

Noisy audio speech enhancement using Wiener filters derived from visual speech

Ben Milner and Ibrahim Almajai

School of Computing Sciences, University of East Anglia, UK

{b.milner, i.almajai }@uea.ac.uk

Abstract

The aim of this paper is to use visual speech information to create Wiener filters for audio speech enhancement. Wiener filters require estimates of both clean speech statistics and noisy speech statistics. Noisy speech statistics are obtained from the noisy input audio while obtaining clean speech statistics is more difficult and is a major problem in the creation of Wiener filters for speech enhancement. In this work the clean speech statistics are estimated from frames of visual speech that are extracted in synchrony with the audio. The estimation procedure begins by modelling the joint density of clean audio and visual speech features using a Gaussian mixture model (GMM). Using the GMM and an input visual speech vector a maximum a posteriori (MAP) estimate of the audio feature is made. The effectiveness of speech enhancement using the visually-derived Wiener filter has been compared to a conventional audio-based Wiener filter implementation using a perceptual evaluation of speech quality (PESQ) analysis. PESQ scores in train noise at different signal-to-noise ratios (SNRs) show that the visually-derived Wiener filter significantly outperforms the audio-Wiener filter at lower SNRs.

Index Terms: Audio-visual, speech enhancement, Wiener filter, MAP, HMM

1. Introduction

The multimodal nature of many communication devices allows not only audio from a speaker to be captured but also video. The video stream provides information that is not present in the audio, such as gesture, but also provides an alternative, visual, representation of some of the information that is present in the audio. One of the main advantages of considering this visual speech representation is that it is unaffected by acoustic noise. This fact has led to many audio-visual speech recognition systems that achieve robust performance in noise [1,2]. The work presented in this paper moves away from audio-visual speech recognition and instead uses visual speech information to enhance noisy audio speech. This is motivated by an analysis of audio-visual speech features that reveals significant correlation to exist between the two streams [3,9].

Many techniques have been proposed for speech enhancement and these typically operate by first estimating the contaminating noise and then removing it from the noisy speech to leave an enhanced speech signal [4,5]. These typically use either a voice activity detector to identify speech inactive periods and update noise model parameters, or minimum statistics methods where the noise model takes on minimum power levels found in the input audio signal [6].

In this work a visually-derived Wiener filter is proposed for speech enhancement. Wiener filters have previously been applied to speech enhancement, although one of the major problems is obtaining clean speech statistics necessary for their implementation. In this work it is proposed to utilise visual speech features, extracted from a speaker's mouth, to provide the clean audio speech statistics needed in Wiener filtering. To be successful, this method relies on correlation existing between the visual features and the audio signal. This is supported by the generation process of speech, which is related to movements of articulators (tongue, lips, etc) and gives rise to correlation between the resulting audio and the visual shape of the mouth [3,7]. Of course a spectrally detailed audio signal cannot be estimated from the mouth shape (for example, source information is not present in mouth shape) but an estimate of spectral envelope can be obtained.

The remainder of this work begins in section 2 by describing the operation of the visually-derived Wiener filter. Section 3 describes a phoneme-specific maximum a posteriori (MAP) method for estimating audio speech features from visual speech features using a network of hidden Markov models (HMMs) to provide phoneme localisation. Section 4 demonstrates the effectiveness of the visually-derived Wiener filter using PESQ analysis and compares this against two conventional audio-only based methods of enhancement.

2. Visually-derived Wiener filter

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

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

$P_{XX}(f)$, $P_{NN}(f)$ 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 be estimated from the input noisy audio speech. Obtaining the power spectra of the clean speech is less straightforward and is one of the main problems in implementing Wiener filters for speech enhancement. In this work the power spectrum of the clean speech is estimated from visual features. Estimating a detailed clean speech power spectrum from visual features is difficult, due limits on audio-visual correlation, so instead the Wiener filter is created initially in the filterbank domain. From this spectrally coarse domain, interpolation is applied to transform the Wiener filter to the dimensionality of the power spectrum where enhancement is applied. Figure 1 illustrates the visually-derived Wiener filter for speech enhancement and the next subsections describe its operation.

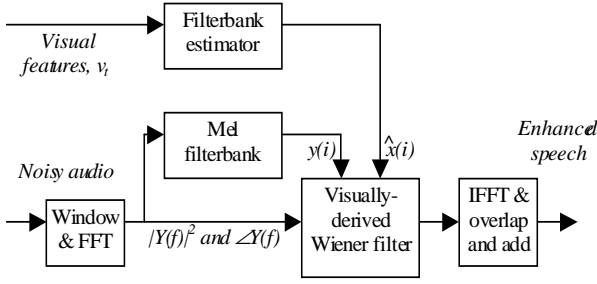


Figure 1: Visually-derived Wiener filter

2.1 Audio and visual speech features

The inputs to the speech enhancement system are the noisy time-domain audio and the visual vectors, \mathbf{v}_t , where t represents frame number. The Wiener filter is initially implemented in the filterbank domain, which is based on the ETSI Aurora distributed speech recognition standard [8]. The noisy audio is segmented into 25ms frames at a rate of 100 frames per second. Following a Hamming window and Fourier transform, a 128 bin power spectrum is calculated and a 23-channel mel-filterbank and log applied [9].

Several visual features have been proposed for audio-visual speech recognition and include active appearance models, 2-D discrete cosine transform (DCT) and cross-DCT [10]. Based on the evaluation of the audio-visual correlation of different audio and visual features in [9], 2-D DCT visual features have been selected due to their high correlation to filterbank features. The features are extracted from 100x100 pixel blocks centred on the speaker's mouth and a 2-D DCT applied. The 24 lowest order coefficients, selected in a zigzag fashion, form the visual feature vector, \mathbf{v}_t . The initial visual frame rate was 25 vectors per second and this was upsampled to 100 vectors per second to equal the audio frame rate.

2.2 Wiener filter

The Wiener filter is first implemented in the filterbank domain, $W^{FB}(i)$, as,

$$W^{FB}(i) = \frac{\hat{x}(i)}{y(i)} \quad 0 \leq i \leq I-1 \quad (2)$$

$\hat{x}(i)$ is the i^{th} channel of the clean filterbank estimated from visual features, with its computation discussed in section 3. $y(i)$ is the i^{th} channel of the filterbank computed from the noisy input speech. For speech enhancement, the 23-dimensional filterbank-domain Wiener filter is transformed into a 128 bin power spectral-domain Wiener filter, $W(f)$, using cubic spline interpolation. The Wiener filter is applied to the power spectrum, $|Y(f)|^2$, of the noisy speech to give an enhanced power spectrum estimate, $|\hat{X}(f)|^2$,

$$|\hat{X}(f)|^2 = |Y(f)|^2 W(f) \quad (3)$$

The power spectrum estimate is combined with the noisy phase, $\angle Y(f)$, and an inverse Fourier transform used to obtain a window of time-domain samples. Overlapping and adding of these windows produces the enhanced time-domain waveform. A critical stage in implementing the visually-

derived Wiener filter is obtaining clean filterbank estimates from visual features and this is discussed in the next section.

3. Clean filterbank estimation

Estimation of clean filterbank vectors from 2-D DCT visual vectors is achieved by modelling the joint density of the audio-visual feature vector space. This begins by defining an audio-visual feature vector, \mathbf{z}_t , as,

$$\mathbf{z}_t = [\mathbf{x}_t, \mathbf{v}_t] \quad (4)$$

where \mathbf{x}_t is the filterbank vector extracted from clean speech at the same time frame, t , as the 2-D DCT visual vector, \mathbf{v}_t . To model the joint density of audio and visual vectors, two approaches have been considered. The first models the audio-visual feature space globally using a single Gaussian mixture model (GMM). To improve the modelling, based on the correlation analysis in [9], the second method models the audio-visual correlation using a set of phoneme-specific GMMs that are linked together using a network of hidden Markov models (HMMs). This leads to a two-stage method of estimating clean filterbank vectors from visual vectors. First, an audio-visual speech recogniser is employed to identify the particular phoneme being spoken. Second, a phoneme-specific Gaussian mixture model (GMM), trained on audio-visual feature vectors specific to that phoneme, is used to make a MAP estimate of the clean filterbank vector from an input visual vector.

3.1 Global GMM

From a training database of joint feature vectors, expectation-maximization (EM) clustering [11] is used to create a GMM that comprises K clusters which localize the correlation between filterbank and visual vectors in the joint feature vector space,

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

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,

$$\boldsymbol{\mu}_k = \begin{bmatrix} \boldsymbol{\mu}_k^x \\ \boldsymbol{\mu}_k^y \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_k = \begin{bmatrix} \boldsymbol{\Sigma}_k^{xx} & \boldsymbol{\Sigma}_k^{xy} \\ \boldsymbol{\Sigma}_k^{yx} & \boldsymbol{\Sigma}_k^{yy} \end{bmatrix} \quad (6)$$

The mean vectors have two components; the mean of the filterbank vector and the mean of the visual vector. The covariance matrices comprise four components; the covariance matrix of the filterbank vectors, $\boldsymbol{\Sigma}_k^{xx}$, the covariance matrix of the visual vectors, $\boldsymbol{\Sigma}_k^{yy}$, and the covariances of the filterbank and visual vectors, $\boldsymbol{\Sigma}_k^{yx}$ and $\boldsymbol{\Sigma}_k^{xy}$.

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

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

which can be expressed as,

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

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 visual vector coming from each cluster to give a weighted MAP estimate of the filterbank vector,

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

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

$$h_k(\mathbf{v}_i) = \frac{\alpha_k p(\mathbf{v}_i | c_k^{\mathbf{v}})}{\sum_{k=1}^K \alpha_k p(\mathbf{v}_i | c_k^{\mathbf{v}})} \quad (10)$$

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

3.2 Phoneme-dependent HMM-GMM

This second method of 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. HMM decoding is applied to find the optimal sequence of HMMs, from which appropriate GMMs are selected for filterbank estimation.

3.2.1 Audio-visual HMM-based phoneme decoding

To localise the region in the audio-visual feature space from where estimation of the clean filterbank is made, a network of phoneme-based audio-visual HMMs are used. Each HMM comprises two-streams with one stream modelling audio features and the other stream modelling visual features. This allows an integration of the audio and visual features to be made with their respective contributions chosen according to the local SNR thereby increasing noise robustness. The audio-visual features are based on those defined in equation (4) but with the filterbank component transformed into an MFCC vector. The audio and visual vectors are also augmented by their velocity and acceleration derivatives. From hand annotation of the training part of the speech database (described in section 4) a set of $W=36$ three-state diagonal covariance matrix monophone HMMs and a three-state diagonal covariance matrix silence HMM are trained to give a set of HMMs, $\Lambda = [\lambda_0, \dots, \lambda_w, \dots, \lambda_W]$.

When decoding a sequence of input audio-visual feature vectors into a set of phonemes, the signal-to-noise ratio is used to adjust the HMM observation probability contributions from the audio and visual streams as,

$$b_j(\mathbf{z}_t) = b_j^{\mathbf{x}}(\mathbf{x}_t)^{\gamma(SNR_t)} b_j^{\mathbf{v}}(\mathbf{v}_t)^{1-\gamma(SNR_t)} \quad (11)$$

$b_j(\mathbf{z}_t)$ is the observation probability of the audio-visual vector, \mathbf{z}_t , in state j . $b_j^{\mathbf{x}}(\mathbf{x}_t)$ and $b_j^{\mathbf{v}}(\mathbf{v}_t)$ are the observation probabilities from the audio and visual streams respectively and $\gamma(SNR_t)$ is a nonlinear function that maps the SNR into a

weight in the range 0 to 1. At low SNRs, $\gamma(SNR_t)$ approaches zero which reduces the observation probability contribution made by the audio features. Specific details are given in [2].

3.2.2 GMM training

Using the audio-visual speech recogniser described in section 3.2.1, forced Viterbi decoding is applied to all training data utterances to determine the phoneme allocation for each audio-visual vector. Therefore, for a training data utterance $\mathbf{Z} = [\mathbf{z}_0, \dots, \mathbf{z}_t, \dots, \mathbf{z}_{T-1}]$ a model allocation $\mathbf{m} = [m_0, \dots, m_t, \dots, m_{T-1}]$ is computed that indicates the model, m_t , that the t^{th} feature vector is allocated. Using the model allocation for every vector in the training database, \mathbf{Z} , a set of phoneme specific audio-visual vector pools, Ω_w are created

$$\Omega_w = \{ \mathbf{z}_t \in \mathbf{Z} : m_t = w \} \quad (12)$$

From each pool a phoneme-specific GMM, $\Phi_w(\mathbf{z})$, is trained using expectation-maximisation (EM) clustering to model the local joint density of audio-visual vectors

$$\Phi_w(\mathbf{z}) = \sum_{k=1}^K \alpha_{k,w} \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_{k,w}, \boldsymbol{\Sigma}_{k,w}) \quad (13)$$

Each of the K clusters in the GMM for phoneme w is represented by a prior probability $\alpha_{k,w}$, a mean audio-visual vector, $\boldsymbol{\mu}_{k,w}$, and an audio-visual covariance matrix, $\boldsymbol{\Sigma}_{k,w}$. The mean vectors have two components; the mean of the filterbank vector and the mean of the 2-D DCT visual vector. The covariance matrices comprise four components; the covariance of the filterbank vectors, $\boldsymbol{\Sigma}_{k,w}^{\mathbf{xx}}$, the covariance of the 2-D DCT vectors, $\boldsymbol{\Sigma}_{k,w}^{\mathbf{vv}}$, and the cross-covariances of the filterbank and 2-D DCT vectors, $\boldsymbol{\Sigma}_{k,w}^{\mathbf{xv}}$ and $\boldsymbol{\Sigma}_{k,w}^{\mathbf{vx}}$.

3.2.3 HMM-GMM filterbank estimation

To estimate clean filterbank vectors from a sequence of input audio-visual vectors $[\mathbf{z}_0, \mathbf{z}_1, \dots, \mathbf{z}_{T-1}]$ first the sequence of phonemes, and hence the sequence of phoneme-dependent GMMs, is determined. This is achieved by decoding the audio-visual vectors into a model sequence, $\mathbf{m} = [m_0, m_1, \dots, m_{T-1}]$ using the network of audio-visual HMMs. This provides, for each visual vector, \mathbf{v}_t , a phoneme-specific GMM, Φ_{m_t} , from where the filterbank will be estimated.

From the 2-D DCT visual vector, \mathbf{v}_t , a MAP estimate of the filterbank vector, $\hat{\mathbf{x}}_t$ can be made from the k^{th} cluster in the associated GMM, Φ_{k,m_t} , as

$$\hat{\mathbf{x}}_t = \arg \max_{\mathbf{x}_i} \left(p(\mathbf{x}_i | \mathbf{v}_t, \Phi_{k,m_t}) \right) \quad (14)$$

which can be expressed as,

$$\hat{\mathbf{x}}_t = \boldsymbol{\mu}_{k,m_t}^{\mathbf{x}} + \boldsymbol{\Sigma}_{k,m_t}^{\mathbf{xv}} \left(\boldsymbol{\Sigma}_{k,m_t}^{\mathbf{vv}} \right)^{-1} \left(\mathbf{v}_t - \boldsymbol{\mu}_{k,m_t}^{\mathbf{v}} \right) \quad (15)$$

Estimates from each of the K clusters in the GMM can be combined according to the posterior probability, $h_{k,m_t}(\mathbf{v}_t)$, of the 2-D DCT visual vector coming from each cluster to give a weighted MAP estimate of the filterbank vector,

$$\hat{\mathbf{x}}_t = \sum_{k=1}^K h_{k,m_t}(\mathbf{v}_t) \left(\boldsymbol{\mu}_{k,m_t}^{\mathbf{x}} + \boldsymbol{\Sigma}_{k,m_t}^{\mathbf{xv}} \left(\boldsymbol{\Sigma}_{k,m_t}^{\mathbf{vv}} \right)^{-1} \left(\mathbf{v}_t - \boldsymbol{\mu}_{k,m_t}^{\mathbf{v}} \right) \right) \quad (16)$$

The posterior probability, $h_{k,m_t}(\mathbf{v}_t)$, is given as,

$$h_{k,m_t}(\mathbf{v}_t) = \frac{\alpha_{k,m_t} p(\mathbf{v}_t | \Phi_{k,m_t}^v)}{\sum_{k=1}^K \alpha_{k,m_t} p(\mathbf{v}_t | \Phi_{k,m_t}^v)} \quad (17)$$

$p(\mathbf{v}_t | \Phi_{k,m_t}^v)$ is the marginal distribution of 2-D DCT vectors for the k^{th} cluster of the GMM associated with phoneme m_t .

3.3 Filterbank estimation accuracy

This section examines filterbank estimation accuracy from 2-D DCT visual vectors using global and phoneme-specific MAP estimation. Training and testing uses the training and test set parts of the database described in section 4.

3.3.1 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 3.1. Visual vectors were then extracted from the 77 test utterances and in combination with the GMM, 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,

$$E_{\%} = \frac{1}{NM} \sum_{i=0}^{N-1} \sum_{m=1}^M \frac{|x_i(m) - \hat{x}_i(m)|}{x_i(m)} \times 100\% \quad (18)$$

$x_i(m)$ and $\hat{x}_i(m)$ represent the clean and estimated amplitude of the m^{th} 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 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.

3.3.2 Phoneme-dependent HMM-GMM estimation

First the accuracy of the audio-visual speech recogniser in noise is examined as this provides the localization of MAP estimation to a particular phoneme region of the audio-visual feature space. Figure 2 shows phoneme classification accuracy of the audio-visual speech recogniser at SNRs from 0dB to 20dB in train noise and for clean speech.

Phoneme classification accuracy reduces from 60% in clean speech to 30% at an SNR of 0dB. At low SNRs the visual features make more contribution to classification than the audio features, which explains the convergence of accuracy to the level attained by the visual features only which is 28% [2].

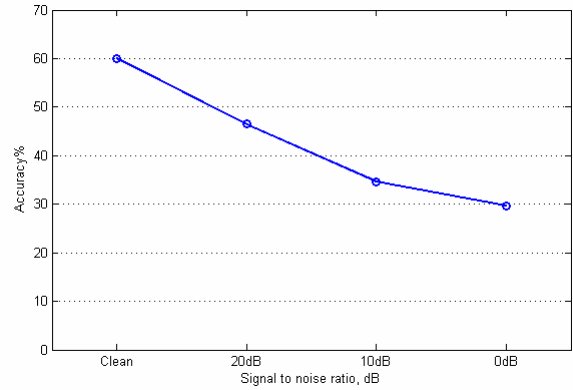


Figure 2: Audio-visual phoneme recognition accuracy from clean to 0dB in train noise

Table 1 shows percentage filterbank estimation error, $E_{\%}$, of the phoneme-specific MAP estimation system using from $K=1$ to 8 clusters in the GMMs with five different configurations analysed. The first column shows estimation error using forced alignment of the input utterance to the correct phoneme sequence to ensure that correct phoneme-specific GMMs are selected for estimation. The remaining columns show the results of tests using an unconstrained phoneme grammar when applied to clean speech and then speech contaminated by train noise at SNRs of 20dB, 10dB and 0dB. Phoneme classification accuracy of these configurations is as shown in figure 2.

	Forced	Clean	20dB	10dB	0dB
K=1	7.11	6.88	8.53	8.84	9.00
K=2	6.87	6.71	8.32	8.78	8.98
K=4	6.71	6.69	8.13	8.74	9.16
K=8	6.37	6.57	8.72	9.02	9.29

Table 2. Percentage filterbank estimation error for forced and unconstrained recognition on clean and noisy speech

In general, increasing the number of clusters in the GMMs reduces estimation error and is attributed to improved modelling of the joint density of audio and visual features. As is expected, lowest estimation errors are obtained when forced alignment provides the sequence of GMMs for filterbank estimation, with a filterbank error of 6.37% being achieved. Moving to unconstrained decoding in clean speech causes only a small change in estimation error – an increase of 0.20%. This is in spite of the phoneme classification accuracy falling from 100% to 60%. When the speech is contaminated by acoustic noise a more significant increase in estimation error is observed. However, it is important to note that the acoustic noise only affects the decoding process of the audio-visual recogniser and hence the selection of phoneme-specific GMMs used for estimation. It does not affect estimation of the filterbank vector once the phoneme-specific GMM has been selected as this uses the visual vector which is unaffected by noise. The results suggest that even though incorrect phoneme-specific GMMs may be selected, their effect on filterbank estimation accuracy is less than may be expected. For example, at 0dB only 30% of filterbank estimates are made from the correct phoneme-specific GMMs, but estimation error increases by only 2.61% over forced alignment where all estimates are made from the correct phoneme-specific GMM.

4. Speech enhancement results

The experiments in this section examine the effectiveness of the visually-derived Wiener filter for enhancing noisy speech. This is achieved through PESQ analysis and comparison with conventional audio-only methods of speech enhancement.

The experiments use an audio-visual speech database comprising 277 sentences of continuous speech spoken by a single male UK English speaker [12]. 200 utterances are used for training and 77 utterances for testing. The audio was sampled at a rate of 8kHz and subsequently processed at a frame rate of 100 audio vectors per second. The video was originally recorded at 25 frames per second and then upsampled to 100 frames per second to give a visual frame rate equal to the audio frame rate. Thirty-six phonemes occur in the database and hand annotation was used to obtain time-aligned phoneme labels for each utterance.

4.1 Listening tests

The experiments in this section use PESQ to measure the effectiveness of speech enhancement provided by the visually-derived Wiener filter. For comparison, two audio-based methods of speech enhancement are also investigated; the Berouti method of spectral subtraction [4] and a conventional audio-only method of Wiener filtering [5]. Implementation details of the two audio enhancement methods are as follows:

Berouti spectral subtraction – this is a nonlinear spectral subtraction method that employs a minimum spectral floor and oversubtraction of the noise based on the local SNR [4].

Audio-only Wiener filter – this implementation adopts a decision directed estimate of the a priori SNR which is used to define the Wiener filter [5].

The speech quality was measured by PESQ on the set of 77 test data files and an average computed across all 77 utterances. Experiments began by contaminating each of the 77 utterances with train noise at SNRs of 20dB, 10dB and 0dB. Average PESQ scores were computed for each of these SNRs, across all test data utterances, by comparing the noisy utterance to the noise-free version. This provided a set of no noise compensation baseline measures. Next, the three different speech enhancement methods were applied and the resulting enhanced speech from each method compared to the noise-free version to provide PESQ scores. Averaged PESQ scores are shown in table 3 for no noise compensation and for the three speech enhancement methods at SNRs of 20dB, 10dB and 0dB.

SNR	No noise comp.	Visual Wiener	Audio Wiener	Spectral subtraction
20dB	2.81	2.88	3.18	2.97
10dB	2.36	2.71	2.22	2.39
0dB	1.95	2.46	1.33	1.89

Table 3. PESQ scores comparing visually-derived Wiener filtering, spectral subtraction and audio-Wiener

With no noise compensation the PESQ scores show a reduction in speech quality as SNR decreases. The visually-derived Wiener filter improves the PESQ scores significantly, particularly at lower SNRs where an improvement of 0.51 is achieved at 0dB. Spectral subtraction operates best at higher SNRs so it outperforms the visually-derived Wiener filter only at 20dB. The audio-only Wiener filter achieves the highest overall PESQ score of 3.18 at an SNR of 20dB.

However, at 0dB, the PESQ score is worse than with no noise compensation. These results show the visually-derived Wiener filter to be significantly more effective at SNRs of 10dB and 0dB than the audio-based methods of enhancement. This may be attributed to more robust estimation of clean speech from visual features than using audio.

Figure 3b shows a spectrogram of the utterance “Sarah argued that I acted as though under his thumb” contaminated with train noise at an SNR of 10dB. Figure 3c shows the same utterance after the application of visually-derived Wiener filtering using the HMM-GMM method of clean filterbank estimation. The Wiener filtering can be seen to have removed large amounts of the noise present in figure 3b. This is particularly evident in non-speech periods, but good noise reduction also occurs in speech periods.

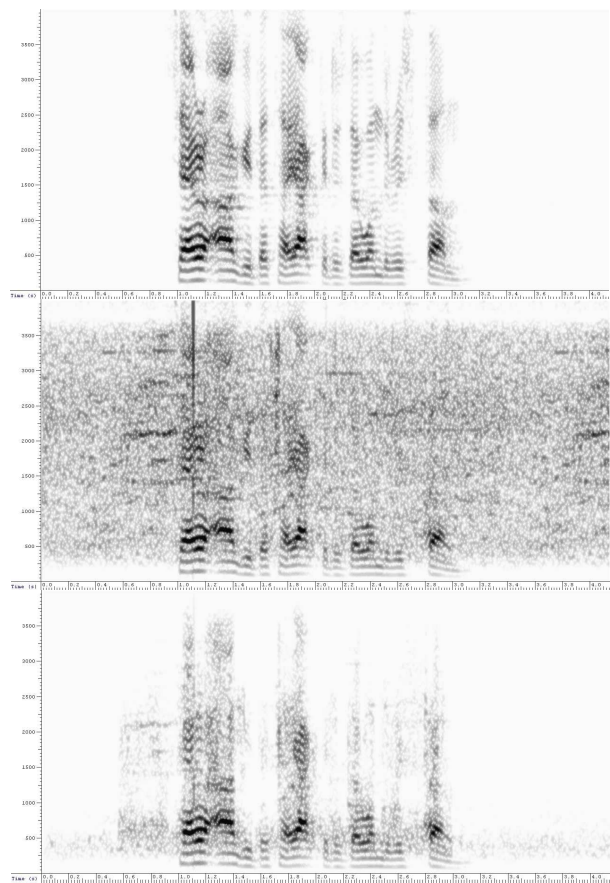


Figure 3: Spectrogram of utterance, a) clean speech, b) contaminated with train noise at an SNR of 10dB, c) after visually-derived Wiener filtering

5. Conclusion

This work has shown that visual speech information can provide estimates of clean speech that can be used to enhance speech within a Wiener filter framework. Estimates of clean speech are made from visual speech using both global and phoneme-specific MAP estimation. Experiments reveal the MAP estimation to be robust to the selection of phoneme from which estimation is made. A comparison of PESQ scores with conventional audio-based enhancement methods shows the visually-derived Wiener filter to perform well, particularly at low SNRs.

6. References

- [1] J. Luettin, G. Potamianos, and C. Neti, "Asynchronous stream modeling for large vocabulary audio-visual speech recognition," in Proc. ICASSP, 2001
- [2] I. Almajai, "Audio-visual speech recognition", PhD upgrade report, University of East Anglia, UK, 2006
- [3] I. Almajai, B.P. Milner and J. Darch, "Analysis of correlation between audio and visual features for clean audio feature prediction in noise", Proc. Interspeech, 2006
- [4] M. Berouti, R. Schwartz, J. Makhoul, "Enhancement of speech corrupted by acoustic noise", Proc. ICASSP, pp. 208-211, 1979
- [5] P. Scalart and J. Vieira-Filho, "Speech enhancement based on a priori signal to noise estimation", Proc. ICASSP, pp. 629-632, 1996
- [6] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," IEEE Trans. SAP, vol. 9, no. 5, pp. 504-512, 2001
- [7] H. Yehia, P. Rubin and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behaviour", Speech Communication, 26(1):23-43, 1998
- [8] A. Sorin and T. Ramabadran, "Extended advanced front end algorithm description, Version 1.1", ETSI STQ Aurora DSR Working Group, Tech. Rep. ES202212, 2003
- [9] I. Almajai and B.P. Milner "Maximising Audio-Visual Speech Correlation", Proc. AVSP, 2007-06-04
- [10] T.F. Cootes, G.J. Edwards and C.J. Taylor, "Active appearance models", IEEE Trans. PAMI, vol. 23, no. 6, pp. 681-685, 2001
- [11] C.W. Therrien, Discrete random signals and statistical signal processing, Prentice-Hall, Englewood Cliffs, NJ, 1992
- [12] B. Theobald. "Visual speech synthesis using shape and appearance models," PhD Thesis, University of East Anglia, Norwich, UK, 2003