This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2010 proceedings

Order Statistics for Voice Activity Detection in VoIP

R. Muralishankar[†], R. Venkatesha Prasad[*‡], Vijay S[*] H. N. Shankar[†]
[*]Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology, The Netherlands
Email: vprasad@ewi.tudelft.nl, V.SathyanarayanaRao@tudelft.nl
[‡]ESQUBE Communication Solutions, Bangalore, India
Email: vprasad@esqube.com
[†]Department of Telecommunication Engineering, PES Institute of Technology, Bangalore - 560085, India
Email: {muralishankar, hnshankar}@pes.edu

***Abstract*— Realtime voice communication over the Internet has rapidly gained popularity. It is indeed essential to reduce the total bandwidth consumption to efficiently use the available bandwidth for the subscribers having low speed connectivity and even otherwise. In this paper we introduce a novel technique to identify the voice and silent regions of a speech stream that is very much suitable for VoIP calls. We use an entropy measure, which is based on the spacings of *order statistics* of speech frames to differentiate the silence zones from the speech zones. We developed an algorithm that uses an adaptive thresholding to minimize the misdetection. The performance of our approach is compared with the built-in VAD of AMR codec. Our approach yields comparatively better saving in bandwidth yet maintaining a good quality of the speech streams. Further, the proposed approach has improved voice detection compared to the AMR schemes under noisy conditions. The ideas presented in this paper has been identified novel during the WIPO international patent search.**

Keywords: VoIP, Voice Activity Detection, Entropy, Order Statistics.

I. Introduction

The traditional voice services using Public Switched Telephone Networks (PSTN) is no more an isolated network cloud. The trend now is to combine the PSTN and the Internet for various value added services to further reduce the cost of immersed communication. With the advent of Skype, Yahoo, Google talk etc., there has been a steady increase in the voice enabled applications over the Internet [1]. Including the type of codecs used, end-to-end delay, and the amount of bandwidth used there are many issues that need to be addressed here. In [2] some issues like jitter, delay, packet drop, etc., have been addressed. The IP suite which is originally built for data traffic works on the principle of best effort delivery. The main advantage of using Internet for the transmission of realtime voice lies in the statistical multiplexing that could be achieved and flexibility in using multitude of codec types such as G.722.1 [3], iLBC [4], or GSM [5] depending on the bandwidth available and the required quality. In fact the Global IP Solutions (GIPS)[6] has many wideband codecs that could be used to enhance the voice quality and it can be better than the toll quality voice. This is because of the usage of higher sampling frequency and in turn higher rate. Thus bandwidth reduction is common for all these codecs and it is all the more important when wideband codecs are used. In fact some of them have built-in Voice Activity Detection (VAD). VAD reduces the bandwidth as well as the computation required for coding the non-speech packets unnecessarily. Thus voice on IP can be economical and better than toll quality as well compared to circuit-switched networks for long distance calls because of coding, lesser bandwidth required due to statistical multiplexing and high rate of compression due to coding.

Along with the benefits there are some challenges as well, the packet delay, packet loss and delay jitter need to be kept under check [2]. One of the simple ways to reduce the delay at the playout buffer is to detect the talk spurts and transmit only those segments that contain speech. This, while reducing the required bandwidth, also avoids building up of packets at the playout buffer since silence packets are never queued at the playout buffer. Thus there is a need for applying VAD algorithms to detect the talk spurts for a voice calls on the Internet which is the central theme of this paper. Bandwidth saving with VAD can be independent of the codecs used. We also note that VAD algorithms should be as simple as possible so that it can be implemented on any simple portable device in real-time since, wireless hand-held devices are preferred by consumers these days.

Speech during conversation is a sequence of contiguous segments of pauses and speech bursts [7]. Typically contribution from each party is less than 50% of the time [8] during a conversation. Kaleed *et al.*, report a 40% activity of speech in VoIP [9]. Even while a person is speaking there are times when sizeable pauses between words and expressions exist [10]. Thus VAD algorithms take recourse to speech pattern classification to differentiate between speech and silence (pause) periods to save the bandwidth.

Usually a speech segment may be classified as an ACTIVE or an INACTIVE (i.e., silent or non-speech) segment based on its energy. The term INACTIVE segment or silent segment refers to a period of incomprehensible sound which may not have zero-energy [11]. This can also be due to low SNR in the ambiance. Therefore VAD algorithms need to be agile enough to tackle periods of having low audible speech and some times at low SNR conditions. If a packet does not contain voice signal it need not be transmitted. The decision by VAD algorithms is always on a packet-by-packet basis.

The *frame* size is determined based on the interactivity and

978-1-4244-6404-3/10/$26.00 ©2010 IEEE

the type of codec. For example, GSM uses 20 ms and iLBC uses 30 ms/20 ms frames. To increase the link throughput some applications use 60 ms of voice in a packet, for example, Skype and Yahoo usually package 60 ms of voice in a packet, thus a packet contains 2 or 3 frames depending on the type of codec. The VAD algorithm for VoIP has to determine whether a frame contains speech information. In this paper we only deal with the decision by VAD algorithm on frame-by-frame basis. Since a packet may contain more than one frame, decision to drop a packet at the application depends on, say, all the frames in a packet being silent or may be based on majority of frames being silent.

In this paper we use the entropy measure on the spacings of the *order statistics* of the speech frames rather than the typical energy based measure and/or Zero Crossing Rate (ZCR) [12] detectors. We use a simple algorithm that invokes an adaptive thresholding to actively adapt to the changing background noise conditions and track the entropy feature of the speech stream in real-time. We provide the performance of our approach in comparison with the AMR codecs on the speech samples of *switchboard* speech corpus [13]. Our algorithm performance better in terms of compression and the speech quality after removing silence segments.

The rest of the paper is organized as follows. We present earlier work on VAD from the literature and a general description of desirable aspects of VAD algorithms for VoIP in the following subsections. In Section II, we discuss the parameters involved in the VAD design. In Sections III and IV a measure based on Order Statistic is defined and then an algorithm based on it is explained respectively. Section V presents the results and related discussion and the conclusions are presented in SectionVI.

A. Related Studies

VAD is widely used in speech recognition systems, compression and speech coding [14], [15], [16], [17]. In speech recognition systems basically it is used to find the beginning and ending of talk spurts. In codecs it is used to reduce the computation and bandwidth. For VoIP applications stringent detection of beginning and ending of talk spurts is not needed. ITU-T ACELP [18] and GSM [5] coding techniques use in-built VAD but they are computationally expensive. Sovka and Pollak have used spectral subtraction [19] and cepstrum [20], [21] mainly for speech enhancement systems. Complex higher order statistics (HOS) was used for VAD in [22]. These are computationally complex and require training and building a model. These algorithms are mainly used in speech recognition and speech enhancements.

Entropy measure is employed in many of the speech recognition solutions. Waheed *et al.* use Entropy for speech segmentation [23] based on Shen's work [14]. Interestingly both of them use this method for the recorded speech samples to effectively filter the speech bursts so that later these bursts can be used to recognize the uttered speech. They use overlapping frames with each frame of size approximately 25 ms with a 25-50% overlap. They construct a histogram with a varying number of bins in the range of 50 to 100. The entropy is calculated and compared with a fixed threshold which is slightly above the mid point of maximum and minimum entropy values. This calls for screening the whole recorded speech file. However this is not a realtime solution. Another entropy based VAD algorithm was proposed in [24] for realtime application. However it was based on spectral entropy which requires higher computation since spectrum coefficients are a must.

Order statistics have been extensively used by statisticians to address the problems related to statistical estimation [25], [26]. A major impact of the application of order statistics has been in the area of signal processing. In particular, a special case of order statistics, the median filtering has been used in speech and image processing effectively to suppress impulsive noise and preserve the sharp discontinuity [27], [28]. The other applications are in the field of spectrum estimation, under water acoustics data normalization and adaptive edge enhancing. In [29], order statistics filters are applied on subband log-energies which significantly reduces error probability when discriminating speech from non-speech in a noisy environment, that in turn improve the performance of speech recognition under noisy conditions. We use an entropy measure which is introduced in the Section III based on the spacings of the order statistics. We also use an adaptive threshold based algorithm explained in Section IV for each frame and in real-time.

B. Requirements of VAD algorithms

Before we go into the details of the VAD algorithms let us first set some requirements. VAD should: (a) have a good decision rule that exploits the properties of speech to consistently classify segments of speech into INACTIVE and ACTIVE segments; (b) have the adaptability to non-stationary background noise to enhance robustness; (c) have low computational complexity to suit real-time applications; (d) achieve toll quality voice even after filtering non-speech frames; (e) maximize the detection of INACTIVE periods to save the bandwidth.

A VAD algorithm should achieve all of the above requirements within a frame period. Some of the assumptions we have made here are [10]: (a) speech is quasi-stationary. Its spectral form changes over short periods, e.g. 20-30 ms. (b) background noise is relatively stationary, changing very slowly with time. (c) energy of the speech signal is *usually* higher than background noise energy.

II. Parameters for VAD Algorithm Design

A. Choice of parameters

ACTIVE frames bundled together are transmitted, i.e., 2-3 frames in a packet, and these packets are queued at the receiver in playout buffer. Playout buffer essentially cancels the delay jitter due to network introduced delay variations. If a buffer of 7-10 packets is used and if the frame size is 10ms then an initial delay of 30-40 ms (3-4 packets) is expected due to initial playout delay. If the frame duration

were to be 50 ms then initial delay would be 150-200 ms, which is undesirable since, maximum round-trip delay should be within 400 ms [30] for an interactive speech. However, the frame size also depends on the codec type used. If the VAD is invariant to the frame size then it can be universally applied. If the frame size used in the VAD algorithm is less, then the decision may not be proper. Thus we use 20 ms frame size so as to cater to the most of the codecs as well as application without increasing the playout delay. The speech is assumed to be quasi-stationary for 20 ms. Thereby the spectral entropy measure is also assumed to be reliable and hence the validity of the decision. Other parameters we have used here are: (a) 8kHz sampling frequency, (b) linear quantization (16bits linear PCM) and (c) single channel (mono) recording. Advantage of using linear PCM data is that the frame once detected as ACTIVE can be coded into any of the codecs such as G.711, G.723, G.729, GSM, iLBC, etc., before sending it on the network. Since we need to only make a decision as to whether the packet has speech information or not, we need to work on the raw samples. This type of VoIP VAD implementation can be seen in Skype [31], VQube [32], etc., where different types of codecs are used depending on the available bandwidth after the VAD block makes a decision whether a frame has speech.

B. Initial Value of the Threshold

The starting value for the threshold is important for the evolution of the threshold, which tracks the background noise. Though an arbitrary initial choice of the threshold can be used in some cases it may result in poor performance. The VAD algorithm can be trained for a small period in the beginning. The initial threshold level can then be computed from these initial samples. We assume that the initial 100 ms of any call does not contain speech. This is a plausible assumption given that, after a call is established, users need some reaction time before they start speaking. These initial 100 ms are always considered INACTIVE.

$$\widehat{H}(r) = \frac{1}{N_b} \sum_{m=0}^{N_b} \widehat{H}(j) \tag{1}$$

The mean initial entropy used for threshold is calculated using Eq.(1) where, $\widehat{H}(j)$ is some entropy measure that will be explained in the next section. $\widehat{H}(r)$ is used as the threshold (reference) and it would be made adaptable (see the Algorithm IV). We find the entropy for the first five frames to initialize the entropy contour. That is $N_b = 5$ which corresponds to 100 ms. We further keep estimating this parameter for each of the later frames in real-time. A fixed threshold would be 'deaf' to varying acoustic environment of the speaker thus we update the threshold on a continuous basis.

III. Order Statistics and Spacings

Here, we present a short review of the order statistic and spacings and also present the estimates of entropy obtained from the density estimates constructed from the spacings. Suppose we have a set of random variables $X = X_1, X_2, ...X_i..., X_N$ in a speech frame and we arrange this set of random variables in ascending order of magnitude such that,

$$X_{(1)} \leq X_{(2)} \leq .... \leq X_{(N)}, \tag{2}$$

where subscript (n) denotes the index of the variable after ordering. Let us denote the elements of the set X after ordering by Y_n, where $n = 1, 2, ..., N$ such that

$$Y_n \equiv X_{(n)}.$$

Then Y_n is called the n^{th} - order statistic. A spacing of order m, or m-spacing, is then defined as,

$$Y_{i+m} - Y_i \quad \text{for} \quad 1 \leq i < i + m \leq N. \tag{3}$$

Based on the spacings it is possible to construct a density estimate [33],

$$f_N(y) = \frac{m}{N} \left(\frac{1}{Y_{im} - Y_{(i-1)m}} \right), \tag{4}$$

if $y \in \left[Y_{im}, Y_{(i-1)m} \right)$. This density estimate is consistent if as $n \to \infty, m_n \to \infty$ and $\frac{m_n}{n} \to 0$. Where m_n is nothing but m in the limiting sense. The estimates of entropy based on sample-spacings can be derived by substituting spacing density estimate in the place of the density function. It was reported in [33] that one can get a consistent spacing based entropy estimate from a non-consistent spacing density estimate. The m-spacing estimate of the entropy for fixed m was considered in [33] and is given below as,

$$\widehat{H}_{m,N}(Y) = \frac{1}{N} \sum_{i=1}^{N-m} \ln \left(\frac{N}{m} \left(Y_{i+m} - Y_i \right) \right) - \psi(m) + \ln(m), \tag{5}$$

where $\psi(x) = -(\ln\Gamma(x))'$ is a *Digamma* function. In order to decrease the asymptotic variance m_n-spacing estimate with $m_n \to \infty$ was considered by Vasicek [34] and is given below as,

$$\widehat{H}_N(Y) = \frac{1}{N} \sum_{i=1}^{N-m_n} \log \left(\frac{N}{m_n} \left(Y_{i+m_n} - Y_i \right) \right). \tag{6}$$

IV. Adaptive Threshold Algorithm

We first provide our algorithm as a pseudo code here. Later we explain the stages in which we have arrived at the specific methodology of adapting entropy measure for VAD in VoIP. We denote N as the number of samples in a frame of 20 ms (which is equal to 320 samples for usual 8 kHz sampling). We define a variable $threshold$ which would follow the entropy curve of the speech in a real-time call. We use a boolean variable $bSpeechFrame$ to denote whether the frame under consideration contains speech or pause. $bSpeechFrame$ is 1 if the current frame is a speech frame and it is 0 otherwise.

Let $nCompression$ denote the running total number of frames declared as speech till that instant. $nCompression$ is counted whenever $bSpeechFrame$ is 1. A constant HANG-OVER_COUNT (HC), which denotes the consecutive number of frames that do not contain speech could be used to allow

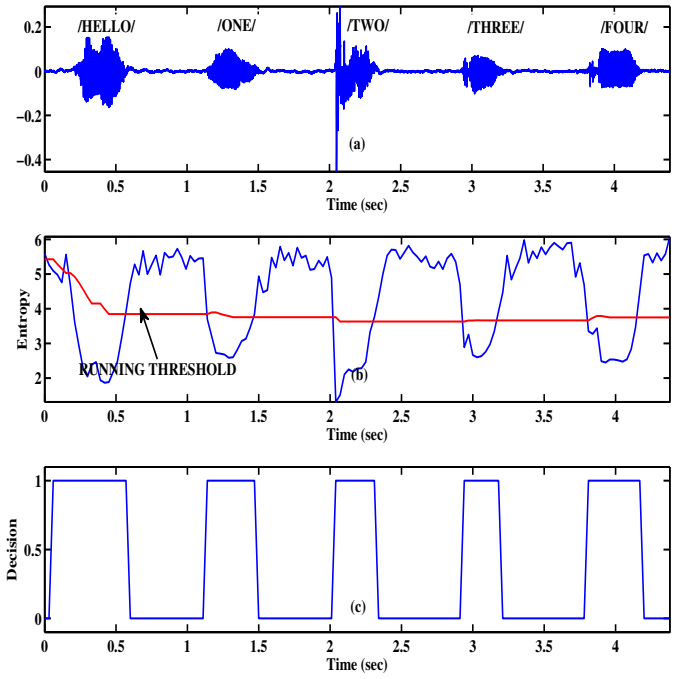

Fig. 1. VAD decisions for clean speech. (a) Original speech waveform (b) Entropy curve with thresholding under 20ms frame size (c) Decision obtained from proposed algorithm

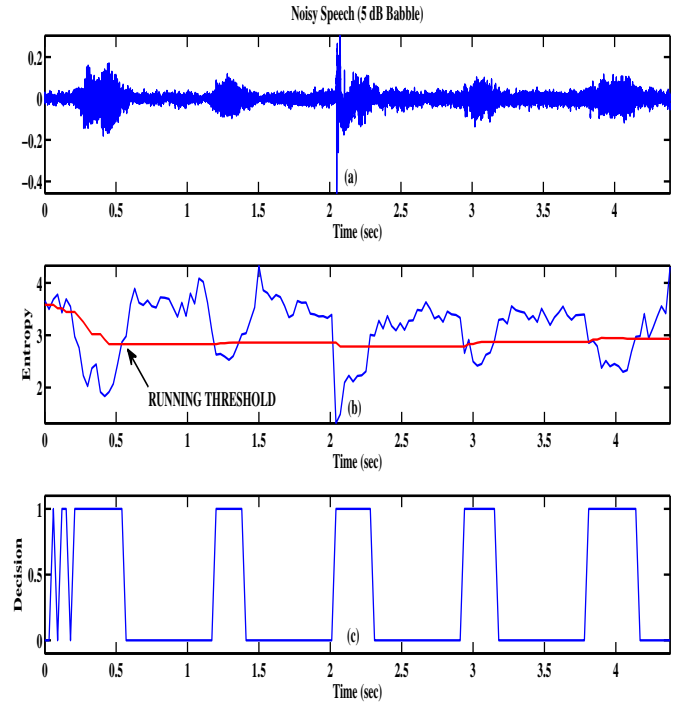

Fig. 2. VAD decision for noisy speech signal. (a) Original speech waveform with 5dB Babble noise (b) Entropy curve with thresholding for the speech (c) Decision obtained from proposed algorithm.

a few more silence packets once a pause is detected and it is not used here to check the algorithm for its robustness. This algorithm is given in Algorithm IV. In Step 1 we find

Algorithm 1 Entropy based VAD Algorithm

1a. *for* $j = 1$ to 5
$\widehat{H}_{m,N}(Y) = \frac{1}{N}\sum_{i=1}^{N-m} \ln\left(\frac{N}{m}\left(Y_{i+m} - Y_i\right)\right)$
$-\psi(m) + \ln(m)$; Eq.5
1b. $maxValue = \max\{\widehat{H}(j)\}\ \forall j = 1$ to 5
2. $threshold = \text{mean}(\widehat{H}(1) : \widehat{H}(5))$
3. $minValue = maxValue;\ nCompression = 0;$
4. *for* $j = 6$ to end of Call
(a) find $\widehat{H}(j)$ (as in Step.1);
(b) $if(maxValue < \widehat{H}(j))\{$
$maxValue = \widehat{H}(j);$
$incr = (maxValue + minValue)/10;$
$threshold = (threshold + incr)/1.25;$
$\}$
(c) $if(\widehat{H}(j) < threshold)\{$
$bSpeechFrame = 1;$
$nCompression = nCompression + 1;$
$if(minValue > \widehat{H}(j))$
$\{$
$minValue = \widehat{H}(j);$
$incr = (maxValue + minValue)/10;$
$threshold = (threshold + incr)/1.25;$
$\}$
$\}$
(d) $else\{$
$bSpeechFrame = 0;$
$minValue = maxValue;$
$\}$

the entropy for the first five frames. In Step 2, we use the mean of the first five calculated entropy values. $threshold$, the contour tracker, is nothing but a moving average and is initialized to the mean of first five entropy values as given in Step 3 which is used to set $maxValue$ and $minValue$. In Step 4 we take each frame starting from the sixth frame and as-and-when a recorded speech frame is available for decision making, and we calculate its entropy. The $threshold$ adapts to the contour by using $maxValue$ and $minValue$ as given in Step 4(b) and Step 4(c). The decision is also made in Step 4(c) and Step 4(d). The decision may also include some guard band (hangover) using HC so that the decision is not made immediately after detecting the first INACTIVE frame to avoid clipping which is not done here to compare the raw decision. In this entropy based solution the guard band can be really small and of the order of even 2-3 frames in contrast with higher number of frames required in energy based solutions because of possibility of misdetections [11],[24]. Percentage of compression can be found by using the expression $(1-(nCompression/j))\times 100$, where j is the running total number of frames at an instant, j. As long as $bSpeechFrame$ is false (or zero) we can withhold

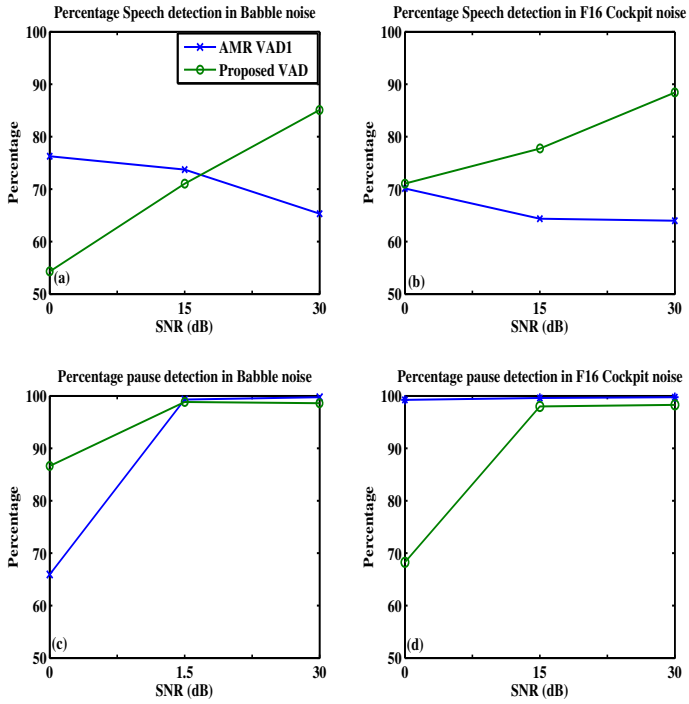

Fig. 3. Performance comparison between AMR VAD1 and the proposed method. (a) and (b) Percentage speech detection performance under Babble and F16 Cockpit noise conditions (c) and (d) Percentage pause detection under Babble and F16 Cockpit noisy conditions. SNR is varied from 0 to 30 dB.

the transmission of speech frames. If the VoIP application is using higher packet size i.e., having more than one frame, then one more level of decision making is needed. For example, the decision can be based on majority frames being INACTIVE; or a packet can be termed as speech even if one frame in that packet is found to be a speech frame.

V. Results and Discussions

We did the experiments with recorded speech samples. We report comparative performances of our algorithm with AMR VAD1 [35]. VAD1 is a better than AMR VAD2 thus we have compared our algorithm with it [36]. We also give some insights into the working of our algorithm based on order statistic spacings.

We first divided incoming signal into a number of frames with frame size equal to 20 ms. We then used order-statistics to compute the entropy using equation (6). Fig. 1(a) shows the speech signal for utterances of /Hello/, /One/, /Two/, /Three/ and /Four/ with deliberate pauses in between the words. Fig. 1(b) shows the Entropy obtained from the spacings of the order statistics using the equation (6). The red line in Fig. 1(b) shows the adaptive threshold values for the respective entropy values. Figure 1(c) shows the decision taken by the algorithm. The speech frame is marked as ACTIVE if decision is 1 and INACTIVE otherwise. The decision is 1 when the entropy value is less than the threshold. Fig. 2(a) shows the speech signal of Fig. 1(a) with the signal being corrupted with additive babble noise and the overall SNR is 5 dB. Even at 5 dB SNR the decision is almost similar to the clean speech. One can also observe that the initial fluctuations in the decision making before adapting to the background noise in Fig. 2(c). We can compare Fig. 2(c) and Fig. 1(c) to infer the robustness of our entropy based algorithm.

Fifty samples were considered from the SWITCHBOARD corpus [13]. The SWITCHBOARD corpus is composed of approximately 2,400 telephone conversations between unacquainted adults. The participants in the conversations vary in age and represent all major U.S. dialect groups. From the corpus we used 50 conversations that were syntactically parsed. Each conversation is of 1 minute duration. We simulate noisy SWITCHBOARD samples using noise samples from the NOISEX database. Here, SNR of 0 dB 15 dB and 30 dB are considered. The reason behind the selection of these intervals of SNR is due the fact that AMR [35] results are available only for these SNR values. For each of those SNR increments, we compute the percentage speech and pause detection. We then compared the results of our algorithm with the standard AMR VAD.

Fig. 3 shows the outcome of speech and pause detection under babble and F16 cockpit noise. Pause detection performance of the proposed algorithm and AMR VAD is well above 90% for SNR between 0 and 5 dB. However, speech detection performance is comparatively low for 0 dB SNR where it starts from 55% and reach the maximum performance of 90% for 30 dB. The not so impressive speech detection performance in the initial part is due to the absence of hangover. This can be suitably incorporated to maximize speech detection percentage. The proposed algorithm is very simple to implement. Nature of the running threshold is such that it can adopt quickly with the changing background conditions. The computational complexity of our algorithm is $O(N/2\log(N))$ per frame. This low delay and the low complexity makes our approach easily feasible to be implemented on many embedded devices too. On Mobile Intel Pentium 4 CPU with 2.20 GHz clock, for 20 ms frame the time required to execute the algorithm is 23 μs. To the best of our knowledge, entropy based on spacings of order statistics for VAD is probably the first. The novelty of this work has been acknowledged by WIPO (under PCT) claim [37].

VI. Conclusions

We proposed a novel algorithm for VAD using entropy derived from the spacings of the order-statistics to find ACTIVE and INACTIVE zones in a speech stream. We compared our scheme with the VAD in relatively new AMR [35] codec in terms of percentage of speech and non-speech detection. We see a better compression rate without any major loss in the subjective quality compared with the AMR scheme. Our VAD scheme is largely invariant to speaker change. We have shown that our approach results in a better speech and matched AMR performance during. It can be seen that the overall delay and computational complexity are minimal. In fact, the complexity is highly due to the sorting algorithm which could be made

small by going for smaller frame sizes. While we find some advantages in our approach, we think there is a long way ahead in terms of applicability of our approach in various situations. The next step is to compare it with the other VAD algorithms available in many of the standard codecs. We have considered only babble and F16 noise in this paper. Babble noise is one of the most common noise type that can affect VoIP calls. However it will be interesting to see the effect of other types of noise such as car noise. Next logical step is to enhance our algorithm to be more effective under different conditions and test it in real environments like [32].

REFERENCES

[1] S. Pracht and D. Hardman, "Agilent technologies voice quality in converging telephony and IP networks," Cisco," Ciscoworld Magazine White Paper, 2001.

[2] J.-C. Bolot and A. Vega-Garca, "Control mechanisms for packet audio in the internet," in *Proc. IEEE INFOCOM'96*, San Francisco, CA, Mar. 1996, pp. 232–239.

[3] "Low-complexity coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss," ITU," Specification, 2001.

[4] "Internet Low Bitrate Codec (iLBC)," IETF," RFC 3951, 2001.

[5] "GSM-Enhanced Full Rate Specifications 06.51, 06.60-63 and 06.82," ETSI," Specification, 2001.

[6] iSAC: Wideband Codec from Global IP Solutions. [Online]. Available: http://www.gipscorp.com/files/english/datasheets/Codecs.pdf

[7] B. Gold and N. Morgan, *Speech and Audio Signal Processing*. New York: John Wiley and Sons, 2000.

[8] J. Natvig, S. Hansen, and J. De Brito, "Speech processing in the pan-european digital mobile radio system (GSM) – system overview," in *Proc. IEEE Global Telecommunications Conference (IEEE GLOBECOM 1989)*, 1989, pp. 1060–1064.

[9] K. El-Maleh and P. Kabal, "Natural quality background noise coding using residual substitution," in *Proc. EUROSPEECH*, vol. 5, Sept. 1999, pp. 2359–2362.

[10] A. M. Kondoz, *Digital Speech*. New York: John Wiley and Sons, 1999.

[11] R. V. Prasad, A. Sangwan, H. S. Jamadagni, and M. C. Chiranth, "Comparison of voice activity detection algorithms for VoIP," in *Proc. IEEE Symposium on Computer and Communications*, July 2002, pp. 530–535.

[12] L. Rabiner and M. Sambur, "An algorithm for determining end-points of isolated utterances," *Bell Syst. Techn. J.*, pp. 297–315, Feb. 1975.

[13] SWITCHBOARD: A User's Manual. [Online]. Available: http://www.ldc.upenn.edu/Catalog/readme_files/switchboard.readme.html

[14] J. L. Shen, J. W. Hung, and L. S. Lee, "Robust entropy based endpoint detection for speech recognition in noisy environments," in *Proc. Int. Conf. on Spoken Lang. Processing*, 1998.

[15] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Sig. Proc. Lett.*, vol. 6, no. 1, pp. 530–535, 1999.

[16] Y. D. Cho, K. Al-Naimi, and A. Kondoz, "Mixed decision-based noise adoption for speech enhancement," *IEE Electr. Lett.*, vol. 6, 2001.

[17] K. El-Maleh and P. Kabal, "Comparison of voice activity detection algorithms for wireless personal communications systems," in *Proc. IEEE Canadian Conference on Electrical and Computer Engineering*, 1997, pp. 470–473.

[18] "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP)," ITU," ITU-T Rec. G.729, 1996.

[19] P. Pollak, P. Sovka, , and J. Uhlir, "The noise suppression system for a car,," in *Proc. EUROSPEECH*, vol. 5, Sept. 1993, pp. 1073–1076.

[20] P. Pollak, P. Sovka, and J. Uhlir, "Cepstral speech/pause detectors," in *Proc. IEEE Workshop on Nonlinear Signal and Image Processing*, 1995, pp. 388–391.

[21] P. Sovka and P. Pollk, "The study of speech-pause detectors for speech enhancement methods," in *Proc. EUROSPEECH*, 1995, pp. 1575–1578.

[22] E. Nemer, R. Goubran, and S. Mahmould, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Trans. Speech and Audio Proc.*, vol. 9, no. 3, pp. 217–231, 2001.

[23] K. Waheed, K. Weaver, and F. M. Salam, "A robust algorithm for detecting speech segments using an entropic contrast," in *Proc. 45th IEEE International Midwest Symposium on Circuits and Systems*, vol. 3, Aug. 2002, pp. 328–331.

[24] R. Venkatesha Prasad, R. Muralishankar, S. Vijay, and H. N. Shankar, "Voice activity detection for voip — an information theoretic approach," in *Proc. IEEE Globecom*, Nov. 2006.

[25] E. H. Lloyd, "Least-squares estimation of location and scale parameters using order statistics," *Biometrica*, vol. 39, pp. 88–95, 1952.

[26] A. E. Sarhan, "Estimation of mean and standard deviation by order statistics," *Ann. Math. Statistics*, vol. 25, pp. 317–328, 1954.

[27] L. R. Rabiner, M. R. Sambur, and C. E. Schmidt, "Applications of a nonlinear smoothing algorithm to speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 552–557, 1975.

[28] L. R. Huang, *Two-dimensional Digital Signal Processing II: Transforms and Median filters*. New York: Springer Verlag, 1981.

[29] J. Ramirez, J. C. Segura, C. Benitez, A. Torre, and A. Rubio, "Two-dimensional digital signal processing II: Transforms and median filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 13, no. 6, pp. 1119–1129, Nov. 2005.

[30] "One-way transmission time," ITU," ITU-T Rec. G.114, 1993.

[31] Skype P2P Internet Telephony. [Online]. Available: http://www.skype.com/

[32] VQube Internet Telephony Application. [Online]. Available: http://www.vqube.com/

[33] J. Beirlant, E. J. Dudewicz, L. Gyorfi, and E. C. Meulen, "Nonparametric entropy estimation: An overview," *International Journal of the Mathematical Statistics and Sciences*, no. 6, pp. 17–39, 2001.

[34] O. Vasicek, "A test for normality based on sample entropy," *Journal of the Royal Statistical Society, Series B*, vol. 38(1), no. 6, pp. 54–59, 1976.

[35] "Digital cellular telecommunications system (phase 2+); universal mobile telecommunications system (UMTS); Adaptive Multi-Rate (AMR) speech codec; (3GPP TS 26.102 version 6.0.0 Release 6)," 3GPP," TS 126 102, ETSI, Jan. 2005.

[36] R. Padmanabhan, P. Sree Hari Krishnan, and Hema A. Murthy, "A pattern recognition approach to VAD using modified group delay," in *Proc. 14th National conference on Communications,*, Feb. 2008, pp. 432–437.

[37] The World Intellectual Property Organization (WIPO). [Online]. Available: http://www.wipo.int/pctdb/en/wo.jsp?WO=2008090564