Voice Activity Detection based on Fusion of Audio and Visual Information

Shin'ichi Takeuchi[1]*, Takashi Hashiba*[2]*, Satoshi Tamura*[3] *and Satoru Hayamizu*[3]

[1]Virtual System Laboratory, Gifu University, Japan
[2]Graduate School of Engineering, Gifu University, Japan
[3]Faculty of Engineering, Gifu University, Japan
{takeuchi@hym.info., hashiba@hym.info., tamura@info., hayamizu}@gifu-u.ac.jp

Abstract

In this paper, we propose a multi-modal voice activity detection system (VAD) that uses audio and visual information. Audio-only VAD systems typically are not robust to (acoustic) noise. Incorporating visual information, for example information extracted from mouth images, can improve the robustness since the visual information is not affected by the acoustic noise. In multi-modal (speech) signal processing, there are two methods for fusing the audio and the visual information: concatenating the audio and visual features, and employing audio-only and visual-only classifiers, then fusing the unimodal decisions. We investigate the effectiveness of these methods and also compare model-based and model-free methods for VAD. Experimental results show feature fusion methods to generally be more effective, and decision fusion methods generally perform better using model-free methods.
Index Terms: voice activity detection, multi-modal, AVVAD

1 Introduction

Automatic speech recognition (ASR) has received increasing interest in recent years, and can now be found in several real-world applications. However in a real environment, ASR performance can significantly be degraded by environmental noise. Therefore, to compensate for the noise, techniques which support ASR are needed as a front-end of recognition.

Voice Activity Detection (VAD) is one such front-end technique. The goal of VAD is to distinguish sections of a signal that contain speech from those that do not. Thus a speech recogniser can then focus effort only on segments of the signal that contain speech. One approach to VAD is to use power of input signal to classify a segment of signal as speech/non-speech. However, a limitation is consonants often contain low power, so words beginning with a consonant might be mis-recognised as the VAD system mis-identifies the onset of the work in the signal. Alternatively, noise reduction methods can be used in combination with VAD as an ASR front-end [1]. Generally the noise compensation are carried out as separate tasks, but often the result from one component is fed back to the other to improve the effectiveness.

Small cameras are becoming ubiquitous on laptop computers, cellular phones, and Personal Digital Assistants (PDAs), which. This makes it easier to capture video and brings forward the possibility of more widespread audio-visual speech recognition (AVASR) and audio-visual voice activity detection (AVVAD). There have been some recent research into AVVAD [2, 3, 4]. For instance, Yamamoto et al. [2] proposed an AVVAD method using a microphone array and a camera for hands-free speech recognition in noisy conditions. In their method, the location of a subject is determined from captured pictures, before acoustic and visual information are integrated using Bayesian networks. Butko [3] combines SVM-based and HMM-based methods for audio information and they are integrated with a video-based system. Almajai[4] fuse 2-D discrete cosine transform (DCT) visual features with acoustic features for AVVAD.

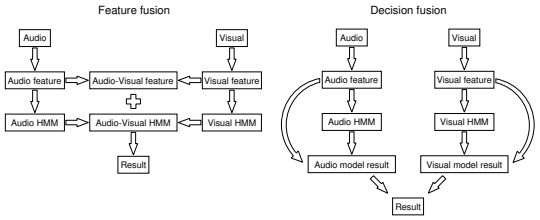

Figure 1: Comparison of fusion methods.

In this paper, we propose a multi-modal voice activity detection that uses optical flow computed from lip images as visual features.

This paper is organised as follows: Section 2 describes two fusion methods for multi-modal VAD and the difference between a HMM model-based and a model-free system. Section 3 describes experiments for VAD and discusses the results. Finally, Section 4 provides the conclusions and describes future work.

2 Proposed Method

How to combine the information from different modalities is one of the main problem for audio-visual speech recognition. One (low-level) method is to combine the audio and the visual features — known as feature fusion. An alternative (high-level) method is to perform uni-modal recognition then fuse the decision, known as decision fusion. Figure 1 contrasts these approaches.

VAD methods can be divided into two categories: model-based methods and model-free methods. Model-based methods use training data to create models, which are used to recognise a specific class. Conversely, model-free method does not utilise class-specific training data directly. The advantage of model-based methods is that they have information about target they must later recognise. It is therefore natural that model-based methods perform better result than model-free methods if training is done correctly. The advantage of model-free methods is there is no need to provide prior class labels to perform training.

This paper compares 4 approaches for AVVAD: a model-based

and a model-free approach that each utilise both feature and decision-level fusion. These are outlined in Table 1.

Table 1: Combination of fusion and model.

	Model training	
	with	without
feature fusion	2.1	2.2
decision fusion	2.3	2.4

2.1 Feature Fusion with Model Training

VAD can be considered a special case of ASR. However, rather than classifying phonemes or words, the task is to distinguish segments of a signal that contain speech from those that do not.

In this work a multi-stream HMM, often used in AVASR, is used as the model-based feature fusion method. Voice and non-voice models are created using clean (i.e. noise-free) data. 3 state left-to-right HMMs with GMMs employing diagonal covariance are used. The number of mixtures is 8 for the audio modality and 4 for visual modality. The acoustic features are either the first 12 Mel-Frequency Cepstral Coefficients (MFCCs), with the Δ and Δ Δ coefficients, or the short term power and its Δ and Δ Δ terms. Optical flow computed from images containing the speakers mouth is used as visual features. The horizontal and vertical mean vector (2D) and their variance vector (2D) are calculated in each frame. In this paper, this method is called **F-w** (**F**eature fusion **w**ith model training).

2.2 Feature Fusion without Model Training

This is a simplistic approach that classifies speech from non-speech segments by combining audio and visual features, then performing a simple thresholding. As acoustic features, the power of the input signal is computed using:

$$feat_A(i) = 10 \cdot \log_{10}\left(\frac{1}{N}\sum_{n=0}^{N-1} s_i^2(n)\right), i = 0, 1, ..., M-1 \quad (1)$$

where $s_i(n)$ is the signal and $feat_A(i)$ is the average power of the signal in the i-th frame, and N means frame length. The visual features are the variance of the optical flow vectors computed from of lip images. The particular algorithm used in the Horn-Schuck algorithm [5]. The advantage of using optical flow is the features are dynamic, describing the movement of the visible articulators from one frame to the next. In the context of VAD, this allows the system to distinguish between a stationary open mouth (likely not speaking) from a moving open mouth (maybe speaking).

Optical flow vectors computed from images in non-speech segments generally equate to small values — there is some small mouth movements, but little to no movement in the cheeks. Conversely, during speech regions there is generally a lot of facial movement and the difference in the variance of the optical flow vectors is significantly greater than non-speech segments.

2.2.1 Feature fusion

Feature fusion involves computing the acoustic and the visual parameters from the audio and video signals respectively, the combining the feature vectors before applying as input to a (single) classifier. In i-th frame, the audio features, $feat_A(i)$, and visual features, $feat_V(i)$, are united using Eq. (2) because they are used to decided voice/non-voice by a threshold, as shown in Eq. (3).

$$feat_{AV}(i) = \beta \cdot feat_A(i) + (1-\beta) \cdot feat_V(i) \quad (2)$$

where β is a parameter weight applied to the audio feature. For $\beta = 1$, $feat_{AV}(i)$ is purely audio features. To classify speech and non-speech segments in the i-th of the signal, Equation (3) is applied.

$$result(i) = \begin{cases} \text{voice} & (feat(i) \geq T) \\ \text{non-voice} & (feat(i) < T) \end{cases} \quad (3)$$

where T is the decision threshold. If $feat(i)$ is equal to or greater than the threshold, i-th frame is decided as voice. In this paper,this method is called **F-w/o** (**F**eature fusion **w**ith**o**ut model training).

2.3 Decision Fusion with Model Training

Model-based decision fusion is a HMM-based VAD, similar to that described in Section 2.1. However, separate model models are trained to classify using acoustic and visual parameters individually. The individual decisions are then combined using logical conjunction, as in Equation (4, or logical disjunction, as in Equation (5.

$$result_{AV}(i) = result_A(i) \cap result_V(i) \quad (4)$$

$$result_{AV}(i) = result_A(i) \cup result_V(i) \quad (5)$$

The result of Equation (4) is the segment contains speech only if both modalities indicate the presence of speech. Equation (5) results in a segment being classified as speech if either modality indicates the presence of speech. In this paper, this method is called **D-w** (**D**ecision fusion **w**ith model training). **D-w(AND)** signifies Equation (4), and **D-w(OR)** signifies Equation (5).

2.4 Decision Fusion without Model Training

In the case of decision fusion without model training, audio feature and visual features are classified as speech/non-speech segments using Equation (3) for each individual modality. These results are then combined using Equation (4) or Eq. (5) as described in Section 2.3. This method is referred to as **D-w/o** (**D**ecision fusion **w**ith**o**ut model training). **D-w/o(AND)** uses Equation (4), and **D-w/o(OR)** uses Equation (5).

3 Experiments

In this section the effectiveness of each of the methods previously described are compared. Speech sequences that are contaminated with various several noise sources are used. To evaluate the performance, the False Acceptance Rate (FAR), False Rejection Rate (FRR), and their average are used.

3.1 Experimental conditions

An existing speech corpus [6] is used in this work. The corpus is formed of $2,750$ utterances spoken by 11 male speakers, each speaking 250 continuous digits. The utterances are formed of both regions of speech ($\approx 40\%$ of the corpus) and silence (the remaining $\approx 60\%$ of the corpus). The sequence were recorded

Figure 2: A sample image in the audio-visual speech database.

in a soundproof recording room using a lapel microphone. The video was captured using a DV camera located roughly 1m away from the speaker(s). Recording and capture conditions are detailed in Table 2 and a sample image from the cirpus is shown in Figure 2.

Table 2: Recording and capture conditions.

Audio sampling rate	16kHz (Downsampled from 48kHz)
Audio quantization bit	16bit
Video size	180x120 (Downsampled from 720x480)
Video frame rate	15fps (Progressive video)
Video color depth	24bit (Truecolor)
Video format	DV

To evaluate noise robustness of proposed methods, two forms of noise were adopted: white additive noise and classical (instrumental) music. Specifically the RWC music database[7] is used. Noise is added to the signals to generate signals with Sound to Noise Ratio (SNR) of 10dB and 0dB respectively. The SNR is is calculated over the entire signal, including the silence regions.

3.2 Results

*3.2.1 Feature Fusion with Model Training (**F-w**)*

Table 3 shows the result using **F-w**: "clean" signifies the noise-free condition, "white 10dB" and "white 0dB" signify white additive noise at 10 and 0dB respectively, and "music 10dB" and "music 0dB" signify speech degraded using music. For the model-based approaches, the mixing weight is added to the mixing weight the GMMs in speech/non-speech model. The error rate is low for hight SNR (i.e. clean, white 10dB, and music 10dB), but FAR increases as the noise increases. Note that in these experiments, a speaker independent approach is adopted. HMMs are trained from 10 of the 11 speakers, and the system is tested on the held out speaker. The process is repeated for each speaker in turn and the mean results, averaged over all speakers, are presented.

*3.2.2 Feature Fusion without Model Training (**F-w/o**)*

Table 4 shows result of **F-w/o**. The audio weights are set experimentally using the best result from previous experiments. The thresholds, T, from Equation (3), is dynamic and is the average of the feature parameters from frame 1 to the current frame. This method shows small FAR except for the condition of music 0dB.

Table 3: Error rate by Fusion-model-based method. (%)

	FAR	FRR	Average	audio weight
clean	2.6	6.6	4.7	0.9
white 10dB	7.7	5.3	6.4	0.4
music 10dB	11.8	5.0	8.4	0.2
white 0dB	30.3	8.6	19.4	0.1
music 0dB	18.5	22.4	20.5	0.0

Table 4: Error rate by Fusion-model-free method. (%)

	FAR	FRR	Average	audio weight
clean	2.7	10.7	6.7	0.9
white 10dB	2.9	10.7	6.8	0.9
music 10dB	3.2	12.7	8.0	0.9
white 0dB	6.6	11.8	9.2	0.9
music 0dB	18.1	21.3	19.7	0.5

*3.2.3 Decision Fusion with Model Training (**D-w**)*

Table 5 to table 7 shows the result of **D-w(AND)** and **D-w(OR)** for audio/visual only models respectively. The results of **D-w** show the error rate increases with increasing noise (specifically white 0dB and music 0dB). In table 5, FAR decreases by combining results. Audio-only classification often mis-classifies "speech", so the FAR is relatively high, yet FRR keeps is relatively low rate.

Table 5: Error rate by Decision(AND)-model-based method. (%)

	FAR	FRR	Average
clean	0.94	24.2	12.5
white 10dB	3.5	23.7	13.6
music 10dB	3.5	23.6	13.5
white 0dB	21.5	23.8	22.6
music 0dB	20.6	22.3	21.4

*3.2.4 Decision Fusion without Model Training (**D-w/o**)*

Table 8 shows the result of **D-w/o(AND)** and Table 9 shows the result of **D-w/o(OR)**. Again the threshold T (Equation (3)) is determined empirically given the best value of previous experiments.

The performance of **D-w/o(AND)** is relatively poor. Generally, the logical disjunction (OR) without prior training over classifies segments as speech, whereas logical conjunction (AND) tends to perform better (see Table 9).

3.3 Discussions

The worst performing systems evaluated here are:

Table 6: Error rate by Decision(OR)-model-based method. (%)

	FAR	FRR	Average
clean	24.7	3.2	14
white 10dB	40.4	2.6	21.5
music 10dB	30.9	2.6	16.7
white 0dB	91.7	0.8	46.2
music 0dB	86.7	1.5	44.1

Table 7: Error rate by individual model-based method. (%)

	FAR	FRR	Average
clean	1.9	8.7	5.3
white 10dB	20.1	7.5	13.8
music 10dB	10.6	7.4	9.0
white 0dB	89.4	5.8	47.6
music 0dB	83.5	5.1	44.3
visual	23.7	18.7	21.2

Table 8: Error rate by Decision(AND)-model-free method. (%)

	FAR	FRR	Average
clean	19.5	32.7	26.1
white 10dB	19.4	32.9	26.2
music 10dB	19.3	33.2	26.2
white 0dB	17.4	36.6	27.0
music 0dB	17.0	37.4	27.1

Table 9: Error rate by Decision(OR)-model-free method. (%)

	FAR	FRR	Average
clean	2.6	11.1	6.9
white 10dB	2.5	11.2	6.9
music 10dB	3.6	13.5	8.5
white 0dB	2.2	13.8	8.0
music 0dB	28.5	28.5	28.2

- **F-w** in low noise condition (shown in Table 3)
- **F-w/o** in high noise condition (shown in Table 4).

Figure 3 shows the result of **F-w** and **F-w/o**. Feature fusion performs better than decision fusion and the experimental results presented here demonstrate the same tendency. The model-based method (**F-w**) performs best when the test data are similar to the training data. For differing training and test conditions, the model-free method (**F-w/o**) performs better. The best result in terms of FAR is **D-w(OR)** (shown in Table 6), and best result in terms of FRR is **D-w/o(OR)** (shown in table 9).

4 Conclusions

In this paper we have investigated AVVAD methods. For the approaches tested here, feature fusion is most effective, as is typical in AVASR. Combining the speech detection of **D-w(OR)** and the non-speech detection of **D-w/o(OR)** will produce, on average, the most accurate system. For future work, we will investigate degradation of the visual modality as well as the acoustic modality.

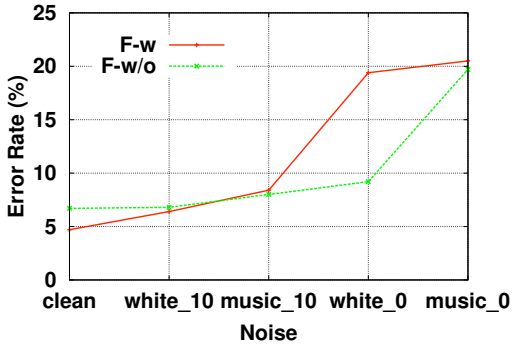

Figure 3: Comparison between model-based and model-free method.

References

[1] M. Fujimoto, K. Ishizuka, and T. Nakatani, "Study of integration of statistical model-based voice activity detection and noise suppression," in *Proceedings of Interspeech*, 2008, pp. 2008–2011.

[2] F. Asano, K. Yamamoto, I. Hara, J. Ogata, T. Yoshimura, Y. Motomura, N. Ichimura, and H. Asoh, "Detection and separation of speech event using audio and video information fusion and its application to robust speech interface," *EURASIP Journal on Applied Signal Processing*, vol. 2004, pp. 1727–1738, 2004.

[3] T. Butko, A. Temko, C. Nadeu, and C. Canton, "Fusion of audio and video modalities for detection of acoustic events," in *Proceedings of Interspeech*, 2008, pp. 123–126.

[4] I. Almajai and B. Milner, "Using audio-visual features for robust voice activity detection in clean and noisy speech," in *Proceedings of EUSIPCO2008*, 2008, pp. 123–126.

[5] B. K. P. Horn and B. G. Schunk, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185–203, 1981.

[6] S. Tamura, K. Iwano, and S. Furui, "Multi-modal speech recognition using optical-flow analysis for lip images," *The Journal of VLSI Signal Processing*, vol. 36, pp. 117–124, 2004.

[7] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music databases," in *Proceedings of the 3rd International Conference on Music Information Retrieval*, 2002, pp. 287–288.