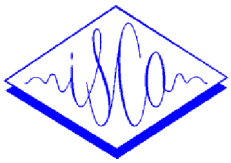

Auditory-Visual Speech Processing
2005 (AVSP'05)
British Columbia, Canada
July 24-27, 2005

3D LIP TRACKING AND CO-INERTIA ANALYSIS FOR IMPROVED ROBUSTNESS OF AUDIO-VIDEO AUTOMATIC SPEECH RECOGNITION

Roland Goecke[1,2]

[1]Autonomous System and Sensing Technologies, National ICT Australia, Canberra, Australia
[2]Department of Information Engineering, Research School of Information Sciences and Engineering, Australian National University, Canberra, Australia

Email: roland.goecke@nicta.com.au

ABSTRACT

Multimodality is a key issue in robust human-computer interaction. The joint use of audio and video speech variables has been shown to improve the performance of automatic speech recognition (ASR) systems. However, robust methods in particular for the real-time extraction of video speech features are still an open research area. This paper addresses the robustness issue of audio-video (AV) ASR systems by exploring a real-time 3D lip tracking algorithm based on stereo vision and by investigating how learned statistical relationships between the sets of audio and video speech variables can be employed in AV ASR systems. The 3D lip tracking algorithm combines colour information from each cameras' images with knowledge about the structure of the mouth region for different degrees of mouth openness. By using a calibrated stereo camera system, 3D coordinates of facial features can be recovered, so that the visual speech variable measurements become independent from the head pose. Multivariate statistical analyses enable the analysis of relationships between sets of variables. Co-inertia analysis is a relatively new method and has not yet been widely used in AVSP research. Its advantage is its superior numerical stability compared to other multivariate methods in the case of small sample size. Initial results are presented, which show how 3D video speech information and learned statistical relationships between audio and video speech variables can improve the performance of AV ASR systems.

1. INTRODUCTION

It is widely accepted these days that the addition of visual speech information improves the recognition rate of an ASR system significantly, in particular in environments with acoustic noise [1]. The series of AVSP conferences is testament to the development of our understanding of the underlying processes in AVSP in humans and of the application of that knowledge to AV ASR systems. However, despite the advances in recent years, we still need to improve the robustness of such systems before they are ready to be used in everyday environments, such as homes and cars. While much research has been done on audio speech signal processing, the quest for robust and real-time methods for automatically processing visual speech information is still ongoing research. In this paper, a real-time 3D lip tracking algorithm based on stereo vision is explored. Given the decreasing cost for cameras, stereo camera systems are feasible in many situations these days.

Another open research area is the issue of how variables from different modalities are related to each other. Questions such as which variables or combination of variables in each modality are related to variables or combination of variables in the other modality arise. The machine learning approach assumes that, given enough training samples, the relationships can be learned from the data. In this paper, the multivariate statistical method of co-inertia analysis is investigated and applied to AVSP [2]. It has better numerical stability properties than other methods, in particular for small sample sizes.

The rest of this paper is structured as follows. Section 2 describes the details of the 3D lip tracking algorithm. Section 3 gives an overview of the AV speech data corpus used for the experiments. In Section 4, co-inertia analysis is described. Section 5 outlines the experiments and presents the experimental results. Finally, Section 6 presents the conclusions and outlines future work.

2. STEREO VISION 3D LIP TRACKING

A calibrated stereo camera system allows the reconstruction of 3D coordinates, because depth information (distance from cameras to object) can be recovered from the stereo disparity. Applied to

facial feature tracking, and in this instance lip tracking, it has the advantage that 3D coordinates of lip feature points can be measured irrespectively of the head pose. Conventional single camera systems measure only 2D image coordinates without separating head pose-related effects from facial movements (Figure 1).

Figure 1. Examples from the AVOZES data corpus showing different head poses which would impact on 2D lip tracking.

Our real-time 3D lip tracker builds on top of a real-time stereo vision face tracking system, which enables non-intrusive tracking of facial features. No facial markers or special make-up are required, yet the system achieves a high degree of accuracy. These properties are highly desirable in an AV ASR system, because artificial tracking aids pose the risk of inhibiting the speaker from speaking naturally.

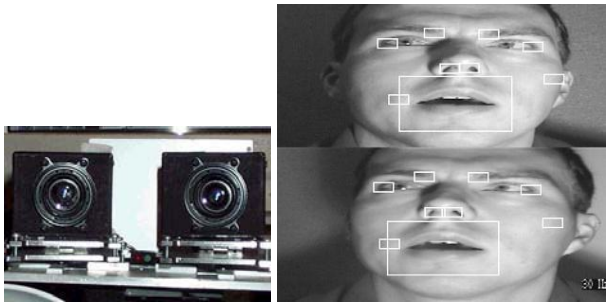

Figure 2. Left: Stereo camera pair. Right: Stereo camera output multiplexed into one video frame at half the vertical resolution. Small rectangles show the templates used for head tracking, while the large rectangle shows the mouth region used in the 3D lip tracking algorithm.

The head tracking system is based on template matching using normalised cross-correlation and is able to track the person's movements at video frame rate. The system consists of two calibrated standard, colour analog NTSC video cameras (Figure 2, left). The camera outputs are multiplexed at half the vertical resolution into a single 512x480 image (Figure 2, right). Details can be found in [3].

The 3D lip tracking algorithm is applied to the mouth regions in each camera's image which are automatically determined during the head tracking based on the head pose (Figure 2, right). The algorithm combines colour information from the images with knowledge about the structure of the mouth region for different degrees of mouth openness. For example, in an open mouth, we often expect to see teeth, so we can specifically look for them to improve the robustness of the lip tracking.

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. 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 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 have difficulty in coping with this lack of contrast, if employed on tracking the outer lip contour. Furthermore, facial hair affects the visibility of the outer lip contour. Given the different appearance of the oral cavity, the inner lip contour does not suffer from these problems.

2.1 A THREE-STEP ALGORITHM

The lip tracking algorithm is a three-step process. 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 is called solving the point correspondence problem and incorrectly identified correspondences lead to incorrect 3D coordinates. As the mouth shape is changing rapidly during speech, static methods such as template matching do not work well. Therefore, a combination of colour information and structural knowledge is used.

The first step determines the general degree of mouth openness. The lip tracking algorithm must be able to handle mouth shapes during speech ranging from a completely closed mouth to a wide open mouth. No single image processing technique would give good results for all possible mouth shapes. By pre-classifying mouth shapes into one of three

categories based on mouth openness (closed, partially open, wide open), specific techniques individually targeted at each category can be applied to give better results.

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 result is then used in the third and final step, in which the positions of the lip midpoints are determined. Figure 3 shows the detailed steps of the algorithm.

3. AVOZES DATA CORPUS

The AVOZES (*A*udio-*V*ideo *OZ*stralian *E*nglish *S*peech) data corpus is used in our experimental work [4]. AVOZES is the first audio-video speech data corpus for Australian English (AuE) and it is available at [5]. The data corpus is also novel in that the stereo camera system described in Section 2 was used for the video recordings. To the best of our knowledge, no other AV speech data corpus with stereo video is available.

The design of the AVOZES data corpus follows a modular framework [6] proposed in which is also in accordance with the design methodology proposed in [7]. A modular approach, where each module contains certain sequences, allows for extensibility in terms of the various factors that need to be addressed in corpus design [6][8]. In summary, the framework suggests that any AV speech data corpus contains three mandatory modules as a minimum, which cover the recording setup without and with speakers, as well as the actual speech material sequences which should contain the phonemes and visemes of a language. Additional optional modules can be added to cover specific issues, e.g. different view angles, different levels of illumination or acoustic noise. The AVOZES data corpus has a total of six modules - one general module and five speaker-specific modules. These six modules are:

- the scene without any speaker;
- the scene with speaker, head turning;
- 'calibration sequences' exhibiting horizontal and vertical lip movements during speech production;
- CVC- and VCV-words in a carrier phrase covering the phonemes and visemes of AuE;
- the digits "0"-"9" in a constant carrier phrase; and
- three sentences as examples of continuous speech.

The CVC- and VCV-words form the core part of the corpus and were also used for the experiments presented here. Each utterance was recorded once. Recordings were made in clean audio conditions.

AVOZES currently contains recordings made from 20 native speakers of AuE (10 female + 10 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 tentatively classified into the three speech varieties of AuE (broad, general, cultivated) 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 accent varieties, it should be noted that individual groups are not gender balanced, and that their size is small for statistical analyses on an individual group basis. It is also worthwhile to remember that the accent varieties are not discrete entities, but rather span a continuum of accent variation, so that some classifications represent a best estimate.

4. CO-INERTIA ANALYSIS

Multivariate statistical analyses deal with data containing observations of p variables measured on a set of n objects. In the AVSP, the measured audio and video speech parameters form the variables, which are measured on the set of phonemes. Some multivariate analyses, such as canonical correlation analysis, can suffer from collinearity in the sample data particularly for small sample sizes, leading to numerical instabilities in the results. To overcome these problems, coinertia analysis (COIA) was developed by Dolédec and Chessel [2], in which the

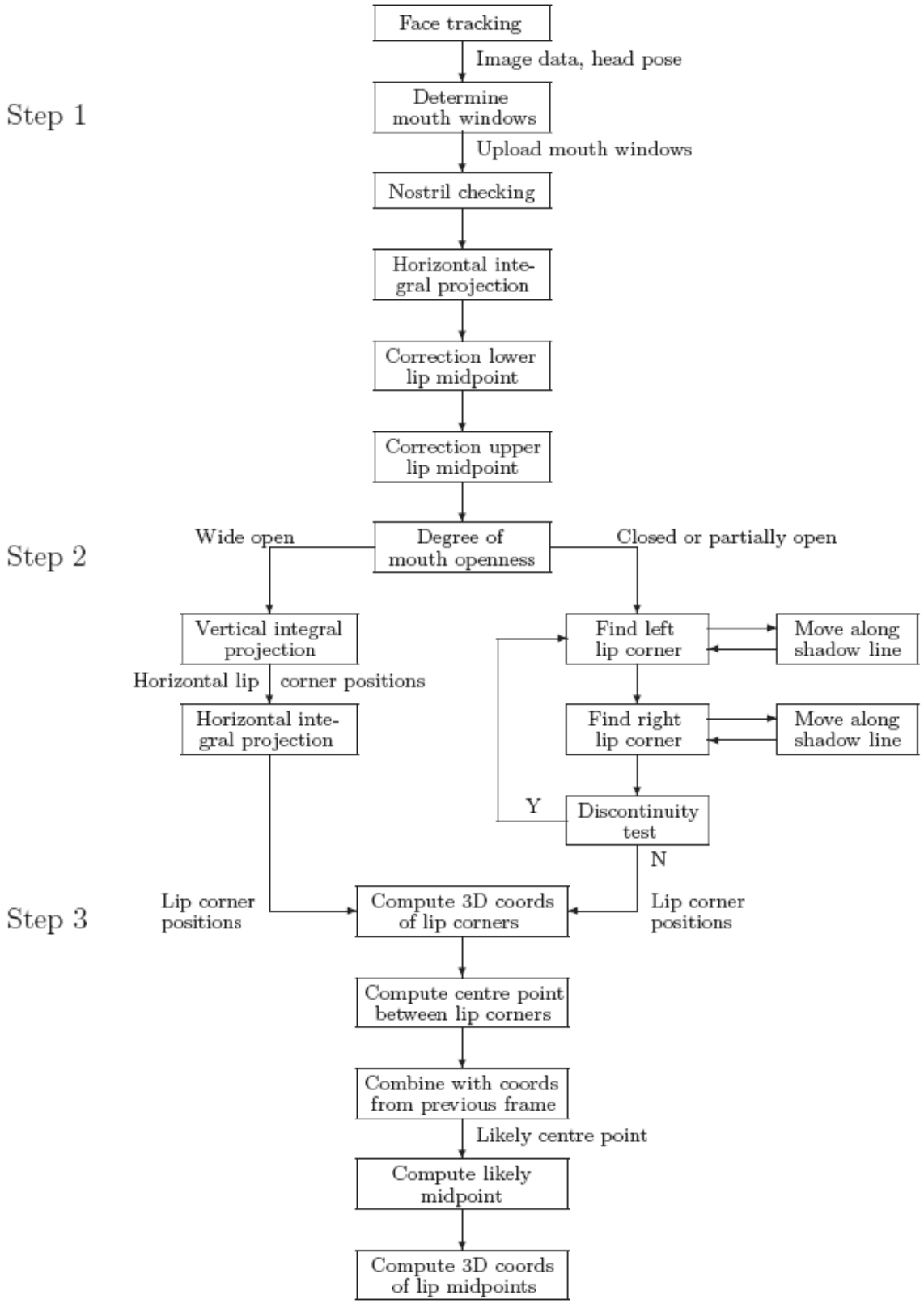

Figure 3. The three-step 3D lip tracking algorithm.

number of parameters relative to the sample size does not affect the accuracy and stability of the results. COIA has so far been used mostly in ecological studies for some years, but we reported on initial results on using it for analysing statistical relationships between audio and video speech variables in [9]. The method is related to other multivariate statistical analyses such as canonical correspondence analysis, redundancy analysis, and canonical correlation analysis [10]. COIA can also

be coupled easily with other statistical methods, e.g. principal component analysis. First, these methods are performed on the data of the two domains separately, and then a COIA follows. In fact, it can be shown that COIA is a generalisation of many multivariate methods [11].

COIA uses the term 'inertia' as a synonym for variance. The method rotates the data to a new coordinate system and the new variables are linear combinations of the variables in each set. COIA maximises the co-inertia (or covariance) which can be decomposed as

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

COIA finds a compromise between the correlation (corr(*A,V*)), the variance in the audio speech variable set *var*(*A*), and the variance in the video speech variable set *var*(*V*). It aims to find orthogonal vectors ("co-inertia axes") in the two sets which maximise the co-inertia value. The number of axes is equal to the rank of the covariance matrix.

COIA provides a number of measures for the analysis of the relationships between two sets of variables. The co-inertia value is a global measure of the co-structure in the two sets. If the value is high, the two sets of variables vary accordingly (or inversely), and if the value is low, the sets vary independently. The correlation value gives a measure of the correlation between the co-inertia axes of both sets of variables. Furthermore, one can project the variance onto the new axes of each set and then compare the projected variance of the separate analyses with the variance from the COIA [2]. 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 set of variables that is taken by each co-inertia axis. In addition, COIA computes the weights (coefficients) of the variables in the linear combinations of each set. They show which variables contribute to the common structure of the two sets. These weights are numerically more stable than the weights that can be obtained from a canonical correlation [2]. An overall value of relatedness of the two modalities is given by the so-called RV coefficient [12].

5. EXPERIMENTS

The goal of the experiments was to see how the 3D lip tracking algorithm and the results of the co-inertia analysis could be employed in an AV ASR system. Data from the AVOZES data corpus is used for these experiments. The initial results presented here are for a subset of the AVOZES data. This subset consists of the 10 female speakers and their CVC-word utterances, i.e. the words containing the 18 vocalic phonemes of AuE. No noise was added.

As a baseline, an audio-only ASR system is used. Using the HTK toolkit, a 3-state left-right HMM with no skips is built for each monophone. 13 MFCC parameters and their delta and delta-delta parameters were used as audio speech variables. From the monophone HMMs, context-dependent triphone HMMs were built by simply cloning the monophone HMMs and re-estimating them using triphone transcriptions. In the experiments, the leave-one-out method was used. For each of the speakers, the recogniser was trained with data from the other nine speakers and then tested on the left-out speaker's data. The word error rate (WER) results are shown in Table 1. The relatively WER for some speakers can be explained by the limited training data. Confusions typically occurred between short and long realisations of the same sound (e.g. /æ/ vs. /ə:/), and among low to mid-low front to central vocalic phonemes (e.g. /ɛ/ vs. /æ/).

Speaker	Audio-Only	AV	AV + COIA
f1	27.8	22.0	21.5
f2	27.8	20.3	19.8
f3	27.8	21.3	20.6
f4	33.3	23.4	22.8
f5	11.1	7.8	7.5
f6	16.7	13.6	13.2
f7	11.1	8.1	7.9
f8	33.3	23.9	23.2
f9	38.9	28.3	27.9
f10	33.3	25.0	24.6

Table 1. Experimental results: WER in % for the audio-only case, the AV case, and for the AV plus co-inertia results case.

In the joint AV recognizer, the mouth width, mouth height, protrusion of upper lip, protrusion of lower lip, and relative teeth count [9] were added as video speech variables, including their delta and delta-delta variables. A feature fusion approach was taken here, i.e. the values of the video speech variables were added to the audio feature vectors. Then, the HMMs were re-trained and again the recognizer was tested using the approach as for the audio-only case. The results are shown in Table 1. The inclusion of visual speech information improves the WER by about 20-30% relative. These results are similar to results reported by others [1]. However, a future direction will be to implement a 2D lip tracker, for example by using AAMs [13], and to compare the results achieved with 2D and 3D lip trackers.

In the final experiment, we want to see if the addition of information from COIA could further improve the performance. The timings of the central phonemes in the CVC-words and VCV-words were hand-labelled and audio and video speech variables extracted as before. Next, COIA was performed on these two variable sets separately for each phoneme (see [9] for some COIA results). Of the various values given by COIA, the co-inertia value was chosen for the experiments here. When re-training the monophone HMMs, this value was added to the previously used set of AV variables based on the word transcriptions. Then, the triphone HMMs were re-trained. For the recogniser, the recorded sequences underwent a COIA and the results stored for each time step, so that the co-inertia value would be at hand during recognition. Table 1 again shows the results. The WER improved on average by about 1.5-2% relative. It needs to be further tested, if adding more values from COIA can further improve the performance. We hope to present some further results at the workshop.

6. CONCLUSIONS

The details of a real-time 3D lip tracking algorithm have been presented. This algorithm is robust to head pose variations and is completely non-invasive, i.e. no artificial markers or make-up is required to track the 3D shape of the lips. Co-inertia analysis has been used to explore the statistical relationships between the sets of audio and video speech variables. Initial results have been presented which show that the word error rate improves by about 20-30% relative when the visual speech information using the 3D coordinates of lip feature points is included. To get a better understanding how these results compare to a single camera system, we will implement a 2D lip tracking algorithm. Further including learned relationships between audio and video speech variables from COIA improved the performance by up to 2% relative. So far, we have only performed experiments on the subset of vocalic phonemes. It needs to be seen, what the results are for consonantal phonemes. Some consonants show strong visible speech articulation, so that improved results are expected for these. We hope to present further results at the workshop.

7. ACKNOWLEDGEMENT

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.

8. REFERENCES

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

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

[3]. Newman, R., Matsumoto, Y., Rougeaux, S., and Zelinsky, A. "Real-time stereo tracking for head pose and gaze estimation", in *Proc. of Automatic Face and* Gesture Recognition FG2000, Grenoble, France, 2000, pp. 122-128.

[4]. Goecke, R., and Millar, J.B., "The Audio-Video Australian English Speech Data Corpus AVOZES," in *Proc. 8th Int. Conf. Spoken Language Processing ICSLP2004*, Jeju, Korea, 2004, vol. III, pp. 2525–2528.

[5]. Goecke, R. "The AVOZES Data Corpus", http://users.rsise.anu.edu.au/~roland/.

[6]. Goecke, R., Tran, Q., Millar, J.B., Zelinsky, A., and J. Robert-Ribes, J., "Validation of an Automatic Lip-Tracking Algorithm and Design of a Database for Audio-Video Speech Processing", *in Proc. 8th Australian Int. Conf. Speech Science and Technology SST2000*, Canberra, Australia, 2000, pp. 92–97.

[7]. Millar, J.B., Wagner, M., and Goecke, R., "Aspects of Speaking-Face Data Corpus Design Methodology", in *Proc. 8th Int. Conf. Spoken Language Processing ICSLP2004*, Volume II, Jeju, Korea, 2004, pp. 1157–1160.

[8]. 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, Digest Reference Number 1996/213, London, UK, November 1996, pp. 7/1–7/7.

[9]. Goecke, R., and Millar, J.B., "Statistical Analysis of the Relationship between Audio and Video Speech Parameters for Australian English," in *Proc. ISCA Tutorial and Research Workshop on Audio Visual Speech Processing AVSP2003*, St Jorioz, France, Sept. 2003, pp. 133–138.

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

[11]. Dray, S., Chessel, D., and Thioulouse, J., "Co-inertia analysis and the linking of ecological data tables," *Ecology*, vol. 84, 2003, pp. 3078–3089.

[12]. Heo, M., and Gabriel, K.R., "A permutation test of association between configurations by means of the RV coefficient", *Communications in Statistics - Simulation and Computation*, vol. 27, 1997, pp. 843–856.

[13]. Matthews, I., Cootes, T.F., Bangham, J.A., Cox, S., and Harvey, R., *Extraction of Visual Features for Lipreading*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(2):198–213, February 2002.