A Bayesian Approach to Voice Activity Detection Using Multiple Statistical Models and Discriminative Training

Tao Yu, John H. L. Hansen

CRSS: Center for Robust Speech System, Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Texas, USA
tao.yu@student.utdallas.edu, john.hansen@utdallas.edu

Abstract

In this study, the problem of voice activity detection (VAD) is formulated in a Bayesian hypothesis testing framework. Unlike traditional VAD schemes that employ a single statistical model, multiple models are assumed to be potentially engaged with a priori probabilities, due to the statical diversity of the environmental noise degrading the speech. Moreover, the optimal a priori probabilities are explored using discriminative training based method, which is suggested to directly reduce the miss-hit rate and false-alarm rate of the VAD. As shown in the evaluations, VAD performance, both in terms of absolute performance and consistency across a diverse set of noise conditions, can be significantly improved using the proposed Bayesian method.

Index Terms: Voice activity detection (VAD), Bayesian hypothesis testing, discriminative training

1. Introduction

An important problem in many speech processing areas is the determination of speech presence periods in a given noisy signal. Effective VAD performance can significantly reduce the insertion errors in ASR, as well as improve speaker and language ID systems employing GMM models. In general, voice activity detection (VAD) belongs to the general category of statistical binary hypothesis testing problem where the VAD classifies the incoming signal into speech-active or speech-pause (noise only). It is assumed that the speech s is degraded by an uncorrelated additive noise n. Under two hypotheses $H_0$ (speech-pause) and $H_1$ (speech-active), the observation $x$ in the short-time Fourier transform (STFT) domain can be written as,

$$H_0(\text{speech-pause}): x_{k,t} = n_{k,t},$$

$$H_1(\text{speech-active}): x_{k,t} = s_{k,t} + n_{k,t},$$

where $k$ and $t$ are the frequency-bin and time-frame index, respectively. For the $t$th time-frame, time-frame-wise statistics for the VAD decision can be obtained using the maximum likelihood (ML) criterion as,

$$l_{k,t} = \log \frac{p(x_{k,t}|H_1)}{p(x_{k,t}|H_0)},$$

where the testing statistic $l_{k,t}$ is the log-likelihood ratio (LLR) of the $k$th frequency-bin in the $t$th time-frame; where

$p(x_{k,t}|H_i)$ with $i \in \{0, 1\}$ is the likelihood based on the probability density function (pdf) modeled for $H_i$. With this, a time-frame-wise decision rule can be established as,

$$l_t = \frac{1}{K} \sum_{k=1}^{K} l_{k,t} \frac{p_l}{\eta_l} \eta_t,$$

where $K$ is the total number of frequency-bins and $\eta$ is the detection threshold that controls the trade-off between the miss-hit rate and false-alarm rate.

An effective model for $p(x_{k,t}|H_i)$ is the focus of this study. A Gaussian statistical model was originally employed in [1]. Later, a Laplacian statistical model was found to be more effective in some noisy environments in [2], and a mixed Laplacian-Gaussian statistical model was also studied in [3]. Recently, [4] evaluated VAD performance for different statistical models across various noisy environments and suggested an environment-oriented model selection scheme. Recent advancement for “Environmental Sniffing” [5] suggested measurable improvement in speech system performance when environmental knowledge is available to the speech system.

Indeed, in any practical speech related applications, it is impossible to obtain an exact statistical model for either the speech-pause hypothesis $H_0$ or the alternative speech-active hypothesis $H_1$, due to the diversity of the noise conditions. In this study, rather than assign a single true statistical model as in conventional VAD, a class of candidate models is considered and assumed to be tentatively viable in the Bayesian hypothesis testing framework, as discussed in Section 2. In Section 3, a discriminative training method is introduced, where an optimal a priori probability for model selection is obtained. Evaluations are conducted in Section 4 and conclusions drawn in Section 5.

2. Bayesian Hypothesis Testing Based on Multiple Statistics Model

2.1. Bayesian Hypothesis Testing

Suppose a set of $M$ statistical models are pre-chosen as candidates, denoted as $\{\Lambda_1, \Lambda_2, \ldots, \Lambda_M\}$, the VAD testing statistics in (2) can be obtained using a Bayesian approach as,

$$l_{k,t} = \log \frac{p(x_{k,t}|H_1)}{p(x_{k,t}|H_0)} = \log \frac{\sum_{m=1}^{M} p(x_{k,t}|\Lambda_m, H_1) p_k(\Lambda_m | H_1)}{\sum_{m=1}^{M} p(x_{k,t}|\Lambda_m, H_0) p_k(\Lambda_m | H_0)}$$

where $p_k(\Lambda_m | H_i)$ is the a priori probability associated with statistical model $\Lambda_m$ under hypothesis $H$, for the $k$th frequency-
bin; \( p(x_{k,t} | \Lambda_{m}, H_t) \) is the likelihood of \( x_{k,t} \) under model \( \Lambda_{m} \) and hypothesis \( H_t \). In a manner different from the non-Bayesian approach, multiple statistical models are simultaneously considered and assigned with \( a \ priori \) probabilities which may vary according to the specific noise conditions (e.g., the potential noise types can be viewed as a set of colors from a painter’s plate, with the specific noise condition reflecting a mixture of noise “colors” at any time block.). The possible statistical models and the optimal \( a \ priori \) probability will be discussed in the following sections.

### 2.2. Candidate Statistical Models

The appropriate statistical models for the acoustic signal are either Gaussian or super-Gaussian. Here, three suitable models (i.e., \( M=3 \)) are directly employed from [4] and listed below:

1. **Gaussian Statistical Model (A1):**
   \[
   p(x_{k,t} | A_1, H_0) = \frac{1}{\pi \lambda_k^2} \exp \left\{ - \frac{|x_{k,t}|^2}{\lambda_k^2} \right\},
   \]
   \[
   p(x_{k,t} | A_1, H_1) = \frac{1}{\pi (\lambda_k^2 + \lambda_0^2)} \exp \left\{ - \frac{|x_{k,t}|^2}{\lambda_k^2 + \lambda_0^2} \right\},
   \]

2. **Laplacian Statistical Model (A2):**
   \[
   p(x_{k,t} | A_2, H_0) = \frac{1}{\lambda_0^2} \exp \left\{ - \frac{2|\lambda_0 x_{k,t}|}{\lambda_0^2} \right\},
   \]
   \[
   p(x_{k,t} | A_2, H_1) = \frac{1}{\lambda_0^2 + \lambda_k^2} \exp \left\{ - \frac{2|\lambda_0 x_{k,t}|}{\lambda_0^2 + \lambda_k^2} \right\},
   \]

3. **Gamma Statistical Model (A3):**
   \[
   p(x_{k,t} | A_3, H_0) = \frac{\sqrt{6}}{8\pi \lambda_k^2 |\lambda_k^{(R)}|^{0.5} |\lambda_k^{(I)}|^{0.5}} \times \exp \left\{ - \frac{\sqrt{3}(|\lambda_k^{(R)}| + |\lambda_k^{(I)}|)}{\sqrt{2} \lambda_k^2} \right\},
   \]
   \[
   p(x_{k,t} | A_3, H_1) = \frac{\sqrt{6}}{8\pi \lambda_0^2 + \lambda_k^2 |\lambda_k^{(R)}|^{0.5} |\lambda_k^{(I)}|^{0.5}} \times \exp \left\{ - \frac{\sqrt{3}(|\lambda_k^{(R)}| + |\lambda_k^{(I)}|)}{\sqrt{2} \lambda_0^2 + \lambda_k^2} \right\},
   \]

where \( \lambda_k^s \) and \( \lambda_k^i \), respectively, denote the estimated power of the speech \( s \) and noise \( n \) in the \( k \)th frequency-bin. Also \( x_{k,t}^{(R)} \) and \( x_{k,t}^{(I)} \) denote the real and imaginary parts of \( x_{k,t} \), respectively.

In this study, superscripts are used to denote the labels of a variable. For the Laplacian model (6) and Gamma model (7), it is more beneficial to substitute \( \lambda_0^2 + \lambda_k^2 \) with the smoothed magnitude spectrum \( |x_{k,t}| \) as done in [4],

\[
|\tilde{x}_{k,t}| = \alpha |x_{k,t}| + (1-\alpha) \sqrt{\lambda_0^2 + \lambda_k^2};
\]

where \( \alpha \) is a smoothing factor and is set to 0.9 as used in [4].

### 3. Discriminative Training for the A Priori Probability

It is clear that the \( a \ priori \) probability for each candidate model (i.e., \( p_k(\Lambda_m | H_t) \)) controls VAD performance and should be environment dependent. Therefore, environment dependent training can be employed to estimate the \( a \ priori \) probability. In this section, discriminative training is formulated to solve for the optimal priors.

#### 3.1. Training Setup

Since a VAD decision is made at each time-frame, a vector, \( \mathbf{x}_t = \{x_{1,t}, x_{2,t}, \ldots, x_{K,t}\}^T \), can be employed to denote all the STFT coefficients in a time-frame \( t \). Suppose a set of time-frame-wise labeled STFT coefficient vectors is used for training, denoted as \( \mathcal{X} = \{X^n, X^p\} \), where \( X^n = \{x^n_t, t=1, 2, \ldots, N^n\} \) and \( X^p = \{x^p_t, t=1, 2, \ldots, N^p\} \) represent the portion of the set containing all the STFT vectors labeled as speech-active or speech-pause, respectively. To simplify the notation, the following definitions are used,

- **Emission Probability:**
  \[
  b_{k,t,m,i} \triangleq p(x_{k,t} | \Lambda_m, H_t),
  \]

- **Model Weights:**
  \[
  w_{k,m,i} \triangleq p_b(\Lambda_m | H_t),
  \]

- **Parameter Space (i.e., the set of model weights):**
  \[
  \mathcal{W} \triangleq \{w_{k,m,i} : k=1, \ldots, K; m=1, \ldots, M; i=0, 1\},
  \]

- **Classification Function:**
  \[
  L(\mathbf{x}_t) \triangleq l_t,
  \]

\[
= \frac{1}{K} \sum_{k=1}^{K} \log \left( \sum_{m=1}^{M} w_{k,t,m,i} b_{k,t,m,i} \right) - \log \left( \sum_{m=0}^{M} \sum_{i=0}^{1} w_{k,t,m,i} b_{k,t,m,i} \right).
\]

#### 3.2. Minimal Classification Error Criterion

In discriminative training, VAD performance is directly associated with a designed objective function, which could be optimized based on the training data. Minimum classification error (MCE) learning [6] is a well known discriminative training approach and has been successfully applied for VAD weight training in [7], which aims to minimize the misclassification errors over the entire training set, and can be defined as,

\[
C(\mathcal{W}, \mathcal{X}) \triangleq \frac{1}{N^p} \sum_{t=1}^{N^p} I(\mathbf{L}(\mathbf{x}_t^n) - \eta) + \frac{1}{N^p} \sum_{t=1}^{N^p} I(\eta - \mathbf{L}(\mathbf{x}_t^p)),
\]

where \( \mathbf{1}(z) \) is an indicator function where for argument \( z \),

\[
\mathbf{1}(z) = \begin{cases} 1, & z > 0 \\ 0, & \text{otherwise} \end{cases}
\]

and can be approximated by a sigmoid function, that can be differentiated during the optimization as,

\[
\mathbf{1}(z) \approx D(\beta z) \frac{1}{1 + \exp(-\beta z)},
\]

where \( \beta \) is the gradient of the sigmoid function.

Basically, the MCE loss function in (13) is an average of two types of detection errors, (e.g., the miss-hit errors and false-alarm errors), and could be minimized through the adjustment of the model weights set \( \mathcal{W} \) in the classifier.

#### 3.3. Optimization Algorithm

Here, an optimization algorithm is derived for MCE training. To begin with, some intuitive constraints are imposed on the weights set \( \mathcal{W} \). Notice that \( \mathcal{W} \) is an set of \( a \ priori \) probabilities; hence we have \( \forall \{k, m, i\}, w_{k,m,i} \geq 0 \) and \( \forall \{k, i\}, \sum_{m=1}^{M} w_{k,m,i} = 1 \). Therefore, a parameter transformation can be employed as follows,

\[
w_{k,m,i} = (\tilde{w}_{k,m,i})^2.
\]
This parameter transform automatically guarantees the non-negativity of entries in $W$ throughout the optimization. Solving for the optimal weights $\tilde{w}_{k,m,i}$ simultaneously leads to optimal weights $w_{k,m,i}$. The optimal weights $w_{k,m,i}$ could be updated with a steepest descent technique as follows,

$$\tilde{w}_{k,m,r}^{t+1} = \tilde{w}_{k,m,r}^t - \frac{\partial C(W,X)}{\partial \tilde{w}_{k,m,r}} \bigg|_{w_{k,m,r}=\tilde{w}_{k,m,r}^t},$$

(17)

$$w_{k,m,r}^{t+1} = \frac{w_{k,m,r}^{t+1}}{\sum_{m=1}^M (\tilde{w}_{k,m,r}^{t+1})^2},$$

(18)

$$w_{k,m,i}^{t+1} = (\tilde{w}_{k,m,i}^{t+1})^2,$$

(19)

where $\epsilon>0$ is the learning rate for updating and $r$ is the iteration index; $\frac{\partial C(W,X)}{\partial w_{k,m,i}}$ is the gradient of the MCE loss function and could be obtained as,

$$\frac{\partial C(W,X)}{\partial w_{k,m,i}} = \frac{(-1)^{i+1}}{K \times N_p} \sum_{t=1}^{N_p} D(L(x_i^t) - \eta)(1-D(L(x_i^t) - \eta)) \times \frac{1}{\sum_{m=1}^M |w_{k,m,i}|^2} b_{k,t,m,i} \tilde{w}_{k,m,i},$$

$$- \frac{(-1)^{i+1}}{K \times N_p} \sum_{t=1}^{N_p} D(\eta - L(x_i^t))(1-D(\eta - L(x_i^t))) \times \frac{1}{\sum_{m=1}^M |w_{k,m,i}|^2} b_{k,t,m,i} \tilde{w}_{k,m,i}.$$  

(20)

The parameter transform step from (16) and (19), along with the normalization step in (18) guarantees that the constraints on $W$ are satisfied throughout the iterations.

4. Evaluations

4.1. Implementation

While the proposed VAD can be employed in many speech applications, here it is embedded for the speech enhancement. The implementation employs an analysis windows of 32ms as a time-frame, with a 50% overlap frame for recordings at an 8kHz sample rate. The STFT is calculated using a Hamming window with an FFT length of 256. The IMCRA algorithm [8] is used for noise power estimation and the decision-directed method [9] is used to estimate the speech power in each frequency-bin (in total of 128). To study the effectiveness of the proposed method, the VAD decision is only based on the current time-frame, (i.e., total of 10 minutes of training data and 10 minutes of test data). The percentage of hand-marked speech-active frames is 53.4%. The signal-to-noise ratio (SNR) is approximately 5dB.

Fig.1 compares VAD performance in terms of the receiver operating characteristics (ROC) curves for the test data. The proposed Bayesian VAD is compared with conventional VAD using a single statistical model. Since local optimal solutions may exist, ROC curves of the proposed VAD are plotted for different initial weights. In all noise conditions, the proposed Bayesian VAD significantly outperforms the conventional VAD. It is interesting to note that VAD performances are greatly improved for the cases where the initial model weights are selected in favor of a single Gaussian model, a single Gamma model, or uniform weights for the three models[1]; however, for the case where weights are initialized in favor of the Laplacian model, the improvements from the MCE training is not obvious for many of the noise conditions (e.g., babble, airport, car, etc.,), due to the local optimality of the Laplacian model.

It is also interesting to plot the optimal model weights (a priori probabilities) for the three candidate statistical models in different noise conditions, as shown in the stacked bar chart of Fig.2 (e.g., this reflects the diverse unique “color” structure of the noise environments). Under the speech-active hypothesis $H_1$, as in Fig.2 (a), the Laplacian model has the largest weights in a majority of the noise conditions, while the Gaussian model and Gamma model have relatively smaller weights, corresponding to our intuitive impression of a Laplacian distribution of the speech spectra. However, it would be expected and also shown in Fig.2 (b), that the Gaussian model is preferred over the other two models under the speech-pause hypothesis $H_0$.

5. Conclusions

The Bayesian hypothesis testing framework was investigated using multiple statistical models for the task of voice activity detection. The power of the proposed VAD is that multiple candidate models are considered and allocated with optimal environment dependent model weights, instead of using a single model as in the conventional VAD. The proposed scheme was evaluated within the noisy AURORA database, with measurable ROC improvement over single model based methods.

6. References


1For the preferred model, an initial weight of 0.99 is assigned, and for the other two models, 0.005 is assigned.

3116
Figure 1: ROC curves for different AURORA noise conditions: (a) babble, (b) airport, (c) car, (d) station, (e) exhibition and (f) restaurant. The proposed Bayesian VAD is evaluated for different initial model weights (i.e., \textit{a priori} probabilities), with MCEG, MCEL, MCEM and MCEU denoting the ROC curves of the proposed Bayesian VAD that is trained using the initial weights in favor of a single Gaussian model, a single Laplacian model and a single Gamma model, and uniform weights for the three models, respectively. As a comparison, ROC curves are also plotted for the conventional VAD that utilizes a single statistical model of either Gaussian, Laplacian or Gamma distribution.

Figure 2: Stacked bar charts of the optimal weights (i.e., \textit{a priori} probabilities, $w_1$, $w_2$ and $w_3$) for the three statistical models (e.g., Gaussian model, Laplacian model and Gamma model) across different AURORA noise conditions, under (a) speech-active hypothesis $H_1$ and (b) speech-pause hypothesis $H_0$, respectively.