Comparative Evaluation of Different Methods for Voice Activity Detection

Hongfei Ding, Koichi Yamamoto and Masami Akamine

Corporate Research & Development Center, Toshiba Corporation
1, Komukai-Toshiba-cho, Saiwai-ku, Kawasaki 212-8582, Japan
koichi10.yamamoto@toshiba.co.jp, masa.akamine@toshiba.co.jp

Abstract

This paper presents a comparative evaluation of different methods for voice activity detection (VAD). A novel feature set is proposed in order to improve VAD performance in diverse noisy environments. Furthermore, three classifiers for VAD are evaluated. The three classifiers are Gaussian Mixture Model (GMM), Support Vector Machine (SVM) and Decision Tree (DT). Experimental results show that the proposed feature set achieves better performance than spectral entropy. In the comparison of the classifiers, DT shows the best performance in terms of frame-based VAD accuracy as well as computational cost.

Index Terms: voice activity detection (VAD), Gaussian Mixture Model (GMM), Support Vector Machine (SVM), Decision Tree (DT)

1. Introduction

The importance of voice activity detection (VAD) has considerably increased due to its application in numerous areas such as mobile communications and speech recognition. In real-life environments, VAD is required to be robust to background noises, insensitive to different input gain and computable in real-time. Generally, VAD consists of two parts: feature extractor and classifier. The feature extractor extracts frame-based acoustic features from input signal. Then, the classifier calculates a speech likelihood of each frame. If the speech likelihood exceeds a threshold, the frame is classified as speech otherwise, it is classified as non-speech.

In order to meet the requirements for VAD, various acoustic features and classifiers have been proposed [1-8]. A spectral entropy feature was firstly used for VAD in [1]. The spectral entropy shows the uncertainty of the spectrum and it is a good feature for extracting the inherent characteristics of the speech spectrum. In the last decade, several methods have been proposed to improve the robustness of the spectral entropy. In [2], the spectral entropy was combined with an energy feature in order to compensate the robustness for non-stationary noises such as babble and background music. In [3][4], the spectral entropy was calculated from the spectrum with “whitening” filter, which resulted in greater robustness for colored noises. However, the performance of the spectral entropy is unsatisfactory in diverse noisy environments.

In this paper, we propose a novel feature set in order to improve the robustness of the spectral entropy. The new feature set consists of the spectral entropy, the signal-to-noise ratio (SNR) and a spectral Mahalanobis distance. Furthermore, their neighboring frames are concatenated. Each component is introduced to compensate the performance of the spectral entropy: The SNR has the information of the relative level of an observed signal compared to the level of the background noise. Taking account of the variance information of the background noise, the Mahalanobis distance shows how far the current frame is from noise frames. The concatenation of the neighboring frames is introduced to take advantage of temporal information.

Moreover, we evaluate three classifiers with the proposed feature set. The three classifiers are Gaussian Mixture Model (GMM), Support Vector Machine (SVM) and Decision Tree (DT). GMM is one of the most common classifiers for VAD [5][6]. In [6], discriminative feature extraction (DFE) was introduced in order to train the parameters with a minimum classification error (MCE) criterion [9]. In [7], SVM was applied for VAD with multiple features. In [8], DT was used to make a final decision using multiple speech/non-speech pre-classifiers. In this paper, we propose to apply a multi-mixture DT [10] for VAD.

The rest of this paper is organized as follows. Section 2 describes a novel feature set. Section 3 gives a general introduction for each classifier and Section 4 presents the experimental results. Finally, conclusions and future work are given in Section 5.

2. Proposed feature set

The proposed feature set consists of three different features based on the signal spectrum: 1) spectral entropy, 2) SNR and 3) spectral Mahalanobis distance. All of these features require an estimation of the noise spectrum. For this purpose, the initial 80ms of signal is assumed to be background noise and an initial noise spectrum is estimated. Thereafter, this noise spectrum is updated as:

\[
\hat{N}_{f,j} = \begin{cases} 2\hat{N}_{f,j-1} + (1-\lambda)S_{f,j} & \text{if } \log_{10}\left(\frac{S_{f,j}}{N_{f,j-1}}\right) < T_s \\ \hat{N}_{f,j-1} & \text{otherwise} \end{cases}
\]  \hspace{1cm} (1)

where, \( f \) and \( t \) are the frequency band index and time index, respectively. \( S \) and \( \hat{N} \) are the amplitude spectrum of observed signal and estimated noise, respectively. \( T_s \) is a predetermined threshold to control update. In the same way, the variance of the noise spectrum can also be estimated.

2.1. Spectral entropy

Spectral entropy [1] is based on the assumption that the noise spectrum is more uniformly distributed and therefore results in higher entropy. For each frame, the spectral entropy \( H \) is calculated as follows:


\[ H = - \sum_{f} P_f \cdot \log P_f , \]  

(2)

\[ P_f = \frac{S_f}{\sum_f S_f} , \]  

(3)

where, \( S_f \) is the amplitude spectrum of the \( f \)th frequency band. The following measures are applied in order to improve the robustness: a) Only the frequency bands from 250Hz to 5000Hz are used to calculate entropy [1]. b) The observed spectrum \( S_f \) is divided by the estimated noise spectrum \( N_f \) given by Eq. (1), which results in whitening the noise spectrum [2]. c) A white noise with small amplitude is added to the input waveform before feature extraction to eliminate the undesired effects of division by very small values [2].

Spectral entropy does not use the information of the relative level of an observed signal compared to the level of the background noise. This information is captured in SNR, as explained in the next section.

2.2. SNR

Signal-to-noise ratio (SNR) is widely used as a feature for VAD because of its simplicity. In this paper, the SNR is calculated as follows:

\[ \text{SNR} = \log_{10} \left( \frac{|\hat{S}_{\text{mel}}|}{|\hat{N}_{\text{mel}}|} \right) , \]  

(4)

where, \( S_{\text{mel}} \) and \( N_{\text{mel}} \) are the observed spectrum and the estimated noise spectrum in mel-filterbank space, respectively.

2.3. Spectral Mahalanobis distance

A Mahalanobis distance feature between the observed spectrum and the estimated noise spectrum in mel-filterbank space is expected to utilize the variance information of the background noise, resulting in greater robustness. This distance \( D_{\text{mf}} \) is computed as:

\[ D_{\text{mf}} = \left( S_{\text{mel}} - \hat{N}_{\text{mel}} \right) W^{-1} \left( S_{\text{mel}} - \hat{N}_{\text{mel}} \right) , \]  

(5)

\[ S'_{\text{mel}} = \frac{S_{\text{mel}}}{|\hat{S}_{\text{mel}}|} , \quad \hat{N}'_{\text{mel}} = \frac{\hat{N}_{\text{mel}}}{|\hat{N}_{\text{mel}}|} , \]  

(6)

where, \( S'_{\text{mel}} \) and \( \hat{N}'_{\text{mel}} \) are the spectra which are normalized onto the unit circle. \( W \) is the diagonal covariance matrix of the normalized spectrum \( N_{\text{mel}} \).

2.4. Frame concatenation

In order to utilize temporal information, three features (spectral entropy, SNR and spectral Mahalanobis distance) from neighboring frames are concatenated. The concatenation is motivated by our preliminary experiments where concatenation showed better performance than using delta and delta-delta features.

3. Classifiers for voice activity detection

3.1. Gaussian mixture model

The distributions of the features are estimated as a weighted combination of \( K \) Gaussian distributions. The likelihood of a feature vector \( x \) is given by:

\[ p(x \mid \Lambda) = \sum_{i=1}^{K} w_i p_i(x \mid \mu_i, \Sigma_i) , \]  

(7)

where, \( \mu_i, \Sigma_i \) and \( w_i \) are the mean vector and covariance matrix of the \( i \)th mixture component and its weight, respectively. For the purpose of VAD, one GMM is trained with speech data and another is trained with non-speech data. The log-likelihood ratio \( L(x) \) of speech and non-speech GMMs is calculated as follows:

\[ L(x) = \log p(x \mid \Lambda_s) - \log p(x \mid \Lambda_n) , \]  

(8)

where, \( \Lambda_s \) and \( \Lambda_n \) represent the parameter sets for speech model and non-speech model, respectively. If \( L(t) \) exceeds a speech threshold, the frame is classified as speech otherwise, it is classified as non-speech.

3.2. Support vector machine

Support Vector Machine (SVM) finds a hyperplane that separates training data with the maximum margin by introducing the slack variable \( \xi \) and error cost \( C \) [11]. For a non-linear SVM, the idea is to map a linear non-separable pattern into a high-dimensional space, which might lead to a linear separable solution with the slack variable \( \xi \). By applying a kernel trick, it can avoid the real mapping. Instead, it uses a kernel function in the original low-dimension space.

In the evaluation, the output of a non-linear SVM is calculated by:

\[ f(x) = \sum y_i \alpha_i K(s_i, x) + b \]  

(9)

where, \( s_i \) is the \( i \)th support vector (SV) and \( y_i \) is its sign. \( \alpha_i \) is a pre-calculated Lagrange multiplier, \( b \) is a bias and \( K(s_i, x) \) is the kernel function. In this paper, the following radial basis function (RBF) kernel is used:

\[ K(x, y) = \exp \left( -\frac{|x - y|^2}{2\sigma^2} \right) . \]  

(10)

In the SVM-based VAD, \( f(x) \) is used as the speech likelihood and it is compared to a speech threshold.

3.3. Decision tree

3.3.1. VAD using DT

DT is trained using the “grow down first then prune back” scheme [10]. A question of every node is found to split the training data such that the total likelihood of positive (speech) frames is maximized. A tree stops growing when it meets a predefined criterion. For each node, a likelihood \( L \) is calculated as follows:

\[ L = \frac{N_{\text{speech}}}{(N_{\text{speech}} + N_{\text{non-speech}}) \cdot P_r} , \]  

(11)

where, \( N_{\text{speech}} \) and \( N_{\text{non-speech}} \) are the counts of the speech and non-speech frames which visit this node, respectively. \( P_r \) is the prior probability of speech frames in the training data set.

In the evaluation phase, each test frame enters at the root node and chooses the child to visit according to the answer of the question being asked until it reaches a leaf node. Finally, DT returns a pre-computed value of the leaf node, which provides the speech likelihood of the feature. In the DT-based VAD, the likelihood is compared to a speech threshold.
3.3.2. Mixture of DTs
Due to the finite number of leaves, DT could be discrete. In other words, DT could be sensitive to a small change in a feature value. In order to solve the problems, a mixture of DTs (DT-forest) was proposed in [10]. The trees are trained using different components of the training data in a similar manner as the training of GMMs. The initial forest is trained by random assignment. Then, EM algorithm is used to reassign the data points and retrain the trees. This process is repeated during a certain number of iterations. In this paper, we propose to apply the mixture of DTs for VAD.

4. Experimental results
4.1. Experimental setups
The training dataset consists of 14000 short utterances. Different noises such as car noise, factory noise, babble noise and white noise were artificially added to clean utterances with various SNR levels from 0dB to 30dB. The test dataset consists of 16000 utterances of Japanese city names. Noises other than those used in the training dataset were added to the test dataset at SNR of 0dB, 5dB, 10dB and 20dB.

The input signal was sampled at 11025Hz and framed using a hamming window. The length of one frame was 23ms with 8ms shift. The number of mel-filterbanks for the SNR and spectral Mahalanobis distance was set to eight. The previous 10th and the subsequent 10th frames were concatenated and the total dimension of feature vector became nine. For the GMM and SVM classifiers, each element of the feature vector was scaled to have zero mean and one variance.

For the GMM classifier, 9×9 projection matrix was applied to decorrelate vector elements. The model complexity for each classifier was tuned according to the performance and computational cost, which are described below.

4.1.1. GMM classifier
The projection matrix was obtained using principal component analysis (PCA). 32-mixture diagonal GMMs were used for both speech and non-speech models. The GMMs were trained by EM algorithm, where the initial mean vectors were obtained using the LBG algorithm and the initial diagonal variances and mixture weights were set to 1 and 1/32, respectively. Moreover, in order to train the parameters of the projection matrix and GMM parameters discriminatively, discriminative feature extraction (DFE) [6][9] was also applied.

4.1.2. SVM classifier
SVMlight [12] was used for training and evaluation. In order to reduce the number of support vectors to a realistic usage, subset data were chosen from the training database. The total number of training frames became 10000. σ in Eq. (10) was set to 1.0 and C’ was set to 0.5, which were tuned based on the error rate of $\hat{\xi}_a$-estimate [13]. The final number of support vectors became approximately 3500.

4.1.3. DT classifier
Both a single and a multi-mixture DT were evaluated. The multi-mixture DT has five trees. A chi-square test with required score of 0.001 was used to validate a split. The “worst node first out” scheme was used to prune the tree back to a demand size. Tree size was tuned based on frame-based VAD accuracy and it resulted in approximately 2000 nodes (1000 leaves) for each tree. For the multi-mixture DT, the number of iterations was 10.

4.2. Frame-based VAD accuracy
Receiver operating characteristic (ROC) curves for 5dB noisy environments were drawn to show comparisons of the frame-based VAD accuracy (Figures 1 to 3). Equal error rate (EER) is given in Table 1. In these Figures and Table, “Entropy” represents the results using the spectral entropy (described in Section 2.1) with simple threshold processing.

“GMM” and “GMM-DFE” show the results of the proposed feature set using GMM classifiers without/with DFE training, respectively. “SVM” shows the results of the proposed feature set with the SVM classifier. “DT-single” and “DT-forest” show the results of the proposed feature set with single DT and 5-mixture DT classifiers, respectively.

The experimental results indicate that the proposed feature set has made a significant improvement over the single spectral entropy in all noisy conditions. Furthermore, it outperforms the spectral entropy for all classifiers. The use of the additional features (SNR and spectral Mahalanobis distance) and frame concatenation contribute to compensate the robustness of the spectral entropy in diverse noisy environments.

As for the comparison of the classifiers, SVM and DT achieve better performance than the GMM classifiers even if DFE was applied to parameter training. GMM assumes the feature distribution to be Gaussian. There is a possibility that the distribution of the proposed feature set did not match the Gaussian distribution. In the comparison of DT classifiers, DT-forest slightly outperforms DT-single. From the experimental results, we can see that DT-forest gives the best performance for the proposed feature set.

4.3. Computational cost and model complexity
The model complexity for each classifier was tuned according to the performance and computational cost. The GMM classifiers with 32 mixtures had 1214 parameters:

\[
( ( 9(\text{feature dim}) \times 2(\text{mean/variance}) \times 32(\text{mixtures}) ) + 31(\text{weight}) ) \times 2(\text{speech/non-speech})
\]

In the preliminary experiments, there was no improvement even if the number of mixtures was increased.

SVM had around 3500 support vectors in the nine-dimensional feature space. It means that SVM had approximately 31500 parameters and needed much bigger computational cost for classification compared to the other classifiers.

For DT-single, the number of parameters was around 2000 (1000 internal node thresholds and 1000 leaf likelihoods). The average height of the tree was about 10. Therefore, only about 10 comparisons were needed for the classification of each frame. DT-forest (5-mixture DT) needed five times more plus some extra computation for combining the likelihood. However, its computational cost was still lower than GMM and SVM.

<p>| Table 1. EER (%) using test data with SNR of 0dB, 5dB, 10dB and 20dB. The “all” is the average EER in all conditions. |
|---|---|---|---|---|---|</p>
<table>
<thead>
<tr>
<th>Feature</th>
<th>Classifier</th>
<th>white</th>
<th>car</th>
<th>babble</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entropy</td>
<td>Threshold</td>
<td>14.21</td>
<td>8.95</td>
<td>16.24</td>
<td>12.13</td>
</tr>
<tr>
<td>Proposed</td>
<td>GMM</td>
<td>2.58</td>
<td>3.95</td>
<td>7.99</td>
<td>4.78</td>
</tr>
<tr>
<td>Proposed</td>
<td>GMM-DFE</td>
<td>2.39</td>
<td>3.50</td>
<td>6.88</td>
<td>4.27</td>
</tr>
<tr>
<td>Proposed</td>
<td>SVM</td>
<td>2.12</td>
<td>2.60</td>
<td>5.40</td>
<td>3.65</td>
</tr>
<tr>
<td>Proposed</td>
<td>DT-single</td>
<td>2.10</td>
<td>2.34</td>
<td>5.86</td>
<td>3.59</td>
</tr>
<tr>
<td>Proposed</td>
<td>DT-forest</td>
<td>2.00</td>
<td>2.28</td>
<td>5.46</td>
<td>3.35</td>
</tr>
</tbody>
</table>
of the classifiers, DT shows the best classification. Therefore, we expect that DT will be able to handle various kinds of categorical features such as speaker and environmental information.

5. Conclusions and future work

This paper reported a comparative evaluation of different methods for voice activity detection (VAD). A novel feature set was proposed in order to improve the robustness in diverse noisy environments. Furthermore, three classifiers for VAD were also evaluated. The three classifiers are Gaussian Mixture Model (GMM), Support Vector Machine (SVM) and Decision Tree (DT). Experimental results have shown that our proposed feature set achieves better performance than the conventional spectral entropy in diverse noisy environments. In the comparison of the classifiers, DT shows the best performance in terms of frame-based VAD accuracy as well as computational cost.

Future work includes the optimization of feature set for DT. DT does not make any assumption about features used for classification. Therefore, we expect that DT will be able to handle various kinds of categorical features such as speaker and environmental information.

6. References