VAD for VoIP Using Cepstrum

R. Venkatesha Prasad[#], H.S. Jamadagni[#], Abhijeet Sangwan[*], Chiranth M.C[*]

[#]CEDT, Indian Institute of Science, Bangalore, India,
[*]Department of E&C, PESIT, Bangalore, India.

Abstract. As telephony services are being supported on Internet the focus is now on multiplexing many speech streams by exploiting the speech characteristics. The multiplexing gain is an important factor when applications such as teleconference service are ported on to the Internet. Here we discuss Voice Activity Detection (VAD) for Voice over Internet Protocol (VoIP) based on Cepstrum. VAD aids in saving bandwidth of a voice session. Such a scheme would be implemented in the application layer thus VAD is independent of the lower layers. The standard codecs would inherently have the VAD algorithms to reduce the bandwidth. However they are costly and computationally complex. In this paper, we compare the quality of speech, level of compression and computational complexity of our method of Cepstrum based VAD with the standard GSM and ITU-T G.729 codecs. Bandwidth reduction is achieved by not transmitting the non-speech packets. Our algorithm adapts to the varying background noise.

1 Introduction

Traditional voice-based communication uses Public Switched Telephone Networks (PSTN) [3]. Such systems are expensive when the distance between the calling and called subscriber is large because of dedicated connection. The current trend is to provide this service on data networks [10]. Data networks work on the best effort delivery and resource sharing through statistical multiplexing. Therefore, the cost of services compared to circuit-switched networks is considerably less. However, these networks do not guarantee faithful voice transmission. Voice over packet or Voice over IP (VoIP) systems have to ensure that voice quality does not significantly deteriorate due to network conditions such as packet-loss and delays. Therefore, providing Toll Grade Voice Quality [5] through VoIP systems remains a challenge. In this paper we concentrate on the problem of reducing the required bandwidth for a voice connection on Internet using Voice Activity Detection (VAD), while maintaining the voice quality.

VAD algorithms find the beginning and end of talk spurts. VAD is used in non real-time systems like Voice Recognition systems, Compression and Speech coding [4,6,11]. VAD is also useful in VoIP, in which stringent detection of beginning and end of talk spurts is not needed.

In VoIP systems the voice data (or payload for packet) is transmitted along with a header on a network. The header size for Real Time Protocol (RTP, [9]) is 12 bytes.

M.M. Freire, P. Lorenz, M.M.-O. Lee (Eds.): HSNMC 2003, LNCS 2720, pp. 522-530, 2003.

The ratio of header to payload size is an important factor for selecting the payload size for a better throughput from the network. Smaller payload helps in a better real-time quality, but decreases the throughput. Normally, 10ms-40ms speech frames are used in VoIP systems. The requirements of VAD algorithms for VoIP applications are:

a) Low computational requirements (not more than one packet time)
b) Toll grade quality voice reproduction
c) Saving in bandwidth to be maximized

1.1 Speech Characteristics

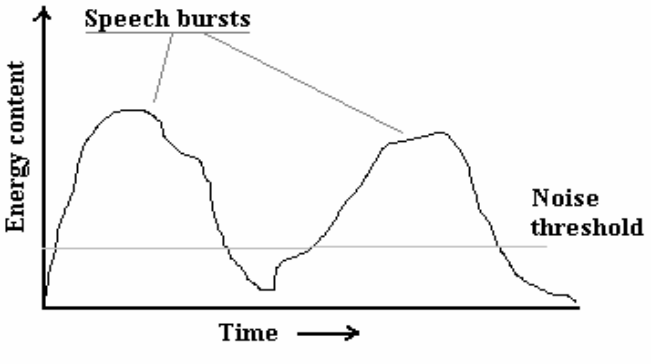

Fig. 1. A typical speech signal

Conversational speech is a sequence of contiguous segments of silence and speech (Fig.1) [20]. VAD algorithms take recourse to some form of speech pattern classification to differentiate between voice and silence periods. Thus, identifying and rejecting transmission of silence periods helps reduce Internet traffic.

1.2 Background Noise

The term 'silence segment' does not refer to a period of zero-energy, but of incomprehensible sound. VAD algorithms have to deal with silence periods having small audible content.

1.3 Desirable Aspects of VAD Algorithms Include:

- A Good Decision Rule: A physical property of speech that can be exploited to give consistent judgement in classifying segments of the signal into silence or otherwise.
- Adaptability to Changing Background Noise: Adapting to non-stationary background noise improves robustness, especially in wireless telephony where the user is mobile.

- Low Computational Complexity: Internet telephony is a real-time application. Therefore the complexity of VAD algorithm must be low to suit real-time applications.

2 Parameters for VAD Design

The differentiation of the voiced signal into speech and silence is done on the basis of speech characteristics. The signal is sliced into contiguous frames. A real-valued non-negative parameter is associated with each frame. If this parameter exceeds a certain threshold, the signal frame is classified as ACTIVE; else it is INACTIVE.

2.1 Choice of Frame Duration

ACTIVE Frames need to be embedded in suitable packets adhering to the network protocol being used for transmission. VoIP receivers queue up incoming packets in a packet-buffer that allows them to play audio even if incoming packets are delayed due to network conditions.

Consider a VoIP system having a buffer of 3-4 packets. Having packet duration of 10ms allows the VoIP system to start playing the audio at the receiver's end after 30 to 40ms from the time the queue started building up. If the packet duration were 50ms, there would be an initial delay of 150-200ms, which is unacceptable. Therefore, the packet duration must be chosen properly. Current VoIP systems use 20-40ms packet sizes.

For ease of DCT computation, frame duration of 8ms, corresponding to 64 samples is used (64 = 26), to avoid padding. An 8ms frame can cover 125Hz and above in cepstrum domain. A 40ms packet will have 5 frames.

2.2 Encoding Specifications

The specifications for encoding speech for VAD algorithms are that of Toll Grade Quality [5]:

- 8 kHz sampling frequency
- 256 levels of linear quantization (8 Bit PCM)
- Single channel (mono) recording.

The advantage of using linear PCM is, the voice data can be transformed to any other coding (such as G711, G723, G729) for compressing the voice data packet.

2.3 Adaptability to Background Noise

A fixed threshold would be ‘deaf’ to varying acoustic environments of the speaker. The scheme must have the wherewithal to adapt the threshold online and in real-time.

An arbitrary initial choice of the threshold is prone to deteriorated performance. For finding the initial threshold of non-speech frames, algorithm may be trained for a small period by a prerecorded speech sample that contains only background noise. Alternately, as users are not active as soon as the call is established, we may assume that the initial 200ms of the sample does not contain any speech. To adapt to the change in the background noise a simple adaptive technique for updating the threshold considering only non-speech frames has been used (as given in [8]).

3 Cepstral VAD

In this paper, we present a Cepstrum based VAD algorithm adopted from [7], wherein the analysis is done for enhancing the speech quality when it is under real car noise. Here the same technique is used for VoIP voice packets to detect the presence of speech. However, we calculated the Cepstral coefficients by using DCT (Discrete Cosine Transform) instead of the usual DFT spectrum that has been newly proposed in [13]. This is a new way of feature extraction and in addition it reduces the computation, as DCT is a real transform. Use of DCT Cepstrum (DCTC) [13] has the advantage of carrying the binary phase information along with the magnitude information when compared to DFT based Cepstrum. This often results in less error in the Cepstral coefficients. The phase of noise is arbitrary where as for speech it is deterministic.

3.1 Background

In a VoIP end terminal (user terminal), the speech is sampled at 8KHz sampling rate from the audio card with 8/16 bit linear quantization. These samples are packetized at 20ms-40ms intervals. The DCT - Cepstrum is computed as shown in Figure 2.

In each packet smaller frames are constructed as discussed, we take frames that contain 64 samples that corresponds to 8ms of speech. The Cepstrum coefficients are calculated for each frame as shown above. For i^{th} frame (with speech signal Xi[n]), Cepstral coefficients is computed as follows,

$$C_i[n] = \text{idct}(\log(\text{dct}(\text{hamming}(X_i[n])))) \ , \quad \textbf{(1)}$$

where $n = 1$ to 64.

We take this 64 dimensional vector to calculate the distance from the origin. We take L2 Norm to find the distance. The basis of the algorithm is that the frames in which the speech is not present have very low values of the cepstral distance around the origin. The Cepstrum peak due to the voiced excitation is more compared to the noise frame (non-speech frame). The Cepstra coming from noise are known to have much lower variances than those for speech especially in the case of low 'quefrency'. We are using in our algorithm, all the coefficients without confining to first few. For want of finding the energy of speech packets, we are considering vocal tract and source information regarding the activity. We are not confining ourselves to modeling vocal tract here. With all the coefficients we are using full information of the signal

provided by Cepstrum. While taking the difference in distance between speech frame and non-speech frame, the energy of the non-speech packet is subtracted. Hence, the higher value of difference cepstrum would imply a speech frame, and consequently a packet.

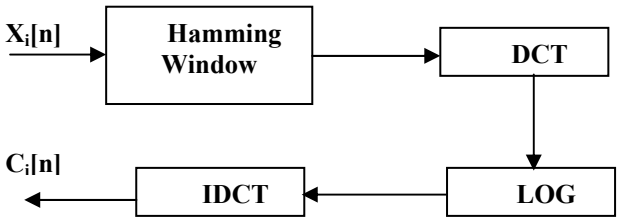

Fig. 2. Computation of Cepstrum Coefficients Ci[n] from the signal speech of one frame Xi[n]

3.2 Algorithm

In a VoIP call as soon as the call is established, the audio card will start recording the speech input. The user will see a message to imply that the call has been established. Usually the response time of the user before speaking into the microphone is at least 200ms. Thus the first *25* frames in our case have the recording of background noise. The Cepstral coefficients of the background noise is thus calculated and averaged.

Step 1 : Calculate the $C_i[n]$ for the i^{th} cepstral coefficient, for *i = 1 to 25, n = 1 to 64*, i.e. for the first *25* frames, *64* cepstral coefficients, using Equation (1).

Step 2 : Compute $C_{th}[n]$

$$C_{th}[n] = \frac{1}{25}\sum_{i=1}^{25}\left(C_i[n]\right) \quad , \tag{2}$$

for *i = 1* to *25, n = 1* to *64*

Step 3: Compute the first approximation of the threshold of non-speech frames (first *25* frames) using,

$$d_{th} = \frac{1}{64}\sum_{i=1}^{64}\left(C_{th}[n]\right)^2 \quad . \tag{3}$$

Step 4: For all i frames, $i > 25$, compute $C_i[n]$,

$$d_i = \frac{1}{64}\sum_{i=1}^{64}\left(C_i[n] - C_{th}[n]\right)^2 \quad . \tag{4}$$

Step 5 : The **decision rule** for detection of speech activity is

If $abs(d_i) > k.abs(d_{th})$, **then** frame is ACTIVE (5)

Else the frame is INACTIVE

where $k = 2$

Step 6: The threshold d_{th} is updated using the following equation only if a frame d_i is identified as a non-speech frame in *Step 5*. (Adaptive Threshold).

$$d_{th(new)} = p.d_{th(old)} + (1-p).d_i \ , \tag{6}$$

where $\boldsymbol{p} = 0.8$

In Step 1 and 2, we compute the threshold cepstral coefficients for the first *25* frames corresponding to 200ms. In Step 3 we compute the average of threshold cepstral coefficients. In Step 4, the relative distance between the current cepstral vector and the threshold is computed. The step 5 makes the decision whether the frame has speech or not. The value of *'**k**'* in Step 5 is experimentally set to *2*. A packet is resolved as having Speech if it carries at least one ACTIVE frame. For example, a packet carrying 40ms speech will have 5 frames (of 8ms each).

At step 6 the background noise is continually updated only upon detection of INACTIVE frames. The parameter *'**p**'* updates the threshold and decides the rate of change of adaptation as well. Sangwan, et. al. [21] have discussed the selection of *'**p**'* as well as second and third-order adaptability.

4 Results & Discussions

MATLAB was used to test the algorithm developed on various sample signals. The test templates used varied in loudness, speech continuity, background noise and accent. Both male and female voices have been used. The performance of the algorithms was studied on the basis of the following parameters:

1 Floating Point Operations (FLOPS) required: The total number of floating point operations is calculated for all algorithms to compare their relative complexity. This parameter is useful in comparing algorithms of their applicability for real-time implementation.

2 Percentage compression: The ratio of total INACTIVE frames detected to the total number of frames formed expressed as a percentage. A good VAD should have high percentage compression.

3 Subjective Speech Quality: The quality of the samples was rated on a scale of 1 (poorest) to 5 (best) where 4 represents toll grade quality. The input signal was taken to have speech quality 5. The speech samples after compression were played to independent jurors randomly for an unbiased decision.

4 Objective Assessment of Misdetection: The number of frames which have speech content, but were classified as INACTIVE and number of frames without speech content but classified as ACTIVE are counted. The ratio of this count to the total number of frames in the sample explored as a percentage is taken as the **%MISDETECTION**. This gives a quantitative measure of VAD performance. Though this number represents the quality of speech after applying a VAD technique, the quality of speech has to be assessed only by the subjective grading of the speech which is equivalent to MoS. This number gives an approximate assessment of the performance of the algorithm. To find an average values of misdetection a long speech file up to 3minutes are used.

An effective VAD algorithm should have high compression and a low number of FLOPS while maintaining an acceptable Speech Quality (and low misdetection)
It is necessary to note that the percentage compression also depends on the speech samples. If the speech signal were continuous, without any breaks, it would be unreasonable to expect high compression levels.

Table 1. The speech sample results [1]

Speech input	% of Compression	FLOPS (Max)	Subjective Quality	% Misdetection of packets
Sample 1	30	40000	4	16
Sample 2	17	40000	4	29
Sample 3	25	40000	4	25
Sample 4	29	40000	4	6
Sample 5	30	40000	4	30

The FLOPS are found using the MATLAB function. These results are compared with G.729B and GSM AMR VAD in Table 2.

Table 2. The VAD features of G.79B and GSM-FR

Speech input	% of Compression	FLOPS (Max)	MoS	% Misdetection of packets
G.729B	64 to 8kbps	3.9×10^6	4	13.4
GSM-AMR (VAD1)	64 to 12.2kbps	3.4×10^6	4	15

[1] Subjective Quality is based on majority decision of 7 listeners.
% of Compression is the saving in bandwidth.
% Misdetection is manually determined.

Table 2. shows an important aspect of our algorithm and the standard codecs. The numbers in the table 2 are compiled from many sources [15, 16, 17, 18,19]. Though the percentage misdetection is less in the standard codecs, the cost of computation is higher. MIPS to FLOPS conversion is taken as 1 MIPS = 3.2MFLOPS[14]. GSM VAD assumes stationary background noise. Therefore, it misses the low energy unvoiced speech [6].

5 Conclusions

We have presented an algorithm for VAD that can be implemented on a PC in real-time. The speech quality was observed to be of 'good' quality. The comparison with standard codecs shows a remarkable saving in the computation costs. Nonetheless, the saving in bandwidth is high for the standard algorithms. These standard algorithms are proprietary and implemented using DSPs. In VoIP systems, normally, the speech activity of a speaker is found to be around 40% [22]. Thus our algorithms would be saving at least 50% of the bandwidth at a lesser cost. With our algorithm the bandwidth reduces approximately to 32kbps where as in standard codecs it can be up to 8kbps. Thus the trade off is between the cost of the codecs and saving in bandwidth.

The advantage of these algorithms is in providing the teleconference applications wherein streams from a large number of participants have to be handled. There is a less possibility of codecs being available at all user terminals and thus forcing all the participants to use linear PCM. If VAD is applied to each stream at an intermediate server, computational capability of the servers cannot scale up for a large number of streams for coding the signals at a server.

Our VAD algorithm applied to PCM coded speech, which is a waveform coding, preserves the speech characteristics compared to parametric coding used in G.729 and GSM. This is an important aspect while mixing many streams in a conference. A word of caution is that the Cepstrum is noise sensitive and computationally expensive compared to energy-based algorithms [12].

References

1. Abhijeet Sangwan, Chiranth M.C, H.S.Jamadagni, Rahul Sah, R. Venkatesha Prasad, Vishal Gaurav, VAD Techniques for Real-Time Speech Transmission on the Internet, The Fifth International Conference on High-Speed Networks and Multimedia Communications HSNMC'02, July 3-5, 2002 - Jeju Island, S. KOREA.
2. A. Sangwan, Chiranth M. C, R. Shah, V. Gaurav, R. Venkatesha Prasad "Voice Activity Detection for VoIP- Time and Frequency domain Solutions", Tenth annual IEEE Symposium on Multimedia Communications and Signal Processing, Bangalore, Nov 2001, 20-24.
3. J.E. Flood, Telecommunications Switching - Traffic and Networks, Prentice Hall India
4. Jongseo Sohn, Nam Soo Kim and Wonyong Sung, "A statistical model-based voice activity detection", IEEE Signal Processing Letters, Vol. 6, no. 1, January 1999
5. Kamilo Feher, Wireless Digital Communications, Prentice Hall India, 2001

6. Khaled El-Maleh and Peter Kabal, "Comparison of Voice Activity Detection Algorithms for Wireless Personal Communications Systems", IEEE Canadian Conference on Electrical and Computer engineering, May 1997, 470-473
7. Petr Pollak, Pavel Sovka, and Jan Uhlir, "Cepstral Speech/Pause Detectors", proc. of IEEE Workshop on Nonlinear Signal and Image Processing, Neos Marmaras, Greece, June 1995, 388-391.
8. Petr Pollak and Pavel Sovka, and Jan Uhlir, "Noise Suppression System for a Car", proc. of the Third European Conference on Speech, Communication and Technology - EUROSPEECH'93, Berlin, Sept 1993, 1073-1076
9. RTP, Real Time Protocol, RFC 1889, http://www.ietf.org/rfc/rfc1889.txt
10. Stefan Pracht and Dennis Hardman, Agilent Technologies - "Voice Quality in Converging Telephony and IP Networks", Ciscoworld Magazine - White Paper 2001
11. Y.D.Cho, K.Al-Naimi and A.Kondoz, "Mixed Decision-Based Noise Adaption for Speech Enhancement", IEEE Electronics Letters Online No. 20010368, 6 Feb 2001.
12. R. Venkatesha Prasad, Abhijeet Sangwan, H.S. Jamadagni, Chiranth M.C, Rahul Sah, VishalGaurav Comparison of Voice Activity Detection Algorithms for VoIP, published at The Seventh IEEE Symposium on Computers and Communications, ISCC'2002, Taormina, Sicily, ITALY, July, 2002.
13. R. Muralishankar and A. G. Ramakrishnan, “DCT Based Pseudo Complex Cepstrum”, ICASSP, Vol. 1, Orlando, Florida, May 2002, 521-524.
14. M.R. Swanson, T. Critchlo, R. Kessler, and L.B. Stoller. "The Design of the Schizophrenic Workstation System," In the Proceedings of the Third Usenix Mach Symposium, April 1993
15. Jerry D. Gibson (ed.), Multimedia Communications - Directions & Innovations, Academic Press, 2001
16. Thomas Enderes Swee Chern Khoo Clare A. Somerville Kostas Samaras Mobile Networks and Applications Volume 7 , Issue 2 (April 2002) . 153-161
17. http://www.nuntius.com/solutions11.html (ITU-T recommended vocoders)
18. http://www.cisco.com/warp/public/788/voip/codec_complexity.html (G729)
19. Jianping Zhang, Wayne Ward and Bryan Pellom, "Phone Based Voice Activity Detection Using Online Bayesian Adaptation with Conjugate Normal Distributions," in ICASSP'2002, Orlando Florida, May 2002.
20. B. Gold and N. Morgan, Speech and Audio Signal Processing, John Wiley Publications.
21. Abhijeet Sangwan, H.S. Jamadagni, Chiranth M.C., Rahul Sah, R. Venkatesha Prasad, Vishal Gaurav, Second and Third Order Adaptable Threshold for VAD in VoIP, ICSP, Beijing, Aug 2002, 1693-1696.
22. Khaled El-Maleh and Peter Kabal, Natural quality background noise coding using residual substitution, EUROSPEECH, Budapest, Sep 1999, Vol. No 5, 2359-2362.