants, Vowels and Diphthongs," *Austral. Teacher of the Deaf*, vol. 18, pp. 46–50, July 1977.

[5] H.P. Graf, E. Cosatto, D. Gibbon, M. Kocheisen, and E. Petajan, "Multi-Modal System for Locating Heads and Faces," in *Proc. IEEE FG'96*, Killington, USA, Oct. 1996, pp. 88–93.

[6] H.A. Rowley, S. Baluja, and T. Kanade, "Neural Network-Based Face Detection," *IEEE Trans. Patt. Anal. Mach. Int.*, vol. 20, no. 1, pp. 23–38, Jan. 1998.

[7] A. Adjoudani and C. Benoît, "On the Integration of Auditory and Visual Parameters in an HMM-based ASR," in *Speechreading by Humans and Machines*, 1996, vol. 150 of *NATO ASI Series*, pp. 461–471.

[8] J. Luettin, N.A. Thacker, and S.W. Beet, "Active Shape Models for Visual Speech Feature Extraction," in *Speechreading by Humans and Machines*, 1996, vol. 150 of *NATO ASI Series*, pp. 383–390.

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

[10] G. Potamianos, J. Luettin, and C. Neti, "Hierarchical Discriminant Features for Audio-Visual LVCSR," in *Pro. IEEE ICASSP'01*, Salt Lake City, USA, May 2001, vol. 1, pp. 165–168.

[11] M.T. Chan, "HMM-Based Audio-Visual Speech Recognition Integrating Geometric- and Appearance-Based Visual Features," in *Proc. IEEE MMSP-2001*, Cannes, France, Oct. 2001, pp. 9–14.

[12] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active Appearance Models," in *Proc. ECCV'98*, Freiburg, Germany, June 1998, vol. 2, pp. 484–498.

[13] S. Dupont and J. Luettin, "Audio-Visual Speech Modeling for Continuous Speech Recognition," *IEEE Trans. Multim.*, vol. 2, no. 3, pp. 141–151, Sept. 2000.

[14] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-Visual Speech Recognition," Wksh. report, Johns Hopkins Univ., Baltimore, USA, 2000.

[15] J. Robert-Ribes, M. Piquemal, J.-L. Schwartz, and P. Escudier, "Exploiting Sensor Fusion Architectures and Stimuli Complementarity in AV Speech Recognition," in *Speechreading by Humans and Machines*, 1996, vol. 150 of *NATO ASI Series*, pp. 193–210.

[16] J.-L. Dugelay, J.-C. Junqua, C. Kotropoulos, R. Kuhn, F. Perronnin, and I. Pitas, "Recent advances in biometric person authentication," in *Proc. IEEE ICASSP'02*, Orlando, USA, May 2002, vol. 4, pp. 4060–4063.

[17] R. Chellappa, C.L. Wilson, and S. Sirohey, "Human and Machine Recognition of Faces: A Survey," *Proc. of the IEEE*, vol. 83, no. 5, pp. 705–741, May 1995.

[18] M. Turk and A. Pentland, "Eigenfaces for recognition," in *Proc. CVPR-91*, 1991, pp. 1–2.

[19] G. Chetty and M. Wagner, ""Liveness" Verification in Audio-Video Authentication," in *Proc. ICSLP2004*, Jeju, Korea, Oct. 2004, vol. III, pp. 2509–2512.

[20] R. Goecke and B. Millar, "The Audio-Video Australian English Speech Data Corpus AVOZES," in *Proc. ICSLP2004*, Jeju, Korea, 2004, vol. III, pp. 2525–2528.

[21] E. Bailly-Baillire, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariethoz, J. Matas, K. Messer, V. Popovici, F. Boree, B. Ruiz, and J.-P. Thiran, "The BANCA Database and Evaluation Protocol," in *Proc. AVBPA2003*, Guildford, UK, June 2003, pp. 625–638.

[22] E.K. Patterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy, "CUAVE: A New Audio-Visual Database for Multimodal Human-Computer Interface Research," in *Proc. IEEE ICASSP2002*, Orlando, USA, 2002, vol. 2, pp. 2017–2020.

[23] C.C. Chibelushi, S. Gandon, J.S. Mason, F. Deravi, and D. Johnston, "Design Issues for a Digital Integrated Audio-Visual Database," in *IEE Colloq. Integrated AV Proc. for Rec., Synth. & Comm.*, London, UK, Digest Ref. No. 1996/213, 1996, pp. 7/1–7/7.

[24] K. Messer, J. Matas, and J. Kittler, "Acquisition of a large database for biometric identity verification," in *Proc. BIOSIGNAL 98*, Brno, Czech Republic, 1998, pp. 70–72.

[25] C. Sanderson and K.K. Paliwal, "Fast Features for Face Authentication under Illumination Direction Changes," *Pattern Recognition Letters*, vol. 24, no. 14, pp. 2409–2419, Oct. 2003.

[26] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The Extended M2VTS Database," in *Proc. AVBPA'99*, Washington (DC), USA, 1999, pp. 72–77.

[27] B. Millar, M. Wagner, and R. Goecke, "Aspects of Speaking-Face Data Corpus Design Methodology," in *Proc. ICSLP2004*, Jeju, Korea, 2004, vol. II, pp. 1157–1160.

[28] R.W. Picard, *Affective Computing*, MIT Press, Cambridge (MA), USA, 1997.

[29] L. ten Bosch, "Emotions, speech and the ASR framework," *Speech Communication*, vol. 40, no. 1–2, pp. 213–225, Apr. 2003.

[30] P. Ekman and E.L. Rosenberg, *What the Face Reveals*, Series in Affective Science. Oxford University Press, Oxford, UK, 1997.