VISUALLY-DERIVED WIENER FILTERS FOR SPEECH ENHANCEMENT

Ibrahim Almajai[1], Ben Milner[1], Jonathan Darch[1] and Saeed Vaseghi[2]

[1]School of Computing Sciences, University of East Anglia, UK
[2]Dept. of Electronic and Computer Engineering, Brunel University, UK
{i.almajai, b.milner, jonathan.darch}@uea.ac.uk, saeed.vaseghi@brunel.ac.uk

ABSTRACT

This work begins by examining the correlation between audio and visual speech features and reveals higher correlation to exist within individual phoneme sounds rather than globally across all speech. Utilising this correlation, a visually-derived Wiener filter is proposed in which clean power spectrum estimates are obtained from visual speech features. Two methods of extracting clean power spectrum estimates are made; first from a global estimate using a single Gaussian mixture model (GMM), and second from phoneme-specific estimates using a hidden Markov model (HMM)-GMM structure. Measurement of estimation accuracy reveals that the phoneme-specific (HMM-GMM) system leads to lower estimation errors than the global (GMM) system. Finally, the effectiveness of visually-derived Wiener filtering is examined.

Index Terms— Audio-visual, Wiener filter, speech enhancement, HMM, GMM

1. INTRODUCTION

The aim of this work is to use visual speech information to enhance noise-contaminated audio speech. This is motivated by the fact that speech is produced by movements of articulators (tongue, lips, etc) which leads to the existence of correlation between the resulting audio and visual shape of the mouth [1,2,3]. Therefore, by knowing the current mouth or lip shape, certain information regarding the audio speech being produced can be inferred. Of course, a spectrally detailed audio signal cannot be estimated from the lip shape (for example source information is not present in lip shape) but a coarse spectral envelope can be obtained and is noise free. This coarse spectral representation of clean speech, obtained from the visual speech, can be incorporated into a Wiener filter and used for audio speech enhancement.

The correlation between audio and visual speech features is analysed in section 2. This considers correlation both globally and within phoneme classes. Section 3 introduces the visually-derived Wiener filter for speech enhancement. Section 4 describes two methods for obtaining the clean power spectrum estimates needed for the Wiener filter; one using global correlation and the other phoneme-based correlation. Experimental results are presented in section 5 that examine clean speech estimation accuracy from visual features and the effectiveness of visual Wiener filtering.

2. CORRELATION ANALYSIS

This section investigates the correlation between audio and visual speech features. The motivation for this comes from the fact that speech is produced by controlled movements of articulators (tongue, lips, etc) which makes it likely that the resultant audio and visual shape of the mouth exhibit correlation. Several studies have shown that correlation exists between the audio speech signal and the visual shape of the mouth and also to articulator parameter positions [1,2]. Previous work [3] investigating the correlation between different audio features (MFCCs, formants) and visual features (AAM, 2-D DCT and cross-DCT) has shown that active appearance model (AAM) visual features and log filterbank audio features exhibit relatively high levels of correlation.

In this work the audio features are 23-D log mel-filterbank vectors, represented as $\mathbf{x}_i$ where i indicates the frame number. These are extracted at a rate of 100 vectors per second from 25ms frames of audio in accordance with the ETSI Aurora DSR front-end [4]. The visual features are 14-D AAM vectors, represented as $\mathbf{v}_i$. These are computed from the mouth region at a rate of 25 vectors per second and subsequently upsampled to 100Hz to match that of audio feature extraction. Details are given in [5].

2.1 Measurement of audio-visual correlation

The correlation between audio and visual features has been measured using 200 training utterances, taken from the database described in section 5. For each of the 23 filterbank channels, correlation has been measured with respect to the 14-D AAM vector using multiple linear regression. The AAM vector is considered the independent variable and each filterbank channel as the dependent variable [6].

Correlation is first measured from the entire set of audio-visual vectors to give the global correlation across all speech sounds. Figure 1 displays the multiple correlation coefficient for each filterbank channel (dashed line). Also shown is the mean phoneme-specific correlation (solid line). This is calculated by considering the audio-visual correlation within each phoneme separately and taking the mean over all phonemes. Phoneme boundaries for this analysis were obtained by hand annotation.

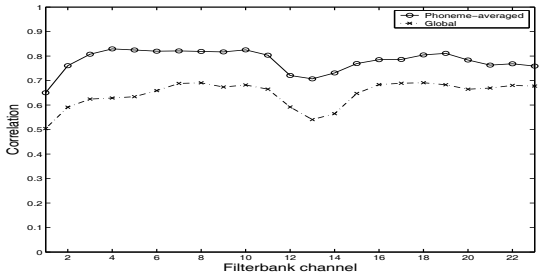

Fig. 1. Audio-visual correlation across 23 channel log mel-filterbank measured globally (dotted) and within phonemes (solid).

Figure 1 shows higher audio-visual correlation when considering the local correlation within each phoneme individually rather than computing correlation globally. The mean correlation, across all filterbank channels, is approximately 0.64 when computed globally and increases to 0.78 when computed within each phoneme. This suggests that when utilising visual information to estimate clean audio it is advantageous to consider the phoneme being spoken.

3. VISUALLY-DERIVED WIENER FILTER

This section proposes a visually-derived Wiener filter for speech enhancement that exploits the audio-visual correlation. In the frequency domain the Wiener filter, *W(f)*, is defined,

$$W(f) = \frac{P_{XX}(f)}{P_{XX}(f) + P_{NN}(f)} = \frac{P_{XX}(f)}{P_{YY}(f)} \tag{1}$$

where $P_{XX}(f)$, $P_{NN}(f)$ and $P_{YY}(f)$ denote the power spectra of the clean speech, noise and noisy speech respectively. The power spectrum of the noisy speech can usually be estimated from the input noisy speech. Obtaining the power spectra of clean speech is less straightforward and is one of the main problems in implementing Wiener filters for speech enhancement. In this work it is proposed to utilise the audio-visual correlation and use visual features to estimate the power spectrum of clean speech. Figure 2 illustrates the visually-derived Wiener filter for enhancement.

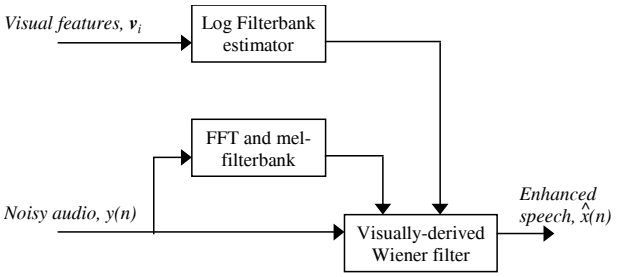

Fig. 2. Visually-derived Wiener filter for speech enhancement

System inputs are the noisy time-domain audio, *y(n)*, and visual features (AAM vectors), $\mathbf{v}_i$, – where *n* and *i* represent sample number and frame number respectively. By utilizing the correlation between clean audio and visual vectors, an estimate of the log filterbank of clean audio, $\hat{\mathbf{x}}_i$, is made from the AAM vector, $\mathbf{v}_i$. Applying an exponential operation transforms this to a linear filterbank estimate, $L_{\hat{X}}(m)$, where *m* represents the filterbank channel. The noisy audio data is also transformed to the linear filterbank domain using the processing described in section 2 to give $L_Y(m)$. Finally, the clean and noisy filterbank vectors provide a filterbank implementation of the Wiener filter, $L_W(m)$,

$$L_W(m) = \frac{L_{\hat{X}}(m)}{L_Y(m)} \tag{2}$$

For speech enhancement, the noisy audio, *y(n)*, is first converted to power and phase spectra, $|Y(f)|^2$ and $\angle Y(f)$. The filterbank domain Wiener filter is interpolated to the dimensionality of the power spectrum to give *W(f)*. Using this, an enhanced speech power spectrum, $|\hat{S}(f)|^2$ can be obtained,

$$|\hat{S}(f)|^2 = |Y(f)|^2 \, W(f) \tag{3}$$

Combination with the noisy phase, followed by an inverse Fourier transform and overlap-and-add returns the enhanced speech to the time-domain. The crucial stage in the visually-derived Wiener filter is obtaining the clean log filterbank estimate from the visual features. This is the subject of the next section.

4. CLEAN LOG FILTERBANK ESTIMATION

Estimation of clean log filterbank vectors from AAM vectors is achieved by modeling the joint density of the audio-visual feature vector space. First an audio-visual feature vector, $\mathbf{z}_i$, is defined as,

$$\mathbf{z}_i = [\mathbf{x}_i \ \mathbf{v}_i] \tag{4}$$

where $\mathbf{x}_i$ is the log filterbank vector extracted from clean speech at the same time frame, *i*, as the AAM visual vector, $\mathbf{v}_i$. To model the joint density of audio and visual vectors, two approaches have been considered. The first models the audio-visual feature space globally with a single Gaussian mixture model (GMM). To improve the modeling, based on the correlation analysis in section 2, the second method models the audio-visual correlation using a set of phoneme-specific GMMs.

4.1 Global GMM

From a training database of joint feature vectors, expectation-maximization (EM) clustering [7] is used to create a GMM that comprises *K* clusters which localize the correlation between log filterbank and AAM vectors in the joint feature vector space,

$$p(\mathbf{z}) = \sum_{k=1}^{K} \alpha_k \, N(\mathbf{z}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{5}$$

Each cluster is represented by a prior probability, α_k, and a Gaussian probability density function (PDF), *N*, with mean vector, $\boldsymbol{\mu}_k$, and covariance matrix, $\boldsymbol{\Sigma}_k$, where,

$$\boldsymbol{\mu}_k = \begin{bmatrix} \boldsymbol{\mu}_k^{\mathbf{x}} \\ \boldsymbol{\mu}_k^{\mathbf{v}} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_k = \begin{bmatrix} \boldsymbol{\Sigma}_k^{\mathbf{xx}} & \boldsymbol{\Sigma}_k^{\mathbf{xv}} \\ \boldsymbol{\Sigma}_k^{\mathbf{vx}} & \boldsymbol{\Sigma}_k^{\mathbf{vv}} \end{bmatrix} \tag{6}$$

The mean vectors have two components; the mean of the log filterbank vector and the mean of the AAM vector. The covariance matrices comprise four components; the covariance matrix of the log filterbank vectors, $\boldsymbol{\Sigma}_k^{\mathbf{xx}}$, the covariance matrix of the AAM vectors, $\boldsymbol{\Sigma}_k^{\mathbf{vv}}$, and the covariances of the log filterbank and AAM vectors, $\boldsymbol{\Sigma}_k^{\mathbf{vx}}$ and $\boldsymbol{\Sigma}_k^{\mathbf{xv}}$.

The GMM can now be used to estimate the log filterbank vector of the i^{th} frame of speech, $\hat{\mathbf{x}}_i$, from its AAM vector representation, $\mathbf{v}_i$. For the k^{th} cluster in the GMM, c_k, the maximum a posterior (MAP) estimate of the log filterbank estimate, $\hat{\mathbf{x}}_i$, is given as,

$$\hat{\mathbf{x}}_i = \arg\max_{\mathbf{x}_i} \left(p\left(\mathbf{x}_i \middle| \mathbf{v}_i, c_k \right) \right) \tag{7}$$

which can be expressed as,

$$\hat{\mathbf{x}}_i = \boldsymbol{\mu}_k^{\mathbf{x}} + \boldsymbol{\Sigma}_k^{\mathbf{xv}} \left(\boldsymbol{\Sigma}_k^{\mathbf{vv}} \right)^{-1} \left(\mathbf{v}_i - \boldsymbol{\mu}_k^{\mathbf{v}} \right) \tag{8}$$

Estimates from each of the K clusters in the GMM can be combined according to the posterior probability, $h_k(\mathbf{v}_i)$, of the AAM vector coming from each cluster to give a weighted MAP estimate of the log filterbank vector,

$$\hat{\mathbf{x}}_i = \sum_{k=1}^{K} h_k(\mathbf{v}_i) \left(\boldsymbol{\mu}_k^{\mathbf{x}} + \boldsymbol{\Sigma}_k^{\mathbf{xv}} \left(\boldsymbol{\Sigma}_k^{\mathbf{vv}} \right)^{-1} \left(\mathbf{v}_i - \boldsymbol{\mu}_k^{\mathbf{v}} \right) \right) \tag{9}$$

The posterior probability, $h_k(\mathbf{v}_i)$, is given as,

$$h_k(\mathbf{v}_i) = \frac{\alpha_k \, p\left(\mathbf{v}_i \middle| c_k^{\mathbf{v}} \right)}{\sum_{k=1}^{K} \alpha_k \, p\left(\mathbf{v}_i \middle| c_k^{\mathbf{v}} \right)} \tag{10}$$

where $p\left(\mathbf{v}_i \middle| c_k^{\mathbf{v}} \right)$ is the marginal distribution of AAM vectors for the k^{th} cluster in the GMM.

4.2 Phoneme-dependent HMM-GMM

To exploit the higher phoneme-dependent audio-visual correlation that was identified in section 2, this second method of log filterbank estimation uses a set of phoneme-specific GMMs that are selected according to a network of HMMs. Associated with each phoneme-based HMM is a GMM that models the local audio-visual correlation from which estimation is made. From a stream of audio-visual vectors, HMM decoding is applied to find the optimal sequence of HMMs, from which appropriate GMMs are selected for log filterbank estimation.

4.2.1 Training of HMM-GMMs

To model the phoneme-dependent joint density of log filterbank vectors and AAM vectors a set of monophone HMMs are required. These are trained from MFCC vectors created from the log filterbank vector component of the augmented feature vector. In this work 36 3-state diagonal covariance matrix monophone HMMs and a 3-state diagonal covariance matrix silence HMM have been used. From a training data utterance, $\mathbf{X}=[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_D]$, and its phoneme sequence, forced Viterbi decoding is applied using the set of HMMs to determine the phoneme allocation, $\mathbf{m}=[m_1, m_2, \ldots, m_D]$ for each feature vector. These indicate the phoneme, m_i, to which the i^{th} feature vector is allocated.

From the phoneme allocation of all training data vectors, vector pools are created for each phoneme, w. Using the vector pools and the training procedure described in section 4.1, a set of phoneme-specific GMMs, c_w, are trained which are represented by mean vectors, $\boldsymbol{\mu}_{k,w}$, covariance matrices, $\boldsymbol{\Sigma}_{k,w}$, and prior probabilities, $\alpha_{k,w}$, corresponding to the k^{th} cluster of the GMM associated with phoneme w.

4.2.2 Estimation of log filterbank vector

To estimate the clean log filterbank vectors from a sequence of audio-visual vectors $[\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_D]$ first the sequence of phoneme-dependent GMMs, from which estimation is made, is determined. This is achieved by decoding the sequence of MFCC vectors into a phoneme sequence, $\mathbf{m}=[m, m_2, \ldots, m_D]$, using the set of HMMs. This provides, for each visual vector, $\mathbf{v}_i$, a phoneme-specific GMM, c_{m_i}, from which the log filterbank vector will be estimated, based on a phoneme specific variant of equation (9). Further details of this HMM-GMM procedure can be found in [8] when applied to fundamental frequency prediction from MFCCs.

5. EXPERIMENTAL RESULTS

These experiments first measure the accuracy of clean log filterbank estimation from AAM vectors using both the global and phoneme-specific approaches. Second, the effectiveness of visually-derived Wiener filtering is examined. Experiments use an audio-visual speech database comprising 277 sentences of continuous speech spoken by a single male speaker [9]. 200 utterances are used for training and 77 utterances are used for testing. The audio data was sampled at a rate of 8kHz. The video was originally recorded at 25 frames per second and was upsampled to 100 frames per second to give a visual frame rate equal to the audio frame rate.

5.1 Clean log filterbank estimation

This section compares the accuracy of log filterbank estimation using first global modeling then phoneme-specific modeling.

5.1.2 Global GMM-based estimation

From the 200 training data sentences, audio-visual vectors were extracted and used to create a GMM as described in section 4.1. AAM vectors were then extracted from the 77 test utterances and in combination with the GMM, log filterbank vectors estimated. To measure the accuracy of estimation, a mean percentage error, $E_{\%}$, is computed by averaging the percentage estimation error across the $M=23$ channels of the $N=38,728$ vectors contained in the 77 test data utterances, where,

$$E_{\%} = \frac{1}{N\,M} \sum_{i=0}^{N-1} \sum_{m=1}^{M} \frac{\left| x_i(m) - \hat{x}_i(m) \right|}{x_i(m)} \times 100\% \tag{11}$$

$x_i(m)$ and $\hat{x}_i(m)$ represent the clean and estimated amplitude of the m^{th} log filterbank channel from the i^{th} vector. Table 1 shows the mean percentage estimation error, $E_{\%}$, computed using from 1 to 16 clusters within the GMM.

Num. clusters	GMM, $E_{\%}$
1	11.76
2	10.04
4	10.48
8	9.92
16	10.54

Table 1. GMM log filterbank estimation errors for 1 to 16 clusters.

The result shows that increasing the number of clusters in the GMM, up to 8 clusters, reduces estimation errors due to the more detailed modeling of the joint density of audio and visual vectors. At 16 clusters an increase in error occurs which is likely to be due to insufficient training data.

5.1.2 Phoneme-dependent HMM-GMM estimation

Log filterbank vectors are now estimated from the phoneme-specific HMM-GMM system of section 4.2. Table 2 shows the mean percentage estimation error, $E_{\%}$, computed from 1 to 8 clusters within each phoneme-specific GMM. Two methods for obtaining the sequence of HMM-GMMs are examined; first using forced alignment from hand annotation of the utterance and second using unconstrained monophone recognition. Finally the effect of noise is considered, whereby the audio used to determine the phoneme sequence is contaminated by white noise at a signal-to-noise ratio (SNR) of 10dB.

Num. clusters	Forced, $E_{\%}$	Unconstrained at clean, $E_{\%}$	Unconstrained at 10dB, $E_{\%}$
1	8.94	9.03	9.87
2	8.15	8.18	9.02
4	7.59	7.64	8.56
8	7.38	7.43	8.59

Table 2. HMM-GMM log filterbank estimation errors for cluster sizes from 1 to 8 for forced and unconstrained recognition.

Using forced alignment to determine the sequence of phoneme-dependent GMMs is artificial, as in practice the correct phoneme sequence would not be known. However, this test provides an upper bound on performance. The result shows a considerable decrease in estimation error when using the phoneme-specific HMM-GMM system in comparison to global GMM estimation. This is expected, based on the analysis of audio-visual correlation made in section 2. Moving to unconstrained monophone recognition (which has an accuracy of 60.0%) shows only a slight increase in estimation error. This is encouraging and shows that obtaining the precise phoneme sequence is not crucial for obtaining good estimation of the log filterbank. The results also suggests that using 36 monophone HMMs may be an over-detailed modeling of the speech sounds and that a reduced set of sound models would be better. When the audio is contaminated by noise, the recognition accuracy falls to 39.3%. Whilst this does cause a slight drop in log filterbank estimation accuracy, the fall is not too great and remains higher than with the GMM-only system.

5.2. Wiener filtering

The effectiveness of visually-derived Wiener filtering is examined using clean power spectrum estimates obtained from the 8 cluster HMM-GMM system. Figure 3a shows a spectrogram of the utterance "*Sarah argued that I acted as though under his thumb*" contaminated with white noise at an SNR of 10dB. Figure 3b shows the same utterance after the application of visually-derived Wiener filtering using the HMM-GMM method of clean log filterbank estimation.

Figure 3b shows that Wiener filtering has successfully removed large amounts of the noise present in figure 3a. This is particularly evident in non-speech periods, but good noise reduction also occurs in speech periods. In a series of informal listening tests, using utterances similar to that in figure 3, noise reduction was judged to be effective. During speech periods, most of the noise was removed at the expense of the introduction of slight distortion that caused a "muffling" of the speech.

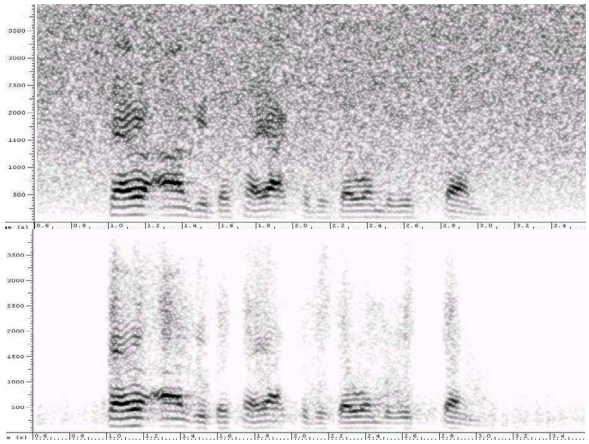

Fig. 3. Spectrogram of utterance, a) contaminated with white noise at an SNR of 10dB, b) after visually-derived Wiener filtering

6. CONCLUSION

This work has shown that it is possible to enhance noisy speech by using visual speech information. To achieve this, correlation between audio and visual speech has been examined and subsequently utilized to make estimates of clean log filterbank vectors from visual AAM vectors. Higher correlation, and subsequently better estimation, has been obtained when considering the individual phonemes being spoken rather than considering all speech globally. To identify the phonemes being spoken a simple unconstrained monophone network has been shown to be effective, even when decoding noisy speech. Further improvements to log filterbank estimation, and subsequently to speech enhancement, may be achieved by utilizing the noisy speech information in combination with the visual information.

7. REFERENCES

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

[2] J. Barker and F. Berthommier, "Estimation of speech acoustics from visual speech features: A comparison of linear and non-linear models" , Proc. AVSP-99, 1999

[3] I. Almajai, B. Milner and J. Darch, "Analysis of correlation between audio and visual speech features for clean audio feature prediction in noise", Proc. ICSLP, 2006

[4] 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

[5] T.F.Cootes, G.J. Edwards and C.J.Taylor. "Active Appearance Models", IEEE PAMI, Vol.23, No.6, pp.681-685, 2001

[6] S. Chatterjee, A.S. Hadi, and B. Price, "Regression analysis by example", John Wiley and Sons, Canada, 2000

[7] C.W. Therrien, Discrete random signals and statistical signal processing, Prentice-Hall, Englewood Cliffs, NJ, 1992

[8] X. Shao and B. Milner, "Predicting fundamental frequency from MFCCs to enable speech reconstruction", J. Acoust. Soc. Am., 118(2):1134-1143, 2005

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