Novel Voice Activity Detection Based on Cepstrum moments

Ali Farzan
Islamic Azad University, Shabestar Branch, East Azerbijan, IRAN
gs22416@mutiara.upm.edu.my & a_farzan@hotmail.com

Syamsiah Bt. Mashohor
University of Putra Malaysia (UPM), Serdang, Malaysia
syamsiah@eng.upm.edu.my

Ali Nourmohammadi
Islamic Azad University, Tehran_ markaz Branch, Tehran, IRAN

Sarvin Sadra
Tabriz University, Tabriz, East Azerbijan, IRAN

Abstract-**Statistical methods for voice activity detection (VAD) have shown impressive performance especially with respect to their ability to be tuned parametrically and adoptability with deferent environments. In this paper we propose a novel statistical VAD algorithm using Cepstrum coefficients and their moments as features for classification. In this method, we use moment ratio of conversation part and silent part to evaluate a threshold measure for differentiating between silent and active (Speech) parts of conversation. To make it robust in noisy environments, we will gradually tune the threshold to adopt it with dynamic background noise. Simulation results show that our proposed method has good performance in noisy environments.**

I. INTRODUCTION

As we know, conversation is a sequence of speech and silent signal parts. Voice Activity detection refers to the ability of separating silent pattern of conversation from speech parts. VAD is a crucial part of speech communication systems such as speech coding, echo cancellation, speech recognition, hands free telephony, audio conferencing, etc. In modern communication systems such as Voice over IP (VOIP), VAD is used to improve bandwidth usage and increase the number of communication channels by detecting silent parts of conversation and preventing them to be transmitted. This can be considered as a kind of voice compression. Similarly, in 2G and 3G cellular communications, VAD is used to facilitate improved usage of available radio frequency spectrum and increase RF-channels with low power consumption [1][2]. The seemingly easy task of voice activity detection becomes difficult when conversation occurs in noisy background, where speech has to be detected in presence of non-stationary and unpredictable real world noise. Further, the difficulty increases if the signal to noise ratio (SNR) of the noisy speech is lowered. Therefore, in practice, the mobile environment of cellular telephone system is the most challenging scenario for voice activity detection as it is least controlled and speech is subject to a variety of acoustical noise and SNR's.

There are various VAD systems in literature. Some of them are based on the energy level difference, zero-crossing rate (ZCR), and spectral difference [3]. Asgari *et al.* used entropy as a measure to evaluate the organization of frames and then classifying them as speech or silence [4]. Shin *et al.* applied weighted likelihood ratios for implementing VAD decision rules [5]. Gauci *et al.* applied the Teager energy Cepstrum coefficients to obtain a noise robust feature extraction method for VAD [6]. Farsinejad *et al.* introduced an efficient probabilistic neural network (PNN) Model-based Voice Activity Detection (VAD) algorithm [7]. Yang *et al. presented a new* voice activity detector based on Mel filter-bank spectral entropy to distinguish speech from noise [8]. Ghiodi applied the Wavelet Packet Transform and the Teager Energy Operation (TEO) for designing a new VAD algorithm [9]. Asgari *et al.* introduced a VAD system based on vector quantization method which is adapted to environments with non-stationary background noise [10]. Venkatesha Prasad *et al.* used cepstrum thresholds to implement a fast VAD system [11].

In this paper we present a new method for voice activity detection using Cepstrum coefficients and their moments as features to classify silent and speech frames of conversation and finally compare its compression ability with standard ITU-T G.729.

II. CEPSTRUM AND MOMENTS

Cepstrum is the result of taking Discrete Cosine Transform (DCT) of speech signal, then find its log and finally apply Inverse DCT (IDCT) to the results. This can be shown mathematically as:

$$C_i[n] = IDCT\left(\log\left[DCT\left(S_i[n]\right)\right]\right)$$

In which $S_i[n]$ is signal element/sample and $C_i[n]$ is Cepstrum coefficient. Conceptually, the Cepstrum can be seen as information about rate of change in the different spectrum bands of signal. Literature shows that Cepstrums can be used as features for classifiers in various VAD systems [6][11]. Herein, we use moments of cepstrum coefficients to distinguish between silent and active parts of conversation signal. In general, moments are used to characterize the probability density function of a random

978-1-4244-5586-7/10/$26.00 ©2010 IEEE

variable. The *expected value* of function $g(x)$ of a discrete random variable is defined as:

$$E[g(x)] = \sum_{i=0}^{n-1} g(x_i) \quad P(x_i)$$

And *n*th *moment about the origin* can be calculated as:

$$E[x^n] = \sum_{i=0}^{n-1} (x_i)^n \quad P(x_i)$$

Accordingly, *n*th *central moment* can be calculated as:

$$E[(x-m)^n] = \sum_{i=0}^{n-1} (x_i - m)^n \quad P(x_i)$$

Where

$$m = E[x] = \sum_{i=0}^{n-1} x_i \quad P(x_i)$$

We use the first and second central moments of cepstrum coefficients in our algorithm.

III. Our Method

In this method we use sampling frequency of 8^{KHz}, PCM quantization level of 8^{bit} and mono voice channel. We divide each voice signal into the sequence of equi-width packets of size 40^{ms} each of which includes 5 frames of size 8^{ms}. Each frame is composed of 64 samples of signal. In order to incorporate the background noise specification into our algorithm, we assumed that the first 200^{ms} of the conversation (25 frames) has no speech part and it is a pure background noise. So we use first central moment of this part to quantify the background noise. To this end, we calculate the cepstrum coefficients and then expected value of their square for each of these 25 frames:

$$T_{fi} = E\left[C_i^{\,2}\right]$$

Where C_i is the cepstrum coefficient of frame *fi.*

After that, the expected value of all 25 T_{fi}s for initial silent part will be used for characterizing background noise.

$$T = E\left[T_{fi}\right]$$

Now we have to define some parameters used in our algorithm. We have "k" initialized by 1 as *silent-factor* which is used to incorporate the specification of varying background noise into the algorithm. Another parameter is $,step,$ initialized by 0.001 which assures the stability of algorithm against impulse background noise which we may encounter seldom and it has not to influence our algorithm tremendously.

Each packet will be classified as silent or active speech. Our rule for this classification is such that: "*A packet is silent, if all of its frames are silent. Otherwise it is active speech*". So we need to classify frames at first as silent or active speech. In order to do this classification for each new frame, it is assumed a silent frame if it qualifies the following condition.

$$E\left[C_i^{\,2}\right] > T * k$$

If we find a new silent frame we have to adopt our algorithm to new background noise. To do this, we have to change k and $step$ as follows:

$$step = abs(step - 0.005)$$

$$k = \min\left\langle (k + Step),\ \mathrm{E}\left[(\mathrm{V} - \mathrm{E}[\mathrm{V}])^2\right] \Big/ \mathrm{E}\left[(\mathrm{S} - \mathrm{E}[\mathrm{S}])^2\right] \right\rangle$$

Where V and S are Cepstrum coefficients of voice and silent parts of conversation. Indeed we use second central moment of voice and silent parts to adopt algorithm.

One important thing in our implementation is the way we dealt with zero DCT parameters in evaluating the Cepstrum coefficients. By examining different ways to handle it, experiments show that replacing them by 1000 ends up in good results.

IV. Experimental Results

To evaluate our algorithm and demonstrate its effectiveness, we used the standard ITU-T G.729 to be compared with. Sound Forge has been used to do voice signal manipulation tasks such as recording sounds, adding specific noises to it, etc. We use "*seven o eight*" template with 3 different additive white noise signals as testing templates. Simulation results show that our algorithm has more compression ratio and clean distinguishable separated speech parts in comparison with ITU-T G.729.

Table I represents compression ratio of our proposed method and G.729.

Table I. Compression Ratio for G.729 and proposed method

Background Noise(db)	G.729	Proposed Method
0	67%	69.2%
5	58%	66%
10	52.4%	59.7%

V. Conclusion

In this paper we have employed the Cepstrum coefficients and their moments as classification features. Simulation results showed the ability of this method in various environments and its compression ability. It seems that investigating more in this area and incorporating higher order momentums for designing VAD systems could be an open area for future works.

References

[1] I. D. Lee, H. P. Stern, and S. A. Mahmoud, "A voice activity detection algorithm for communication systems with dynamically varying background acoustic noise," in Proc. 48th IEEE Conf. Vehicular Technology, May 1998, pp. 1214–1218.

[2] F. Beritelli, S. Casale, and A. Cavallaro, "A low complexity speechpause detection algorithm for communication in noisy environments," Eur. Trans Telecommun., vol. 15, pp. 33–38, 2004.

[3] ITU-T: 'A silence compression scheme for G.729 optimised for terminals conforming to ITU-T V.70". ITU-T. Rec. G. 729, Annex B, 1996

[4] Asgari, M.; Sayadian, A.; Farhadloo, M.; Mehrizi, E.A. "Voice Activity Detection Using Entropy in Spectrum Domain", Telecommunication Networks and Applications Conference, 2008. ATNAC 2008. Australasian, 7-10 Dec. 2008 Page(s):407 – 410

[5] Jong Won Shin, Joon-Hyuk Chang, Nam Soo Kim, "Voice activity detection based on statistical models and machine learning approaches", Computer Speech & Language, march,2009

[6] Oliver Gauci, Carl J. Debono, Paul Micallef, "A Maximum Log-Likelihood Approach to Voice Activity Detection", ISCCSP 2008, Malta, 12-14 March 2008

[7] M. Farsinejad, M.Mohammadi, B.Nasersharif, A.Akbari, "A Model-based Voice Activity Detection Algorithm using probabilistic neural networks", Proceedings of APCC2008 2008 IEICE 08 SB 0083

[8] Jianjun Lei, Jiachen Yang, Jian Wang, Zhen Yang, "A Robust Voice Activity Detection Algorithm in Nonstationary Noise", IEEE 2009 International Conference on Industrial and Information Systems.

[9] Roberto CHIODI and Daniel MASSICOTTE, "Voice Activity Detection Based on Wavelet Packet Transform in Communication Nonlinear Channel", IEEE 2009 First International Conference on Advances in Satellite and Space Communications.

[10] Meysam Asgari, Abolghasem Sayadian, Farhad Tehranipour, "Novel Voice Activity Detection Based on Vector Quantization", UKSim IEEE 2009: 11th International Conference on Computer Modeling and Simulation.

[11] R. Venkatesha Prasad, H.S. Jamadagni, Abhijeet Sangwan, Chiranth M.C, 2003. " VAD for VoIP Using Cepstrum". HSNMC 2003, LNCS 2720, pp. 522-530, 2003. Springer-Verlag Berlin Heidelberg