Analysis of Correlation between Audio and Visual Speech Features for Clean Audio Feature Prediction in Noise

Ibrahim Almajai, Ben Milner and Jonathan Darch

School of Computing Sciences
University of East Anglia, Norwich, UK
{i.almajai, b.milner, jonathan.darch}@uea.ac.uk

Abstract

The aim of this work is to examine the correlation between audio and visual speech features. The motivation is to find visual features that can provide clean audio feature estimates which can be used for speech enhancement when the original audio signal is corrupted by noise. Two audio features (MFCCs and formants) and three visual features (active appearance model, 2-D DCT and cross-DCT) are considered with correlation measured using multiple linear regression. The correlation is then exploited through the development of a maximum a posteriori (MAP) prediction of audio features solely from the visual features. Experiments reveal that features representing broad spectral information have higher correlation to visual features than those representing finer spectral detail. The accuracy of prediction follows the results found in the correlation measurements.

Index Terms: audio-visual speech, correlation, AAM, formants.

1. Introduction

It is well known that humans utilize visual information to enhance their perception of speech especially when the audio is corrupted by noise [1]. This audio-visual bimodal nature of speech production and perception has attracted much research over the past two decades and has been applied to both robust speech recognition and enhancement. For speech recognition, work has shown that combining audio and visual features can significantly improve accuracy in noise [2,3]. For speech enhancement, some systems re-synthesize the speech using clean speech parameters extracted from a combination of noisy audio and clean visual features. For example, [4] synthesizes speech using excitation extracted from the audio signal and a vocal tract filter obtained from a combination of audio and visual features. Alternatively, [5] derives a Wiener filter from visual features for removing noise from the audio speech signal. The Wiener filter is created from clean spectral envelope estimates obtained by utilizing the correlation between audio and visual information.

To effectively utilise visual information for speech enhancement there must exist sufficient correlation between audio and visual features. Several studies [6,7] have measured the correlation between acoustic (audio), face movement (visual) and articulatory features. Rather than using visual features directly, these investigations have used markers positioned around the face. The experiments have revealed that correlation exists between face movement and acoustic features, although less than that which exists between articulatory features and audio features.

The aim of this work is to examine the correlation between a range of different audio and visual speech features and subsequently to use this to examine the effectiveness of predicting clean audio features from visual features. The analysis will determine which visual feature offers highest correlation to audio features. The motivation for this is to identify audio-visual feature sets that can provide clean spectral envelope estimates for the purpose of enhancing noise corrupted speech. Three different visual features and two different audio features are used in the correlation measurement. These are discussed in sections 2 and 3. Multiple linear regression is used as the method of measuring correlation between the visual features and components of the audio features. This is discussed in section 4. A maximum a posteriori (MAP) method for predicting clean audio features from visual features is also developed in this section as a way of exploiting the correlation. Section 5 presents experimental results that compare the correlation of the two audio features with the visual features. Results are also presented on the accuracy of formant frequency prediction from visual features.

2. Visual feature extraction

Visual feature extraction techniques fall into two main categories; model-based (shape) and pixel-based (appearance) [2]. This section briefly discusses three different visual features that will be used in the correlation analysis. Two pixel-based features are considered, 2-D discrete cosine transform (DCT) and cross-DCT, while the third is derived from an active appearance model (AAM) and is a combination of shape and appearance. For consistency all three visual features are represented by $M=14$ dimensional vectors.

2.1. Active appearance model (AAM)

The AAM is a frequently used method of visual feature extraction and statistically models shape and appearance information [8]. From a set of training images, labeled with landmark points, the AAM uses image warping to deform each image to a mean shape and then builds a statistical model combining shape and appearance across the object. Given a test image the AAM minimizes the difference between its synthesized image and the actual image by varying the model parameters as well as incurring some displacements in position, scale, and orientation. Features of the final synthesized image are extracted at the end of the search process to generate a

M=14 dimensional visual vector, $\mathbf{x}^{\mathrm{AAM}}$. Further details are given in [8].

To determine the best part of the face from which to extract visual features, AAM features were extracted from different facial regions of interest (ROI) and visual speech recognition accuracy measured. These preliminary results established that extracting AAM features from the lower half of the face gave best performance. Figure 1a shows an example image from the database (discussed in section 5) while figure 1b shows the distribution of landmark points on the lower half of the speaker's face. Figure 1c shows the warped image of the lower face generated by the visual vector extracted from the image.

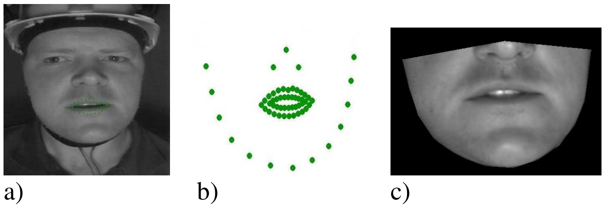

Figure 1: a) Original image, b) distribution of landmark points on lower face, c) warped image of lower face.

2.2. Two-dimensional DCT

The 2-D DCT is a common method of extracting pixel-based visual features from an image of a speaker's mouth [3]. In this work a 2-D DCT is applied to a 200x200 pixel median-filtered ROI centralized around the mouth. Following the 2-D DCT, the energy from the image is concentrated in the lower coefficients of the resulting matrix. The final visual vector, $\mathbf{x}^{\mathrm{2DDCT}}$, is obtained from fourteen 2-D DCT coefficients extracted in a zigzag order,

$$\mathbf{x}^{2DDCT} = [c_{0,0}, c_{0,1}, c_{1,0}, c_{2,0}, c_{1,1}, c_{0,2}, c_{0,3}, c_{1,2} \\ c_{2,1}, c_{3,0}, c_{4,0}, c_{3,1}, c_{2,2}, c_{1,3}\ ] \qquad (1)$$

where $c_{i,j}$ is the coefficient of the i^{th} row and j^{th} column.

2.3. Cross DCT

As a simple alternative to the 2-D DCT visual feature, some audio-visual systems extract a horizontal row and vertical column of pixels from the center of the mouth [3]. One-dimensional DCTs are applied to the horizontal and vertical vectors of pixels. The resulting vectors are truncated to include the 7 lowest-order coefficients and the two vectors concatenated to produce a single 14-D visual vector, $\mathbf{x}^{\mathrm{CrossDCT}}$.

3. Audio feature extraction

Two different audio features are considered in this work. The first are mel-frequency cepstral coefficients (MFCCs) as frequently used in speech recognition. MFCCs are a perceptually motivated representation of speech and through inversion can provide a spectral envelope estimate. In fact, recent work has shown that intelligible speech can be reconstructed solely from MFCC vectors [9]. The second audio feature comprises the first four formant frequencies. Formants are a useful acoustic parameter that measure resonant frequencies in the vocal tract.

3.1. MFCC feature extraction

MFCC features have been extracted according to the ETSI Aurora distributed speech recognition standard to give a 14-D audio vector, $\mathbf{y}^{\mathrm{MFCC}}$, comprising MFCCs 0 to 12 and log energy [10]. Each MFCC vector is computed from a 25ms frame of speech at a rate of 100 vectors per second.

3.2. Formant frequency extraction

The first four formant frequencies were extracted from 25ms frames of speech using linear predictive analysis. These initial formant frequency estimates are then refined through Kalman filtering [11] to create a four dimensional audio feature vector, $\mathbf{y}^{\mathrm{Formant}}$=[F1, F2, F3, F4], at a rate of 100 vectors per second.

4. Analysis of audio-visual features

This section first explains how multiple linear regression is used to measure correlation between audio and visual speech features. Secondly, based on this correlation, a MAP prediction of audio features from visual features is developed.

4.1. Measurement of audio-visual correlation

The audio and visual features are analysed by measuring the correlation between each element of the audio feature vector and the entire visual feature vector using multiple linear regression [12]. A linear model is developed which describes the relation between the visual features (the independent variables) and the audio feature (the dependent variable). Multiple linear regression enables each element of the audio feature vector, $y(j)$, to be represented in terms of the elements in the visual vector, $\mathbf{x}$, using a set of $M+1$ regression coefficients, $\{b_{j,0}, .., b_{j,m}, .., b_{j,M}\}$, which are specific to the j^{th} element of the audio feature vector,

$$y(j) = b_{j,0} + b_{j,1}\,x(1) + b_{j,2}\,x(2) + \ldots + b_{j,M}\,x(M) + \varepsilon\ ,\ 1 \leq j \leq M \qquad (2)$$

where ε is an error term. Using a set of training data, least squares estimation can be used to determine the regression coefficients [12]. These regression coefficients can be used to make a prediction of the j^{th} element of the i^{th} audio feature vector, $\hat{y}_i(j)$, from the i^{th} visual vector, $\mathbf{x}_i$,

$$\hat{y}_i(j) = b_{j,0} + b_{j,1}\,x_i(1) + b_{j,2}\,x_i(2) + \ldots + b_{j,M}\,x_i(M) \qquad (3)$$

The amount of correlation between the j^{th} element of the audio vector and the visual vector, $\mathbf{x}$, can finally be determined from the R-squared term which is defined as,

$$R^2 = 1 - \frac{\sum_i (y_i(j) - \hat{y}_i(j))^2}{\sum_i (y_i(j) - \bar{y}(j))} \qquad (4)$$

$\bar{y}(j)$ is the mean of the j^{th} element of the audio vector [12].

4.2. Estimation of audio features from visual features

Assuming correlation exists between audio and visual features, it is possible to predict the audio feature from the visual feature using equation 3. Instead, however, a Gaussian mixture model (GMM) is developed to model the joint density of the audio and visual feature vector space. A GMM is selected as the model, as its individual clusters allow a localized modeling of the joint density. First an audio-visual vector, **z**, is defined,

$$\mathbf{z} = \left[\mathbf{x}, \mathbf{y}\right] \tag{5}$$

where **x** is a visual feature vector (AAM, 2-D DCT or cross DCT) and **y** an audio feature vector (MFCCs or formants). From a set of audio-visual training data, unsupervised clustering is implemented using the expectation-maximisation (EM) algorithm to produce a GMM, $\Phi^{\mathbf{z}}$, which comprises a set of *K* clusters that localize the correlation between the audio and visual vectors in the joint feature space,

$$\Phi^{\mathbf{z}}(\mathbf{z}) = \sum_{k=1}^{K} \alpha_k \, f\left(\mathbf{z}; \boldsymbol{\mu}_k^{\mathbf{z}}, \boldsymbol{\Sigma}_k^{\mathbf{z}}\right) \tag{6}$$

Each cluster, or mixture component, is represented by a prior probability, α_k, and Gaussian probability density function (PDF), *f*, with mean vector, $\boldsymbol{\mu}_k^{\mathbf{z}}$, and covariance matrix, $\boldsymbol{\Sigma}_k^{\mathbf{z}}$.

Prediction of the audio vector, $\hat{\mathbf{y}}_i$, can be made from a visual vector, $\mathbf{x}_i$, using the GMM by making a maximum a posteriori (MAP) estimate [9],

$$\hat{\mathbf{y}}_i = \arg\max_{y_i} \left\{ p\left(\mathbf{y}_i \middle| \mathbf{x}_i, \Phi^{\mathbf{z}} \right) \right\} \tag{7}$$

Audio feature predictions from each cluster are weighted by the posterior probability, $h_k(\mathbf{x}_i)$, of $\mathbf{x}_i$ belonging to the k^{th} cluster. This gives the prediction of the audio feature as,

$$\hat{\mathbf{y}}_i = \sum_{k=1}^{K} h_k(\mathbf{x}_i) \left(\boldsymbol{\mu}_k^{\mathbf{y}} + \boldsymbol{\Sigma}_k^{\mathbf{yx}} \left(\boldsymbol{\Sigma}_k^{\mathbf{xx}} \right)^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{\mathbf{x}} \right) \right) \tag{8}$$

where $\boldsymbol{\mu}_k^{\mathbf{y}}$ and $\boldsymbol{\mu}_k^{\mathbf{x}}$ are the means of the audio and visual vectors in cluster *k*, while $\boldsymbol{\Sigma}_k^{\mathbf{xx}}$ is the covariance of the visual vector and $\boldsymbol{\Sigma}_k^{\mathbf{yx}}$ is the cross covariance of the audio and visual vectors. The posterior probability is given by,

$$h_k(\mathbf{x}_i) = \frac{\alpha_k \; p\left(\mathbf{x}_i \middle| \Phi_k^{\mathbf{x}} \right)}{\sum_{k=1}^{K} \alpha_k \; p\left(\mathbf{x}_i \middle| \Phi_k^{\mathbf{x}} \right)} \tag{9}$$

where $p\left(\mathbf{x}_i \middle| \Phi_k^{\mathbf{x}} \right)$ is the marginal distribution of the visual vector for the k^{th} cluster of the GMM.

5. Experimental results

The experiments in this section first examine the correlation between the audio and visual features. Second, the accuracy of predicting formant frequencies from visual features is evaluated.

The experiments are performed on an audio-visual speech database which comprises 277 sentences of continuous speech spoken by a single male speaker [13]. Of these, 200 utterances were used for training and 77 for testing. The audio data was sampled at a rate of 8kHz. The video was originally recorded at 25 frames per second. This was upsampled to 100 frames per second to give a visual frame rate equal to the audio frame rate.

5.1. Correlation analysis

The aim of this section is to examine the correlation that exists between individual elements of audio features and the three different visual features. The correlation has been measured by pooling together audio-visual data from all utterances in the training set. Multiple regression has then been applied and the R-squared term computed as described in section 4.1. Figure 2 shows the multiple correlation (square root of R-squared term) of MFCCs 0 to 12 and log energy with the AAM, 2-D DCT and cross-DCT features. Similarly, figure 3 shows the multiple correlation of the four formant frequencies (F1, F2, F3, F4) with the AAM, 2-D DCT and cross-DCT features.

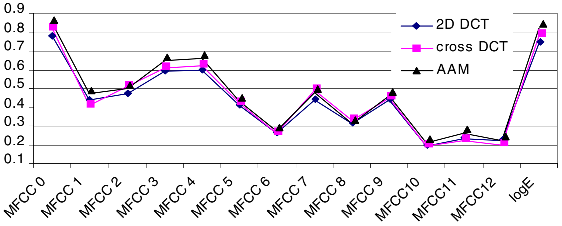

Figure 2: Multiple correlation coefficients for MFCCs

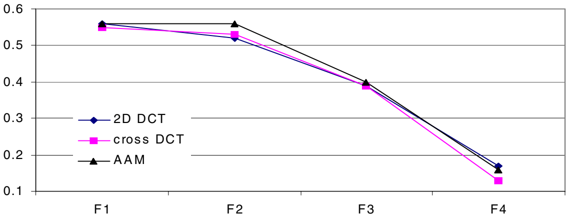

Figure 3: Multiple correlation coefficients for formants

All three visual features exhibit very similar levels of correlation to either the MFCCs or formants. The AAM feature exhibits slightly higher correlation to both MFCCs and formants than the other visual features. Of the individual audio features shown in figure 2, MFCC 0 and log energy have highest correlation with the visual features – giving correlations of *R=0.85* and *R=0.82* respectively with the AAM features. Higher order MFCCs, such as MFCCs 10, 11 and 12, exhibit much lower correlation with the visual features. The MFCCs which exhibit highest correlation with the visual features are based on measurements such as energy or the broad spectral structure of the speech which can adequately be observed from the mouth shape encoded by the visual features. Higher-order MFCCs represent finer spectral structure which is much more difficult to determine from mouth shape, hence the lower correlation values. The visual to formant correlation is considerably lower than the visual to MFCC correlation. Of the four formants, F1 and F2 have highest correlation of *R=0.56* while for higher frequency formants the correlation reduces – to *R=0.16* for F4.

5.2. Formant frequency prediction

This section presents the result of predicting the formant frequencies of a frame of speech solely from the AAM visual feature representation. Formant frequencies are selected as the audio feature for prediction, rather than MFCCs, as they are a more intuitive representation of speech which can be visualized through spectrogram overlay. AAM features are selected as the visual feature due to their higher correlation to audio features.

The training data utterances have been used to create the GMM described in section 4.2, while evaluation of formant prediction has been made from the test utterances which comprise 24,297 vectors. The accuracy of formant frequency prediction is measured by percentage formant frequency estimation error, E_p,

$$E_p = \frac{1}{N}\sum_{i=0}^{N-1}\frac{1}{4}\sum_{j=1}^{4}\frac{\left|F_i(j)-\hat{F}_i(j)\right|}{F_i(j)}\times 100\% \qquad (10)$$

Where $F_i(j)$ and $\hat{F}_i(j)$ are the reference and predicted frequency of the j^{th} formant at time instant i. Figure 4 shows the percentage formant frequency error as the number of clusters in the GMM is increased from 1 to 32.

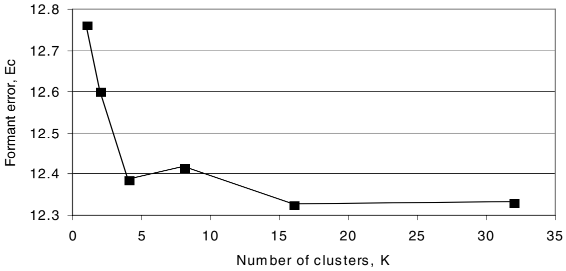

Figure 4: Percentage formant frequency error for varying numbers of clusters

As the number of clusters in the GMM is increased up to *K=16* the accuracy of formant prediction increases due to more accurate modeling of the joint distribution of formants and AAM features. Increasing the number of clusters to 32 gives no further increase in accuracy. To illustrate the accuracy of formant frequency prediction, figure 5 shows a 2 second spectrogram of the sentence "*Sarah argued that I acted as though under his thumb*" taken from the test set with predicted and reference formants overlaid.

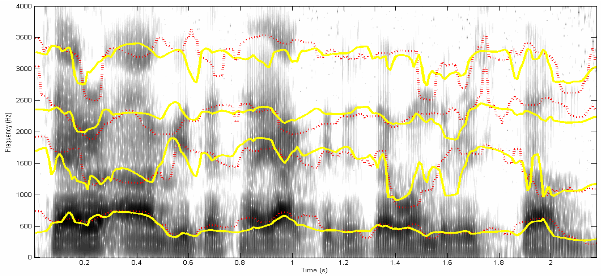

Figure 5: Spectrogram of utterance "*Sarah argued that I acted as though under his thumb*" showing reference formants (dotted line) and predicted formants (solid line).

The figure illustrates that the predicted F1 contour follows closely the reference F1 contour. F2 also follows reasonably closely, better than F3, but worse than F1. F4 is considerably worse, although this formant is unclear even in the spectrogram. The accuracy of the predicted formants follows the trend in figure 3 which shows higher correlation for F1 and F2 and lower correlation for F3 and particularly for F4.

6. Conclusion

This analysis has shown that correlation exists between audio and visual representations of speech. Audio features that represent broad spectral shape, such as log energy, MFCCs 0 to 4, and F1 and F2, exhibit higher levels of visual correlation than features representing finer spectral structure. It is also interesting that all three visual features exhibit very similar levels of correlation to the audio features even though their methods of computation are very different. Tests to predict formant frequencies from AAM features using a GMM showed that F1 and F2 could be predicted more accurately than F3 and F4, which is consistent with the correlation analysis. These tests demonstrated that clean spectral envelope information can be extracted from visual speech features which has importance when audio features have been contaminated by noise. The use of these visually predicted audio features for speech enhancement is the subject of further work.

7. References

[1] W.H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise", *JASA,* 26(2):212–215, 1954

[2] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari and J. Zhou, "Audio-visual speech recognition", Technical Report, Center for Language and Speech Processing, Baltimore, Maryland, 2000

[3] G. Meyer, J. Mulligan and S. Wuerger, "Continuous audio-visual digit recognition using N-best decision fusion", Information Fusion, 5:91-101, 2003

[4] L. Girin, J.L. Schwartz and G. Feng, "Audio-visual enhancement of speech in noise", JASA, 6(109):3007-3020, 2001

[5] F. Berthommier, "Audiovisual speech enhancement based on association between speech envelope and video features", Proc. Eurospeech, 2003

[6] H. Yehia, P. Rubin and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behaviour", Speech Communication, 26(1):23-43, 1998

[7] J. Jiang, A. Alwan, P.A. Keating, E.T. Auer and L.E. Bernstein, "On the relationship between face movements, tongue movements and speech acoustics", EURASIP Journal on Applied Sig. Proc., 11:1174-1188, 2002

[8] T.F. Cootes and C.J. Taylor, "Statistical models of appearance for computer vision," Draft report, University of Manchester, UK, 2001, http://www.isbe.man.ac.uk

[9] X. Shao and B.P. Milner, "Predicting fundamental frequency from MFCCs to enable speech reconstruction", JASA 118(2):1134-1143, 2005

[10] A. Sorin and T. Ramabadran, "Extended advanced front end algorithm description, Version 1.1", ETSI STQ Aurora DSR Working Group, Tech. Rep. ES 202 212, 2003

[11] Q. Yan, E. Zavarehei, S. Vaseghi and D. Rentzos, "A formant tracking LP model for speech processing in car/train noise", Proc. ICSLP, 2004 .

[12] S. Chatterjee, A.S. Hadi, and B. Price, "Regression Analysis By Example", John Wiley and Sons, Canada, 2000

[13] B. Theobald. "Visual speech synthesis using shape and appearance models," PhD Thesis. School of Computing Sciences, University of East Anglia, Norwich, UK, 2003.