Front. Electr. Electron. Eng. China (2006) 4: 425–430
DOI 10.1007/s11460-006-0081-5

RESEARCH ARTICLE

LIU Peng, WANG Zuo-ying

Audio-visual voice activity detection

Abstract In speech signal processing systems, frame-energy based voice activity detection (VAD) method may be interfered with the background noise and non-stationary characteristic of the frame-energy in voice segment. The purpose of this paper is to improve the performance and robustness of VAD by introducing visual information. Meanwhile, data-driven linear transformation is adopted in visual feature extraction, and a general statistical VAD model is designed. Using the general model and a two-stage fusion strategy presented in this paper, a concrete multimodal VAD system is built. Experiments show that a 55.0 % relative reduction in frame error rate and a 98.5 % relative reduction in sentence-breaking error rate are obtained when using multimodal VAD, compared to frame-energy based audio VAD. The results show that using multimodal method, sentence-breaking errors are almost avoided, and frame-detection performance is clearly improved, which proves the effectiveness of the visual modal in VAD.

Keywords speech recognition, voice activity detection, multimodal

1 Introduction

Traditionally, audio-based VAD is performed by using some features of the audio stream, for example, frame-energy [1] or entropy [2]. However, we should pay attention to some inherent disadvantages of these audio approaches. The frame-energy based approach is designed for audio detection rather than voice detection substantially due to the poor consistence of the frame-energy, so it may suffer from the background noises in non-speech segments and there will be some periods with low energy in speech segments, for example, in the case of consonants or pauses between phrases. Although the spectral characteristic of voice is taken into consideration in the entropy-based approach, it still cannot deal with colored noises with similar spectral characteristic of voice, for example, cross-talking noises. With the help of the visual modal, we expect that the problem can be solved more satisfactorily because the appearances of lips reflect the speaker's expression directly and cannot be affected by background noises.

Recently, a lot of attention has been paid to the multimodal interactive systems [3, 4] that employ visual information (especially the movement of the speaker's lips) to supplement traditional audio information in speech recognition and speech synthesis. It is shown that with fusion of the visual modal with the audio modal, the accuracy and robustness of speech recognition systems can be enhanced significantly, and the output from the visual modal in speech synthesis will lead to more comprehensible results. However, there have not been many results reported on audio-visual information-based VAD, which is an important front-end of many kinds of speech signal processing systems. We focus on this problem in this paper.

There are several key problems that have to be addressed for audio-visual VAD:

1) How can visual features that can represent the shape and movement of lips be extracted?

2) How can models for voice and non-voice in visual feature space be built?

3) How can binary decision rules and the system framework for audio-visual VAD be designed? These issues will be theoretically discussed and experimentally tested in this paper.

In Sect. 2, the problem of visual feature extraction is discussed. In Sect. 3, we design appropriate models and decision rules for audio-visual VAD. In Sect. 4, we focus on how to define meaningful evaluation matrices for VAD. In Sect. 5, experimental results are shown and discussed. Finally, in Sect. 6, we present our conclusions.

Translated from *Journal of Tsinghua University (Science and Technology)*, 2005, 45(7): 896–899 (in Chinese)

LIU Peng, WANG Zuo-ying(✉)
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
E-mail: wzy-dee@mail.tsinghua.edu.cn

2 Visual feature extraction

The procedure of visual feature extraction for audio-visual

VAD is illustrated in Fig. 1. Given a video frame sequence includes the rough area of the speaker's lips, the basic steps for extracting visual feature may include the following:

1) A lip-tracking algorithm is applied to the frame for the region of interest (ROI).

2) An intra-frame visual feature extraction algorithm is applied to ROI to obtain the original frame visual feature $\boldsymbol{V}^{\mathrm{I}}$.

3) To synchronize with audio VAD, $\boldsymbol{V}^{\mathrm{I}}$ is interpolated at the same sample rate as the audio stream to get $\boldsymbol{V}^{\mathrm{II}}$.

4) The features from several adjacent frames are combined to form a large vector and are processed by another dimension reduction transform to produce the final visual feature $\boldsymbol{V}^{\mathrm{III}}$, which has the same dimension as $\boldsymbol{V}^{\mathrm{II}}$ but contains more information about the inter-frame correlation.

5) Combine $\boldsymbol{V}^{\mathrm{III}}$ with its first order difference for the final visual feature $\boldsymbol{V}=\{\boldsymbol{V}^{\mathrm{III}}, \Delta\boldsymbol{V}^{\mathrm{III}}\}$.

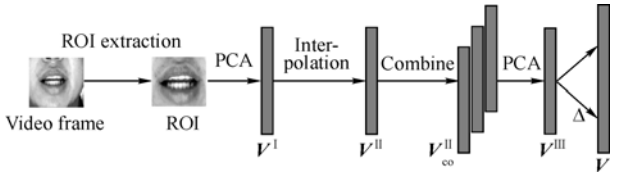


Fig. 1 Block diagram of visual feature extraction

Primary component analysis (PCA) [5] is a widely adopted linear transformation for feature extraction. It makes use of prior knowledge to build a linear subspace that describes the data distribution well. We apply this approach to ROI extraction and dimension reduction, and the two steps are performed in one pass. We obtain the ROI and extract the frame feature in a video frame using model matching. The model is built by finding several main variation modes of ROI using PCA. Given a training frame labeled with their ROI rectangles, we re-sample the rectangle with A columns and B rows to get the ROI vector $\boldsymbol{R}$ with dimension AB. From such vectors, the mean vector $\boldsymbol{\mu}_{\boldsymbol{R}}$ and the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{R}}$ are calculated. Then the eigenvectors corresponding to the largest d eigenvalues are selected as the column vectors of $\boldsymbol{P}_{\boldsymbol{R}}(AB\times d)$, which is the projection matrix to a d-dim subspace that best describes the variation of the ROI vector. The degree of fitting between a candidate ROI vector $\boldsymbol{R}$ and the subspace is measured by calculating the Euclidean distance between them:

$$D^2(\boldsymbol{p}) = \left\|\boldsymbol{R}(\boldsymbol{p}) - \boldsymbol{\mu}_{\boldsymbol{R}}\right\|^2 - \left\|\boldsymbol{P}_{\boldsymbol{R}}\left(\boldsymbol{R}(\boldsymbol{p}) - \boldsymbol{\mu}_{\boldsymbol{R}}\right)\right\|^2 \quad (1)$$

where $\boldsymbol{p}$ is the vector that consists of the parameters determining the ROI. We assume that $\boldsymbol{p} = (x, y, s, \theta)$, where (x, y) are the central coordinates of ROI, s is the scale and θ is the rotation angle, as illustrated in Fig. 2. Model matching can be formulated as the optimization problem:

$$\boldsymbol{p}_{\mathrm{opt}} = \arg\max_{p} D(\boldsymbol{p}) \quad (2)$$

Since the derivative of distance function cannot be calculated analytically, we use downhill simplex method [6] for numerical solution of the optimization problem.

Once the ROI is extracted, the coordinates of $\boldsymbol{R}(\boldsymbol{p}_{\mathrm{opt}})$ in the main variation subspace are selected as the components of the d-dim visual frame feature vector $\boldsymbol{V}^{\mathrm{I}}_{\mathrm{PCA}} = (c_1, c_2, \cdots, c_{\mathrm{d}})$, where $c_i = (\boldsymbol{R}(\boldsymbol{p}_{\mathrm{opt}}) - \boldsymbol{\mu}_{\boldsymbol{R}}, \boldsymbol{p}_{\boldsymbol{R}_i})$ is the ith component of the coordinates ($(\bullet,\bullet)$ denotes inner product), and $\boldsymbol{p}_{\boldsymbol{R}_i}$ is the ith column vector of the projection matrix $\boldsymbol{P}_{\boldsymbol{R}}$.

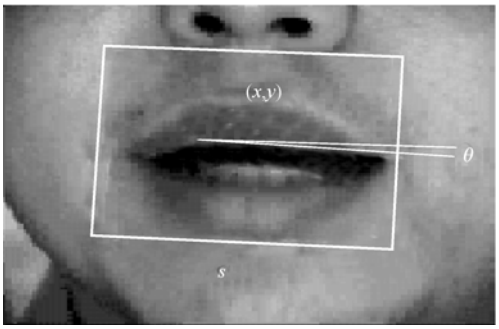


Fig. 2 An example of ROI extraction (the ROI is marked with the white rectangle)

To extract ROI, the feature dimension required is no more than three in practice. However, the feature dimension can be increased to describe the ROI more precisely. We rebuild the image of ROI with extracted frame visual feature $\boldsymbol{V}^{\mathrm{II}}$ in corresponding PCA subspace to give an intuitionist demonstration. By giving the original image and the rebuilt image of ROI, Fig. 3 shows examples that demonstrate the effectiveness of $\boldsymbol{V}^{\mathrm{II}}$ in reconstructing the image in the ROI with extracted ten-dimensional PCA subspace. It is observed from the demonstration that the most useful information for describing the lip's shape is reserved. In our experiments, the visual feature spaces with more than six dimensions are not considered.

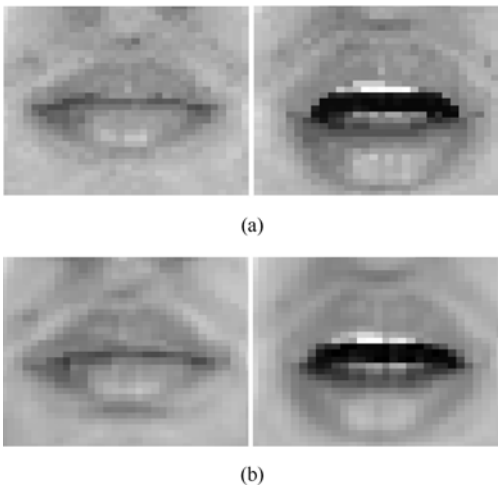


Fig. 3 Examples of ROI rebuilding in ten-dimensional PCA subspace. (a) Original ROI image; (b) Rebuilt ROI image

In the final step, the combination of the frame features of

$2K$+1 adjacent frames are placed into a large vector $\boldsymbol{V}_{\mathrm{co}}^{\mathrm{II}}(t) = [\boldsymbol{V}^{\mathrm{II}}(t\text{-}K), \cdots, \boldsymbol{V}^{\mathrm{II}}(t), \cdots, \boldsymbol{V}^{\mathrm{II}}(t\text{+}K)]$. Then we perform PCA on it to reduce its dimension back to d for the inter-frame feature $\boldsymbol{V}^{\mathrm{III}}$. Although the process does not introduce more information, it is expected to obtain features that characterize the lips' movement better.

3 Audio-visual VAD system

3.1 Models and decision rules

In general, VAD can be regarded as a bi-class pattern recognition problem in a d-dimensional feature space. The two classes to be distinguished are voice class S and non-voice class N. According to pattern recognition theories, some mathematical models of the two classes should be built in the feature space. Based on these models, appropriate classifiers (decision rules) for voice detection can be derived. Unlike the general problem of the bi-class pattern recognition, it is noticeable that voice segments and non-voice segments are both of long duration correspondingly. In other words, the states of the frames within a short interval are highly correlated. This characteristic can be utilized for better performance.

In this paper, statistical models are employed to describe voice and non-voice classes. First, the conditional probability density $p(\boldsymbol{F}|c)$ of the frame feature vector $\boldsymbol{F}$ is estimated for each given class $c = S$ or N. Then, the difference $\Delta d(\boldsymbol{F}) = \ln p(\boldsymbol{F}|S) - \ln p(\boldsymbol{F}|N)$ in logarithmic domain is used as the judging function for determining which class a frame belongs to. The classification can be done by comparing $\Delta d(\boldsymbol{F})$ with a threshold. The threshold can be set to zero simply. However, there may be some abnormal frames in both classes that cannot be described by the model properly or even more closely by the opposite model. For reducing the influence of these frames, we employ a couple of thresholds $T_S < 0 < T_N$ [7] to accommodate the long-term aforementioned characteristics: during the frame-synchronous VAD process, if current frame is classified to be in S state, then T_S will be used in the next frame, otherwise T_N will be used. Moreover, we can set a minimal length of each segment.

In frame-energy based audio VAD module, the feature is simply $F_A = \ln E$, where E is the frame-energy. The conditional probabilities $p(F_A|c)$ are modeled as Gaussian distributions, i.e.:

$$p\left(F_A \middle| c\right) = N\left(F_A \middle| \mu_{Ac}; \sigma_{Ac}^2\right),\ c \in \{S,N\} \tag{3}$$

where $N(\bullet|\mu;\sigma^2)$ denotes the probability density function of the one-dimensional Gaussian distribution with mean μ and variance σ^2. Assuming $\sigma_{AS} = \sigma_{AN} = \sigma_A$, we obtain $\Delta d_A(F_A) = K[F_A - (\mu_{AN} + \mu_{AN}) / 2]$, here we use the constant K as a concise expression. Because the paused frames in voice segments and frames disturbed by non-stationary noises in non-voice segments are all abnormal frames as mentioned earlier, the bi-threshold method should be used. We select the two thresholds as $T_S = \Delta d_A(\mu_{AN} + \Delta\mu_1) / K$ and $T_N = \Delta d_A(\mu_{AS} + \Delta\mu_2) / K$, where $\Delta\mu$ are also constants. Then the corresponding thresholds in feature domain can be given by:

$$F_S = \mu_{AN} + \Delta\mu_1,\ F_N = \mu_{AS} - \Delta\mu_2 \text{ (see Fig. 4)} \tag{4}$$

The thresholds lead to the traditional audio VAD rule. Here the key problem is how to estimate μ_{AN} and μ_{AS}, because these parameters change considerably with the volume of the speaker and the level of background noises. In our experiments, a fuzzy clustering method [7] is used to estimate the two parameters online for environment adaptation.

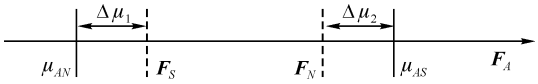


Fig. 4 A demonstration of audio models and decision rule for VAD

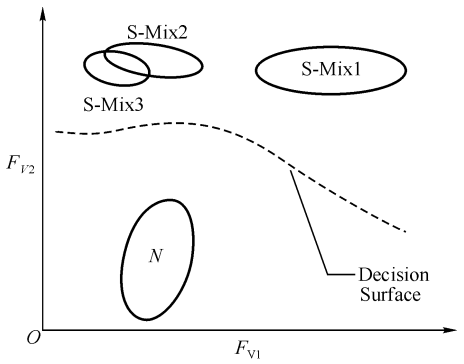
In visual VAD module, we represent the visual feature that can properly describe the static and dynamic characteristics of the lips' shape as $\boldsymbol{F}_{\mathrm{V}}$. Generally, the feature dimension is more than one. Note that the shape of the speaker's lips distorts only slightly during periods of silence but may appear in diverse modes when speaking. Therefore, for non-voice class, we use a multi-dimensional single Gaussian distribution:

$$p\left(\boldsymbol{F}_{\mathrm{V}} \middle| N\right) = N\left(\boldsymbol{F}_{\mathrm{V}} \middle| \boldsymbol{\mu}_{\mathrm{VN}}; \boldsymbol{\Sigma}_{\mathrm{VN}}\right) \tag{5}$$

where $N(\bullet|\boldsymbol{\mu};\boldsymbol{\Sigma})$ denotes Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Therefore, for non-voice class, we use a multi-dimensional single Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and for the voice class, we use a multi-dimensional Gaussian mixture distribution:

$$p\left(\boldsymbol{F}_{\mathrm{V}} \middle| S\right) = \sum_{m=1}^{M} w_m N\left(\boldsymbol{F}_{\mathrm{V}} \middle| \boldsymbol{\mu}_{\mathrm{VS}m}; \boldsymbol{\Sigma}_{\mathrm{VS}m}\right) \tag{6}$$

where M is the number of mixtures and w_m is the weight of the mth Gaussian mixture. Since the consistence of the visual features of both the two classes is much better, we can use $T_S = T_N = 0$ as the thresholds to distinguish the two classes. An example of the visual models and the corresponding decision surface in a two-dimensional visual feature space are shown in Fig. 5.



(S-Mix$i$: the ith Gaussian mixture of the model of voice; N: the model of non-voice model; $F_{\mathrm{V}i}$: the ith coordinate of the visual feature)

Fig. 5 A demonstration of visual models and decision rule for VAD

The visual models used in VAD are estimated offline in our work. Note that the parameters of the visual models are independent of the environments, so it is reasonable to estimate them only one time in advance. Correspondingly, the parameters of audio models must be estimated online because the volume and level of background noises may change momentarily in a long period of multimodal stream. To estimate the parameters of the visual models, we do VAD by audio modal on the training stream, then the VAD results are used as labels for training the visual models. Since there are not many parameters of visual models to be estimated, a little training data is enough.

3.2 System framework

The audio-visual fusion strategies in multimodal systems can be classified into two broad categories: feature fusion and decision fusion. The former puts the audio and visual features together for the combined features used in decision, while the latter fuses the results of the two modes at an appropriate level. To build an audio-visual VAD system, we should analyze the strengths and weaknesses of the two modes first. Considering that the lips' movement is tightly correlated with the state whether the speaker is keeping silence, while the frame-energy may be very low even in the voice segments, one expects that the sentences can be extracted from the multimodal stream more precisely by using the visual modal than by using the audio modal; so the sentence-breaking ability of the visual modal is expected to be better. However, better frame-detection performance may be achieved by audio modal than by visual modal because the frame-energy changes more dramatically than the lip's shape at sentence boundaries. To fully utilize the strength of each modal, we adopt a two-stage decision-fusion strategy in our audio-visual VAD system, as illustrated in Fig. 6.

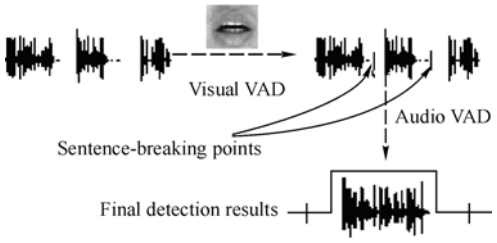


Fig. 6 Framework of audio-visual VAD

In the first stage, the sentence-breaking points, defined as the midpoint of the end of one detected sentence and the start of the next detected sentence, are extracted by visual VAD. In the second stage, a bi-direction audio VAD is performed on each segment between one sentence-breaking point and the next one in order to obtain more precise sentence boundaries. The audio VAD process is performed not only in the forward direction from the start of the segment but also in the backward direction from the end of the segment. In both directions, the process stops once the boundary is detected since we believe that there is only one sentence in one segment extracted by visual VAD. In this way both sentence-breaking performance and frame-detection performance are improved.

4 VAD performance evaluation

In our experiments, two measures are defined for testing the two most important aspects of the VAD performance. They are frame error rate and sentence-breaking error rate.

When used in speech communication or speech coding, the most important measure of VAD performance is the frame error rate P_{FE}. It is defined as the percentage of the false-classified frame number relative to the total frame number. It can be calculated by $P_{\mathrm{FE}} = P_{\mathrm{FF}} + P_{\mathrm{FM}}$, where P_{FF} is a false alarm rate defined as the percentage of the number of the frames that is detected to be voice but non-voice actually, and P_{FM} is a missed alarm rate defined as the percentage of the number of the frames that is detected to be non-voice but voice actually, both relative to the total frame number (see Fig. 7).

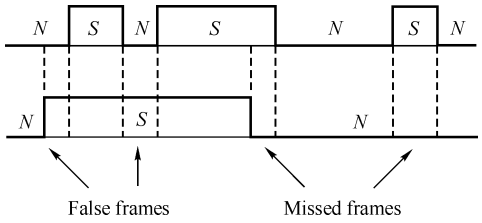


Fig. 7 A demonstration of frame errors. Top: labeled sentence; Bottom: the VAD result

When used in automatic speech recognition (ASR) system, another important purpose of VAD is to break the audio stream into sentences for recognition. The sentence-breaking performance will infect the understanding ability of the entire system evidently. Therefore, for evaluating the breaking performance, sentence-breaking error rate P_{BE} is used. It is defined as the percentage of the number of false sentence-breaking points relative to the number of the total sentence-breaking points. P_{BE} can be calculated by $P_{\mathrm{BE}} = P_{\mathrm{BD}} + P_{\mathrm{BI}}$, where P_{BD} is breaking deletion error rate, and P_{BI} is breaking insertion error rate. To clarify the concept, we treat the multimodal stream as a series of segments, which are voice and non-voice alternates. Given the sentence label, correct segments are obtained: $\{\mathrm{RN}_1, \mathrm{RS}_1, \mathrm{RN}_2, \cdots \mathrm{RS}_{K-1}, \mathrm{RS}_K\}$, where K is the total non-voice segment number. Considering the VAD result, the midpoint of each detected non-voice segment is called sentence-breaking point: $\{\mathrm{BP}_l \mid 1 \leqslant l \leqslant L\}$, where L is the number of non-voice interval in VAD result. A break insertion error appears if there is more than one sentence-breaking point detected in a non-voice interval or more than zero sentence-breaking points detected in a voice

interval. A break insertion error appears if there is no sentence-break point detected in a non-voice interval (see Fig. 8). We can calculate the deletion error number N_{BD} and insertion error number N_{BI} as follows:

$$\begin{aligned} N_{\mathrm{BD}} &= \sum_{k=1}^{K} N_{\mathrm{BD}}^{k} = \sum_{k=1}^{K} \prod_{l=1}^{L} \left[1-\delta\left(\mathrm{RN}_k, \mathrm{BP}_l\right)\right] \\ N_{\mathrm{BI}} &= \sum_{k=1}^{K-1} \sum_{l=1}^{L} \delta\left(\mathrm{RS}_k, \mathrm{BP}_l\right) \\ &\quad + \sum_{k=1}^{K} \left\{\left(1-N_{\mathrm{BD}}^{k}\right)\left[\sum_{l=1}^{L} \delta\left(\mathrm{RN}_k, \mathrm{BP}_l\right)-1\right]\right\} \end{aligned} \tag{7}$$

where $\delta(R,\mathrm{BP}) = \begin{cases} 1 & \text{if BP in } R \\ 0 & \text{otherwise} \end{cases}$. The sentence-breaking error rate can then be calculated by:

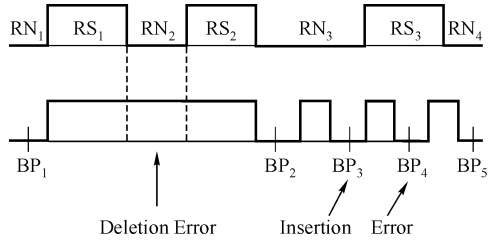


Fig. 8 A demonstration of sentence-breaking errors. Top: labeled sentence; Bottom: VAD result

$$P_{\mathrm{BE}} = \frac{N_{\mathrm{BD}} + N_{\mathrm{BI}}}{K} \tag{8}$$

5 Experiments

Our audio-visual database is collected for Mandarin audio-visual large vocabulary continuous speech recognition (LVCSR). It consists of video and audio of male speakers uttering a script of 1 560 sentences for five times. The video is captured in color at a frame rate of 29.97 Hz (NTSC). The audio is captured at a rate of 16 KHz using 16 bit quantization. One hundred sentences are used for model training and the rest for testing.

Our experiments have two purposes: one is to select the best visual feature and model in audio-visual VAD by optimizing several parameters; the other is to test and compare the performance of several VAD approaches.

The goal of our first experiment is to find out the optimal value of several parameters, including the dimension of visual feature and the number of Gaussian mixtures for voice class in the visual VAD module.

First, we build both the non-voice model and the voice model with one mixture, i.e., single Gaussian distribution model, in variant visual feature spaces for selecting the best visual feature dimension. The results are listed in Table 1. As we can see, the frame error rate and sentence-breaking error rate are both minimized by three-dimensional visual features without delta information.

Subsequently, the number of Gaussian kernels in the voice model is selected experimentally by considering both the frame error rate and the sentence-breaking error rate using three-dimensional visual features. The results are listed in Table 2. As we can see, the voice model of two Gaussian kernels leads to the best performance. It is noticeable that the more complicated model does not always lead to better performance.

Table 1 Visual feature dimension selection

Feature	Dim	P_{FE}/%	Feature	Dim	P_{BE}/%
$\boldsymbol{V}^{\mathrm{III}}$	1	15.51	$\boldsymbol{V}$	2	4.36
	2	3.99		4	9.71
	3	3.87		6	7.11
	4	3.93		8	7.02
	5	3.92		10	6.80
	6	3.90		12	6.69

Table 2 Number of kernels selection in visual voice model

M	P_{FE}/%	P_{BE}/%
1	3.87	0.08
2	3.37	0.05
3	3.38	0.13
4	3.44	0.15
5	3.40	0.14

In the second experiment, we focus on comparing the performances of several VAD systems. In audio VAD, we apply fuzzy clustering based upon Bayesian information criterion to estimate the two energy thresholds online [8]. The results are listed in Table 3. We can observe that visual VAD performs significantly better than audio VAD in sentence-breaking error rate. However, the frame error rate of audio VAD is not better than visual VAD as expected. It can be due to too many sentence-breaking errors. Hence, the fusion framework for audio-visual VAD system is still appropriate. As shown in the last row of Table 3, the frame error rate of bi-modal VAD is P_{FE} = 2.13 %, and the corresponding relative reduction is 55.0 % compared with audio VAD.

Table 3 Comparison of several VAD approaches (relative reductions are listed in the brackets)

Approach	P_{FE}/%	P_{BE}/%
Audio VAD (baseline)	4.73	3.24
Visual VAD	3.37 (28.8)	0.05 (98.5)
Audio-visual VAD	2.13 (55.0)	0.05 (98.5)

6 Conclusions

The research on multimodal system shows us a new way of

resolving the VAD problem. Based upon the research on the general framework of VAD, we can build a novel VAD system using visual features characterizing static and dynamic natures of lips. It is shown by experiment results that the visual model performs significantly better than audio VAD, especially in terms of sentence-breaking error rate. Consequently, by fusing audio and visual information, a bi-model VAD system even outperforms any uni-modal system in frame error rate, which shows that the system can break audio-visual stream into sentences and obtain the corresponding boundaries precisely.

References

1. Lamel L. F., Rabiner L. R., Rosenberg A. E. et al. An improved endpoint detector for isolated word recognition, IEEE Trans. Acoust., Voice, Signal Processing, 1981, 29(8): 777–785
2. Shen J. L., Hung J. W., Lee L. S., Robust entropy based endpoint detection for voice recognition in noisy environments, Proc 4th Int Conf on Spoken Language Processing (ICSLP'96), Philadelphia, 1996: 881–884
3. Chen Tsuhan, Audiovisual speech processing, IEEE Signal Processing Magazine, 2001, 18(1): 9–21
4. Liu Peng, Wang Zuo-ying, Visual information assisted Mandarin large vocabulary continuous speech recognition, Proc. NIP-KE'03, 2003
5. Kirby M., Sirovich L., Application of the Karhunen-Loeve procedure for the characterization of human faces, IEEE Trans. Pattern Analysis and Machine Intelligence, 1990, 12(1): 103–108
6. Nelder J. A., Mead R., A simplex method for function optimization, Comput. J., 1965, 7(4): 308–313
7. Tanyer S. G., Ozer H., Voice activity detection in nonstationary noise, IEEE Trans Acoust, Voice, Signal Processing, 2000, 8(7): 478–482
8. Tian Ye, Wu Ji, Wang Zuo-ying et al., Fuzzy clustering and Bayesian information criterion based threshold estimation For robust voice activity detection, Proc 2003 IEEE Int. Conf. on Acoustic, Speech, and Signal Processing (ICASSP'03), Hong Kong: 2003, 444–447