

PHONE BASED VOICE ACTIVITY DETECTION USING ONLINE BAYESIAN ADAPTATION WITH CONJUGATE NORMAL DISTRIBUTIONS

Jianping Zhang, Wayne Ward and Bryan Pellom

Center for Spoken Language Research
University of Colorado at Boulder
Boulder, Colorado 80309-0594, USA

<http://cslr.colorado.edu>

ABSTRACT

In this paper, we developed a highly efficient frame-level on-line adaptive voice activity detection (VAD) algorithm for the telephone-based CU Communicator spoken dialog system. The adaptive algorithm uses prior speaker and channel statistics as well as acoustic features of current sample frames to update model parameters. The algorithm achieved .05xRT in contrast to .7xRT of a compared VAD algorithm using 5-state HMMs. We detail the adaptive algorithm and address some real-time implementation issues. Experiments on live collected data show that there is a 23% error reduction compared with G.729B VAD.

1. INTRODUCTION

Voice activity detection (VAD) is an important part of human-machine dialog systems. Previous algorithms vary in feature and model selection, hang-over scheme, choice of statistics and methods of noise processing. The energy-based approach [1] is a classic one and works well under high SNR conditions. In [2] high order statistics (HOS) are calculated from the Linear Prediction (LP) residual to distinguish speech from noise frames. This approach is based on findings that properties of HOS of speech are distinct from those of Gaussian-like noise. However, the HOS may not be effective in non-Gaussian noise environments. Other authors have considered a method of the fusion of a least-square periodicity estimator and a geometrically adaptive energy threshold. Results came from isolated speech corrupted by white noise [3]. In [4], a two-state (speech and noise) hang-over scheme was proposed in which the probability of transitioning from speech to noise was set empirically (e.g., $P(\text{speech} \rightarrow \text{noise}) = 0.1$). Parallel model combination (PMC) and speech enhancement (SE) methods were proposed in [5]. The estimate of corrupted-speech model in PMC may be obtained from clean speech model and noise model according to a mismatch function. In practical non-stationary noise environments, it is difficult to separate noise from noisy speech and the model combination procedure is computationally complex, so it is more suited for off-line applications.

In this paper, we evaluate our VAD algorithms in a telephone environment using a spoken dialog system, the CU Communicator [6, 7]. From our system requirements, the following elements must be considered in implementing a VAD algorithm: 1) Computational complexity. Since spoken dialog systems require real-time

performance, it needs to be efficient, stable and accurate. 2) Adaptation. The algorithm needs to account for varieties of speaker, channel and noisy environments. 3) General purpose. The algorithm should be general purpose and have strong theoretic background.

We focus on the Bayesian adaptive VAD algorithm because of its mathematical attractiveness as well as its successful use in other recognition tasks [8]. The adaptive VAD will be discussed in Section 2. Other VAD algorithms are implemented for comparison and introduced in Section 3. Evaluation setups are described in Section 4. Experimental results are given in Section 5 followed by a short conclusion.

2. BAYESIAN ADAPTIVE VAD

2.1. Principles of Bayesian adaptation

Suppose there are Q phone models $\lambda_0, \lambda_1, \dots, \lambda_{Q-1}$. Assume the probability density function (pdf) of the L -element observation vector \mathbf{x} conditioned on the q^{th} phone model parameter $\lambda_q = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ can be expressed as

$$f(\mathbf{x}|\lambda_q) = \frac{1}{(2\pi)^{\frac{L}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\} \quad (1)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean vector and covariance matrix of the model λ_q . If the components of \mathbf{x} , x_i and x_j ($i \neq j$), are statistically independent, $\boldsymbol{\Sigma}$ is a diagonal matrix. To deal with varieties of speaker, channel characteristics and noise conditions, a general purpose model $\lambda \in \{\lambda_i\}$ must be adapted according to observed samples and prior statistics which characterize differences among speakers and channels. An adaptive procedure mainly consists of two steps: decision and adaptation. During the decision step, the most likely phone is selected in the maximum a posteriori (MAP) sense by calculation of model likelihoods for all possible models. Assume \mathbf{x} an arbitrary observation vector, then the optimum phone q^* in the MAP sense is:

$$q^* = \arg \max_q P(\lambda_q|\mathbf{x}). \quad (2)$$

By Bayes rule, the term $P(\lambda_q|\mathbf{x})$ can be written as:

$$P(\lambda_q|\mathbf{x}) = \frac{f(\mathbf{x}|\lambda_q)P(\lambda_q)}{\sum_{i=1}^Q f(\mathbf{x}|\lambda_i)P(\lambda_i)}. \quad (3)$$

Since \mathbf{x} is an additional observation of the model λ_{q^*} , we can use it to update the model. For simplicity, consider one component

This work supported by DARPA through SPAWAR under grant # N66001-00-2-8906

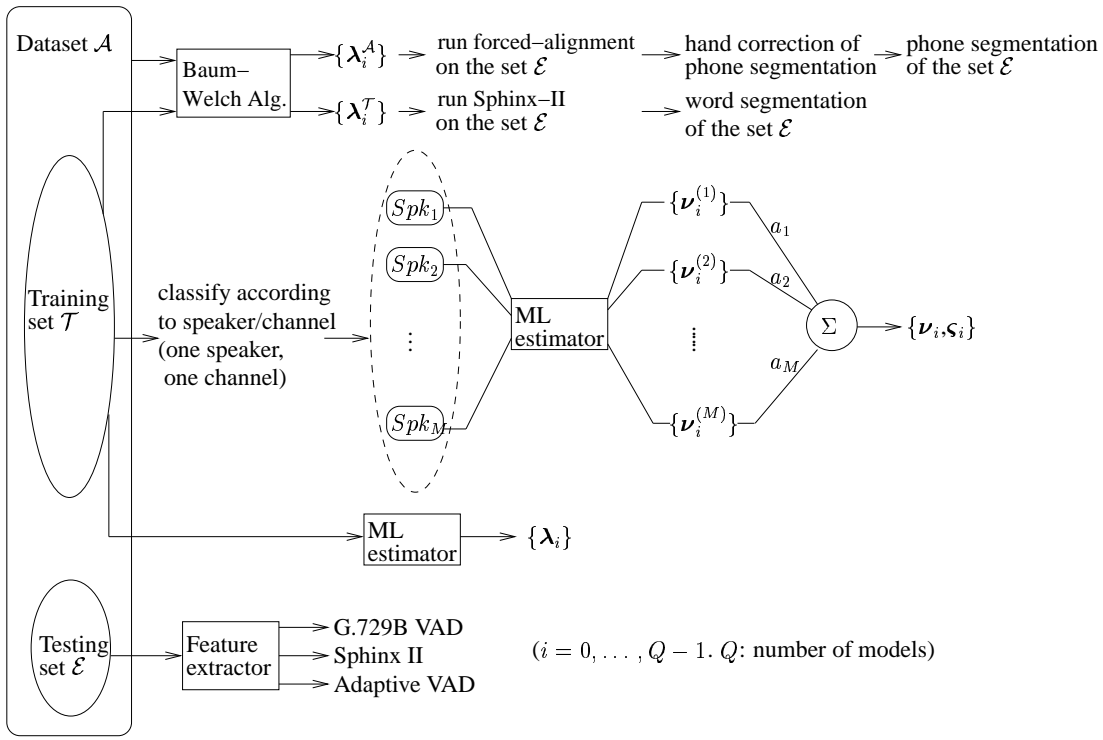


Fig. 1. Block diagram of VAD experimentation

λ of the model vector, and denote $\lambda = \{\mu, \sigma^2\}$ where μ and σ^2 are the mean and covariance of a component of an observation distribution. If the phone q^* has occurred n times, denote the n -observation sequence as $\mathbf{o} = o(0) \cdots o(n-1)$ and $\bar{\mathbf{o}}$ the mean of the n samples. Let $f(x|\mathbf{o})$ stand for $f(x|\lambda, \mathbf{o})$, then,

$$f(x|\mathbf{o}) = \int f(x|\mu) f(\mu|\mathbf{o}) d\mu. \quad (4)$$

Since

$$f(\mu|\mathbf{o}) = \alpha^{-1} f(\mathbf{o}|\mu) f(\mu)$$

where

$$\alpha = \int f(\mathbf{o}|\mu) f(\mu) d\mu,$$

use conjugate priors, that is, the prior pdf $f(\mu)$ and posterior pdf $f(\mu|\mathbf{o})$ belong to the same family of distributions for any number of n and any value of x , assuming $f(\mu) \sim \mathcal{N}(\nu, \varsigma^2)$ (the estimate of ν and ς^2 will be discussed later), then,

$$f(\mu|\mathbf{o}) \sim \mathcal{N}\left(\frac{\sigma^2 \nu + n \varsigma^2 \bar{\mathbf{o}}}{\sigma^2 + n \varsigma^2}, \frac{\sigma^2 \varsigma^2}{\sigma^2 + n \varsigma^2}\right),$$

and

$$f(x|\mathbf{o}) \sim \mathcal{N}\left(\frac{\sigma^2 \nu + n \varsigma^2 \bar{\mathbf{o}}}{\sigma^2 + n \varsigma^2}, \sigma^2 + \frac{\sigma^2 \varsigma^2}{\sigma^2 + n \varsigma^2}\right).$$

So the new mean value of the model λ is

$$\hat{\mu} = \frac{\sigma^2 \nu + n \varsigma^2 \bar{\mathbf{o}}}{\sigma^2 + n \varsigma^2}. \quad (5)$$

As $n \rightarrow \infty$, $\hat{\mu} \rightarrow \bar{\mathbf{o}}$. $\hat{\mu} \rightarrow \nu$ as $n \rightarrow 0$. From our experiments, ν is about equal to μ while ς^2 is smaller than σ^2 .

We only perform Bayesian adaptation on the Gaussian mean μ while the variance σ^2 is fixed. To estimate ν and ς^2 given the training set \mathcal{T} , we classified our training set into M subsets according to caller's identifier. Individual subset is used to train a speaker dependent model. Then ν and ς^2 are given, respectively, by

$$\nu = \sum_{m=1}^M a_m \nu^{(m)}, \quad (6)$$

and

$$\varsigma^2 = \sum_{m=1}^M a_m (\nu^{(m)} - \nu)^2 \quad (7)$$

where a_m is a weight associated with the m^{th} training subset.

2.2. Algorithm of adaptive VAD

Based on above discussion, a Bayesian adaptive VAD algorithm is performed in four steps:

Bayesian adaptive VAD algorithm:

1. Initialization

Set initial models to the general models $\{\lambda_i\}$. Reset the sum and number of observation vectors for phone i : $sum_o[i]$ and $num_o[i]$, i from 0 to $Q - 1$.

2. Decision

For each frame, calculate its MFCCs and other features. Decide the optimum phone q^* it belongs to in the MAP sense using Eq. (2).

Let d_N be the normalized distance between first two candidates and β a threshold (see Section 2.3).

if ($d_N < \beta$) /* not reliable decision */

go to step 4.

Add the feature values to the corresponding $sum_o[q^*]$ and increment corresponding $num_o[q^*]$ by 1.

3. Adaptation

According to above decision, obtain a new model (mean vector) for the phone using Eq. (5).

4. State machine

Add the corresponding phone to a phone history buffer.

a) *noise state*

if (current phone is speech)

if (duration in speech state $> T_1$)

switch to speech state.

b) *speech state*

if (current phone is noise)

if (duration in noise state $> T_0$)

switch to noise state.

In above algorithm, T_0 and T_1 are minimum frame numbers required to enter noise and speech state respectively. β will be explained later.

2.3. Some considerations

- Choice of T_0 and T_1 . In practical systems, T_1 affects the time delay of the system response. The larger T_1 is, the slower the system responds. In a telephone-based dialog system with barge-in capabilities [6], a larger T_1 causes longer echoes from system prompts to go into the recorded signal since the system will not stop playing until the beginning of speech is declared by a VAD module. T_0 controls the delay of the end-point detection. Generally speaking, the system is not sensitive to T_0 .
- Threshold β . β is an important parameter in a system. It controls the model update. The distance between first and second candidates is defined as:

$$d = \log \max_q P(\lambda_q | \mathbf{x}) - \log \max_{q \neq q^*} P(\lambda_q | \mathbf{x}) \quad (8)$$

where q^* is defined in Eq. (2). The normalized distance d_N is defined as $d_N = d/d_{max}$ where d_{max} is estimated from the training data set. In our implementation, $d_{max} = 10.0$ and $\beta \simeq 0.3$. The meaning of the equation is since not every decision is correct, we do not update the model if current decision is not reliable enough. The larger the distance between first two candidates becomes, the more accurate the detection is.

3. COMPARATIVE VAD ALGORITHMS

The following VAD techniques were implemented for assessing the proposed algorithm.

1. Energy-based VAD

VAD using short-term energy [1] is simple, efficient, and has reasonable performance under high SNRs. The energy and all acoustic features throughout the paper are computed on a frame at the length of 25.6 ms and the interval of 10 ms.

2. G.729B VAD

The ITU-T G.729B VAD [9] uses full-band energy E_f , low-band energy E_l , zero-crossing rate, and line spectral frequencies $LSF_i, i = 1, \dots, p$ where p is the order of an inverse filter. In our implementation, $p = 12$ and E_l is defined to be

$$E_l = 10 \log_{10} \left(\frac{1}{N} \mathbf{h}^T \mathbf{R} \mathbf{h} \right) \quad (9)$$

where N is the frame length, \mathbf{R} is a 13×13 Toeplitz auto-correlation matrix, and \mathbf{h} the impulse response of a 13-tap FIR filter with cutoff frequency at 1 kHz. Full-band energy is $E_f = 10 \log_{10} \left[\frac{1}{N} R(0) \right]$. We implemented the FIR filter by Parks-McClellan algorithm [10]. The algorithm realized an optimum equiripple approximation of an ideal low-pass filter. It is a linear, causal and stable filter with the group delay $\tau = 6$ samples and the Nyquist frequency $f = 4$ kHz. The line spectral frequencies are computed using the algorithm introduced in [11].

3. Combined Method

The combined method uses short-term energy, zero-crossing rate and MFCCs as features. An L -element MFCC vector \mathbf{x} is composed of 12 cepstra: $x_i, 1 \leq i \leq 12$ and one power coefficient x_0 . The vector mean and covariance matrices for the Q base phones are trained from the set \mathcal{T} using maximum likelihood (ML) estimation.

4. EVALUATION SETUPS

The overall data set \mathcal{A} is collected from live calls. \mathcal{A} is divided into two disjoint sets: training set \mathcal{T} and testing set \mathcal{E} . The training set \mathcal{T} includes calls from 286 users totaling about 15.3 hours of recordings. The set \mathcal{E} consists of twenty one-minute dialogues of which half is from male and half from female. 50% calls of both genders are from cellular phones. The evaluation set \mathcal{E} is selected so that experimental results would not favor any case.

Examination of the testing set revealed the diversity and complexity of noisy environments that any practical dialog system may experience. In addition to common noises (e.g., clicks, breath, vehicle noise), noises from TV or radio broadcasting, cross talking, baby crying, laughing, hesitating, phone going on/off hook were also found. The SNRs vary from 5 dB to 40 dB with an average of 20 dB within the set \mathcal{E} .

In order to obtain accurate results, each utterance in the set \mathcal{E} is segmented by aligning acoustic features against models $\{\lambda_i^{\mathcal{A}}\}$ which are trained from the set \mathcal{A} using Baum-Welch algorithm. Then segmentation results are hand corrected. The block diagram of the overall VAD experimentation is depicted in Fig. 1. Throughout our experiments, 50 base phones (e.g., AA, BD, IX) and 12 filler (noise) phones (e.g., SIL, BACKGROUND, BREATH) are used. An example of hand-corrected utterance is illustrated in Fig. 2

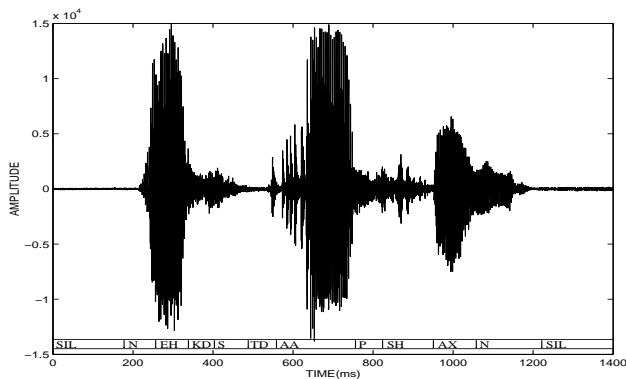


Fig. 2. Waveform and hand-labeled phone sequence for the utterance “Next option.”

5. EXPERIMENTAL RESULTS

Since the testing data has been segmented, we can calculate error detection rates for each phone. False alarm rate P_f , missing detection rate P_m and total error rate P_e are defined as $P_f = N_{n \rightarrow s} / N$, $P_m = N_{s \rightarrow n} / N$, and $P_e = P_f + P_m$ in which $N_{n \rightarrow s}$ and $N_{s \rightarrow n}$ are number of noise or speech frames falsely detected as speech or noise frames and N the total frame number of the testing data. Our testing data has 118984 frames, of which 68% are noise frames and 32% are speech frames.

1. Comparison of phone models with different states

In the above discussion, we focus on features and pdfs of various vectors. We trained a phone model λ for each of the base phones and filler phones from training corpus disregarding the fact that phone features may have fundamental changes during one phone duration length. From the point of HMMs, we use one state to represent the whole phone duration. It is reasonable to further improve the model λ by using multi-state and mixture Gaussian distribution HMMs. We resort to Sphinx-II [12] and test how five-state HMMs can improve the detection accuracy. Results are shown in Table 1. It is worth to note that using multi-state HMMs can greatly decrease the detection error from 17.7% to 13.2%. Note no adaptation and no features other than MFCCs are used in these cases. Note that the VAD using five-state HMMs takes .7xRT on a PIII 600 MHz machine.

It is well known that language model (LM) is a must for a recognizer. So a language model trained from the training set \mathcal{T} is applied to the recognizer. Again, the error rate decreased further to 9.7%. By incorporating the LM, we have provided about 6 bits/word of information for the \mathcal{E} .

Table 1. RESULTS OF DIFFERENT STATES (%)

# HMM states	P_f	P_m	P_e
1-state HMM	15.4	2.3	17.7
5-state HMM	10.3	2.9	13.2
5-state HMM + LM	6.7	3.0	9.7

2. Comparison with other VAD algorithms

Results for (A) energy-based VAD, (B) G.729B VAD, (C) the combined VAD and (D) Bayesian adaptive VAD are shown in Table 2. From the table, we find that methods (B) and (C) have some improvements over (A). However, due to the diversity and complexity of the practical non-stationary noise environments, the gain of using additional features is limited. The Bayesian adaptive method updates acoustic models at the frame level so that the adapted models can better match actual channel conditions. Results show there is a 23% error reduction compared with G.729B VAD. Also it is worth to note that the adaptive VAD takes 0.05xRT which should satisfy the requirement of real-time spoken dialog systems.

Table 2. RESULTS OF VARIOUS VAD APPROACHES (%)

Method	P_f	P_m	P_e
(A) Energy based	7.3	6.4	13.7
(B) G.729B VAD	5.6	7.8	13.4
(C) Combined method	6.1	6.4	12.5
(D) Bayesian adaptive	5.0	5.3	10.3

6. CONCLUSIONS

In this paper, we present our recent development on voice activity detection. We have tried various acoustic features and models. Bayesian adaptation was exploited to perform on-line adaptation and account for various noisy environments. Linguistic information plays an important role in voice activity detection. We can predict that a robust voice activity detector which incorporates short-term energy, zero-crossing rate, MFCCs and on-line adaptation techniques may meet many real-time dialog system requirements.

7. REFERENCES

- [1] L. R. Rabiner and M. R. Sambur, “An algorithm for determining the endpoints of isolated utterances,” *Bell Syst. Tech. J.*, 54(2):297–315, 1975.
- [2] E. Nemer, R. Goubran and S. Mahmoud, “Robust Voice Activity Detection Using Higher-Order Statistics in the LPC Residual Domain,” *IEEE Trans. Speech and Audio Proc.*, 9(3):217–231, 2001.
- [3] S. Tanyer and H. Özer, “Voice Activity Detection in Nonstationary Noise,” *IEEE Trans. Speech and Audio Proc.*, 8(4):478–482, 2000.
- [4] J. Sohn, N. Kim and W. Sung, “A Statistical Model-Based Voice Activity Detection,” *IEEE Signal Proc. Lett.*, 6(1):1–3, 1999.
- [5] R. Sarikaya and J. Hansen, *Robust Speech Activity Detection in the Presence of Noise*, ICSLP, Sydney, 1998.
- [6] J. Zhang, W. Ward, B. Pellom, X. Yu and K. Hacioglu, *Improvements in Audio Processing and Language Modeling in the CU Communicator*, Eurospeech, 3:2209–12 Denmark, 2001.
- [7] <http://communicator.colorado.edu>
- [8] C. Lee, C. Lin and B. Juang, “A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models,” *IEEE Trans. Signal Processing*, 39(4):806–814, 1991.
- [9] A. Benyassine, E. Shlomot and H. Su, “ITU-T Recommendation G.729 Annex B: A Silence Compression Scheme for Use with G.729 Optimized for V.70 Digital Simultaneous Voice And Data Applications,” *IEEE Commun. Mag.*, 64–72, Sept. 1997.
- [10] T. Parks and J. McClellan, “A Program for the Design of Linear Phase Finite Impulse Response Digital Filters,” *IEEE Trans. Audio and Electroacoustics*, 20(3):195–199, 1972.
- [11] F. Soong and B. Juang, *Line Spectrum Pair (LSP) and Speech Data Compression*, ICASSP 1.10.1-1.10.4, 1984.
- [12] M. Ravishankar, *Efficient Algorithms for Speech Recognition*, Ph.D. Dissertation, Carnegie Mellon University, 1996.