INTERSPEECH 2009 BRIGHTON

Enhancing Audio Speech using Visual Speech Features

Ibrahim Almajai and Ben Milner

School of Computing Sciences, University of East Anglia, Norwich, UK

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

Abstract

This work presents a novel approach to speech enhancement by exploiting the bimodality of speech and the correlation that exists between audio and visual speech features. For speech enhancement, a visually-derived Wiener filter is developed. This obtains clean speech statistics from visual features by modelling their joint density and making a maximum a posteriori estimate of clean audio from visual speech features. Noise statistics for the Wiener filter utilise an audio-visual voice activity detector which classifies input audio as speech or nonspeech, enabling a noise model to be updated. Analysis shows estimation of speech and noise statistics to be effective with human listening tests measuring the effectiveness of the resulting Wiener filter.

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

10.21437/Interspeech.2009-576

1. Introduction

Modern communication devices enable not just an audio signal to be captured from a speaker but also a visual signal. The visual signal provides both additional information that is not present in the audio and also a visual representation of some of the information that is present in the audio. One of the main advantages of this visual speech is that it is unaffected by acoustic noise. This has led to speech recognition systems that combine audio and visual speech features to achieve robust performance in noise [1]. However, the work presented in this paper moves away from audio-visual speech recognition and instead proposes an audio-visual method of speech enhancement.

Many audio-only methods of speech enhancement have been proposed [2, 3]. However, this work aims to exploit clean speech information that is available in visual features to create a novel method of speech enhancement. This relies on correlation existing between 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 speech and visual shape of the mouth [4, 5]. Therefore, from knowledge of mouth shape, information regarding the audio speech being produced can be inferred. Of course, spectrally detailed audio signals cannot be estimated from mouth shape (source information is not present in lip shape) but a spectral envelope can be estimated.

Several studies have examined previously the use of visual speech features for audio enhancement, particularly within a Wiener filtering framework [6, 7, 8]. One of the main problems in Wiener filtering is obtaining clean speech statistics and these methods exploit visual features to obtain them. In this work, previous work is extended by first deriving the Wiener filter more formally by making separate estimates of the clean speech statistics and noise statistics directly from audio-visual features. Noise statistics are estimated by employing an audio-visual voice activity detector that identifies nonspeech frames that update a noise model. Clean speech estimation uses a network of phoneme-specific estimators that provide clean speech statistics from audio-visual features. A detailed analysis of the resulting speech quality is also made through a series of human listening tests that explicitly measure speech distortion, noise intrusiveness and overall speech quality.

The remainder of this work begins in section 2 with a description of the proposed visually-derived Wiener filter. For implementation, this requires both a clean speech estimate and a noise estimate. Section 3 describes how the clean speech estimate is obtained from visual speech vectors. Section 4 introduces an audio-visual VAD which is used to obtain the noise estimate. Experimental results are presented in section 5 which first examine the accuracy of clean speech and noise estimation. Secondly, the effectiveness of the visually-derived Wiener filter is analysed through of a series of human listening tests.

2. Visually-derived Wiener filter

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

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

Two challenges in Wiener filtering are to obtain: i) – the clean speech power spectrum $P_{XX}(f)$, and ii) – the noise power spectrum $P_{NN}(f)$. Obtaining the clean speech power spectrum is one of the main problems in Wiener filtering and many methods have been proposed to achieve this [3]. This work proposes estimating the clean speech statistics from visual speech features by exploiting the audio-visual correlation of speech. Estimating a clean speech power spectrum from visual features is difficult due to the limit of audio-visual correlation, so instead a filterbank-domain Wiener filter is estimated, $\hat{W}_t^{FB}(i)$

$$\hat{W}_t^{FB}(i) = \frac{\hat{a}_t(i)}{\hat{a}_t(i) + \hat{n}_t(i)} \tag{2}$$

$\hat{a}_t(i)$ is the ith channel of the tth clean speech filterbank, estimated from visual features, and $\hat{n}_t(i)$ is the ith channel of the noise filterbank estimate. The filterbank-domain noise estimate is obtained by averaging filterbank frames from periods of nonspeech, which are identified by an audio-visual VAD [9].

For speech enhancement, the I-dimensional filterbank-domain Wiener filter, $\hat{W}_t^{FB}(i)$, is transformed into a 128-dimensional power spectral-domain Wiener filter, $\hat{W}_t(f)$, using cubic spline interpolation. The Wiener filter is applied to the power spectrum of the noisy input signal, $|Y_t(i)|^2$, to give an enhanced power spectrum estimate, $|\hat{X}_t(i)|^2$

$$|\hat{X}_t(f)|^2 = |Y_t(f)|^2 \hat{W}_t(f) \tag{3}$$

The power spectrum estimate is combined with the phase of the noisy speech, $\angle Y_t(f)$, and an inverse Fourier transform applied to obtain a frame of time-domain samples. Overlap and adding of these frames gives the enhanced time-domain waveform.

The next two sections describe estimation of the clean speech and noise filterbanks required by the Wiener filter.

3. Clean speech filterbank estimation

Estimation of clean filterbank vectors from visual vectors is achieved by modelling the joint density of audio-visual features. This begins by defining an I+J-dimensional audio-visual feature vector, $\boldsymbol{z}_t$

$$\boldsymbol{z}_t = [\boldsymbol{a}_t, \boldsymbol{v}_t] \tag{4}$$

$\boldsymbol{a}_t$ and $\boldsymbol{v}_t$ are the audio and visual features, which in this work are log filterbank audio vectors and 2-D DCT visual features, as these were found to exhibit high levels of audio-visual correlation [4]. Given a model of the joint density of the audio-visual vectors, Φ^{av}, a MAP estimate of the audio vector, $\hat{\boldsymbol{a}}_t$, can be made from a visual vector, $\boldsymbol{v}_t$

$$\hat{\boldsymbol{a}}_t = \arg\max_{\boldsymbol{a}} \left(p\left(\boldsymbol{a} | \boldsymbol{v}_t, \Phi^{av}\right)\right) \tag{5}$$

Previous work established that audio-visual correlation was maximised when measured within individual phonemes rather than measured across all speech [4]. Therefore, the MAP estimation is constrained to make a localised estimate from a set of phoneme-specific models of the joint density of audio-visual features. Localising prediction in this way is achieved by a network of audio-visual HMMs which first decode input audio-visual vectors into a phoneme sequence, from which estimates of clean audio features are made.

3.1. Modelling phoneme-specific audio-visual vectors

Phoneme-specific joint densities of audio and visual vectors are modelled by first creating a set of monophone HMMs. These are trained on audio-visual vectors, based on equation 4, but with the filterbank component transformed into an MFCC vector augmented by velocity and acceleration. A set of W=36 3-state, 16-mode diagonal covariance matrix monophone HMMs are created, together with a 3-state silence HMM.

Using the set of phoneme HMMs, forced Viterbi decoding is applied to each training data utterance to determine the phoneme allocation for each audio-visual feature vector. Therefore, for a training data utterance, $\boldsymbol{Z}^{'} = [\boldsymbol{z}_0^{'}, ..., \boldsymbol{z}_t^{'}, ..., \boldsymbol{z}_{T-1}^{'}]$ (the dash indicates the transform of the filterbank to an MFCC vector and inclusion of temporal derivatives), a phoneme allocation, $\boldsymbol{m} = [m_0, ..., m_t, ..., m_{T-1}]$ is computed that indicates the phoneme, m_t, that the tth feature vector is allocated, where $m_t \in \{0, ..., W\}$. Using the phoneme allocation for each feature vector in the training database, Z, a set of phoneme specific audio-visual vector pools, Ω_w, are created for each phoneme w

$$\Omega_w = \{\boldsymbol{z}_t \in Z : m_t = w\} \tag{6}$$

For the audio-visual vectors in each phoneme pool, expectation-maximisation (EM) clustering is applied to create a K cluster Gaussian mixture model (GMM), Φ_w^{av}, that models the joint density of the audio-visual vectors for phoneme w,

$$\Phi_w^z(\boldsymbol{z}) = \sum_{k=1}^{K} \alpha_k \phi_{k,w}^z = \sum_{k=1}^{K} \alpha_k \, \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_k^z, \boldsymbol{\Sigma}_k^z) \tag{7}$$

The kth cluster is represented by a prior probability, α_k, Gaussian probability density function (PDF), $\phi_{k,w}^z$ with mean vector, $\boldsymbol{\mu}_k^z$, and covariance matrix, $\boldsymbol{\Sigma}_k^z$

$$\boldsymbol{\mu}_{k,w}^z = \begin{bmatrix} \boldsymbol{\mu}_{k,w}^a \\ \boldsymbol{\mu}_{k,w}^v \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_{k,w}^{zz} = \begin{bmatrix} \boldsymbol{\Sigma}_{k,w}^{aa} & \boldsymbol{\Sigma}_{k,w}^{av} \\ \boldsymbol{\Sigma}_{k,w}^{va} & \boldsymbol{\Sigma}_{k,w}^{vv} \end{bmatrix} \tag{8}$$

Each mean vector comprises an I-dimensional mean filterbank audio vector, $\boldsymbol{\mu}_k^a$, and a J-dimensional mean 2-D DCT visual vector, $\boldsymbol{\mu}_k^v$. The covariance matrices comprises four components; the IxI-dimensional covariance matrix of the audio vector, $\boldsymbol{\Sigma}_k^{aa}$, the JxJ-dimensional covariance of the visual vector, $\boldsymbol{\Sigma}_k^{vv}$, and the IxJ and JxI-dimensional cross-covariances of the audio and visual vectors, $\boldsymbol{\Sigma}_k^{av}$ and $\boldsymbol{\Sigma}_k^{va}$.

3.2. Estimation of clean speech filterbank vectors

Estimating clean speech filterbank vectors, using the phoneme-specific GMMs, begins by using the audio-visual speech recogniser to determine the phoneme sequence from the input noisy audio and visual features. Within the audio-visual speech recogniser, the signal-to-noise ratio (SNR) is used to adjust the contribution made by the audio and visual observation probabilities

$$b_\rho(\boldsymbol{z}_t) = b_\rho^{\boldsymbol{a}}(\boldsymbol{x}_t)^{\gamma(SNR_t)} b_\rho^{\boldsymbol{v}}(\boldsymbol{v}_t)^{1-\gamma(SNR_t)} \tag{9}$$

$b_\rho(\boldsymbol{z}_t)$ is the observation probability of the audio-visual vector, $\boldsymbol{z}_t$, in state ρ. $b_\rho^{\boldsymbol{x}}(\boldsymbol{x}_t)$ and $b_\rho^{\boldsymbol{v}}(\boldsymbol{v}_t)$ 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 $0 \leq \gamma(SNR_t) \leq 1$. At low SNRs, $\gamma(SNR_t)$ approaches zero which reduces the contribution made by the audio features. Specific details are given in [10]. From an input sequence of audio-visual vectors, $\boldsymbol{Z} = [\boldsymbol{z}_0, ..., \boldsymbol{z}_t, ..., \boldsymbol{z}_{T-1}]$, a phoneme sequence $\boldsymbol{m} = [m_0, ..., m_t, ..., m_{T-1}]$ is computed using an uncontrained phoneme grammar which provides, for each audio-visual vector, $\boldsymbol{z}_t$, the phoneme-specific GMM, Φ_{m_t}, from where the clean filterbank vector, $\hat{\boldsymbol{a}}_t$, will be estimated. A weighted MAP estimate of the filterbank vector, $\hat{\boldsymbol{a}}_t$, is made using each of the K clusters, $\phi_{k,m_t}^{\boldsymbol{z}}$, in the GMM, $\Phi_{k,m_t}^{\boldsymbol{z}}$

$$\hat{\boldsymbol{a}}_t = \sum_{k=1}^{K} h_{k,m_t}(\boldsymbol{v}_t) \arg\max_{\boldsymbol{a}_t} \left\{ p\left(\boldsymbol{a}_t | \boldsymbol{v}_t, \phi_{k,w}^{\boldsymbol{z}}\right)\right\} \tag{10}$$

$h_{k,m_t}(\boldsymbol{v}_t)$ is the posterior probability of the visual vector, $\boldsymbol{v}_t$, in the kth cluster of the GMM associated with phoneme m_t

$$h_{k,m_t}(\boldsymbol{v}_t) = \frac{\alpha_{k,m_t} \; p\left(\boldsymbol{v}_t | \phi_{k,m_t}^{\boldsymbol{v}}\right)}{\sum_{k=1}^{K} \alpha_{k,m_t} \; p\left(\boldsymbol{v}_t | \phi_{k,m_t}^{\boldsymbol{v}}\right)} \tag{11}$$

where $p(\boldsymbol{v}_t | \phi_{k,m_t}^{\boldsymbol{v}})$ is the marginal distribution of visual vectors for the kth cluster in the GMM specific to phoneme m_t.

4. Noise filterbank estimation

The noise filterbank estimate for the Wiener filter is obtained by averaging nonspeech filterbank vectors preceeding speech. To identify these nonspeech vectors an audio-visual VAD is proposed. This classifies input vectors as speech or nonspeech using a pair of GMMs, one modelling speech and the other nonspeech. Audio-visual vectors from a set of training data, Z, are

first pooled into two sets, one corresponding to speech, $\Psi^{s,z}$, and the other nonspeech, $\Psi^{ns,z}$

$$\begin{aligned}\Psi^{s,z} &= \{\boldsymbol{z}_t \in Z : c_t = speech\} \\ \Psi^{ns,z} &= \{\boldsymbol{z}_t \in Z : c_t = nonspeech\}\end{aligned} \tag{12}$$

c_t is a reference label associated with each feature vector and indicates whether the signal is speech or nonspeech. From the two vector pools, EM clustering is applied to create two GMMs each comprising D clusters, one modelling speech, $\Theta^{s,z}$, and the other modelling nonspeech, $\Theta^{ns,z}$.

Classification of audio-visual vectors as speech or nonspeech could be made from the GMM probabilities directly. However, in noisy speech, audio features becomes less reliable which increases classification errors. To improve robustness, the GMMs are decomposed in a similar way to the audio-visual observation probabilities in equation 9, and the relative contribution of the audio and visual components adjusted by a function, β, of the SNR. Thus the scaled probabilities of an audio-visual vector, $\boldsymbol{z}_t$ being speech or nonspeech can be computed

$$\begin{aligned}p(\boldsymbol{z}_t|s) &= \Theta^{s,a}(\boldsymbol{a}_t)^{\beta(SNR_t)}\Theta^{s,v}(\boldsymbol{v}_t)^{1-\beta(SNR_t)} \\ p(\boldsymbol{z}_t|ns) &= \Theta^{ns,a}(\boldsymbol{a}_t)^{\beta(SNR_t)}\Theta^{ns,v}(\boldsymbol{v}_t)^{1-\beta(SNR_t)}\end{aligned} \tag{13}$$

The terms $\Theta^{s,a}$ and $\Theta^{ns,a}$ represent the audio components of the speech and nonspeech GMMs, while $\Theta^{s,v}$ and $\Theta^{ns,v}$ represent the visual components. Classification of an audio-visual vector, $\boldsymbol{z}_t$, as speech or nonspeech can then be made from the two scaled GMM probabilities

$$\hat{c}_t = \begin{cases} speech & p(\boldsymbol{z}_t|s) \geq p(\boldsymbol{z}_t|ns) \\ nonspeech & p(\boldsymbol{z}_t|s) < p(\boldsymbol{z}_t|ns) \end{cases} \tag{14}$$

For audio-visual vectors labelled as nonspeech, the filterbank components are extracted and averaged to provide the filterbank-domain noise estimate, $\hat{\boldsymbol{n}}_t$, used in the Wiener filter of equation 2. The SNR estimate in equation 13 uses a noise estimate taken from the first few frames of the signal which are assumed noise. The noise estimate produced by the audio-visual VAD is used to provide the SNR estimate in equation 9.

5. Experimental results

The experiments first examine the accuracy of clean filterbank estimation and noise filterbank estimation, which are used by the Wiener filter. The quality of the enhanced speech from the Wiener filter is then analysed using human listening tests. For all experiments a set of 277 utterances, spoken by a UK male speaker have been used, with 200 utterances used for training and 77 for testing, providing $T = 38,728$ test vectors. The audio was sampled at a rate of 8kHz and processed at a rate of 100 vectors per second. The video was originally recorded at 25 frames per second and was upsampled to 100 vectors per second to be equal to the audio frame rate. A total of 36 phonemes occur in the database which was manually aligned to provide phoneme reference labels for each utterance.

5.1. Clean speech and noise filterbank estimation

This section examines the accuracy of clean speech filterbank estimation and noise filterbank estimation from audio-visual speech features. Estimation accuracy is measured using the root mean square (RMS) error. As noise is added artificially in these experiments, reference clean speech and noise values are available for the RMS error calculation. Figure 1 shows clean filterbank and noise filterbank RMS errors at SNRs of 20dB, 10dB and 0dB in car noise.

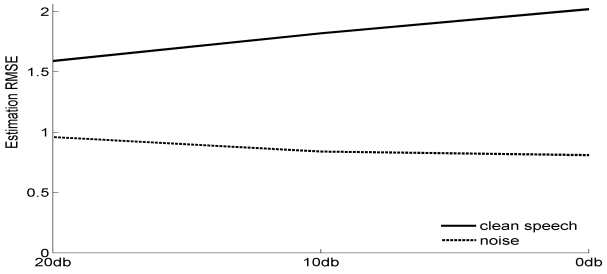

Figure 1: *RMS speech and noise filterbank estimation error*

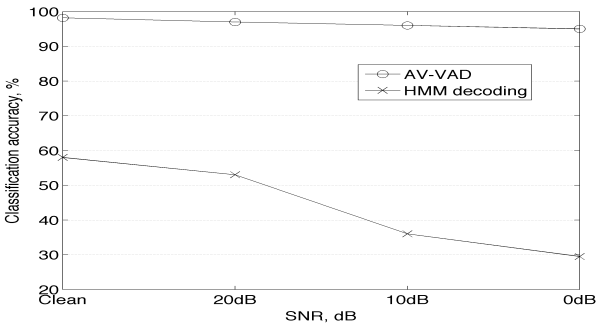

Figure 2: *VAD classification accuracy and HMM phoneme decoding accuracy in clean speech and noisy speech*

Clean speech filterbank estimation is relatively stable, even in low SNRs. This is attributed to significant use of the visual component of the audio-visual vector being used in clean speech filterbank estimation (equation 10) which is independent of noise. Audio features are used in the decoding to determine the phoneme sequence used in estimation, but the effect of noise is minimised by the SNR-dependent adjustment of the audio and visual observation probability components (equation 9). To illustrate this, figure 2 shows audio-visual phoneme decoding accuracy, which reduces from 58% in clean conditions to 30% at 0dB. As SNRs reduce, the visual features make more contribution than audio features and hence performance stabilises at the visual-only performance of 30%.

Figure 1 also shows noise filterbank estimation to be stable across SNRs with lower RMS error than clean speech filterbank estimation. The noise estimates are obtained by averaging audio frames identified as being noise-only by the AV-VAD, thus making VAD accuracy important in noise estimation. Figure 2 shows AV-VAD accuracy to reduce from 98% in clean speech to 95% in noisy speech, which is equal to visual-only VAD performance. As with phoneme decoding, the lower limit on performance is visual-only performance, which avoids the typical breakdown in performance that audio features suffer in noise.

5.2. Speech enhancement

This section uses a series of human listening tests to analyse speech quality following visually-derived Wiener filtering. Test set utterances were first contaminated by car noise at SNRs of 20dB, 10dB and 5dB. The speech was then passed through

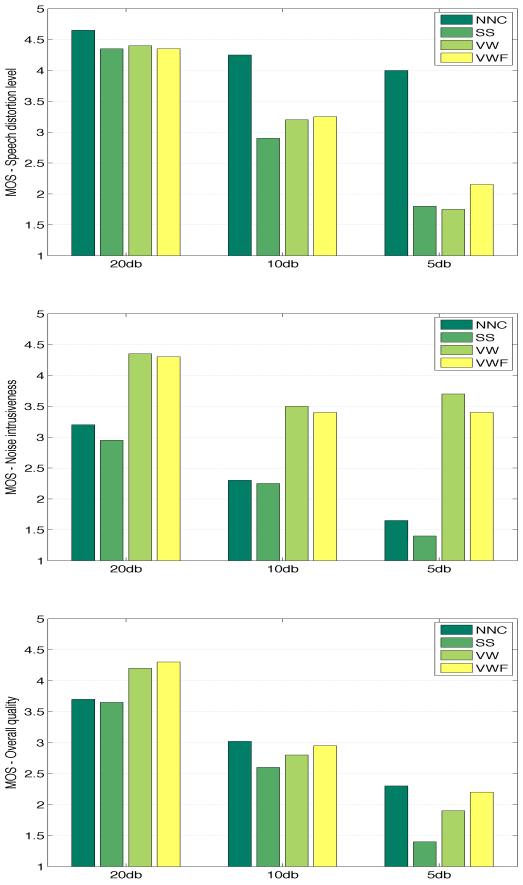

Figure 3: MOS for a) speech distortion, b) noise intrusiveness and c) overall quality, for: no noise compensation (NNC), spectral subtraction (SS), visual Wiener filter (VW) and visual Wiener filter forced alignment (VWF).

the visually-derived Wiener filter for enhancement. For further analyse, the process was repeated but with the audio-visual vectors forced to the correct phoneme sequence for visually-derived Wiener filtering. As a comparison, the spectral subtraction method of speech enhancement was also applied to the speech [2]. Thirty listeners took part in the tests which were carried out in a soundproof room and each listener was played 60 speech utterances. The listeners were asked to make three ratings on a scale of 1 to 5 for each utterance: the level of signal distortion, the level of background noise intrusiveness, and the overall quality. For the three ratings a high score indicates a better signal [11]. Figure 3 shows Mean Opinion Scores of the three measures for no noise compensation (NNC), spectral subtraction (SS), visual Wiener filtering (VW) and visual Wiener filtering with forced phoneme alignment (VWF).

At 20dB, scores for signal distortion rate no noise compensation as being the most undistorted with the noise compensation methods adding small amounts of distortion, leading to a reduction of about 0.3 points. In terms of noise intrusiveness, the visual Wiener methods perform equally well, scoring about 1.1 points higher than NNC. Spectral subtraction perform slightly worse and this is attributed to it introducing musical noise. Overall quality scores rate the two Wiener filtering methods about 0.5 points higher than NNC and spectral subtraction. However, as SNRs fall, the speech distortion imposed by the Wiener filtering increases, resulting in reduced scores. Noise removal by the Wiener filtering remains robust with only small variations even as SNRs reduce to 5dB. This is attributed to the robustness of the visual features. Overall quality also reduces as SNRs fall. For both the speech distortion and overall speech quality, the forced alignment Wiener filter (VWF) outperforms slightly the unconstrained decoding Wiener filter (VW) which is attributed to the more accurate clean filterbank estimates that derive the Wiener filter, although the gain in having 100% phoneme accuracy compared to accuracies between 60% and 30% (figure 2) is small and suggest the system is inherently robust to phoneme decoding errors.

6. Conclusion

A visually-derived Wiener filter has been proposed for speech enhancement. This utilises both audio-visual correlation and the robustness of visual features to noise, to provide estimates of the clean speech and noise statistics needed by the Wiener filter. RMS error analysis shows the estimation of both of these to be relatively robust to noise. Listening test analysis of the enhanced speech revealed the visually-derived Wiener filter to be effective at reducing noise levels, but at the expense of introducing distortion onto the speech signal during the filtering operation. However, this represents initial work at utilising visual speech information for enhancing audio speech and several directions to reduce this distortion can be identified, centring primarily on obtaining more accurate clean filterbank estimates from audio-visual features.

7. References

[1] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," in *Issues in Visual and Audio-visual Speech Processing*. MIT Press, 2004.

[2] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *ICASSP*, vol. 4, Washington, DC, USA, Apr. 1979, pp. 208–211.

[3] Y. Ephraim, D. Malah, and B.-H. Juang, "On the application of hidden markov models for enhancing noisy speech," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 37, pp. 1846–1856, Dec. 1989.

[4] 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.

[5] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal tract and facial behavior," *Speech Communication*, vol. 26, no. 1, pp. 23–43, 1998.

[6] L. Girin, J.-L. Schwartz, and G. Fang, "Audio-visual enhancement of speech in noise," *JASA*, vol. 109, no. 6, pp. 3007–3019, June 2001.

[7] F. Berthommier, "Characterization and extraction of mouth opening parameters available for audiovisual speech enhancement," in *ICASSP*, 2004.

[8] I. Almajai, B. Milner, J. Darch, and S. Vaseghi, "Visually-derived Wiener filters for speech enhancement," in *ICASSP*, vol. 4, Honolulu, Hawaii, USA, Apr. 2007, pp. 585–588.

[9] I. Almajai and B. Milner, "Using audio-visual features for robust voice activity detection in clean and noisy speech," *Proc. EUSIPCO*, 2008.

[10] I. Almajai, "Audio-visual speech recognition," 2006, mPhil to PhD upgrade report, University of East Anglia.

[11] ITU, "P.835: Subjective test methodology for evaluating speech communication systems that include noise suppression algorithms," 2003.