Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2007, Article ID 43218, 7 pages
doi:10.1155/2007/43218

Research Article

A Semi-Continuous State-Transition Probability HMM-Based Voice Activity Detector

H. Othman and T. Aboulnasr

School of Information Technology and Engineering, Faculty of Engineering, University of Ottawa, Ontario, Canada K1N 6N5

Received 15 December 2005; Revised 13 November 2006; Accepted 28 November 2006

Recommended by Thippur V. Sreenivas

We introduce an efficient hidden Markov model-based voice activity detection (VAD) algorithm with time-variant state-transition probabilities in the underlying Markov chain. The transition probabilities vary in an exponential charge/discharge scheme and are softly merged with state conditional likelihood into a final VAD decision. Working in the domain of ITU-T G.729 parameters, with no additional cost for feature extraction, the proposed algorithm significantly outperforms G.729 Annex B VAD while providing a balanced tradeoff between clipping and false detection errors. The performance compares very favorably with the adaptive multirate VAD, option 2 (AMR2).

1. INTRODUCTION

Actual speech activities normally occupy 60% of the time of a regular conversation in a telecommunication system [1]. Voice activity detection (VAD) enables reallocating resources during the periods of speech absence. In modern telecommunication systems, VADs, in conjunction with comfort noise generator (CNG) and discontinuous transmission (DTX) modules, play a critical role in enhancing the system performance.

A VAD distinguishes between speech and nonspeech frames in the presence of background noise. In general, VAD errors can be categorized into two main types of errors, notably clipping errors and false detection errors. Clipping errors occur when speech frames are misclassified as noise frames, which is intolerable in speech encoders due to its effect on speech intelligibility, while false detection errors are due to misclassifying noise frames as speech frames. Echo cancellation systems are normally sensitive to this type of errors because it results in incorrect parameter adaptation.

Traditional VAD algorithms rely on legacy features such as frame energy and zero-crossing rate (ZCR). In recent VAD algorithms, more features are used in different schemes. Among those are likelihood ratio (LR) that is based on complex Gaussian distribution of the signal discrete Fourier transform (DFT) in [2, 3], Higher-order statistics (HOS) of the LPC residuals of the signal that include skewness and kurtosis in [4], power envelope dynamics in [5], and fractals in [6].

In this paper, we focus on voice activity detection in one of the popular standards in voice and multimedia communications, namely G.729. This voice coding standard was introduced by the International Telecommunication Union (ITU) along with a recommended VAD algorithm in G.729-Annex B [7] (G.729B) and was tested by Rockwell International in [1]. The reason we chose G.729 is that it is one of the first coder standards that implement line spectral frequencies. This facilitates integrating the proposed work in any of the newer coders that adopt the same features.

G.729B VAD is based on a simple piecewise linear decision boundary between the set of differential parameters and their respective long-term values. The advantage of the G.729B VAD is that it works in the parameter domain of the underlying coder with no extra load for feature extraction. However, the performance of the G.729B VAD is lower than many other VAD algorithms including the fuzzy logic VADs (FVAD) that have been recently introduced for the G.729 environment in [8, 9]. FVAD provides 43% and 25% in improvement of clipping and false detection errors, respectively, compared with G.729 VAD.

HMM-based VADs have shown good performance when applied to speech signal in the discrete cosine transform

(DCT) domain in [10]. DCT-based coders normally target high voice quality applications, while today's low-bit-rate telecommunication voice coders, such as G.729, prefer line spectral frequencies representation of speech. We continue in the same direction and introduce a hidden Markov model (HMM)-based VAD algorithm that works in the domain of the G.729 parameters and provides a balanced improvement to the traditional G.729B VAD. We also examine the case of multivariate distribution in the HMM states, which eliminates the need for laying an assumption of independency among the distribution components. In order to keep the model simple, we assume that the voice frames are dominated by speech. This assumption is acceptable in nonnegative SNR levels.

The proposed VAD differs from the VAD in [10] on two points, notably, (i) the proposed VAD works in the compressed domain of the line spectral frequencies that are adopted by low-bit-rate speech coders, for example, G.729, while the VAD in [10] works on DCT feature vectors which are adopted by high-quality speech coders, (ii) the proposed VAD assumes that the voice frames are dominated by speech while the VAD in [10] considers a noise distribution within speech. In brief, the proposed VAD targets a class of speech coders that is different than that in [10]. Thus, we compare the performance of the proposed VAD with the performance of the G.729B VAD and the performance of the popular adaptive multirate, option 2 (AMR2) VAD [11].

The proposed VAD softly merges the state conditional likelihood of the frame to be speech/noise (irrespective of past frames) with a dynamic behavioral model across consecutive frames. This choice of avoiding HMM training, for example, Viterbi and Baum-Welch, is consciously taken to avoid excessive complexity of the VAD, which has to remain simple enough to allow for real-time applicability.

The structure of the proposed VAD system is given in Section 2 while the proposed algorithm is described in Section 3. The performance of the proposed VAD is studied and compared with the G.729B VAD and with the adaptive multirate VAD, option 2 (AMR2) in Section 4. A summary is given is Section 5.

2. THE STRUCTURE OF THE PROPOSED VAD

Modern VAD algorithms, in general, consist of two major parts. The main part produces a preliminary decision as for the current frame being a speech or a nonspeech frame. This preliminary decision depends on the difference between the characteristics of speech and noise in a certain domain using a certain criterion of comparison. Due to being far from ideal, the main part of the VAD does not always provide the correct decision, for example, clippings may happen at areas of change from noise to speech and vice versa. In order to compensate for this shortcoming, the second part of VAD modifies the preliminary decision based on the previous decision(s). For example, some VAD algorithms use a discrete Markov chain while others modify the current frame status into *speech* frame if the preliminary decision of the previous frame is speech, regardless of the current frame characteristics. This part of the VAD is often known as the *hangover* scheme. Applying a hangover scheme reduces clipping error rate at the expense of an increase in false detection error rate. A hangover scheme is acceptable as long as the overall performance is improved.

In the proposed VAD, we adopt a semi-continuous state-transition probability HMM-based algorithm. The structure of the HMM provides an integrated probabilistic framework where the main VAD stage and the hangover stage are softly combined. One decision is produced (per frame) based on the interaction between the two system components, namely the hidden layer and the observation layer. The state-transition layer serves as a dynamic hangover while the observation layer takes care of the comparison of the frame features.

2.1. The state-transition layer (hidden layer)

The proposed model assumes two states, S_0 and S_1, representing the noise and speech frames, respectively, as indicated in Figure 1. The probability of being in a certain state given the immediate previous state is defined by a state-transition matrix $\mathbf{A} = \{a_{ij}\}$, where a_{ij} is the probability of a state transition from state S_i to state S_j, subject to the constraint

$$\sum_j a_{ij} = 1, \quad i, j = 0, 1. \tag{1}$$

To reflect the higher likelihood of remaining in the same state, a_{00} and a_{11} are expected to be generally larger than a_{01} and a_{10}, respectively. Both interstate transition probabilities a_{01} and a_{10} play an important role when the conditional state probabilities of the current frame mismatch the actual frame classification. This would happen when the current speech frame appears to better fit in the noise state or vice versa. In such cases, the role of the transition probability from the noise state to the speech state, a_{01}, is to avoid clipping at the inset of the speech, that is, at the beginning of a phrase, whereas the role of the transition probability from the speech state to the noise state, a_{10}, is to avoid clipping in the outset of the speech, that is, at the end of a phrase, in addition to avoiding clipping within a speech phrase. We focus on the latter and adopt a dynamic scheme in which the probability of making such transition, a_{10}, exponentially decreases starting from the beginning of a phrase down to a limit $a_{10\min}$. In other words, a_{10} is inversely proportional to the time spent continuously in a speech state, given that the conditional probability of the current frame $\mathbf{x}_t$ to be produced by state S_1, $b_1(\mathbf{x}_t)$, is higher than the conditional probability of the current frame $\mathbf{x}_t$ to be produced by state S_0, $b_0(\mathbf{x}_t)$. Otherwise, a_{10} exponentially increases to its idle value $a_{10\max}$. The exponential decay rule is used to retain the computational requirements of the VAD as low as possible. Carrying out the HMM computations in the log-domain makes this choice very appealing. Making a transition from one state to the other is not only governed by the transition probabilities but also by the conditional probabilities, which reduces the possibility of incorrect transitions based on only one of

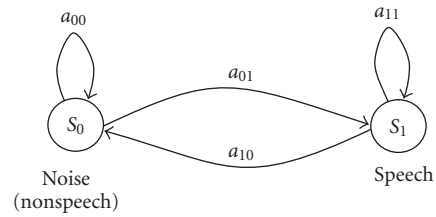

Figure 1: Two-state Markov chain.

them (if it were used individually). Another alternative that could have been used is a uniform transition penalty, which corresponds to a constant transition probability matrix.

The continuous transition probability HMM (CHMM) has a transition matrix that is given by

$$\mathbf{A} = \begin{bmatrix} 1 - f_{01}(t) & f_{01}(t) \\ f_{10}(t) & 1 - f_{10}(t) \end{bmatrix},$$

$$f_{ij}(t) = \begin{cases} \max\left(f_{ij}(t_i) \cdot e^{-(t-t_i)/\tau_i}, a_{ij,\min}\right), & b_i(\mathbf{x}_t) > b_j(\mathbf{x}_t), \\ \min\left(f_{ij}(t_i') \cdot e^{(t-t_i')/\tau_i}, a_{ij,\max}\right), & b_i(\mathbf{x}_t) \le b_j(\mathbf{x}_t), \end{cases} \quad i \neq j, \tag{2}$$

where t_i is time index of the frame where the condition $b_i(\mathbf{x}_t) > b_j(\mathbf{x}_t)$ was first met in the most recent segment, t_i' is time index of the frame where the condition $b_i(\mathbf{x}_t) \le b_j(\mathbf{x}_t)$ was first met in the most recent segment, assuming the first frame is noise, and $b_i(\mathbf{x}_t)$ is the conditional probability of the tth frame whose parameter set is $\mathbf{x}_t$ to be generated by a state S_i, that is: $b_i(\mathbf{x}_t) = P(\mathbf{x}_t \mid S_i)$. The proposed VAD is designed with an aim of adding a minimal extra computational load to the underlying coder. Consequently, it adopts some heuristics in determining the probability of transition from speech to noise and vice versa. Although being rarely used in pattern recognition systems that are mainly composed of HMM such as automatic speech recognition (ASR) and optical character recognition (OCR) systems, these heuristics are not uncommon in VADs that are built specially for telecommunication applications. The reason behind this is that the encoders and decoders in telecommunication applications are designed to be as simple as possible in order to meet the requirements of the hardware implementation, for example, mobile computing limitations and handset battery recharge time. The heuristics we adopt include setting the parameter τ_0 to infinity in order to avoid lingering in the noise state at the beginning of a speech phrase, while $a_{01\max}$, $a_{10\max}$, and τ_1 are set to an empirically chosen value of 0.1. These heuristics reduce the number of free parameters in the system while maintaining emphasis on transitions from the speech state. Thus, $a_{10\min}$ becomes the system parameter that controls the system bias for/against speech. A bias factor β is defined as $\beta = -\log(a_{10\min})$, subject to the constraint $\beta > 0$. In our simulation, we set the bias factor β to an arbitrary value of 10. It should be noted that the higher the bias factor β is, the more difficult it is to leave the speech state, that is, less clipping and more false speech detection may result.

Setting τ_0 to infinity results in a constant a_{00} and a constant a_{01}, and the transition matrix $\mathbf{A}$ becomes

$$\mathbf{A} = \begin{bmatrix} a_{00} & a_{01} \\ f_{10}(t) & 1 - f_{10}(t) \end{bmatrix}. \tag{3}$$

The model is thus a semi-continuous transition probability HMM. This should not be confused with the semi-continuous HMM, where the "semi-continuous" term refers to the probability density function of the HMM.

2.2. The observation layer

The observation layer is the part of the system that is concerned with computing the likelihood of a frame being a speech or a noise frame given a certain state. This conditional likelihood is estimated based on a distribution associated with each state, which takes the form of a probability density function (PDF) for continuous-probability HMMs. A state PDF is normally approximated by a weighted sum of a set of prototype distributions. For simplicity, we approximate the state PDFs in the proposed HMM by one p-dimensional distribution per state PDF. We adopt the generalized multivariate Gaussian distribution in [9, 12] with $\kappa = 0.5$ for Laplacian case:

$$p(\mathbf{x} \mid S_i) = f(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \kappa) = \frac{p\Gamma(p/2)}{\pi^{p/2}\sqrt{|\boldsymbol{\Sigma}_i|}\,\Gamma(1 + p/2\kappa)2^{(1+p/2\kappa)}} \times \exp\left\{\frac{-\left[(\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)\right]^\kappa}{2}\right\}, \tag{4}$$

where $\Gamma(\cdot)$ is the Gamma function, p is the size of the feature vector $\mathbf{x}$, and $\boldsymbol{\Sigma}$ is a nonnegative definite $p \times p$ matrix that is given by

$$\boldsymbol{\Sigma} = \frac{p\Gamma(p/2\kappa)}{2^{1/\kappa}\Gamma(p + 2/2\kappa)}\,\mathrm{cov}(\mathbf{x}), \tag{5}$$

where $\mathrm{cov}(\mathbf{x})$ is the covariance matrix of $\mathbf{x}$.

One has to pay attention to the number of feature vectors that is used to estimate the covariance matrix of $\mathbf{x}$, since insufficient number may reduce the estimation accuracy. Choosing Laplacian distribution to represent the state PDF is motivated by our statistical observations on a set of 32 000 frames from voice streams of two male and two female speakers given in [13].

3. THE PROPOSED ALGORITHM

An initial estimate of noise state PDF is obtained from the first 16 frames from 12 different voice streams assuming that the first 16 frames are nonspeech frames. We believe that this is just about the minimum number of feature vectors to build an initial estimate. A smaller number of vectors would yield insufficient estimates, whereas a larger number of feature vectors may violate the assumption above. The rest of the frames from the voice streams are used in a real-time

adjustment (adaptation) process to enhance the initial estimate of the state PDFs, that is, virtually all the feature vectors in the voice streams (about 9600 in total) are involved in the state PDF estimation and adaptations processes. The initial parameters of the speech state PDF are assumed to be the same except for the variance. The initial variance of the speech state PDF is assumed to be 10 times larger than that of the noise state PDF. This assumption, which is important to compensate for the absence of prior information about speech statistics, seems acceptable in a wide range of SNR (down to 0 dB). However, this assumption is expected to have a negative impact on the system performance at extremely low SNR levels (−5 dB and below) due to the fact that at such a low SNR, the background noise variance becomes extremely large invalidating the assumption of noise variance being 0.1 of the speech variance.

A VAD flag of a frame is set to 1 if the probability of the speech state is larger than or equal to the probability of the noise state at any given frame, and is set to 0 otherwise. We use $\gamma_t(j)$ the a posteriori probability of a state S_j at a time t, given the previous and the current observations, that is, frames, which is given by

$$\gamma_t(j) = P(q_t = S_j \mid \mathbf{x}_{\{t_0,\ldots,t\}}, \lambda), \quad t = t_0, \ldots, T, \tag{6}$$

where q_t is the effective state at the tth frame, t_0 is the index of the first frame, T is the total number of frames in the stream, $\mathbf{x}_t$ is the feature, that is, observation, vector at time t, which consists of zero-crossing rate, frame energy, frame energy in the low-frequency band, and 10 line spectral frequencies (LSF), and λ is the set of HMM model parameters.

This a posteriori probability can be written as

$$\gamma_t(j) = \frac{P(q_t = S_j, \mathbf{x}_{\{t_0,\ldots,t\}} \mid \lambda)}{P(\mathbf{x}_{\{t_0,\ldots,t\}} \mid \lambda)}, \quad t = t_0, \ldots, T. \tag{7}$$

The probability term in the denominator is the same for all the states at a given time t, thus the a posteriori probability can be reduced to the forward probability $\alpha_t(j)$, which represents the likelihood of a state S_j to generate a frame t, whose feature vector is $\mathbf{x}_t$, and the frame sequence up to the time t:

$$\begin{aligned} P(q_t = S_j, \mathbf{x}_{\{t_0,\ldots,t\}}) &= \sum_{i=0}^{1} \left[P(q_{t-1} = S_i, \mathbf{x}_{\{t_0,\ldots,t-1\}}) \cdot P(q_t = S_j \mid q_{t-1} = S_i)\right] \\ &\quad \cdot P(\mathbf{x}_t \mid q_t = S_j), \quad t = t_0, \ldots, T, \end{aligned} \tag{8}$$

where

$$P(q_t = S_j \mid q_{t-1} = S_i) \equiv a_{ij}(t), \quad i, j = 0, 1, \tag{9}$$

q_t is the effective state at the tth frame, t_0 is the number of frames used to initialize the state PDFs, T is the total number of frames in the stream, and the model parameter set λ is not written explicitly for simplicity.

To improve the estimation of the PDF parameters and to compensate for the (presumably) slowly varying changes, we adopt an adjustment scheme by which the parameters of state PDFs are updated as follows:

$$\begin{aligned} \hat{\boldsymbol{\mu}}^{(j)} &= (1-\rho)\boldsymbol{\mu}^{(j)} + \rho \mathbf{x}_t, \\ \hat{\mathrm{cov}}^{(j)}(\mathbf{x}) &= (1-\rho)\,\mathrm{cov}^{(j)}(\mathbf{x}) + \rho\left(\mathbf{x}_t - \boldsymbol{\mu}^{(j)}\right)\left(\mathbf{x}_t - \boldsymbol{\mu}^{(j)}\right)^T, \end{aligned} \tag{10}$$

where

$$j = \arg\max_{r=1,\ldots,N} \left(P(q_t = S_r, \mathbf{x}_{\{t_0,\ldots,t\}})\right) \tag{11}$$

and $\rho = 1/n^{(j)}$, where $n^{(j)}$ is the number of past visits to a state S_j.

Small values of ρ are better from stability point of view but result in slower adjustment. To avoid starting with a large adaptation value at the beginning of a data stream, ρ is initially set a value that is less than 1. There is no minimum value for ρ, thus, this learning process come to a soft end after efficiently large number of frames. An implicit assumption is made here that the environment is stationary. This argument is particularly important in low-performance VAD conditions (e.g., very low SNR), where the correct detection rate is lower than 50%. The complexity of the proposed algorithm is about three folds of that of the G.729 VAD, that is, very small compared with the overall G.729 encoder complexity.

4. RESULTS AND DISCUSSION

The proposed VAD works on top of the G.729 encoder and is applied to a set of 12 voice streams (about 96 seconds) from 4 different speakers; two males and two females with 3 streams/speaker from [13], with almost 58% speech versus 42% silence. The G.729 encoder runs on 100 frame/s (80 samples/frame) and provides the values of energy, low-band energy, zero-crossing rate, and ten line spectral frequencies (LSFs) for each frame. Those are the same set of raw features used by the G.729B VAD and the proposed VAD algorithm as well. The voice streams are corrupted by three types of background noises, white noise, babble noise, and car noise at different average SNR levels between 20 dB and 0 dB. The performance of the VAD is evaluated in terms of the probability of clipping Pc, and the probability of false detection Pe, where (i) Pc is the ratio of the number of speech frames that is mistakenly classified as noise to the total number of speech frames and (ii) Pe is the ratio of the number of noise frames that is mistakenly classified as speech to the total number of noise frames.

The performance of G.729B is given in Section 1 in both Tables 1 and 2 for reference. In order to identify independently the advantage of using *multivariate* state PDFs and the *semi-continuous* state-transition probability scheme in the proposed HMM-based VAD, we first present the performance of an HMM-based VAD with univariate state PDFs and discrete-state-transition probabilities (UDHMM) in Section 2 of Table 1. The univariate state PDFs are constructed as the product of one-dimensional PDFs of each element in the observation vector assuming those elements

Table 1: The performance of univariate discrete and semi-continuous HMM-based VADs against the performance of G.729B VAD. The performance is evaluated in terms of (1) the probability of clipping Pc, and the probability of false detection Pe, (2) the improvement in Pc, which is given by $-(\text{Pc}|_{\text{AMR2/HMM}} - \text{Pc}|_{\text{G.729}}) \times 100/\text{Pc}|_{\text{G.729}}$, and (3) the improvement in Pe, which is given by $-(\text{Pe}|_{\text{AMR2/HMM}} - \text{Pe}|_{\text{G.729}}) \times 100/\text{Pe}|_{\text{G.729}}$.

Noise type	SNR (dB)	G.729B		Univariate discrete HMM VAD				Univariate semi-continuous HMM VAD			
		Pc (%)	Pe (%)	Pc (%)	Pe (%)	Improvement in Pc (%)	Improvement in Pe (%)	Pc (%)	Pe (%)	Improvement in Pc (%)	Improvement in Pe (%)
Babble	20	14.49	28.14	9.54	4.50	34.16	84.01	1.18	10.60	91.86	62.33
	10	25.92	27.21	19.98	3.37	22.92	87.61	5.60	7.99	78.40	70.64
	0	42.12	27.51	33.33	1.89	20.87	93.13	13.68	4.57	67.52	83.39
Car	20	16.16	10.49	6.20	7.09	61.63	32.41	0.40	15.92	97.52	−51.76
	10	27.62	10.42	13.60	4.99	50.76	52.11	1.48	13.86	94.64	−33.01
	0	39.14	10.23	31.80	2.43	18.75	76.25	7.53	7.74	80.76	24.34
White	20	17.99	10.30	18.06	0.21	−0.39	97.96	5.86	2.59	67.43	74.85
	10	30.35	10.42	31.04	0.25	−2.27	97.60	14.11	1.59	53.51	84.74
	0	48.30	10.51	43.46	0.30	10.02	97.15	25.12	0.83	47.99	92.10
Average improvement over G.729B				—	—	24.05	79.80	—	—	75.51	45.29
Section 1				Section 2				Section 3			

Table 2: The performance of the proposed multivariate semi-continuous HMM-based VAD and AMR2 VAD against the performance of G.729B VAD. The performance is evaluated in terms of (1) the probability of clipping Pc, and the probability of false detection Pe, (2) the improvement in Pc, which is given by $-(\text{Pc}|_{\text{AMR2/HMM}} - \text{Pc}|_{\text{G.729}}) \times 100/\text{Pc}|_{\text{G.729}}$, and (3) the improvement in Pe, which is given by $-(\text{Pe}|_{\text{AMR2/HMM}} - \text{Pe}|_{\text{G.729}}) \times 100/\text{Pe}|_{\text{G.729}}$.

Noise type	SNR (dB)	G.729B		AMR2				Multivariate semi-continuous HMM-based VAD			
		Pc (%)	Pe (%)	Pc (%)	Pe (%)	Improvement in Pc (%)	Improvement in Pe (%)	Pc (%)	Pe (%)	Improvement in Pc (%)	Improvement in Pe (%)
Babble	20	14.49	28.14	0.28	61.08	98.07	−117.06	1.02	6.91	92.96	75.44
	10	25.92	27.21	0.08	66.60	99.69	−144.76	5.77	3.81	77.74	86.00
	0	42.12	27.51	0.08	65.12	99.81	−136.71	14.27	2.40	66.12	91.28
Car	20	16.16	10.49	0.49	14.48	96.97	−38.04	0.38	9.54	97.65	9.06
	10	27.62	10.42	0.91	12.40	96.71	−19.00	2.35	6.26	91.49	39.92
	0	39.14	10.23	14.42	4.27	63.16	58.26	12.35	2.22	68.45	78.30
White	20	17.99	10.30	0.49	11.25	97.28	−9.22	6.85	2.01	61.92	80.49
	10	30.35	10.42	1.08	11.00	96.44	−5.57	15.42	0.90	49.19	91.36
	0	48.30	10.51	5.27	7.28	89.09	30.73	26.88	0.05	44.35	99.52
Average improvement over G.729B				—	—	93.02	−42.37	—	—	72.21	72.37
Section 1				Section 2				Section 3			

are independent random variables, whereas the multivariate state PDF is constructed with one multidimensional PDF.

We then include the performance of the univariate semi-continuous state-transition probability HMM (USCHMM) VAD in Section 3 of Table 1 to show the gain from using the semi-continuous state-transition probability scheme alone. (Some of these results are also found in [14, 15].) It can be seen that the UDHMM VAD provides a reasonable improvement over the G.729B VAD in Section 1 of Table 1 in terms of clipping probability (24.05%) and a significant improvement in terms of false detection rate (79.80%). This imbalance in improvement is reversed by introducing the semi-continuous state-transition probability scheme to the discrete PDF HMM as it appears in Section 3 of Table 1. The improvement in clipping probability and false detection probability becomes 75.51% and 45.29%, respectively. Obviously the semi-continuous state-transition probability scheme introduces a bias towards speech. Combining the multivariate state PDF representation and the semi-continuous state-transition probabilities results in a balanced improvement over G.729B in clipping and false detection probabilities of 72.21 and 72.37%, respectively, as given in Section 3 of Table 2.

Table 2 provides the performance of the G.729B VAD as a reference in Section 1 while the performance of the adaptive multirate VAD, option 2 (AMR2) [16] is represented in

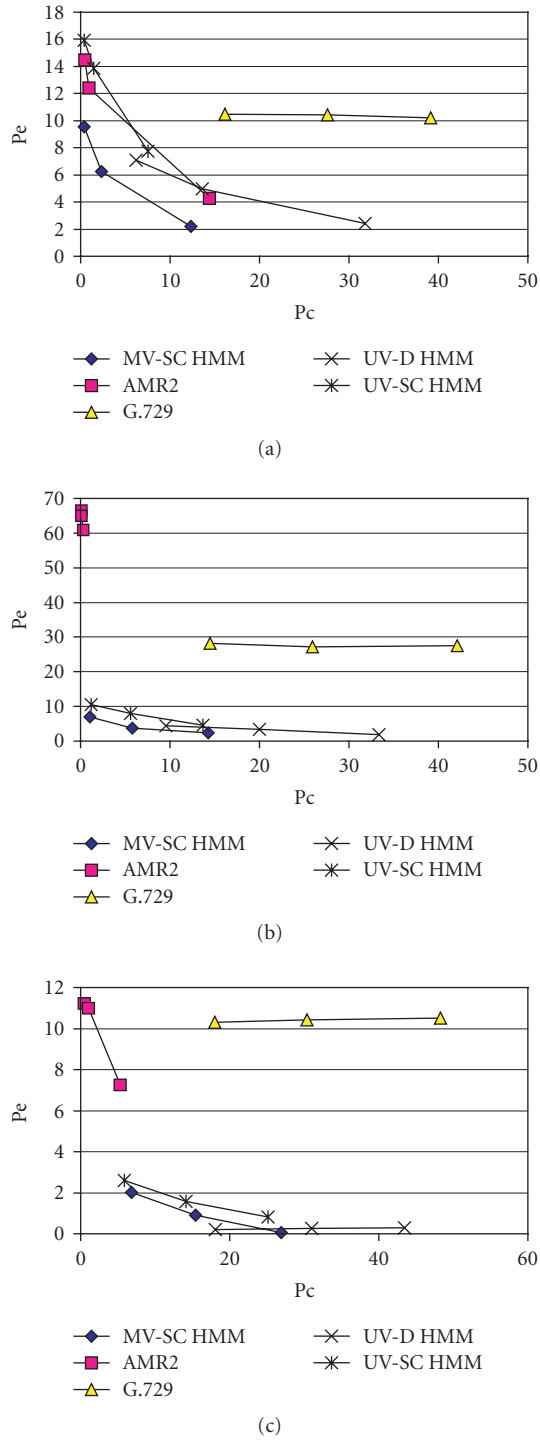

Figure 2: The probability of clipping Pc, and the probability of false detection Pe, for (a) car noise, (b) babble noise, and (c) white noise.

Section 2 in the same table. In general, AMR2 VAD provides the lowest clipping probability over G.729B VAD and the HMM VAD (with 93.02% improvement over G.729B VAD). This happens at the cost of higher false detection probability (42.37% average degradation), specially in the case of babble noise. On the contrary, the proposed multivariate semi-continuous HMM VAD provides a balanced, yet significant, improvement to G.729B for clipping and false detection probabilities; 72.21, and 72.37%, respectively.

Figure 2 shows the relative locations of the different VADs on the clipping versus false detection plane. An ideal VAD, if exists, would be located at the lower-left corner of the graph. The curve that represents the multivariate semi-continuous HMM VAD is always located to the lower-left side of the curves that represent the other VADs, which indicates its ability to deliver low clipping and false detection jointly.

5. SUMMARY

In this paper, we propose an efficient VAD algorithm to work with G.729-compliant encoders in their parameter domain with minimal additional computational load for feature extraction. The proposed VAD is a semi-continuous state-transition probability HMM-based with a Laplacian observation layer, with no need for offline learning process. The proposed VAD provides a robust performance with regard to accurate detection of speech frames and noise frames.

REFERENCES

[1] A. Benyassine, E. Shlomot, H.-Y. Su, D. Massaloux, C. Lamblin, and J.-P. Petit, "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 Communications Magazine*, vol. 35, no. 9, pp. 64–73, 1997.

[2] Y. D. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Processing Letters*, vol. 8, no. 10, pp. 276–278, 2001.

[3] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[4] E. Nemer, R. Gourbran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 217–231, 2001.

[5] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 109–118, 2002.

[6] S. Yang, Z.-G. Li, and Y.-Q. Chen, "A fractal based voice activity detector for internet telephone," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '03)*, vol. 1, pp. 808–811, Hong Kong, April 2003.

[7] ITU-T G.729 Annex B, "A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70," 1996.

[8] F. Beritelli, S. Casale, G. Ruggeri, and S. Serrano, "Performance evaluation and comparison of G.729/AMR/fuzzy voice activity detectors," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 85–88, 2002.

[9] F. Beritelli, S. Casale, and A. Cavallaro, "A robust voice activity detector for wireless communications using soft computing," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 9, pp. 1818–1829, 1998.

[10] S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 498–505, 2003.

[11] ETSI EN 301 708 v7.1.1 (1999-12), "European Standard (Telecommunications series), Digital cellular telecommunications system (Phase 2+); Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels; General description," (GSM 06.94 version 7.1.1 Release 1998).

[12] G. E. Kelly and J. K. Lindsey, "Models for estimating the change-point in gas exchange data," in *Proceedings of the 22nd Conference on Applied Statistics in Ireland (CASI '02)*, Antrim, Ireland, May 2002.

[13] ITU-T Series P, Supplement 23, "ITU-T coded-speech database," February 1998, http://www.itu.int.

[14] H. Othman and T. Aboulnasr, "A Gaussian/Laplacian hybrid statistical voice activity detector for line spectral frequency-based speech coders," in *Proceedings of the 46th IEEE International Midwest Symposium on Circuits and Systems (MWSCAS '03)*, vol. 2, pp. 693–696, Cairo, Egypt, December 2003.

[15] H. Othman and T. Aboulnasr, "A semi-continuous state transition probability HMM-based voice activity detection," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '04)*, vol. 5, pp. 821–824, Montreal, Quebec, Canada, May 2004.

[16] Y. Tian, J. Wu, Z. Wang, and D. Lu, "Fuzzy clustering and Bayesian information criterion based threshold estimation for robust voice activity detection," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '03)*, vol. 1, pp. 444–447, Hong Kong, April 2003.