A SEMI-CONTINUOUS STATE TRANSITION PROBABILITY HMM-BASED VOICE ACTIVITY DETECTION

H. Othman and T. Aboulnasr

School of Information Technology and Engineering,
University of Ottawa, Canada
hisham@uottawa.ca aboulnasr@genie.uottawa.ca

ABSTRACT

In this paper 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 Adaptive MultiRate VAD, phase 2 (AMR2).

1. INTRODUCTION

Actual speech activities normally occupy 60% of the time on a regular conversation in a telecommunication system [1]. Voice Activity Detection (VAD) enables reallocating system resources during the periods of speech absence. In modern telecommunication systems, VAD, in conjunction with Comfort Noise Generator (CNG) and Discontinuous Transmission (DTX) modules, play an important role in enhancing the utilization of system resources.

VAD distinguishes between speech and non-speech 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 a speech frame is misclassified as a noise frame, which is intolerable in speech encoders due to its effect on speech intelligibility. While false detection errors are due to misclassifying a noise frame into a speech frame. Echo cancellation systems are normally sensitive to this type of errors because it results in incorrect parameter adaptation.

In this paper, we focus on voice activity detection of one of the popular communications standards, 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 [3] and was tested by Rockwell International in [1]. G.729B VAD is based on a simple piecewise linear decision boundary between a 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. Fuzzy logic VADs (FVAD) [3] have been recently introduced for the G.729 environment. FVAD provides 43% and 25% an improvement on clipping and false detection errors, respectively compared with G.729 VAD.

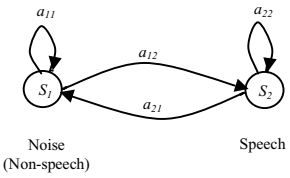

Figure 1. Two-State Markov Chain.

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 with minimal additional complexity. We will compare the performance of the proposed VAD with the performance of the G.729B VAD. We will also compare its performance with performance of the popular Adaptive Multi Rate, option 2 (AMR2) VAD [6], although the latter works in the FFT domain which is different than the G.729 feature domain.

The proposed VAD softly merges the state conditional likelihood of the frame parameters to be speech/noise (irrespective of past frames) with a dynamic behavioral model across consecutive frames. It requires no prior offline learning as opposed to FVAD.

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, phase 2 (AMR2) in Section 4 and 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 non-speech 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, e.g. clippings may happen at areas of change from noise to speech and vise 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. As a rough analogy, 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_1 and S_2, representing the noise and speech frames, respectively. 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:

$$\sum_j a_{ij} = 1,\ i, j = 1,2. \qquad (1)$$

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. The transition probability from the speech state to the noise state, a_{10}, is more important for a communication system VAD than the transition from the noise state to the speech state, a_{01}. Incorrect transition from the speech state to the noise state should be discouraged in order to avoid misclassifying parts of speech, e.g. offset speech, as an outcome of the noise state. We adopt a dynamic scheme in which the probability of making such transition, a_{10}, decreases starting from the beginning of a phrase down to a limit a_{10min}. 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}, gradually returns to its idle value a_{10max}. This form of continuous transition probability HMM (CHMM) has a transition matrix that is given by:

$$\mathbf{A} = \begin{bmatrix} 1 - f_{01}(t) & f_{01}(t) \\ f_{10}(t) & 1 - f_{10}(t) \end{bmatrix} \qquad (2)$$

where $f_{ij}(t)$ is defined as:

$$f_{ij}(t) = \begin{cases} \max(f_{ij}(t_i) \cdot e^{\frac{-(t-t_i)}{\tau_i}}, a_{ij,\min}) & , b_i(\mathbf{x}_t) > b_j(\mathbf{x}_t) \\ \min(f_{ij}(t_i') \cdot e^{\frac{-(t-t_i')}{\tau_i}}, a_{ij,\max}) & , b_i(\mathbf{x}_t) \le b_j(\mathbf{x}_t) \end{cases} \qquad (3)$$

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, and $b_i(\mathbf{x}_t)$ is the conditional probability of the t^{th} frame whose parameter set is $\mathbf{x}_t$ to be generated by a state S_i, i.e.:

$$b_i(\mathbf{x}_t) = P(\mathbf{x}_t \mid S_i)$$

For simplicity, τ_0 is set to infinity while a_{01max}, a_{10max} and τ_1 are set to .1, reducing the number of free parameters in the system while maintaining emphasis on transitions from the speech state. Thus, a_{10min} becomes the system parameter that controls the system bias for/against speech. A bias factor β is defined as $\beta = -\log(a_{10min})$, 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 the more difficult to leave the speech state, i.e. less clipping and more false speech detection may result.

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

$$\mathbf{A} = \begin{bmatrix} a_{00} & a_{01} \\ f_{10}(t) & 1 - f_{10}(t) \end{bmatrix} \qquad (4)$$

The proposed 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 multivariate distribution per state PDF. We adopt a generalized multivariate Gaussian distribution in [4] with κ= 0.5 for Laplacian case:

$$p(\mathbf{x} \mid S_i) = f(\mathbf{x};\boldsymbol{\mu}_i,\mathbf{S}_i,\kappa)$$

$$= \frac{p\Gamma(\frac{p}{2})}{\pi^{p/2}\sqrt{|\mathbf{S}_i|}\Gamma(1+\frac{p}{2\kappa})2^{(1+\frac{p}{2\kappa})}} \exp\left\{\frac{-\left[(\mathbf{x}-\boldsymbol{\mu}_i)^T\mathbf{S}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right]^{\kappa}}{2}\right\} \tag{5}$$

where $\Gamma(.)$ is the Gamma function, p is the size of the feature vector $\mathbf{x}$, and $\boldsymbol{\Sigma}$ is a non-negative definite pxp matrix that is given by:

$$\mathbf{S} = \frac{p\Gamma(\frac{p}{2\kappa})}{2^{1/\kappa}\Gamma(\frac{p+2}{2\kappa})}\mathrm{cov}(\mathbf{x}) \tag{6}$$

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

Choosing Laplacian distribution to represent the state PDF is motivated by our statistical observations on a set of 32000 frames from voice streams of two male and two female speakers [5].

3. THE PROPOSED ALGORITHM

An initial estimate of noise state PDF is obtained from the first 16 frames. 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 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 0dB). 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 .1 of the speech variance.

A VAD flag of a frame is set to 1 if the likelihood of the speech state is larger than or equal to the likelihood of the noise state at any given frame, and is set to 0 otherwise. 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 is given by:

$$P(q_t = S_j, \mathbf{x}_{\{t_0,\ldots,t\}}) = \sum_{i=1}^{N} [P(q_{t-1} = S_i, \mathbf{x}_{\{t_0,\ldots,t-1\}}) \cdot P(q_t = S_j \mid q_{t-1} = S_i)] \cdot P(\mathbf{x}_t \mid q_t = S_j), \quad t = t_0,\ldots,T, \tag{7}$$

where $P(q_t = S_j \mid q_{t-1} = S_i) \equiv a_{ij}(t) \quad , i,j = 1,2,$

q_t is the effective state at the t^{th} frame, t_0 is the number of frames used to initialize the state PDFs and T is the total number of frames in the stream.

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

$$\hat{\boldsymbol{\mu}}^{(j)} = (1-\rho)\boldsymbol{\mu}^{(j)} + \rho\mathbf{x}_t \tag{8}$$

$$\hat{\mathrm{cov}}^{(j)}(\mathbf{x}) = (1-\rho)\mathrm{cov}^{(j)}(\mathbf{x}) + \rho(\mathbf{x}_t - \boldsymbol{\mu}^{(j)})(\mathbf{x}_t - \boldsymbol{\mu}^{(j)})^T \tag{9}$$

$$\text{where} \quad j = \arg\max_{r=1,\ldots,N}(P(q_t = S_r, \mathbf{x}_{\{t_0,\ldots,t\}})) \tag{10}$$

and ρ=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. We note that this adjustment scheme may not be highly robust at large values of ρ where error accumulation may result from wrong decisions. This argument is particularly important in low performance VAD conditions (e.g. very low SNR), where the correct detection rate is lower than 50%. In order to ensure the stability at the beginning of the call where the number of visits to both states is small, we limit the adjustment factor ρ to .1%.

The complexity of the proposed algorithm is about three folds of that of the G.729 VAD, which 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 [5]. The G.729 encoder runs on 100 frame/sec (80 samples/frame) and provides the values of energy, low-band energy, zero crossing rate, and ten Line Spectral Frequencies (LSFs) for each frame. The voice streams are corrupted by three different types of background noise; white noise, babble noise and car noise at different average SNR levels between 20 dB and 0 dB. Table 1 shows a comparison between the performance of the proposed HHM VAD and Adaptive MultiRate VAD, phase 2 (AMR2) [6] against the performance of ITU G.729 B VAD.

Table 1. The performance of the proposed HMM VAD and AMR2 VAD against the performance of G.729B VAD. The performance is evaluated in terms of:

- the probability of clipping, Pc,
- the probability of false detection, Pe,
- the improvement in Pc, which is given by $-(Pc|_{AMR2/HMM}-Pc|_{G.729})x100/\ Pc|_{G.729}$, and
- the improvement in Pe, which is given by $-(Pe|_{AMR2/HMM}-Pe|_{G.729})x100/\ Pe|_{G.729}$.

Noise Type	SNR (dB)	G729B		AMR2				The proposed 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

The performance is evaluated in terms of the probability of clipping, Pc, and the probability of false detection, Pe, where:

- Pe is the ratio of the number of noise frames that are mistakenly classified as speech to the total number of noise frames.
- Pc is the ratio of the number of speech frames that are mistakenly classified as noise to the total number of speech frames.

In general, AMR2 VAD provides the lowest clipping rate over G.729B VAD and the proposed HMM VAD (with 93.02% improvement over G.729B VAD). This happens at the cost of higher false detection rate (42.37% average degradation), specially in the case of Babble noise. On contrary, the proposed HMM VAD provides a balanced, yet significant, improvement to G.729B for clipping rate and false detection rate; 72.21 and 72.37%, respectively.

We note that, the improvement of the proposed system in the false detection rate is better than the improvement of the clipping rate in the case of white noise. This is because the noise is more stationary and thus easier to track. On the other hand, in the case of car noise the improvement in the clipping rate of the proposed system is better compared to the improvement of the false detection rate because the noise is less stationary.

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 probabilities HMM-based with a Laplacian observation layer, with no need for offline learning. The proposed VAD provides a significant improvement to G.729B with a good balance between the drop in clipping rate and in the false detection rate compared with that of the G.729 B VAD.

6. REFERENCES

[1] Adil Benyassine, Eyal Shlomot, and Huan-Yu Su, "ITU 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 Comm. Mag.*, pp. 64-73, September 1997.

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

[3] Francesco Beritelli, Salvatore Casale, and Alfredo 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, December 1998.

[4] G.E. Kelly and J.K. Lindsey, "Models for Estimating the Change-point in Gas Exchange Data," *Proc. Conf. Applied Statistics* in Ireland, CASI 2002.

[5] ITU-T Series P, Supplement 23, "ITU-T Coded Speech Database," Feb. 98.

[6] 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).