Maximising Audiovisual Correlation with Automatic Lip Tracking and Vowel Based Segmentation

Andrew Abel[1], Amir Hussain[1], Quoc-Dinh Nguyen[2], Fabien Ringeval[2],
Mohamed Chetouani[2], and Maurice Milgram[2]

[1] Dept. of Computing Science, University of Stirling, Scotland, UK,
[2] Institute of Intelligent Systems and Robotics, University Pierre and Marie Curie
(Paris 6) 4 Place Jussieu, Paris, France
aka@cs.stir.ac.uk, ahu@cs.stir.ac.uk, quocdinh.nguyen@upmc.fr,
fabien.ringeval@upmc.fr, mohamed.chetouani@upmc.fr,maurice.milgram@upmc.fr

Abstract. In recent years, the established link between the various human communication production domains has become more widely utilised in the field of speech processing. In this work, a state of the art Semi Adaptive Appearance Model (SAAM) approach developed by the authors is used for automatic lip tracking, and an adapted version of our vowel based speech segmentation system is employed to automatically segment speech. Canonical Correlation Analysis (CCA) on segmented and non segmented data in a range of noisy speech environments finds that segmented speech has a significantly better audiovisual correlation, demonstrating the feasibility of our techniques for further development as part of a proposed audiovisual speech enhancement system.

1 Introduction

The multimodal nature of both human speech production and perception is well established. The relationship between audio and visual speech has been investigated in literature, demonstrating that speech acoustics can be estimated using visual information. Almajai et al. [1] recently investigated correlation between audio and visual features using Multiple Linear Regression (MLR), and expanded upon this to devise a visually derived Wiener filter for speech enhancement [2]. Sargin [3] also performed correlation analysis of multimodal speech, but used Canonical Correlation Analysis (CCA) [4] as part of a speaker identification task. However, one aspect of speech processing that has not been researched much is the detailed analysis of multimodal correlation with noisy speech. Since pioneering work by Girin et al. [5], which uses an Independent Component Analysis (ICA) approach to estimate "cleaned" spectral parameters, little additional correlation work in noisy environments has been published.

The most relevant visual speech information is contained in the lip region, and so it is desirable to focus on lip features alone. Speakers very rarely remain motionless when talking, and so it is impractical to extract visual features from the same location of a frame for a whole speech sequence. Visual feature extraction by manually labelling all frames in an image sequence is unsuitable for an integrated speech processing system. A novel Semi Adaptive Appearance Model (SAAM) lip tracking approach has been developed by the authors, and is used here to automatically track lip features in image sequences.

In addition to our automated lip tracking approach, speech segmentation is utilised. Almajai et al. found improved audiovisual correlation when individual phonemes within a sentence were processed rather than a whole sentence, implying that a more nuanced approach to audiovisual data can improve correlation. Therefore, the authors have adapted a vowel based speech segmentation system [6], which detects vowels in speech and segments the data accordingly.

In this paper, we combine our automatic vowel based speech segmentation and lip tracker to perform CCA on speech from the VidTIMIT Corpus [7] and maximise correlation between visual and audio speech signals. The performance of CCA on segmented vowels and automatically tracked images is compared to that of complete spoken sentences [3]. Additionally, the performance of this approach in a range of noisy environments is assessed by adding noise to speech from the corpus. The results produced with our tracking and segmentation techniques demonstrate the feasibility of extending this initial work and integrating these approaches into a future proposed audiovisual speech enhancement system.

The rest of this paper is divided up as follows. Section 2 describes our feature extraction methods, firstly describing the vowel segmentation approach, and then the automated lip tracking method. Section 3 outlines the CCA used on audiovisual data in this work, with results discussed in section 4. Finally, section 5 summarises this paper and outlines future research directions.

2 Multimodal Feature Extraction

2.1 Audio Feature Extraction

In this work, our original vowel detection method [6] has been adapted to deal with noisy speech. Our new system is completely automatic and is both speaker and language independent. According to the source-filter model of speech production, vowels are characterised by a particular spectral envelope qualified as formantic, which reveals the positions of the formants. The detection of vowel-like segments is based on the characterisation of the spectral envelope. A spectral measure termed the "Reduce Energy Cumulating" (REC) function (equation 1) has been proposed for vowel spectrum characterisation [8] by comparing the energy computed from Mel bank filters. The speech is segmented into overlapping frames, N spectral energy values E_i are extracted for each k frame, and those that are higher than their respective mean value $\bar{E}$ are cumulated and weighted by the energy ratio from low E_{LF} and total E_T frequency bands.

$$Rec\,(k) = \frac{E_{LF}\,(k)}{E_T\,(k)} \sum_{i=1}^{N} \left(E_i(k) - \bar{E}(k)\right)^+ \tag{1}$$

For a given sentence, peak detection on the smoothed REC curve (with a Simple Moving Average filter) allows vocalic nucleus detection. To reject low energy peaks, which can be either related to spectral noises or low energy vowels, only those higher than the half mean of the REC values are kept. Due to local spectral variability, successive detected peaks from the REC curve can be very close. If two vowels are detected closer than 150ms, only the single highest peak is kept. Segmental borders are then found depending on the REC curve's local peak configuration: first coming REC values located at the half below their corresponding peaks are used for delimiting the localisation of the detected nucleus. Borders are set at 100ms away from the nucleus to avoid overlapping vowels.

Sentences from VidTIMIT are segmented into 32ms frames with an overlap ratio equal to 50%. 22 MFCC coefficients (Mel Frequency Cepstral Coefficients) are computed on these frames, producing MFCC matrices for full sentences. To extract vowel only data, vowel localisation results are used to group vowel only segments together into a single MFCC file, providing the feature vector f_y.

2.2 Visual Feature Extraction with SAAM

Visual lip features are extracted by using our newly developed semi adaptive appearance models (SAAMs) [9]. Lip tracking essentially deals with non-stationary data, as the appearance of a target object may alter drastically over time due to factors like pose variation and illumination changes. Our lip tracking framework is based on Adaptive Appearance Models (AAMs) [10], which allow us to update the mean and Eigen vectors of d-dimensional observation vectors $x \in R^d$. First, we extend the AAMs by inserting a supervisor model [11] that verifies AAM performance at each frame in the sequence, by using a Support Vector Machine (SVM) to filter the AAM result for an individual frame, as shown in (2) next:

$$f\left(x\right) = sgn\left(\sum_i \alpha_i y_i K\left(x_i, x\right) + b\right) \tag{2}$$

Where $f(x) \in \{-1, 1\}$ signifies whether x is a good or bad result. α, b are trained offline with the SVM in [12], $K\left(.\right)$ is the Gaussian kernel function and x_i,x are trained and observation vectors respectively. Each y_i represents the desired output of each example x_i from the offline training dataset.

Secondly, shape models are constructed to allow our SAAM to track feature points in video sequences. To model deformation, we form a shape model: $S^\circ = S^\circ + P_s b$ where $S^\circ = (x_1^\circ, y_1^\circ, \ldots, x_n^\circ, y_n^\circ)$ is a normalized shape and n represents a number of feature points. To track these, it is sufficient to find the parameters $p = [b_1, \ldots, T_x, T_y, \theta, s]$ where b_i is the coefficient to deform S° and T_x, T_y, θ, s represent translations, rotation and scale parameters.

To track a target object, the aim is to maximize the cost function given in (3) as follows:

$$p^* = \arg\max_p (d_e) \tag{3}$$

Where d_e is a negative exponential of projection error between x_t and the Principal Component Analysis (PCA) subspace created by earlier observations, defined by equation 4 as follows:

$$d_e = exp\left(-\left\|\left(x_t - \bar{x}\right) - UU^T\left(x_t - \bar{x}\right)\right\|^2\right) \tag{4}$$

Note that the distance d_e is a Gaussian distribution, with Eigen vectors U and mean $\bar{x}$, $d_e = p\left(x_t|p\right) \propto N\left(x_t; \bar{x}, UU^T + \epsilon I\right)$ as $\epsilon \rightarrow 0$, and the inverse matrix can be solved by applying the Woodbury formula[11], given in equation 4 as follows:

$$\left(UU + \epsilon I\right)^{-1} = \epsilon^{-1}\left(I - \left(1 + \epsilon\right)^{-1} UU^T\right) \tag{5}$$

The optimal parameter p^* is found with a number of iterations. Here, we use empirical gradient, since we evaluate the cost function in the neighbourhood of the current parameter vector value. Our tracking algorithm works as follows:

1. Manually locate target object in the first frame (t=1). Eigen vectors U are initialized as empty. Our tracker initially works as a template based tracker.
2. At the next frame, find the optimal parameters $p^* = argmax\left(\left\{d_e\left(p_i^k *\right)\right\}\right)$ over a number of iterations:
   - For each parameter p_i.
   - For each Δp and $k \in \{-1, 1\}$, compute $p_i^k\left(p_1, \ldots, p_i + k\Delta p, \ldots, p_{k+4}\right)$
   - Compute $i^* = max\left\{d_e\left(p_i^k\right)\right\}$
   - Do $p \leftarrow p_i *$, store $d_e\left(p_i^k *\right)$
3. Check the observation vector: $x = x\left(W\left(S_e^{'}, p^*\right)^{-1}\right)$
   where W is a transformation matrix, with result estimation phase as shown in equation 2
4. If $f\left(x\right) = 1$, this signifies a good result to add to the model. When the desired number of new images has been accumulated, perform an incremental update.
5. Return to step 2.

This technique is used to find the 2D-DCT vector $f_x = 2D - DCT\left(x\right)$. VidTIMIT contains a number of image sequences of sentences recorded at 25 fps. The first 30 2D-DCT components of each image are vectorised in a zigzag order to produce the vector for a single frame in an image sequence. The resulting 2D-DCT sequence is then interpolated to match the equivalent MFCC matrices.

3 Canonical Correlation Analysis

In this paper, CCA [4] is used to analyse the linear relationships between multidimensional audio and visual speech variables by attempting to identify a basis vector for each variable that then produces a diagonal correlation matrix. CCA maximises the diagonal elements of the correlation matrix. The main difference between CCA and other forms of correlation analysis is the independence of analysis from the coordinate system describing the variables. Ordinary correlation analysis can produce different results depending on the coordinate system used, whereas CCA finds the optimal coordinate system.

Let f_x and f_y represent multidimensional visual and audio signal variables, with projection matrices u_x and u_y, calculated using the QR Decomposition method [11], that mutually maximises the projections of f_y and f_x onto their respective basis vectors. When the linear combinations $\hat{f_x} = u_x^T f_x$ and $\hat{f_y} = u_y^T f_y$ are considered, we aim to maximise ρ as defined in equation 6:

$$\rho = \frac{E\left[\hat{f}_x \hat{f}_y{}^T\right]}{\sqrt{E\left[\hat{f}_x \hat{f}_x{}^T\right] E\left[\hat{f}_y \hat{f}_y{}^T\right]}} \tag{6}$$

With the total covariance block matrix given in equation 7 as follows:

$$C = \begin{bmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{bmatrix} = E\left[\begin{bmatrix} f_x \\ f_y \end{bmatrix} \begin{bmatrix} f_x \\ f_y \end{bmatrix}^T\right] \tag{7}$$

C_{xx} and C_{yy} represent the within sets covariance matrices and $C_{xy} = C_{yx}^T$ represents the between set matrix. In order to find the canonical correlations between f_x and f_y, it is necessary to solve the Eigen value equations in (8):

$$\begin{cases} C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{yx} u_x = \rho^2 u_x \\ C_{yy}^{-1} C_{yx} C_{xx}^{-1} C_{xy} u_u = \rho^2 u_y \end{cases} \tag{8}$$

Where the Eigen values ρ^2 represent the squared canonical correlations. It is only necessary to solve one Eigen value equation due to the two components of equation 8 being related as shown in (9):

$$\begin{matrix} C_{yx} u_x = \rho \lambda_y C_{yy} u_y \\ C_{xy} u_y = \rho \lambda_x C_{xx} u_x \end{matrix} \qquad \text{Where:} \qquad \lambda_x = \lambda_y^{-1} = \sqrt{\frac{u_y^T C_{yy} u_y}{u_x^T C_{xx} u_x}} \tag{9}$$

4 Results

4.1 Comparison of CCA with Sentences and Segments

Initially, synchrony between audio and visual signals was assessed. Existing work [3] found maximum audiovisual correlation with an audio delay of 40ms. To corroborate this, CCA was applied to a 24 sentence dataset from VidTIMIT. The canonical correlations of each sentence were calculated, and the correlation measure used is defined as τ, where:

$$\tau = \sum_{i=1}^{N} \lambda^2 \tag{10}$$

Where N represents the number of canonical correlations found. τ was taken when shifting the visual data in relation to the equivalent audio data. The mean synchronisation results are shown in fig.1(a), confirming that audiovisual correlation is maximised when there is a small degree of asynchrony, in line with results found by Sargin et al. [3]. Accordingly, subsequent experiments are shifted by three frames.

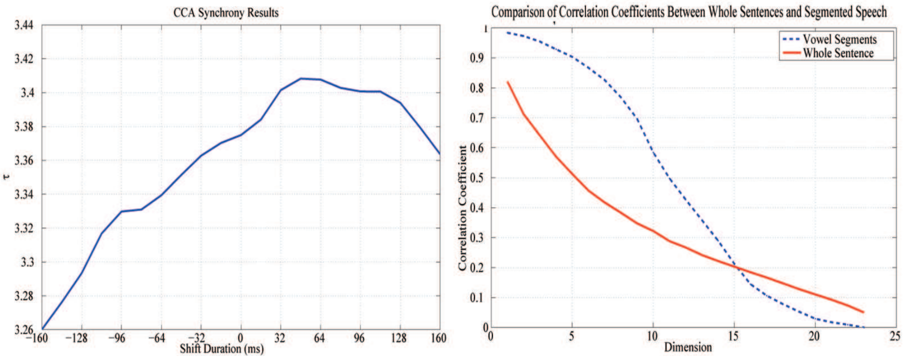


Fig. 1. (a) CCA feature synchrony results, showing maximum τ with a shift of approx 48ms. (b) Comparison of mean canonical correlations of complete VidTIMIT sentences and vowel segments.

The performance of CCA with segmented speech was then assessed. Our proposed vowel segmentation technique was used to extract speech sentence segments containing vowels. CCA was then performed on these segments, and compared to results from full sentences. The process used is shown in fig.2.

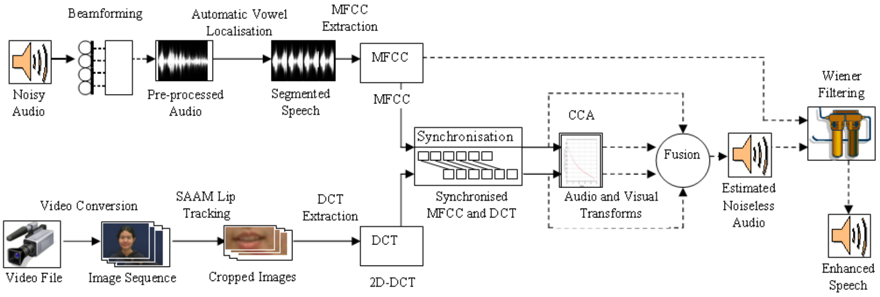


Fig. 2. Proposed multimodal speech enhancement system, showing role of lip tracking, speech segmentation, and CCA transforms described in this paper. Proposed fusion, speech estimation and wiener filtering components are also shown. Solid lines indicate components used in this work, and dashed lines represent proposed future work.

The mean canonical correlations of sentences and segments are shown in fig.1(b), which shows two lines. The solid line indicates the mean canonical correlation coefficients for whole sentences, while the dashed line represents mean coefficients for vowel segments only. This shows a clear difference in correlation values and behaviour, with segmented speech producing a significantly higher multimodal correlation, as it can be seen from fig.1(b) that vowel segments produce much stronger initial canonical correlations. This is proven by comparing the squared sum of canonical correlations (τ) for sentences and segments, with results of 3.41 and 8.48 respectively, confirming that speech segmentation significantly increases correlation.

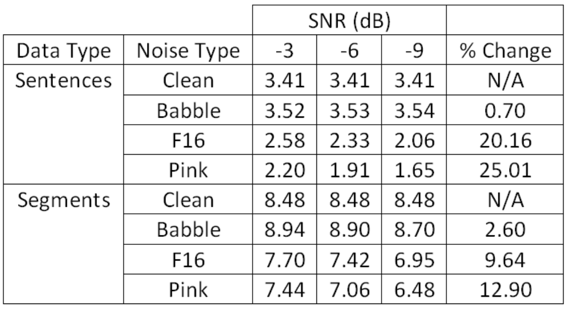
4.2 Noisy Speech Investigation

The previous section was extended by performing CCA in a variety of noisy environments. A number of noises (filtered pink noise, F16 aircraft noise, and incoherent speech babble) were added to the dataset at SNRs of -3, -6, and -9dB. Vowel segmentation was then carried out on the resulting noisy sentences, and the results of CCA on noisy sentences and segments are shown in table 1.

Table 1 shows that two of the noisy speech mixtures, pink and F16, have a much lower correlation than clean speech, which is expected. However, due to the similarities between the incoherent human babble and the target speech, CCA appears to find inaccurate relationships between audio and visual data. In all cases though, there is a significant correlation increase when vowel specific information is used. With the exception of babble, the table shows that segmented speech produces a lower percentage drop in audiovisual correlation when the noise level is increased, showing that our speech segmentation approach is effective for increasing audiovisual speech correlation in noisy environments. It should be noted that these results suggest that our approach functions best in environments where the noise is suitably different from the target speech, such as aircraft or automobiles. In environments dominated by human babble, changing the SNR makes little difference, as similar inaccurate relationships between babble and visual speech features are still found irrespective of the SNR value, explaining the smaller change in correlation produced by babble in table 1.

5 Conclusion

In this paper, we presented work that maximised audiovisual speech correlation by successfully making use of our SAAM approach for automatic visual feature extraction and our vowel based segmentation technique to segment speech. To assess the performance of these techniques, CCA was used to investigate multimodal correlation. It was found that in noisy environments, segmented speech produced much higher correlation than whole sentences, showing the potential of these techniques for future use as part of an integrated multimodal speech enhancement system, as shown in fig.2. This diagram shows the proposed role of the lip tracking, speech segmentation, and CCA transforms discussed in this



		SNR (dB)			
Data Type	Noise Type	-3	-6	-9	% Change
Sentences	Clean	3.41	3.41	3.41	N/A
	Babble	3.52	3.53	3.54	0.70
	F16	2.58	2.33	2.06	20.16
	Pink	2.20	1.91	1.65	25.01
Segments	Clean	8.48	8.48	8.48	N/A
	Babble	8.94	8.90	8.70	2.60
	F16	7.70	7.42	6.95	9.64
	Pink	7.44	7.06	6.48	12.90

Table 1. Segment and Sentence τ comparison of noisy speech at -3, -6, -9 dB SNR

paper in such a proposed speech enhancement system, as well as a proposed audiovisual Wiener filtering approach, and the use of a beamformer (tested in previous work by the authors [13]) for pre-processing the noisy audio signal.

References

1. Almajai, I., Milner, B.: Maximising Audio-Visual Speech Correlation. In AVSP 2007 (2007)
2. Almajai, I., Milner, B., Darch, J., Vaseghi, S.: Visually-Derived Wiener Filters for Speech Enhancement. In ICASSP 2007, vol. 4, pp. 585–588 (2007)
3. Sargin, M.E., Yemez, Y., Erzin, E., Tekalp, A.M.: Audiovisual Synchronization and Fusion Using Canonical Correlation Analysis. Mult., IEEE Trans. on, vol. 9, no. 7, pp. 1396–1403 (2007)
4. Hotelling, H.: Relations between two sets of variates. Biometrika, vol.28, pp. 321–377 (1936)
5. Girin, L., Feng, G., Schwartz, J.L.: Fusion of Auditory and Visual Information For Noisy Speech Enhancement: A Preliminary Study of Vowel Transition. In ICASSP 1998, vol. 2, pp. 1005–1008 (1998)
6. Ringeval, F., Chetouani, M.: A Vowel Based Approach For Acted Emotion Recognition. In Proc. Interspeech, 2008, pp. 2763–2766 (2008)
7. Sanderson, C.: Biometric Person Recognition: Face, Speech and Fusion. VDM-Verlag (2008)
8. Pellegrino, F., Andr-Obrecht,R.: Automatic Language Identification: An Alternative Approach to Phonetic Modelling. Sig. Proc., vol. 80, no. 7, pp. 1231–1244 (2000)
9. Nguyen,Q.D., Milgram, M.: Semi Adaptive Appearance Models For Lip Tracking. Submitted to ICIP 2009 (2009)
10. Levy, A., Lindenbaum, M.: Sequential Karhumen-Loeve basis extraction and its application to images. Image Proc., IEEE Trans., vol. 9, no. 8, pp. 1371–1374 (2000)
11. Golub, G.H., Van Loan, C.F.: Matrix Computations(3rd Edition). John Hopkins Uni. Press (1996)
12. Cauwenberghs, G, Poggio, T.: Incremental and Decremental Support Vector Machine Learning. *NIPS*, pp. 409–415 (2000)
13. Cifani, S., Abel, A., Hussain, A, Squartini, S., Piazza, F.: An Investigation Into Audiovisual Speech Correlation In Reverberant Noisy Environments. (LNCS): Cross-Modal Analysis of Speech, Gest., Gaze and Facial Expr. In Press(2008)