Voice Activity Detection for VoIP—An Information Theoretic Approach

R. V. Prasad*, R. Muralishankar†, Vijay S.‡, H. N. Shankar†, Przemysław Pawełczak* and Ignas Niemegeers*
*Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology, Mekelweg 4, 2600 GA Delft, The Netherlands
Email: {vprasad, p.pawelczak, i.niemegeers}@ewi.tudelft.nl
‡ESQUBE Communication Solutions, Bangalore - 560094, India
Email: vijay@esqube.com
†Faculty of Telecom Engineering
PES Institute of Technology, Bangalore - 560085, India
Email: {muralishankar, hnshankar}@pes.edu

***Abstract*— Voice enabled applications over the Internet are rapidly gaining popularity. Reducing the total bandwidth requirement can make a non-trivial difference for the subscribers having low speed connectivity. Voice Activity Detection algorithms for VoIP applications can save bandwidth by filtering the frames that do not contain speech. In this paper we introduce a novel technique to identify the voice and silent regions of a speech stream very much suitable for VoIP calls. We use an information theoretic measure, called spectral entropy, for differentiating the silence from the speech zones. Specifically we developed a heuristic approach that uses an adaptive threshold to minimize the miss detection in the presence of noise. The performance of our approach is compared with the relatively new 3GPP TS 26.194 (AMR-WB) standard, along with the listeners' intelligibility rating. Our algorithm yields comparatively better saving in bandwidth, yet maintaining good quality of the speech streams.**

I. Introduction

Long distance calls are expensive while transported over Public Switched Telephone Networks (PSTN). Thus the current trend is to provide this service on data networks, especially in the light of popular applications like Skype, Yahoo, etc. IP suite, originally built for data traffic, works on the best effort delivery principle. Since resource sharing through statistical multiplexing is used in data networks, the total number of calls supported may be enhanced at the gateways, which bridge PSTN to the IP networks through speech compression techniques or codecs like G.722.1 [1], iLBC [2], or GSM [3]. Therefore such services can be economical compared to circuit-switched networks for long distance calls.

However, data network does not guarantee a faithful voice transmission and reproduction as in PSTN. We need to keep the packet delay, loss and delay jitter under check. One of the simple ways to reduce the delay at the playout buffer is to detect the talk spurts and transmit only those segments. This, while reducing the bandwidth, also avoids building up of playout buffer. Thus there is a need for applying Voice Activity Detection (VAD) algorithms to detect the talk spurts for a voice call 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 algorithm should be as simple as possible so that it can be implemented on many simple portable devices in real-time.

Conversational speech is a sequence of contiguous segments of pauses and speech bursts [4]. The strategy is to use the fact that no speech channel is continuously active. In a conversational setup, typically contribution from each party is less than 50% of the time [5]; Kaleed *et al.*, report a 40% activity of speech in VoIP [6]. In fact even when one party is speaking, there are times when sizeable pauses between words and expressions exist [7]. Thus VAD algorithms take recourse to speech pattern classification to differentiate between speech and silence (pause) periods to save bandwidth.

A speech segment may be classified as an *active* or an *incative* (silent or non-speech), based on its energy. The term incative segment refers to a period of incomprehensible sound which may not have zero-energy [8]. Therefore VAD algorithms are to be agile enough to tackle periods of having low audible speech, sometimes at low SNR conditions. Simply, 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. In VoIP systems the voice data (or payload of the packet) is transmitted along with headers of the different layers of the network. The header size for Real Time Protocol (RTP) [9] is 12 bytes, UDP/IP adds up another 28 bytes. The ratio of header to payload size is one of the factors for selecting the payload size for a better throughput from the network. Smaller payload yields better interactivity for the conversation, but decreases the throughput of the network. Alternately, a higher payload size increases the throughput but performs poorly in terms of interactivity.

A constant sized samples representing a segment of speech are referred as a *Frame* in this paper, while its size is determined not only based on the above considerations but also based on the phoneme size, codecs used, etc. For example, Skype and Yahoo usually package 60 ms of voice in a packet.

The decision by VAD algorithm in this paper is always on a frame-by-frame basis. Since a packet may contain more than one frame, decision to drop a packet depends on, say, all the frames in a packet being silent. Also in this paper we use spectral entropy based measure instead of the typical energy based measure and Zero Crossing Rate (ZCR) [10] detectors. We use a heuristic approach which invokes a variable median threshold on the spectral entropy feature of a speech stream in real-time. We provide the performance of our approach in comparison with the 3GPP TS 26.194 Adaptive Multi Rate Wide Band (AMR-WB) [11], that employs frame energy, SNR estimation along with threshold adaptation. Our algorithm performs better in terms of compression and the speech quality after removing silence segments.

The rest of the paper is organized as follows. We first present earlier works on VAD 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 and in Section III, the heuristic approach is developed. Section IV presents the results and related discussion, while the conclusions are presented in Section V.

A. Previous VAD Schemes

VAD was used first in Speech Recognition systems, Compression and Speech coding [12]–[15] to find the beginning and ending of talk spurts. For VoIP applications stringent detection of beginning and ending of talk spurts is not needed. Coding techniques like [3], [16] use built-in VAD but they are computationally expensive. Sovka and Pollak have reported their work using spectral subtraction [17] and cepstrum [18], [19], mainly for speech enhancement systems. VAD involving complex higher order statistics (HOS) is proposed in [20]. These are computationally complex and require training and building a model.

Entropy measure is employed in many of the speech recognition solutions. Waheed *et al.* use entropy for speech segmentation [21] based on Shen's work [12]. 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 approximately 25 ms length with a 25-50% overlap. Also 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 considered. This calls for screening the whole recorded speech file.

We use a different approach in contrast to theirs. In this paper a threshold based on variable median, tracking the actual entropy for each frame in real-time is used. The measure is same as in [12], [21], however our approach is considerably different while adopting it for real-time frame by frame decision. Also we take into account the varying nature of speech characteristics vis-á-vis background noise, as well as speaker independent methodology. We tune entropy feature to suit the VoIP voice packets.

We emphasize that VAD for VoIP application is less demanding than the applications mentioned above, and is more in favor of decision making in real-time. It can be noted here that there is no significant increase in bandwidth if a few silent packets marked as speech segments are transmitted after the speech burst to avoid abrupt cuts and to improve quality.

B. Desirable Aspects of VAD Algorithms

We list some desirable aspects of good VAD algorithms for VoIP applications.

- VAD algorithm must implement a good decision rule that exploits the properties of speech to consistently classify segments of speech into inactive and *active*.
- It should adapt to non-stationary background noise to enhances robustness.
- The computational complexity of VAD algorithm must be low to suit real-time applications.
- Toll quality voice should be achieved after applying VAD algorithm.
- VAD must maximize the detection of inactive periods to save the bandwidth.

The assumptions on the VAD algorithm proposed here is based on the following characteristics [7].

- Speech is quasi-stationary. Its spectral form changes over short periods, e.g. 20-30 ms.
- Background noise is relatively stationary, changing very slowly with time.
- Energy of the speech signal is usually higher than background noise energy; else speech will be unintelligible.

II. Parameters for VAD Design

A. Choice of Frame Duration

Active frames bundled together are transmitted and queued up in a packet-buffer at the receiver. This allows to play audio even if incoming packets are delayed due to network conditions. Consider, a VoIP system having a buffer of 7-10 packets. Having a packet size equivalent to 10 ms allows the VoIP system to start playing the audio at the receiver's end after 30-40 ms from the time the queue started building up. If the frame duration were 50 ms, an initial delay would be of 150-200 ms, which is unsuitable since, maximum roundtrip delay should be within 400 ms [22] for a good quality speech. Therefore, the frame duration must be chosen properly. VoIP systems may use 5-60 ms frame sizes and we see many popular VoIP applications use 60 ms packet size. 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. We thus use in our algorithm 20 ms speech frames. A packet may contain many of these frames depending on the design of the application as discussed above.

In our VAD algorithm we assume each speech frame to be (a) 16kHz frequency sampled, (b) linear quantized (8/16 bits linear PCM) and (c) single channel (mono) recorded. Advantage of using linear PCM is that the voice data can be transformed to any other compressed code like 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, we need to work on the raw samples. This type of VoIP implementation can be seen in

VQube [23] VoIP engine, where different types of codecs are used, depending on the available bandwidth after the VAD block makes a decision.

B. Energy of a Frame

The energy of a frame indicates possible presence of speech information and is a useful parameter for VAD algorithms. Let $s(n)$ be the nth sample of speech. If the length of the frame were N samples, then the energy for the jth frame can be represented as

$$E(j) = \sum_{n=0}^{N-1} s^2(n),$$

where $j = 1, \ldots, N_f$ and N_f is the number of frames for a speech stream.

C. Spectral Entropy of a speech stream

The Discrete Fourier Transform (DFT) of $s(n)$ for the jth frame is given by,

$$S_j(k) = \sum_{n=0}^{N-1} s(n) \exp\left(-j\frac{2\pi kn}{N}\right),$$

where $S_j(k)$ is the kth DFT co-efficient in the jth frame. This DFT spectrum for the jth frame can be viewed as a vector of coefficients of the orthonormal basis. In order to handle variations due to different speakers with respect to their pitch frequencies, simultaneously minimizing the noise interference, we consider the mid-frequency band coefficients ranging from 350 Hz to 3000 Hz in the entropy evaluation, e.g., the rest of the coefficients are forced to zero when

$$350\,\text{Hz} > \frac{kF_s}{N} > 3000\,\text{Hz}.$$

The PMF for the spectrum for jth frame can be estimated by normalizing over all the frequencies as

$$p_j(k) = \frac{|S_j(k)|}{\sum_{m=0}^{N-1} |S_j(m)|},$$

where $k = 1, ..., N-1$. Finally the spectral entropy is given by [12]

$$H(j) = -\sum_{k=0}^{N-1} p_j(k) \log(p_j(k)).$$

D. Initial Value of the Threshold

The starting value for the threshold is important for its evolution, 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. Two methods are proposed for finding a starting threshold value [8].

Method 1: The VAD algorithm is trained for a small period using a prerecorded speech sample that contains only background noise. The initial threshold level for various parameters then can be computed from this speech sample. For example, the initial estimate of energy is obtained by taking the mean of the entropy of each sample frame as

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

where $H(r)$ an initial threshold estimate and N_b is the number of frames in prerecorded speech sample. In contrast to [21] the threshold based on entropy is always fixed, by taking into consideration the whole prerecorded speech sample as

$$\gamma = \frac{\max\{H(j)\} - \min\{H(j)\}}{2} + c\min\{H(j)\}.$$

This method can not be used for VoIP, since the calls can be longer and the background noise can vary with time. Further we can not assume that the user would be able to provide some samples often. Thus we would use the second method given below.

Method 2: Though similar to the previous method, here we assume that the initial 100 ms of any call does not contain any speech. This is a plausible assumption given that users need some reaction time before they start speaking. These initial 100 ms are considered inactive. Their mean energy is calculated using (1). We set $N_b = 5$. For entropy based algorithm, we find the initial entropy for the first 5 frames to initialize the entropy contour and assume that it is from the inactive frames. We further keep estimating this parameter for each of the later frames in real-time. A fixed threshold would be 'deaf' to varying acoustic environments of the speaker. Since we try to adopt to the changing background condition here we use this method.

III. Entropy Based VAD Algorithm for VoIP

We first provide our algorithm as a pseudo code. 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 for usual 16 kHz sampling). We define a variable CT (*Contour Tracker*), which would follow the entropy curve of the speech in a real-time call. A constant DB is the decision band and is initialized to an empirical value of 0.04. We use a boolean variable $bSpeechFrame$ to denote whether the frame under consideration contains speech or pause.

Let $nSilentFrames$ denote the number of continuous pauses encountered; a constant HC (*Hangover Count*), which denotes the consecutive number of frames that do not contain speech. Let $nCompression$ denote the running total number of frames declared as silent/pause untill that instant. Full description is given in Algorithm 1.

In Step 1 we find the entropy for the first five frames. In Step 2, we use the median filter of order 5 on the calculated entropy. Contour tracker, CT, is nothing but a moving average

Algorithm 1 Entropy based VAD.

```
1. for $j = 1$ to $5$
   $H(j) = -\sum_{k=0}^{N-1} p_j(k) \log(p_j(k))$;
2. for $j = 1$ to $5$
   $H(j) = \text{medianfilter}(H(j), 5)$;
3. $CT = \text{mean}(H(1) : H(5))$;
4. for $j = 6$ to end of Call
   $(a)$ findH(j);
   $(b)$ $H(j) \leftarrow \text{medianfilter}(H(j), 5)$;
   $(c)$ $CT = \text{mean}(H(1) : H(5))$;
   $(d)$ if$(H(j) < (CT - DB) \; || \; H(j) > (CT + DB))$ {
        $nSilentFrames$ = 0;
        $bSpeechFrame$ = 1;
        $nCompression$ = $nCompression$ + 1;
        }
      else if ($nSilentFrames < HC$) {
        $nSilentFrames$ = $nSilentFrames$ + 1;
        $bSpeechFrame$ = 1;
        $nCompression$ = $nCompression$ + 1;
        }
      else
        $bSpeechFrame = 0$;
```

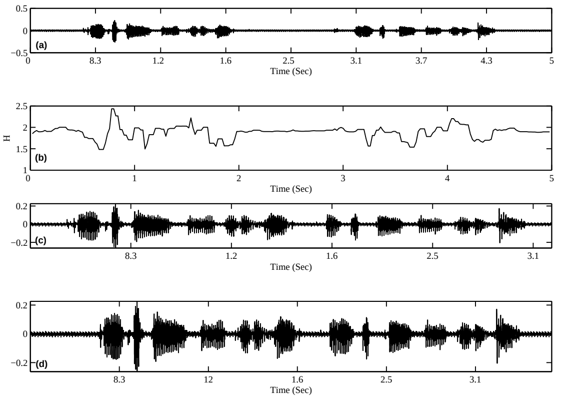

Fig. 1. From top to bottom: (a) Input Speech, (b) Spectral Entropy (c) VAD output using our approach (d) VAD output using AMR-WB.

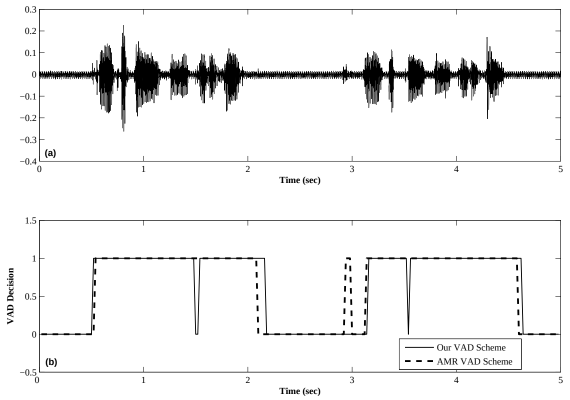

Fig. 2. VAD Decisions under clean speech. From top to bottom: (a) Input speech (b) VAD Decisions.

and is initialized to the mean of first five entropy values as given in Step 3. 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. We use the median filter to avoid large variations of $H(j)$. The CT adapts to the contour by taking the mean between the current CT and the $H(j)$ (Step 4(c)). The decision is made in Step 4(d). This decision also includes some guard band (hangover) using HC, so that the decision is not made immediately after detecting the first inactive frame to avoid clipping (Step 4(d)).

In the entropy based solution the guard band can be really small and of the order of even 2-3 frames, in contrast with a higher number of frames required in energy based solutions because of possibility of miss detections [8]. Percentage of compression can be found by using the variable, $nCompression$, as $1 - (nCompression/N_f)$, where N_f is the total number of frames. As long as $bSpeechFrame$ is false (or zero) we can withhold the transmission of speech frames. If the VoIP application is using higher packet size than frame size, then one more level of decision making is needed. For example, the decision can be based on majority frames being inactive.

IV. Results and Discussion

We used Matlab to run our proposed algorithm on 50 sample files with varying SNR and duration, which were recorded with a PC microphone, generally used in a VoIP setup. We compare our algorithm with AMR-WB based VAD algorithm presented in [11]. Fig. IVa shows the input speech, Fig. IVb spectral entropy of Fig. IVa, Fig. IVc the resultant speech with silence stripped off from Fig. IVa based on entropy approach, and Fig. IVd using AMR-WB based algorithm. The SNR of signal Fig. IVa is around 20 dB. From these figures, we can demonstrate the effectiveness of our approach in terms of inactive zone detection.

The speech zones have been detected effectively as can be seen in Fig. IVb and Fig. IVb. The better performance here is due to the following reasons:

1) Insusceptibility to loudness variations in our approach;
2) The frequency domain filtering of the speech as given in II-C helps to remove speaker variations by eliminating the pitch information in the spectral entropy measure. In fact, it becomes invariable to speaker change in the middle of the conversation. This cannot be expected from other algorithms unless a speech contour coding is used;
3) High frequency part of the spectrum is in general more susceptible to noise and eliminating these portions in the

TABLE I

Comparison of AMR-WB and Entropy based VAD with respect to various parameters.

SNR	Intelligibility		Insertion		Deletion		Compression	
	AMR-WB	Entropy	AMR-WB	Entropy	AMR-WB	Entropy	AMR-WB	Entropy
Clean Speech	98%	98%	11	13	2	3	34.4%	32.1%
15 dB Babble noise	93%	93.5%	18	14	3	5	29.2%	33.9%
10 dB Babble noise	90%	92%	23	10	3	7	24.8%	34.3%
5 dB Babble noise	88%	91.6%	30	9	3	6	20.2%	33.7%

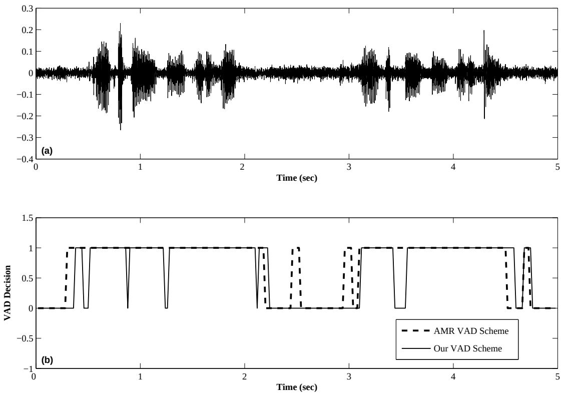

Fig. 3. VAD Decisions under 10 dB Babble noise. From top to bottom: (a) Input speech (b) VAD Decisions.

spectral entropy measure would enhance our detectability. The region from 350-3000 Hz is mainly dominated by the first three formant frequency information and is very useful and reliable in identifying the speech activity;

Figure IV shows an detection results of our algorithm in noisy condition. We added Babble noise to the input speech as shown in Fig. IVa. The SNR of the noisy speech in Fig. IVa is 10 dB. From the decision results, one can infer effectiveness of our algorithm in identifying speech zones compared to the AMR-WB based method. Here the quality of speech is not of concern, because the identified speech zones is already been corrupted by the noise. The important thing to notice is compression obtained with speech zone information. Though we see some miss detection in Fig. IVb, the noise does not influence in VAD decision using our approach. This is because of enhanced spectral entropy feature by frequency domain filtering before taking the decision. Thus we can see robustness of our approach under noisy conditions.

We also use median filter for five past samples which we would have anyway started within the beginning. We want to highlight that taking median filter will not introduce any extra delay in our algorithm. It only requires us to keep a buffer of five units to store the past $H(n)$.

We evaluated performance of our VAD scheme with AMR-WB by considering parameters such as Intelligibility, Insertion, Deletion and compression. We considered 15 unbiased listeners to grade output speech intelligibility of both the schemes. Intelligibility is defined as understanding the speech. A packet insertion is said to be done if an inactive region is miss detected as speech by the algorithm. Similarly, a packet deletion is miss detection of speech packet as inactive. Continuous deletion of the packets results in loss of speech and hence lower intelligibility while a single packet loss goes undetected by the listener. Compression ratio is defined as the ratio of total pause duration to the total duration of the signal. This is represented as percentage. The results are tabulated in Table I. All the results projected are the average values of the parameters obtained from these sample files.

In a VoIP setup, the inactive regions are replaced by the comfort noise during playout at the receiver. So, the intelligibility is expected to vary slightly under noisy conditions provided inactive detection is accurate. From Table I, we can see better intelligibility of our VAD scheme with respect to AMR-WB. This can be related to the lesser packet insertions in our scheme. A $HC = 3$ (60 ms) is used in our algorithm which increases the packet insertions to include final speech periods during transitions from speech to inactive zone. Our scheme uses the information theoretic approach which is higher, whenever there is an increase in randomness in the signal. As the signal moves towards lower SNR, randomness increases in the inactive regions and hence detectability increases due to higher entropy. This improves compression ratio. We can also see that our scheme delivers higher compression compared to AMR-WB VAD. Packet deletions are less for both the schemes under varying SNR condition and the effect is minimal due to non-contiguous packet losses.

When users employ external speakers there would be annoying echo due to acoustic feedback. Another important use of VAD that was not mentioned here until now is that it can also be used for echo suppression, since echo cancelation is difficult to implement and use. It also consumes more computation at the terminals. The idea is to increase the microphone sensitivity only when the user is speaking and decrease the microphone sensitivity when user is listening. This cannot be done easily with energy based VAD, since the VAD is indirectly dependent on the microphone sensitivity. With entropy based VAD we will be reducing the dependence of VAD algorithm on the energy in the frame, and thus we can efficiently implement echo suppression. The computational complexity of our algorithm is $O(N/2\log(N))$ per frame. This is due to the DFT complexity for spectrum calculation. In fact DFT is the basic block in many of the codecs therefore we can reuse it, reducing overall system complexity.

V. Conclusions

We proposed a novel algorithm for VAD using spectral entropy based measure to find active and inactive zones in a speech stream. We compared our scheme with the VAD in relatively new AMR-WB [11] codec in terms of detectability of active and inactive zones, intelligibility, insertions, deletions and compression rate achieved. We have shown higher detectability of our approach compared with AMR-WB scheme. We have also shown better detectability of our approach under noisy conditions. The intelligibility of our approach in many cases has been rated better than the AMR-WB approach. We see better compression rate without any major loss in the subjective quality compared with the AMR-WB scheme. Our VAD scheme is largely invariant to speaker change and to some extent it will enhance the noisy speech stream due to the adoption of frequency domain filtering and adaptive threshold of spectral entropy. In our design, the overall delay is 40 ms from the time of packetization to delivery to the lower layer for transmission on the network. This low delay and the low complexity makes our approach feasible to be implemented on many embedded devices.

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 with the other VAD algorithms available in many of the standard codecs. We have considered only babble noise in this paper since which is the most common noise that can affect VoIP calls. However it will be interesting to see the effect of other types of noise such car noise, etc. Next logical step is to enhance our algorithm to be more effective under different conditions and test it in real environments, like [23].

Acknowledgments

Part of this research was carried out in the Adaptive Ad-Hoc Free Band Wireless Communications (AAF) project funded by the Freeband program of the Dutch Ministry of Economic Affairs. R. Venkatesha Prasad would like to thank EU-funded Magnet Beyond project. The work presented here does not necessarily reflect the views of Magnet Beyond and AAF.

References

[1] "Low-complexity coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss," International Telecommunication Union, Recommendation G.722, 2001.

[2] S. Andersen, A. Duric, H. Astrom, R. Hagen, W. Kleijn, and J. Linden, "Internet low bit rate codec (iLBC)," Internet Engineering Task Force, RFC 3951, 2001.

[3] "GSM-enhanced full rate specifications," European Telecommunications Standards Institute, Specification 06.51, 06.60-63 and 06.82, 2001.

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

[5] 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)*, vol. 2, Nov. 1989, pp. 1060–1064.

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

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

[8] 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*, vol. 5, July 2002, pp. 530–535.

[9] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A transport protocol for real-time applications," Internet Engineering Task Force, RFC 1889, 2001.

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

[11] "Adaptive multi-rate wideband speech transcoding," 3rd Generation Partnership Project, TS 26.194, 1993.

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

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

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

[15] 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*, vol. 2, May 1997, pp. 470–473.

[16] "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP)," International Telecommunication Union, Recommendation G.729, 1996.

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

[18] ——, "Cepstral speech/pause detectors," in *Proc. IEEE Workshop on Nonlinear Signal and Image Processing*, June 1995, pp. 388–391.

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

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

[21] 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.

[22] "One-way transmission time," International Telecommunication Union, Recommendation G.114, 1993.

[23] (2006) VQube internet telephony application. [Online]. Available: http://www.vqube.com/