16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

# USING AUDIO-VISUAL FEATURES FOR ROBUST VOICE ACTIVITY DETECTION IN CLEAN AND NOISY SPEECH

*Ibrahim Almajai and Ben Milner*

School of Computing Sciences, University of East Anglia, UK
Phone: +44 1603 593220, fax: +44 1603 593345, email: i.almajai@uea.ac.uk
Phone: +44 1603 593339, fax: +44 1603 593345, email: b.milner@uea.ac.uk
Web: http://fizz.cmp.uea.ac.uk/Research/speechgroup/

## ABSTRACT

*The aim of this work is to utilize both audio and visual speech information to create a robust voice activity detector (VAD) that operates in both clean and noisy speech. A statistical-based audio-only VAD is developed first using MFCC vectors as input. Secondly, a visual-only VAD is produced which uses 2-D discrete cosine transform (DCT) visual features. The two VADs are then integrated into an audio-visual VAD (AV-VAD). A weighting term is introduced to vary the contribution of the audio and visual components according to the input signal-to-noise ratio (SNR). Experimental results first establish the optimal configuration of the classifier and show that higher accuracy is obtained when temporal derivatives are included. Tests in white noise down to an SNR of -20dB show the AV-VAD to be highly robust with accuracy remaining above 97%. Comparison with the ETSI Aurora VAD shows the AV-VAD to be significantly more accurate.*

## 1. INTRODUCTION

Voice activity detection (VAD) is the process of distinguishing speech from non-speech and as such VADs are found in many different speech processing applications [1]. Some of the most common uses of VADs are in noise estimation and reduction. For noise estimation, the VAD is used to identify regions of non-speech from which estimates of noise are computed. In noise reduction, the VAD identifies speech regions which are then filtered to remove the noise. Traditionally, VADs extract features from the input audio and use these features to determine whether or not speech is present. Typical VAD features include signal energy, spectral tilt and zero crossing rate. VAD accuracy can be very high in clean speech, but at lower signal-to-noise ratios (SNRs) accuracy deteriorates significantly. At these lower SNRs the majority of errors incurred are when noise falsely causes the VAD to classify non-speech as speech.

In this work it is proposed to improve VAD robustness in noise by extracting from the speaker both audio and visual speech features and using this joint information for classification. Visual speech features have the significant advantage that they are not distorted by acoustic noise and have been used in a visual-only VAD previously which gave good accuracy [2]. However, a disadvantage with visual speech features is that they are not as discriminative, in terms of speech or sound classes, as audio speech features. This fact is reflected in the lower speech recognition accuracies obtained for visual speech features than for audio speech features in audio-visual speech recognition [3]. Therefore, the aim of the proposed audio-visual VAD (AV-VAD) in this work is to exploit the strengths of both the audio and visual speech features by varying their contribution to speech/non-speech classification according to the SNR. At high SNRs more emphasis in the VAD will be given to the audio features. At lower SNRs the visual features will make more contribution as they become more discriminative than the audio features which are distorted by the acoustic noise.

The range of applications that may benefit from audio-visual VADs is increasing as the availability of cheap cameras becomes more widespread. Multimedia applications such as video conferencing, audio-visual speech recognition and visually-derived speech enhancement will be able to exploit the improved robustness of the proposed AV-VAD to give more accurate speech/non-speech classification [3][4].

The remainder of this paper is arranged as follows. Section 2 describes the audio and visual speech features used in the AV-VAD. The AV-VAD is explained in section 3, first in terms of audio-only and visual-only VADs before being integrated into an audio-visual-VAD. Section 4 presents experimental results that first determine the optimal feature and model configurations. Speech/non-speech classification results are then presented for both clean and noisy speech using both speaker-dependent and speaker-independent AV-VADs.

## 2. AUDIO AND VISUAL FEATURES

Many different audio and visual speech features have been proposed for use in audio-visual speech processing. Mel-frequency cepstral coefficients (MFCCs) are one of the most successful audio features used in speech recognition [5]. As such MFCCs have been selected as the audio feature for the AV-VAD. Suitable visual features include active appearance models, 2-D discrete cosine transform

(DCT) features and cross-DCT features. An investigation into visual features revealed that a good compromise between information content and computation is given by 2-D DCT features which leads to their selection for the AV-VAD [6]. The remainder of this section briefly describes audio and visual feature extraction.

### 2.1 MFCC audio features

MFCC features are extracted according to the ETSI Aurora standard [5]. Input audio is segmented into 25ms duration frames at 100 frames per second. Following a Hamming window and Fourier transform, a power spectrum is calculated and a 23-D mel-scale filterbank applied followed by a log operation. A DCT is then applied, followed by truncation to give a 12-D MFCC vector comprising coefficients zero to twelve,

$$\mathbf{x}_t = \left[ x_t(1), x_t(2), \cdots, x_t(12) \right] \quad (1)$$

where $t$ indicates the time index. It is also usual to augment the feature vector with its velocity and acceleration temporal derivatives, $\Delta\mathbf{x}_t$ and $\Delta\Delta\mathbf{x}_t$, [7] and this is investigated in section 4.1.

### 2.2 Two-dimensional DCT visual features

Visual features are extracted from a UxV matrix of pixel intensities, **P**, centred around the speaker's mouth, which was tracked using the AVCSR tracker [8]. First a 2-D DCT is applied,

$$c_{m,n} = W_m W_n \sum_{u=0}^{U-1} \sum_{v=0}^{V-1} p_{u,v} \cos\left( \frac{m\pi(2u+1)}{2U} \right) \cos\left( \frac{n\pi(2v+1)}{2V} \right) \quad 0 \le m \le U\text{-}1,\ 0 \le n \le V\text{-}1 \quad (2)$$

where

$$W_n = \begin{cases} \sqrt{1/V} & if \quad n = 0 \\ \sqrt{2/V} & otherwise \end{cases} \quad \text{and} \quad W_m = \begin{cases} \sqrt{1/U} & if \quad m = 0 \\ \sqrt{2/U} & otherwise \end{cases}$$

$p_{u,v}$ refers to the intensity of the pixel in the $u^{th}$ row and $v^{th}$ column of matrix **P** and the resulting 2-D DCT coefficients are given by $c_{n,m}$. After the 2-D DCT, the energy from the image is concentrated in the lower coefficients of the resulting matrix. A visual vector at time $t$, $\mathbf{v}_t$, is obtained by extracting the 2-D DCT coefficients in a zigzag order located in the lower coefficient region of the matrix,

$$\mathbf{v}_t = \left[ c_{0,0}, c_{0,1}, c_{1,0}, c_{2,0}, c_{1,1}, c_{0,2}, c_{0,3}, c_{1,2}, \ldots\ldots \right] \quad (3)$$

From previous work, a suitable size of visual vector was found to be 14 [6].

## 3. AUDIO-VISUAL VOICE ACTIVITY DETECTION

This section introduces the audio-visual VAD for classifying vectors as either speech or non-speech. Operation of the VAD is based on training two models, one on speech and the other on non-speech, and using these to classify input feature vectors. This section first discusses the design of audio-only and visual-only VADs which lead to the design of the audio-visual VAD. Audio and visual VADs are considered separately to allow them to be optimized before integration into the audio-visual VAD.

### 3.1 Audio-only VAD

The audio-only VAD (A-VAD) uses the MFCC vectors, $\mathbf{x}_t$, for classification. MFCC vectors from a set of training data, Z, are first pooled into two sets, one corresponding to speech, $\Psi^{s,\mathbf{x}}$, and the other corresponding to non-speech, $\Psi^{ns,\mathbf{x}}$,

$$\Psi^{s,\mathbf{x}} = \left\{ \mathbf{x}_t \in Z : c_t = speech \right\} \quad (4)$$

$$\Psi^{ns,\mathbf{x}} = \left\{ \mathbf{x}_t \in Z : c_t = non-speech \right\} \quad (5)$$

$c_t$ is a reference label associated with each feature vector and indicates whether the vector represents speech or non-speech.

From the two vector pools, expectation-maximisation (EM) clustering is used to create two Gaussian mixture models (GMMs), one modeling MFCC vectors from speech, $\Phi^{s,\mathbf{x}}$, and the other modeling MFCC vectors from non-speech, $\Phi^{ns,\mathbf{x}}$,

$$p(\mathbf{x}_t|speech) = \Phi^{s,\mathbf{x}}(\mathbf{x}_t) = \sum_{k=1}^{K^{s,\mathbf{x}}} \alpha_k^{s,\mathbf{x}} \phi_k^{s,\mathbf{x}}(\mathbf{x}_t) = \sum_{k=1}^{K^{s,\mathbf{x}}} \alpha_k^{s,\mathbf{x}} N\left(\mathbf{x}_t; \mu_k^{s,\mathbf{x}}, \Sigma_k^{s,\mathbf{x}}\right) \quad (6)$$

$$p(\mathbf{x}_t|non-speech) = \Phi^{ns,\mathbf{x}}(\mathbf{x}_t) = \sum_{k=1}^{K^{ns,\mathbf{x}}} \alpha_k^{ns,\mathbf{x}} \phi_k^{ns,\mathbf{x}}(\mathbf{x}_t) = \sum_{k=1}^{K^{ns,\mathbf{x}}} \alpha_k^{ns,\mathbf{x}} N\left(\mathbf{x}_t; \mu_k^{ns,\mathbf{x}}, \Sigma_k^{ns,\mathbf{x}}\right) \quad (7)$$

The speech GMM, $\Phi^{s,\mathbf{x}}$, comprises $K^{s,\mathbf{x}}$ clusters. Each cluster, $\phi_k^{s,\mathbf{x}}$, is represented by a prior probability, $\alpha_k^{s,\mathbf{x}}$, and a Gaussian probability density function, $N$, with mean vector $\mu_k^{s,\mathbf{x}}$ and covariance matrix $\Sigma_k^{s,\mathbf{x}}$. The non-speech GMM, $\Phi^{ns,\mathbf{x}}$, uses a similar set of parameters denoted by the superscript $ns$.

Classification of an MFCC vector, $\mathbf{x}_t$, as being speech or non-speech utilizes the two GMMs to make a audio VAD estimate, $\hat{c}_t^{A-VAD}$,

$$\hat{c}_t^{A-VAD} = \begin{cases} speech & p(\mathbf{x}_t|speech) \ge p(\mathbf{x}_t|non-speech) \\ non-speech & p(\mathbf{x}_t|speech) < p(\mathbf{x}_t|non-speech) \end{cases} \quad (8)$$

### 3.2 Visual-only VAD

The visual-only VAD (V-VAD) operates in a similar way to the audio-only VAD, except visual 2-D DCT vectors, $\mathbf{v}_t$, replace the audio MFCC vectors. Speech and non-speech visual vector pools are created and EM clustering is applied to create speech and non-speech GMMs of visual vectors, $\Phi^{s,\mathbf{v}}$ and $\Phi^{ns,\mathbf{v}}$,

$$p(\mathbf{v}_t|speech) = \Phi^{s,\mathbf{v}}(\mathbf{v}_t) = \sum_{k=1}^{K^{s,\mathbf{v}}} \alpha_k^{s,\mathbf{v}} \phi_k^{s,\mathbf{v}}(\mathbf{v}_t) = \sum_{k=1}^{K^{s,\mathbf{v}}} \alpha_k^{s,\mathbf{v}} N\left(\mathbf{v}_t; \mu_k^{s,\mathbf{v}}, \Sigma_k^{s,\mathbf{v}}\right) \quad (9)$$

$$p(\mathbf{v}_t|non-speech) = \Phi^{ns,\mathbf{v}}(\mathbf{v}_t) = \sum_{k=1}^{K^{ns,\mathbf{v}}} \alpha_k^{ns,\mathbf{v}} \phi_k^{ns,\mathbf{v}}(\mathbf{v}_t) = \sum_{k=1}^{K^{ns,\mathbf{v}}} \alpha_k^{ns,\mathbf{v}} N\left(\mathbf{v}_t; \mu_k^{ns,\mathbf{v}}, \Sigma_k^{ns,\mathbf{v}}\right) \quad (10)$$

Similar to equation 8, given a visual vector, $\mathbf{v}_t$, the two GMMs can be used to make a visual VAD classification estimate, $\hat{c}_t^{V-VAD}$, as to whether the frame represents speech or non-speech,

$$\hat{c}_t^{V-VAD} = \begin{cases} speech & p(\mathbf{v}_t|speech) \geq p(\mathbf{v}_t|non-speech) \\ non-speech & p(\mathbf{v}_t|speech) < p(\mathbf{v}_t|non-speech) \end{cases} \quad (11)$$

### 3.3 Audio-visual VAD

The audio-visual VAD (AV-VAD) uses both audio and visual features to classify vectors as speech or non-speech. An audio-visual feature vector, $\mathbf{z}_t$, is first defined as,

$$\mathbf{z}_t = [\mathbf{x}_t, \mathbf{v}_t] \quad (12)$$

Vector pools comprising speech and non-speech audio-visual vectors are created and audio-visual GMMs trained for speech and non-speech, $\Phi^{s,\mathbf{z}}$ and $\Phi^{ns,\mathbf{z}}$. Classification of audio-visual vectors, $\mathbf{z}_t$, as being speech or non-speech, $\hat{c}_t^{AV-VAD}$, takes place as before in equations 8 and 11, but is now based on the audio-visual feature vector,

$$\hat{c}_t^{AV-VAD} = \begin{cases} speech & p(\mathbf{z}_t|speech) \geq p(\mathbf{z}_t|non-speech) \\ non-speech & p(\mathbf{z}_t|speech) < p(\mathbf{z}_t|non-speech) \end{cases} \quad (13)$$

In low noise conditions it is likely that the audio component of the joint feature vector will be more accurate than the visual component in speech/non-speech classification. However, as noise power increases, the audio component will become less accurate and at some signal-to-noise ratio, the visual component will become more accurate. To exploit this variation in classification accuracy of the audio and visual components, the signal-to-noise ratio (SNR) is used to adjust the contribution made by the audio and visual components within the speech and non-speech GMMs for classification (assuming diagonal covariance matrices),

$$p(\mathbf{z}_t|speech) = \Phi^{s,\mathbf{z}}(\mathbf{z}_t) = \sum_{k=1}^{K^{s,\mathbf{z}}} \alpha_k^{s,\mathbf{z}} N\left(\mathbf{x}_t; \mu_k^{s,\mathbf{x}}, \Sigma_k^{s,\mathbf{x}}\right)^{\gamma(SNR_t)} N\left(\mathbf{v}_t; \mu_k^{s,\mathbf{v}}, \Sigma_k^{s,\mathbf{v}}\right)^{1-\gamma(SNR_t)} \quad (14)$$

$$p(\mathbf{z}_t|non-speech) = \Phi^{ns,\mathbf{z}}(\mathbf{z}_t) = \sum_{k=1}^{K^{ns,\mathbf{z}}} \alpha_k^{ns,\mathbf{z}} N\left(\mathbf{x}_t; \mu_k^{ns,\mathbf{x}}, \Sigma_k^{ns,\mathbf{x}}\right)^{\gamma(SNR_t)} N\left(\mathbf{v}_t; \mu_k^{ns,\mathbf{v}}, \Sigma_k^{ns,\mathbf{v}}\right)^{1-\gamma(SNR_t)} \quad (15)$$

$\gamma(SNR_t)$ 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 of the audio features, while at higher SNRs the contribution of visual features is reduced. The function $\gamma$ is determined experimentally using training data to find, for a particular SNR, the value of $\gamma$ that maximises classification accuracy over the training data. This is discussed further in section 4.2. The approach of varying the contribution made by audio and visual streams according to SNR has also been successfully applied to audio-visual speech recognition in noise [3].

## 4. EXPERIMENTAL RESULTS

The aim of the experiments in this section is to examine speech/non-speech classification accuracy provided by the VADs proposed in section 3. First the accuracies of the audio and visual VADs are optimised in terms of the number of clusters in the GMMs and the features used. Secondly, audio-visual VAD performance is investigated at SNRs from clean down to -20dB and a comparison made against the VAD used in the ETSI Aurora standard [5]. Evaluations are made for training and testing on a single speaker and training and testing on different speakers.

The audio-visual data used for these experiments is taken from the Grid database which comprises 34 male and female talkers, each saying 1000 three-second phrases [9]. Audio was originally recorded at 50kHz but has been downsampled to 8kHz for these experiments. The video was originally recorded at 25 frames per second and has been upsampled to 100 frames per second to give a visual frame rate equal to the audio frame rate. For each speaker 800 sentences have been used for training and the remaining 200 used for testing. This gives a total of 60,000 vectors for testing.

### 4.1 Optimising features and the number of clusters

This section examines the effect of increasing the number of GMM clusters in the audio-only and visual-only VADs and also examines the effect of including temporal features. Both the audio-only and visual-only VADs are trained and tested using data from s6 of the Grid database. For both the audio-only VAD and visual-only VAD, the number of clusters in the GMMs is varied from 1 to 16. Tests are also presented using only static features (audio or visual) and also with velocity, $\Delta$, and acceleration, $\Delta\Delta$, temporal derivatives augmented onto the feature vector. Table 1 shows VAD classification accuracy for both audio-only (A-VAD) and visual-only (V-VAD) VADs in clean speech and at SNRs down to -20dB in white noise using from 1 to 16 clusters, with and without temporal derivatives. Visual-only VAD accuracy is reported at the top of the table where the SNR is indicated as being not applicable (NA).

Considering first the visual-only VAD, the results show that increasing the number of clusters gives a substantial increase in accuracy due to the improved modeling of visual features by the GMM. Augmenting the static visual

features by the temporal derivatives also increases performance which suggests that neighboring frames influence VAD accuracy.

| *SNR* | *Feature* | *K=1* | *K=2* | *K=4* | *K=8* | *K=16* |
|---|---|---|---|---|---|---|
| NA | **v** | 92.0 | 92.7 | 93.0 | 94.9 | 95.2 |
| | **v**+Δ**v**+ΔΔ**v** | 91.4 | 94.7 | 95.1 | 96.2 | 96.7 |
| Clean | **x** | 94.8 | 94.5 | 94.7 | 94.7 | 95.2 |
| | **x**+Δ**x**+ΔΔ**x** | 97.3 | 97.6 | 97.9 | 98.0 | 98.2 |
| 20dB | **x** | 74.5 | 71.7 | 71.7 | 65.9 | 72.0 |
| | **x**+Δ**x**+ΔΔ**x** | 88.0 | 85.8 | 86.9 | 87.4 | 87.8 |
| 10dB | **x** | 67.6 | 66.5 | 67.0 | 62.2 | 66.6 |
| | **x**+Δ**x**+ΔΔ**x** | 73.4 | 70.2 | 71.6 | 71.2 | 73.1 |
| 0dB | **x** | 57.3 | 57.7 | 56.7 | 54.1 | 56.0 |
| | **x**+Δ**x**+ΔΔ**x** | 51.8 | 50.2 | 49.5 | 49.8 | 52.7 |
| -10dB | **x** | 43.9 | 45.8 | 43.1 | 42.8 | 42.3 |
| | **x**+Δ**x**+ΔΔ**x** | 37.7 | 37.8 | 37.4 | 37.6 | 38.1 |
| -20dB | **x** | 40.1 | 41.9 | 39.7 | 40.1 | 39.1 |
| | **x**+Δ**x**+ΔΔ**x** | 37.2 | 37.2 | 37.2 | 37.2 | 37.2 |

Table 1 – Audio VAD and visual VAD accuracy for varying numbers of clusters at signal-to-noise ratios from clean to -20dB using static and temporal features.

For the audio-only VAD, as is expected, reducing the SNR leads to substantial reductions in VAD accuracy due to more non-speech frames being incorrectly classified as speech. Including temporal derivatives of the audio features improves performance in clean speech and at SNRs down to 10dB, below this static-only performance is higher. Increasing the number of clusters in the GMMs improves accuracy in clean speech but leads to a slight deterioration in performance in noisy speech. In noisy speech, highest classification accuracy is given by the 1 cluster GMM.

Based on this analysis the optimal configuration for the visual-only VAD is selected as the 16 cluster GMM with visual features comprising both static and temporal components. For the audio-only VAD, the 16 cluster GMM and audio features comprising both static and temporal components are selected for further experiments. Even though for noisy speech this configuration is not optimal, it is considered more important to have high accuracy in clean speech. As will be shown in the next section, VAD performance in noise benefits from the visual-only VAD.

### 4.2 Speaker-dependent audio-visual VAD accuracy

This test compares the accuracy of the audio-only, visual-only, audio-visual and ETSI Aurora VADs in conditions ranging from clean speech down to noisy speech at an SNR of -20dB in white noise. For the audio-visual VAD, performance is shown for the simple equal weighting of audio and visual features and also using the SNR-dependent weighting in equations 14 and 15. Training and testing uses speaker s6 of the Grid database. Figure 1 shows the classification accuracies of the five different VADs in clean speech and in white noisy at SNRs from 20dB to -20dB.

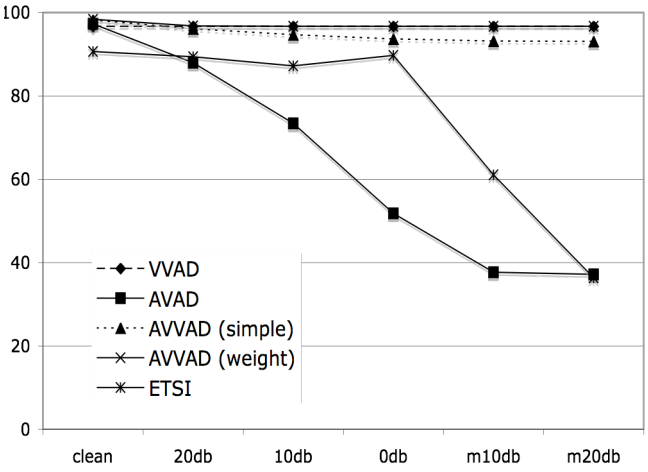


Figure 1 - Audio-only (AVAD), visual-only (VVAD), audio-visual simple (AVVAD simple) and weighted (AVVAD weight) and ETSI Aurora VAD (ETSI) accuracies in clean and noisy speech for training and testing on speaker s6.

As seen in the previous section, the audio-only VAD gives highest performance in clean speech but deteriorates rapidly in noise. The visual-only VAD is slightly less accurate in clean speech than the audio-only VAD, but is unaffected by the noise and so outperforms audio-only in noisy conditions. A simple combination of audio and visual features gives classification accuracy above visual-only in clean conditions and slightly lower in noise. However, adjusting the contribution of the audio and visual features according to the SNR gives substantially better performance. In clean speech and at 20dB, the AV-VAD is more accurate than both audio-only and visual-only VADs. At lower SNRs, AV-VAD accuracy converges on visual-only VAD accuracy as the classification accuracy offered by the audio-VAD deteriorates. The dashed line in figure 2 shows optimal values of the weighting function, $\gamma$, as a function of SNR.

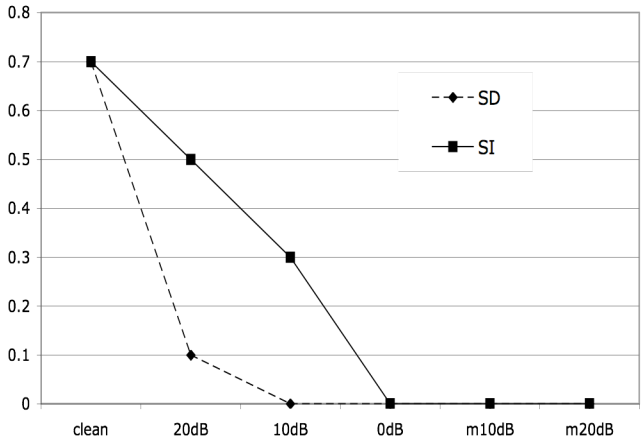


Figure 2 - Weighting function, $\gamma$, as a function of SNR for speaker-dependent and speaker-independent speech.

In clean speech both the audio and visual components make

significant contributions to VAD classification. As noise increases the visual component rapidly dominates due to the deterioration in accuracy of the audio component.

Tests using the ETSI Aurora VAD reveal it to be less accurate in clean speech than the audio-only VAD, but better able to maintain accuracy down to SNRs of 0dB.

### 4.3 Speaker-independent audio-visual VAD

The performance of the audio-only, visual-only and audio-visual VADs are now analysed for speaker-independent training and testing. In the previous section a single speaker (speaker s6) was used for training and testing. In this section VAD testing still uses speaker s6, but VAD training uses data from two different speakers from the database – speakers s12 and s19. The aim of this test is to examine the sensitivity of the VADs to testing on unseen speakers. Figure 3 shows the performance of the five VADs in clean and noisy speech down to an SNR of -20dB.

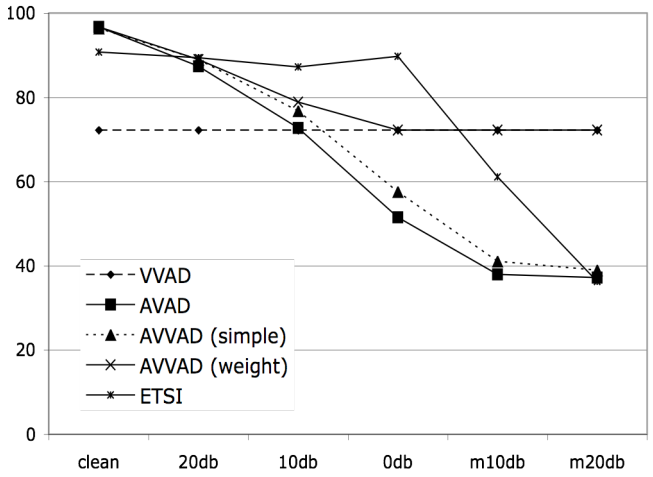


Figure 3 - Audio-only (AVAD), visual-only (VVAD), audio-visual simple (AVVAD simple) and weighted (AVVAD weight) and ETSI Aurora (ETSI) VAD accuracies in clean and noisy speech for training on speakers s12 and s19 and testing on speaker s6.

Comparing the speaker-independent results in figure 3 to the speaker-dependent results in figure 1 shows visual-only VAD accuracy to fall from 97% to 72%. For the audio-only VADs, the accuracy for speaker-dependent and speaker-independent training/tests is virtually equal over both clean and noisy speech. This is attributed to the audio-only VAD operating on energy levels which are similar for training on speaker s6 and for training on speakers s12 and s19. For visual-only VADs, the modeling of the visual features for speech and non-speech is more complex and testing on an unseen speaker leads to a mismatch that reduces classification accuracy. The poorer performance of the visual-only VAD leads to worse AV-VAD performance in comparison to the speaker-dependent AV-VAD. This is further reflected by the SNR-dependent weighting function which is shown by the solid line in figure 2. For the speaker-independent VAD, more weight is given to the audio-VAD due to the poorer classification given by the visual-VAD.

## 5. CONCLUSION

This paper has shown that audio and visual speech information can be successfully combined to make a highly noise robust VAD. An SNR-dependent weighting term increases the contribution made by audio features at high SNRs, while reducing the contribution at lower SNRs, where visual features are more robust. For a system trained and tested on the same speaker, VAD accuracy remains above 97% at all SNRs. However, when testing the VAD on an unseen speaker, performance drops. This is primarily due to poor visual performance because of the mismatch between the visual speaker characteristics of the test and training speakers. However, the results presented here used a very small training data set of only 2 speakers. It is expected that increasing the range of speakers will increase visual VAD accuracy for unseen speakers.

## 6. REFERENCES

[1] J.H. Chang, N.S. Kim and S.K. Mitra, “Voice activity detection based on multiple statistical models”, IEEE Trans. Signal Processing, vol. 54, no. 6, pp. 1965-1976, 2006

[2] D. Sodoyer, B. Rivet, L. Girin, J.-L.. Schwartz and C. Jutten, “An analysis of visual speech information applied to voice activity detection”, Proc. ICASSP, 2006

[3] I. Matthews, T.F. Cootes, J.A. Bangham, S.J. Cox and R.W. Harvey, “Extraction of visual features for lipreading”, IEEE Trans. PAMI, vol. 24, no. 2, pp. 198-213, February 2002

[4] I. Almajai and B. Milner, “Visually-derived Wiener filters for speech enhancement”, Proc. ICASSP, 2007

[5] 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

[6] 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

[7] B.A. Hanson and T.H. Applebaum, "Robust speaker-independent word recognition using static, dynamic and acceleration features: experiments with Lombard and noisy speech", Proc. ICASSP, 1990

[8] Intel AVCSR Toolkit – http://sourceforge.net/projects/opencvlibrary

[9] M. Cooke, J. Barker, S. Cunningham and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition”, JASA, vol. 120, no. 5, pp. 2421-2424, Nov. 2006