VOICE ACTIVITY DETECTION USING VISUAL INFORMATION

Peng Liu and Zuoying Wang

Department of Electronics Engineering
Tsinghua University, Beijing 100084, P.R.China
liupeng@thsp.ee.tsinghua.edu.cn

ABSTRACT

In traditional voice activity detection (VAD) approaches, some features of audio stream, for example frame-energy features, are used for voice decision. In this paper, we present a general framework of visual information based VAD approach in multi-modal system. Firstly, The Gauss mixture visual models of voice and non-voice are designed, and the decision rule is discussed in detail. Subsequently, the visual feature extraction method for VAD is investigated. The best visual feature structure and the best mixture number are selected experimentally. Our experiments show that using visual information based VAD, prominent reduction in frame error rate (31.1% relatively) is achieved, and the audio-visual stream can be segmented into sentences for recognition much more precisely (98.4% relative reduction in sentence break error rate), compared to frame-energy based approach in clean audio case. Furthermore, the performance of visual based VAD is independent of background noise.

1. INTRODUCTION

Voice activity detection (VAD) is of great importance in many speech-processing systems. Traditionally, it is performed based on some kinds of features of audio stream, for example, frame-energy [1] or entropy [2]. But the frame-energy based approach is sensitive to noise because it is designed for audio detection, not voice detection substantially. The entropy based approach pays more attention to the spectral characteristic of voice, but it still cannot deal with the environments with non-stationary noise, for example, crosstalk noise.

Here we present a novel VAD approach using visual information. Recently, more and more attention is abstracted by the multimodal interaction system, which uses visual modal, especially the movement of the speaker's lips, in addition to the audio modal for better perception performance or more intuitive expression. When an audio-visual front-end is available, we can also use the visual modal for VAD. The mainly advantage is that visual information is independent of background noise and tightly correlated with the speaker's expression.

In this paper, several key problems in visual based VAD are investigated: The visual features extraction method for features that describing the lips movement properly, the models for voice and non-voice in visual feature space, and the decision rule. Furthermore, we define the frame error rate and the sentence break error rate for evaluating the performance of VAD. Then experiments are performed for selecting the best visual features and models. Finally, the performance of visual information based approach is compared with frame-energy based approach.

2. VISUAL MODELS AND DECISION RULE

In general, VAD can be regarded as a two-class pattern recognition problem in a d-dim feature space. Firstly, the model of the voice class C_V and the non-voice class C_N are built. Then the decision rule should be selected. Furthermore, we can take advantage of the characteristic of that the voice or non-voice segment is of long term relatively for better detection performance.

In the frame-energy based approach, the dimension of feature E is one. The classes are represented by their mean values μ_{ES} and μ_{EN}. Given an audio frame, the distances from the two classes to the frame-energy are calculated in logarithmic domain for decision: $d_{Ec} = \left|\ln E - \ln \mu_{Ec}\right|, c \in \{N, V\}$. Then the simplest decision can be made by compare $\Delta d_E = d_{EN} - d_{EV}$ with zero. However, the frame-energy can be very low even in the voice segment, so the threshold-based method [3], which uses high and low thresholds for decision, is often adopted in practice. The most important problem is to estimate the levels of noise and speech energy to compute the decision thresholds. In [4], fuzzy clustering is used for estimating the energy of voice and background noise online, which leads to perfect VAD performance.

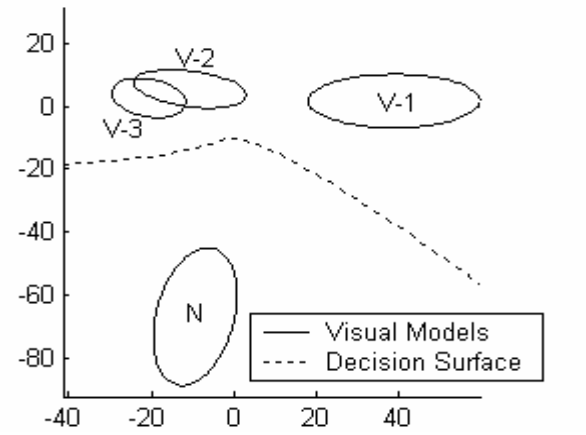

Figure 1. Visual models and decision rule for VAD

When the visual information is used, the visual feature $\mathbf{V}$ should be selected to represent the static or dynamic characteristic of the lips' shape. (Visual feature extraction method will be discussed in section 3.) The feature dimension is more than one generally. Voice or non-voice model can be represented with some forms of probability distribution in the visual feature space. In our work, the non-voice model is represented with single Gauss distribution (SGD), and the voice model is represented with Gauss mixture distribution (GMD). It is because that the lips of the speaker distort slightly when keeping silence, but may appear several kinds of shape modes when speaking. The distances from the two classes to the visual feature can be calculated as follows:

$$d_{\mathbf{V}N} = -\ln\left[p_N(\mathbf{V})\right] = -\ln\left[N\left(\mathbf{V}\middle|\boldsymbol{\mu}_N;\mathbf{S}_N\right)\right]$$

$$d_{\mathbf{V}V} = -\ln\left[p_V(\mathbf{V})\right] = -\ln\left[\sum_{m=1}^{M} w_m N\left(\mathbf{V}\middle|\boldsymbol{\mu}_{Vm};\mathbf{S}_{Vm}\right)\right]$$

where M is the mixture number of voice model, $N\left(\bullet\middle|\boldsymbol{\mu};\mathbf{S}\right)$ represents Gauss distribution function with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{S}$, and w_m is the weight of the m-th Gauss mixture. For another approach, we can replace the summation in GMD by maximization:

$$d'_{\mathbf{V}V} = -\ln\left[\max_{m=1}^{M} N\left(\mathbf{V}\middle|\boldsymbol{\mu}_{Vm};\mathbf{S}_{Vm}\right)\right]$$

This approach is inclined to make decision between the non-voice model and the nearest mixture to it in the voice model. Detection performance of these two approaches is almost the same in practice. Then we calculate the difference value $\Delta d_{\mathbf{V}} = d_{\mathbf{V}N} - d_{\mathbf{V}V}$ for decision. Here we use the same high threshold and low threshold of zero because the visual modal can distinguish between voice and non-voice perfectly. A demonstration of visual models and decision surface is shown in Figure 1.

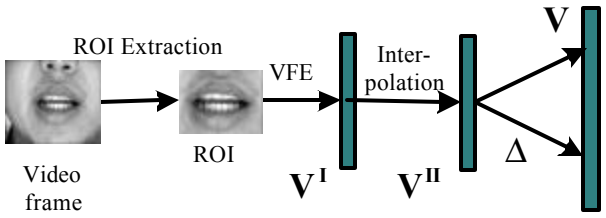

Figure 2. Block diagram of visual feature generation

3. VISUAL FEATURE EXTRACTION

A general framework of visual feature extraction is illustrated in Figure 2. Given a video frame including the rough area of mouth, a cascade of algorithm steps is applied to it for the final visual features. Firstly, a lip-tracking algorithm is applied to the video frame for the region of interest (ROI). Secondly, some kind of feature extraction method is applied to ROI for the original visual feature $\mathbf{V}^{\mathbf{I}}$. For the comparison of frame error rates with audio based VAD, we interpolated the original visual feature for the same sample rate as the audio stream to get the interpolated visual feature $\mathbf{V}^{\mathbf{II}}$. Finally, $\mathbf{V}^{\mathbf{II}}$ is complemented with its first order deference for the final visual feature $\mathbf{V} = \left\{\mathbf{V}^{\mathbf{II}}, \Delta\mathbf{V}^{\mathbf{II}}\right\}$. The feature extraction framework is mainly designed for audio-visual speech recognition. Here we use it for system consistency.

Primary component analysis (PCA) [5] is a widely used data-driven linear transform in feature selection. It takes advantage of prior knowledge for better data describing. Here we use PCA for ROI extraction and feature extraction in a single step.

For finding the ROI in the rough area of lips, we seek in it using model matching. The model is built by finding the several main variation modes of ROI using PCA. Given the training frames labeled with ROI rectangles, we re-sample the rectangle with M columns and N rows to get the vector $\mathbf{R}$ with a dimension of MN. Firstly, the mean vector $\boldsymbol{\mu}_{\mathbf{R}}$ and the covariance matrix $\mathbf{S}_{\mathbf{R}}$ are calculated. Then the eigenvectors corresponding to the largest d eigenvalues is selected as the column vectors of $\mathbf{P}_{\mathbf{R}}\left(MN \times d\right)$, which is the projection matrix to a d-dim subspace that best describes the variation of the ROI vector. The fitness between a candidate ROI vector $\mathbf{R}$ and the subspace is measured by calculating the Euclidean distance between them:

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

where $\mathbf{p}$ is the vector that consists of the parameters determining the ROI. We assume that $\mathbf{p} = (x, y, s, \theta)$, the

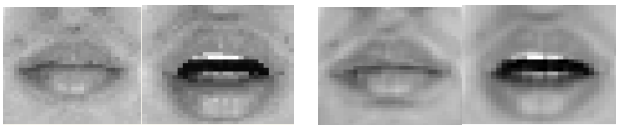

Figure 3. Examples of PCA feature rebuilding. The first two: Original ROI images. The last two: ROI images rebuilt with 6-dim PCA features

variables are: x, y the central coordinates, s the scale and θ the rotation angle. Then the model matching can be represented as the following optimization problem.

$$\mathbf{p}_{opt} = \arg\max_{\mathbf{p}} D(\mathbf{p})$$

Because the derivative of dis tance function cannot be calculated analytically, we use downhill simplex method [6] for numerical solution.

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

Practically, the feature dimension required in ROI extraction is no more than 2. However, the dimension can be increased for describing the ROI more precisely. For giving an intuitive demonstration of PCA features, we rebuild the image of ROI with extracted feature in 6-dimensional PCA subspace. Some examples of PCA feature rebuilding are shown in Figure 3. We can see that most useful information for distinguish the lip's movement is carried by the 6-dim PCA visual feature.

The time complexity of visual feature extraction is very low because the ROI extraction and feature selection is accomplished in one step.

4. PERFORMANCE EVALUATION

The performance of VAD frond-end can be evaluated from several aspects according to different applications. In our work, two measures are defined for evaluation.

When used in speech communication or speech coding, the most important measure of VAD performance is the frame error rate:

$$P_{FE} = P_{FF} + P_{FM}$$

where P_{FF} is false alarm rate defined as the percentage of the frames which detection to be voice but non-voice actually, and P_{FM} is missed alarm rate defined as the percentage of the frames which detection to be non-voice but voice actually, both relative to the total frames.

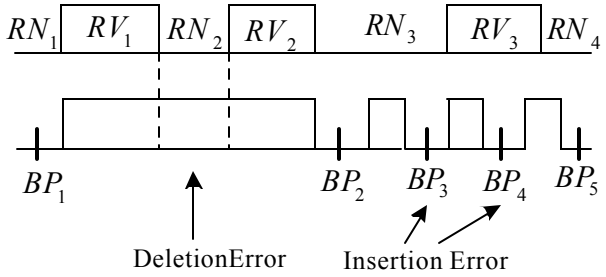

Figure 4. A demonstration of sentence break errors (Top row: labeled sentence. Bottom row: detected result)

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 break performance will infect the understanding ability of the entire system evidently. Therefore, sentence break error rate is defined to evaluate the break performance:

$$P_{BE} = P_{BD} + P_{BI}$$

where P_{BD} is break deletion error rate, and P_{BI} is break insertion error rate (See Figure 4). The audio stream can be label into a series of segments, which are voice and non-voice alternant, according to the correct labeled voice: $\{RN_1, RV_1, RN_2, \ldots, RV_{K-1}, RN_K\}$, where K is the total non-voice interval number. Considering the VAD result, the midpoint of each non-voice segment is called sentence break point: $\{BP_l | 1 \le l \le L\}$, where L is the number of non-voice interval in VAD result. Break insertion errors appear if there is more than one sentence break points detected in a non-voice interval or more than zero sentence break points detected in a voice interval, and break insertion errors appear if there is not any sentence break point detected in a non-voice interval. We can calculate the deletion error number N_{BD} and insertion error number N_{BI} as follows:

$$N_{BD} = \sum_{k=1}^{K} N_{BD}^{k} = \sum_{k=1}^{K} \prod_{l=1}^{L} [1 - \delta(RN_k, BP_l)]$$

$$N_{BI} = \sum_{k=1}^{K-1} \sum_{l=1}^{L} \delta(RV_k, BP_l) + \sum_{k=1}^{K} [(1 - N_{BD}^{k})][\sum_{l=1}^{L} \delta(RN_k, BP_l) - 1]]$$

where $\delta(R, BP) = \begin{cases} 1 & if\ (BP\ \ in\ \ R) \\ 0 & otherwise \end{cases}$. P_{BE} can then be represented as $(N_{BD} + N_{BI})/K$. A demonstration is shown in Figure 4.

5. EXPERIMENTS

Our bi-modal database in collected for Mandarin audio-visual large vocabulary continuous speech recognition

Feature	$\mathbf{V}_{\mathbf{PCA}}^{\mathbf{II}}$ (Interpolated PCA feature)					
Dim	1	2	**3**	4	5	6
P_{FE}(%)	15.51	3.99	**3.87**	3.93	3.92	3.90
Feature	$\mathbf{V}_{\mathbf{PCA}}\left(\mathbf{V}_{\mathbf{PCA}}^{\mathbf{II}}+\Delta\mathbf{V}_{\mathbf{PCA}}^{\mathbf{II}}\right)$ (Final feature)					
Dim	2	4	6	8	10	12
P_{FE}(%)	4.36	9.71	7.11	7.02	6.80	6.69

Figure 5. VAD performance comparison of different visual feature structures with SGD voice model

M	1	2	3	4	5
P_{FE}(%)	3.87	**3.37**	3.38	3.44	3.40
P_{BE}(%)	0.076	**0.051**	0.126	0.152	0.139

Figure 6. VAD performance comparison of different voice mixture numbers

research. It consists of video and audio of male speakers uttering 863 training scripts (1560 sentences) for five times. The video is captured in color at a frame rate of 29.97Hz (NTSC). The audio is captured at a rate of 16 KHz using 16 bits quantization. In our VAD experiments, 10% of the sentences are used for model training and the rest for testing.

The first experiment is performed for selecting the best visual feature structure and the best mixture number of voice model. Firstly, we build both the non-voice model and the voice model with one mixture (single Gauss model) in variant visual feature spaces for selecting the best visual feature structure. The results are listed in Figure 5. In this experiment, only frame error rates are compared. We can see that the frame error rate is minimized when 3-dim interpolated PCA feature is used. Subsequently, the mixture number of voice model is selected by testing both the frame error rate and the sentence break error rate using 3-dim interpolated PCA feature, the results are listed in Figure 6.We can see that the 3-dim interpolated PCA feature and the voice model of 2 Gauss mixtures leads to the best performance. It is noticeable that more complex model or feature doesn't always leads to better performance because that more disturb is also introduced.

The second experiment is performed to compare the performance of our visual information based VAD and traditional frame-energy based VAD. In the latter approach, fuzzy clustering and Bayesian criterion based threshold estimation [3] is used. The results are listed in Figure 7. We can see that the frame error rate and sentence break error rate of visual based VAD are lower than those of frame-energy based VAD prominently.

6. CONCLUSIONS

Method	Frame energy based	Visual Based	Relative reduction
P_{FE}(%)	4.73	3.37	**31.1%**
P_{BE}(%)	3.235	0.051	**98.4%**

Figure 7 VAD performance comparison of frame energy based method and visual based method

The introduction of visual information shows us a completely new and effective approach in VAD because of the information carried by the lips' movement is highly correlated with the voice producing process of the speaker. The visual information based VAD framework discussed in this paper can be implemented with many kinds of visual features and corresponding decision rule. Experiments show that visual information based VAD works better than energy-based method in both the frame detection performance and the sentence break performance. It is remarkable that the sentence break error is almost avoided, which means that the multi-modal stream can be segmented into sentences for recognition very closely to the speaker's expression units.

Furthermore, visual information cannot be affected by the background noise, which leads to perfect robustness. Future work can be focused on the combined VAD using both audio and visual information.

7. REFERENCES

[1] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilson, "An improved endpoint detector for isolated word recognition", *IEEE Trans. Acoust., Voice, Signal Processing*, v29, pp. 777–785, Aug. 1981.

[2] J. L. Shen, J. W. Hung, and L. S. Lee, " Robust entropy based endpoint detection for voice recognition in noisy environments" , in *Proc. ICSLP'96*, 1996.

[3] S. G. Tanyer and H. Ozer, " Voice activity detection in nonstationary noise", *IEEE Trans. Acoust., Voice, Signal Processing*, v8, pp. 478–482, Jul. 2000.

[4] Y. Tian, J. Wu, Z. Wang, et al. " Fuzzy Clustering And Bayesian Information Criterion Based Threshold Estimation For Robust Voice Activity Detection", in *Proc. ICASSP'03,* 2003.

[5] M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces," *IEEE Trans. Pattern Analysis and Machine Intelligence",* vol. 12, no. 1, pp. 103-108, Jan. 1990.

[6] J. A. Nelder and R. Mead, "A simplex method for function optimization," *Comput. J.,* vol. 7, no. 4, pp. 308–313, 1965.