LONG-TERM AUTO-CORRELATION STATISTICS BASED VOICE ACTIVITY DETECTION FOR STRONG NOISY SPEECH

Wei Shi[12], *Yuexian Zou*[2*], *Yi Liu*[1]
[1]Shenzhen Key Laboratory of Intelligent Media and Speech, PKU-HKUST Shenzhen-HongKong Institution, China
[2]ADSPLAB/ELIP, School of Electronic Computer Engineering, Peking University, Shenzhen, China
wei.shi@imsl.org.cn, *zouyx@pkusz.edu.cn, yi.liu@imsl.org.cn

ABSTRACT

*This paper proposes a voice activity detection (VAD) algorithm based on a novel long-term metric. By assuming that the most significant difference between noisy speech and non-speech is the harmonicity of the noisy speech spectrum caused by human nature, the long-term auto-correlation statistics (LTACS) measure is designed to be shown as a powerful metric used in VAD. The LTACS measure is calculated among several successive frames around the concerned frame and it represents the significance of harmonics of the signal spectrum over a long term rather than a short term. A novel LTACS-based VAD algorithm is derived by jointly making use of the minimum operator to reduce non-speech variability and of then calculating variance to detect speech. Simulative comparisons with four standardized VAD algorithms (ETSI adaptive multi-rate option 1 and 2, ETSI advanced front-end and G.729 Annex B) as well as three former proposed VAD algorithms show that the proposed LTACS-based VAD algorithm achieves the best performance under all SNR conditions, especially in strong noisy environments (e.g., SNR is -5dB or -10dB).

Index Terms— long-term auto-correlation statistics, voice activity detection, strong noisy speech

1. INTRODUCTION

Voice activity detection (VAD), or speech endpoint detection, refers to the task of discriminating speech segments from non-speech segments. VAD is an important frontend in many speech-related applications, such as mobile communication system [1], echo cancellation [2], speech enhancement [3], speech coding [4], automatic speech recognition [5], etc. The accuracy of VAD is quite critical to the overall performance of those applications. Researchers have proposed a variety of features in time, frequency or other transform domains to detect speech segments in noisy signals. Typical VAD algorithms are mostly based on short-term features like energy and zero crossing rate [6], auto-correlation [7], speech cepstrum coefficients [8], spectrum entropy or negentropy [9], speech periodicity characters [10], which are developed by taking the advantages of speech short-time stationarity and vocal modeling. Those feature-based methods can achieve relatively good performance under certain circumstances, but the performance will decline rapidly as the SNR of the noisy speech decreases. Moreover, Sohn *et al* proposed a statistical mode based VAD [11] in late 1990s, which showed a significant improvement comparing with feature-based VADs. In this approach, VAD problem is formed as a likelihood ratio test (LRT) with statistical models of speech and noise. Different assumptions about the statistics of noise were proposed to improve the robustness of VAD [12]-[14]. However, since those assumptions do not always hold in practice, this kind of approach cannot work well in low SNR conditions as the speech is strongly polluted by noise.

All of the methods mentioned above can be summarized as short-term VAD algorithms, as the VAD decisions are made at each frame. Multiple features or complex decision rules or hangover post-processing are often essential to enhance the accuracy and robustness. Compared to the short-term frame-level based VAD method, Ramirez *et al* [15] proposed the long-term spectral divergence (LTSD) as the discriminative metric for VAD. LTSD is calculated and smoothed over a long analysis window, consisting of several successive frames. The VAD decision is assigned to the frame in the middle of the analysis window. The idea of calculating measurement through long analysis window inspires many follow-up researches. Experiments show that the long-term based VAD is more robust than its short-term based counterpart, especially in strong noisy environments, such as when the global SNR is lower than 0dB [16]-[21].

In [15], Ramirez *et al* assumed that the most significant information for detecting voice activity in a noisy speech signal remains on the time-varying signal spectrum magnitude. However, this assumption is not true in non-stationary cases. In [16] the calculation of LTSD requires the estimation of noise magnitude spectrum, which may be difficult to carry out in practice. In this paper, we propose that the most significant difference between noisy speech and non-speech is the harmonic structure information of

*This work is partially supported by National Natural Science Foundation of China (No: 61271309) and the Shenzhen Science & Technology Fundamental Research Program (No: JC201105170727A)

noisy speech spectrum. It is caused by the nature of human vocal cords and vocal tract. Typically it appears significantly in the spectrum of voiced speech, and can be observed even if noise presents. The normalized auto-correlation coefficient (NACC) is proposed to represent the significance of the harmonicity in the noisy speech spectrum. NACC is calculated for each frame and minimized over a few successive frames. The variance of minimized NACC is obtained and smoothed in another range of consecutive frames. The final highly discriminative measurement is named as long-term auto-correlation statistics (LTACS) and proposed to be used in VAD decision rule. This method is evaluated intensively in the context of the TIMIT test corpus [22] in different noisy conditions with SNR ranging from 10dB to -10dB. Referenced VADs include Sohn's algorithm (LRT-VAD) [11], Ramirez's method (LTSD-VAD) [16], Ma's long-term spectral flatness measure based VAD (LFSM-VAD) [21] as well as four standard VADs [23]-[25].

2. LONG-TERM AUTO-CORRELATION STATISTICS MEASURE

The proposed method is relied on the assumption that the most significant difference between noisy speech and non-speech is that only the spectrum noisy speech has harmonic structures. This is true in most cases except when music signal or multi-tone signal, like dual tone multi frequency, is presented. The proposed VAD is based on the LTACS measure, which is assigned to each frame but calculated over a long window of frames. The calculation of LTACS is described as follows:

Step 1. Decompose the observed signal $x(t)$ into overlapped frames and calculate the auto-correlation of each frame. The NACC of the windowed *l-th* frame signal $a(l,t)$ is expressed as

$$r_a(l,\tau) = r_a(l,-\tau) = \frac{\sum_{t=0}^{N_w-\tau} a(l,t)a(l,t+\tau)}{\sum_{t=0}^{N_w} a^2(l,t)} \quad (1)$$

where N_w is the window length and τ is the lag. The windowed signal is achieved by

$$a(l,t) = \big(x(t-T(l)) - \mu(l)\big)w(t) \quad (2)$$

where $T(l) = (l-1)N_{sh}$ and $\mu(l)$ is the start time and mean of the l-th frame, respectively. N_{sh} is the frame shift. $w(t)$ is the window function and in this paper it is chosen to be a *Hann window*, which is given by

$$w(t) = 0.5 - 0.5\cos(2\pi t / N_w) \quad (3)$$

A correction is made to compensate the low-frequency suppression in the lag domain due to the use of window function [26]. The estimation of the auto-correlation $r_x(\tau)$ of the original signal segment can be expressed as

$$r_x(l,\tau) \approx \frac{r_a(l,\tau)}{r_w(\tau)} \quad (4)$$

where $r_w(\tau)$ is the NACC of a *Hann window*. According to (3) it is obtained by

$$r_w(\tau) = \left(1 - \frac{|\tau|}{N_w}\right)\left(\frac{2}{3} + \frac{1}{3}\cos\frac{2\pi\tau}{N_w}\right) + \frac{1}{2\pi}\sin\frac{2\pi|\tau|}{N_w} \quad (5)$$

In order to improve robustness, the first and last η % coefficients are removed from the estimated auto-correlation, which gives

$$r_x'(l,\tau) = r_x\left(l, \tau > N_w * \eta\% \text{ and } \tau < N_w * (1-\eta\%)\right) \quad (6)$$

Those coefficients are removed because they are less reliable and may bring negative influence to the discriminative ability of LTACS.

Step 2. Compute the variance of the smoothed auto-correlation. The auto-correlation is smoothed by minimizing the coefficients at each lag over a long window of frames, which gives

$$M(l,\tau) = \min\left\{r_x'(l+j,\tau)\right\}_{j=-R1}^{j=+R2} \quad (7)$$

where $\tau = 1,2,\cdots K, K = N_w - 2\eta\%$, $R1$ and $R2$ are the number of frames before and after the l-th frame. From the long-term perspective, different methods are proposed for smoothing. Ramirez smoothed the magnitude spectrum by maximization [15]. Ma smoothed the power spectrum by obtaining arithmetic and geometric mean [21]. Ghosh simply computed the entropy of power spectrum over a long window of frames [20]. In this paper, smoothing is done by minimization. This is based on the observation that in non-speech segments the coefficients' magnitude at the same lag may vary a lot, so minimization of the auto-correlation can minimize its variance. While in speech regions, the coefficients of successive frames at the same lag that correspond to the same harmonic may have small differences, thus the variance of auto-correlation can remain large. Obviously $R1$ and $R2$ cannot be set to too large as harmonics may change significantly during a relatively long period. After minimizing the auto-correlation, the variance is given by

$$\xi(l) = \frac{1}{K}\sum_{\tau=1}^{K}\Big(M(l,\tau) - \overline{M(l,\tau)}\Big)^2 \quad (8)$$

where $\overline{M(l,\tau)} = \frac{1}{K}\sum_{\tau=1}^{K} M(l,\tau)$.

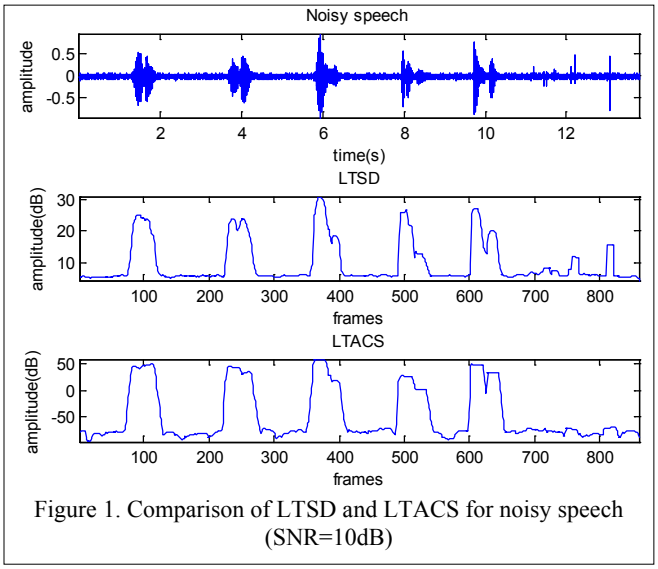

Figure 1. Comparison of LTSD and LTACS for noisy speech (SNR=10dB)

Step 3. Determine the LTACS measure. As mentioned above, $R1$ and $R2$ should not be set to very large values, $\xi(l)$ may be small in short pauses between voiced segments. To overcome this problem, $\xi(l)$ is smoothed over a larger analysis window, and the proposed LTACS measure is given by

$$\mathcal{L}(l) = 10\log_{10}\left(\frac{1}{R3+R4+1}\sum_{n=l-R3}^{l+R4}\left(\xi(n)-\overline{\xi(n)}\right)^2\right) \quad (9)$$

where $\overline{\xi(n)} = \frac{1}{R3+R4+1}\sum_{n=l-R3}^{l+R4}\xi(n)$.

The LTACS measure is developed to represent the significance of the harmonic structure of the signal spectrum in the auto-correlation domain. Several signal processing tricks are introduced to maximize the discriminative capability of LTACS, including compensating NACC in the lag domain, removing unreliable NACC, using a minimum operator to reduce non-speech variability, and calculating local variance to detect speech. An intuitive demonstration for the discriminative ability of LTACS is shown in Fig. 1. In Fig. 1, the speech is consisted of five short utterances and at the end of the recording some electronic and mechanical noise is introduced due to pushing buttons of the recorder to end it. White noise is added to the original signal with a 10dB global SNR. The LTACS and LTSD curves of the noisy speech are shown. From Fig. 1 we can see that both LTSD and LTACS have high discriminative ability to distinguish speech from noise. Two major advantages of LTACS over LTSD are concluded: 1) LTACS is less influenced by the energy variability in an utterance, which makes it a more reliable measurement to detect unvoiced speech; 2) LTACS is less sensitive to electronic and mechanical noise than LTSD. It is noted that the LTSD curve is smoother than LTACS in pure noise regions. This is because that the additive noise is stationary and the LTSD utilizes the average noise spectrum magnitude information, which is estimated from the initial part of the noisy signal and updated in each non-speech segment.

3. THE DECISION RULE OF THE LTACS-BASED VOICE ACTIVITY DETECTION

The VAD decision for the l-th frame is made by comparing $\mathcal{L}(l)$ and an adaptive threshold $\lambda(l)$. To update $\lambda(l)$, the last 100 realizations of $\mathcal{L}(l)$ of frames marked as speech and non-speech are stored in buffers $\chi_{S+N}(l)$ and $\chi_N(l)$, respectively. The subscript "S" refers to speech, while "N" refers to noise. With observations of $\chi_{S+N}(l)$ and $\chi_N(l)$, $\lambda(l)$ is updated by the following expression:

$$\lambda(l) = \alpha\min\{\chi_{S+N}(l)\} + (1-\alpha)\max\{\chi_N(l)\} \quad (10)$$

where α is the convex combination parameter. Experiments on the TIMIT training set show that α=0.25 gives maximum accuracy.

To initialize the algorithm, the first 100 frames are assumed to be pure noise. Thus we obtain 100 realizations of $\mathcal{L}_N$. Let μ_N and ω_N denote the mean and maximum of the first 100 $\mathcal{L}_N$. Then λ is initialized by $\lambda = \mu_N + \beta(\omega_N - \mu_N)$, where β is selected to be 1.05.

4. EVALUATION

The TIMIT test corpus was used to evaluate the performance of the proposed LTACS-based VAD algorithm. The test section of TIMIT corpus is consisted of 1680 phonetically balanced sentences from eight different dialects. Each sentence has an average length of 3.09s, and on average more than 87.5% of each sentence is labeled as speech. This is not suitable for a VAD test. Thus 2-seconds long silence segments were padded before and after each utterance. White noise from the NOISEX-92 database was added to all 1680 padded sentences at 5 different SNR levels (10dB, 5dB, 0dB, -5dB and -10dB). As a result, the final test set consisted of 198.44 minutes of noisy signal, of which 38.1% was noisy speech. The input signal is windowed by *Hann window* with a frame of 20 msec long and of 50% overlap. The sample rate is 16kHz, so N_w was set to be 320 and N_{sh} 160. The LTACS measure $\mathcal{L}(l)$ for the l-th frame is computed using the previous $R1$+$R3$ and following $R2$+$R4$ frames as well as the l-th frame. Parameters $R1$, $R2$, $R3$, $R4$ and η were fixed for simplicity. They were empirically selected by means of grid search on the train set of TIMIT corpus. More specifically, when $R1$=$R2$=3, $R3$=$R4$=9 and η=8, the proposed method achieved the best overall performance in different noisy conditions.

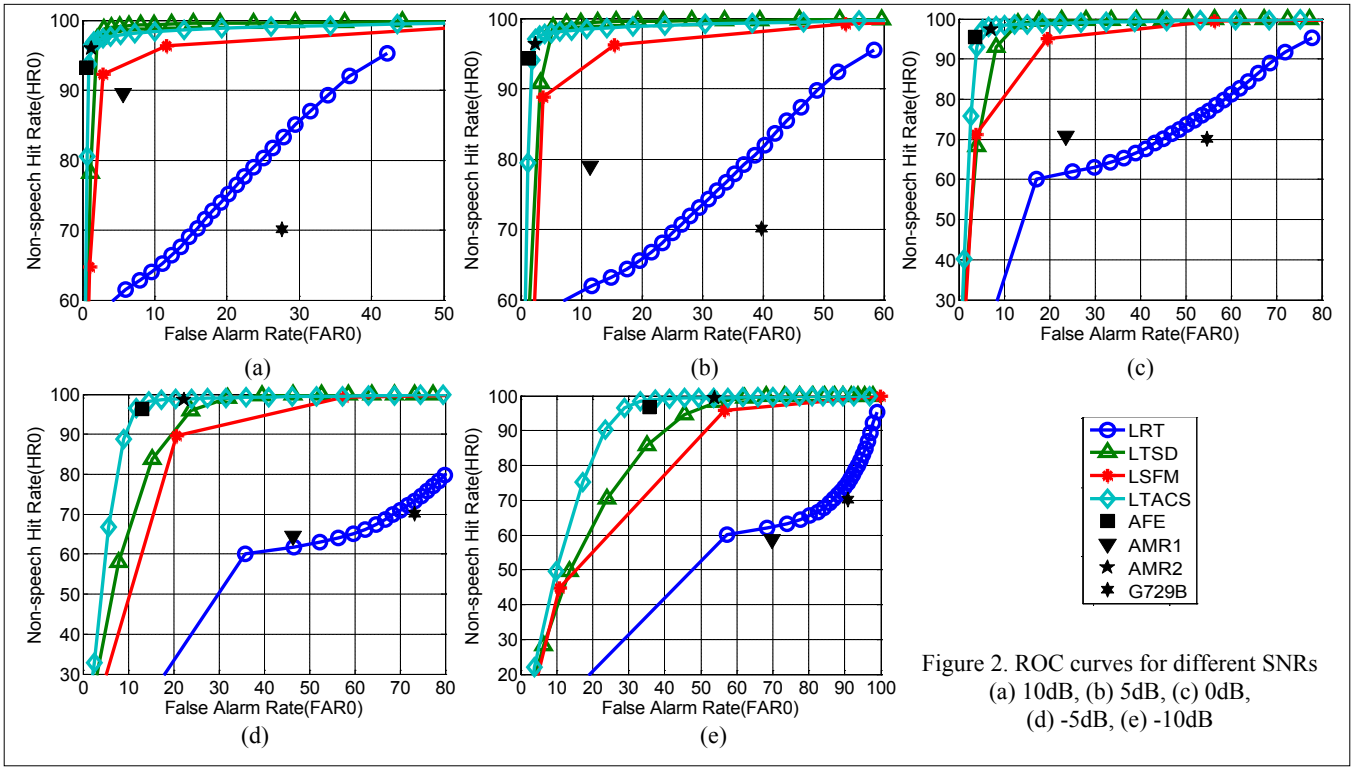

Figure 2. ROC curves for different SNRs (a) 10dB, (b) 5dB, (c) 0dB, (d) -5dB, (e) -10dB

Three recently reported VAD algorithms as well as four modern standardized VADs were compared with the proposed method. The implementation of LRT-VAD was extracted from VOICEBOX [27]. And the LTSD- and LSFM-VAD were implemented according to the original papers [16][21]. For standard VAD schemes, their implementations were taken from [28], [29] and [30]. The accuracy of each algorithm was calculated by comparing the detected results and the referenced results obtained from the TIMIT transcription, and then summarized over all 1680 silence-padded sentences.

The receiver operating characteristic (ROC) curve [31] was used to completely describe the VAD error rate. ROC curve was a plot of non-speech hit rate (HR0) and false alarm rate (FAR0) for varying the threshold λ. LTACS bigger than that threshold indicates the presence of speech. Fig. 2 shows the ROC curves of the LRT-, LTSD-, LSFM- and the proposed LTACS-VAD under different SNR levels. The working points for the AFE, AMR 1 & 2 and G729B VADs are also included. From Fig. 2 one can conclude that: 1) The AFE VAD and AMR VAD option 2 are well-designed schemes. They can achieve the best working points under all considered SNR conditions. 2) When SNR is larger than 0dB, the AMR VAD 2 is superior to AFE VAD, as it gives higher non-speech hit rate with nearly the same false alarm rate. 3) The proposed LTACS has the best discriminative power among long-term measures like LTSD and LSFM and model-based likelihood ratios. The proposed VAD based on LTACS yields the lowest false alarm rate for a fixed non-speech hit rate, and also the highest non-speech hit rate for a given false alarm rate. Especially in conditions when SNR is lower than 0dB (Fig. 2 (d) and (e)), the LTACS-VAD show significant improvement compared with the other three algorithms, which proves that the proposed method is robust and efficient for strong noisy speech. 4) Generally the LTACS-VAD can work as well as, if not better than the AFE and AMR2 VADs under all SNR conditions.

5. CONCLUSIONS

In this paper, we presented a novel long-term auto-correlation statistics (LTACS) based voice activity detection algorithm. The proposed method is intended to mitigate the performance decline suffered by most speech applications when the SNR is low. By exploiting the harmonic structure observed in noisy speech spectrum, the LTACS measure is constructed and its discriminative capability is maximized. An intensive experiment was carried out on a large amount of data (198.44 min). The comparison to three former reported VAD algorithms as well as four standard VADs showed that the proposed method could achieve the least VAD error rate. Moreover, the significant improvement under conditions when SNR was -5dB and -10dB proved that the LTACS-VAD is robust and suitable for strong noisy speech.

REFERENCES

[1] Freeman D. K., Southcott C. B., Boyd I., and Cosier G., "A voice activity detector for pan-European digital cellular

mobile telephone service," *Proc. IEEE ICASS*, vol. 1, pp 369-372, 1989.

[2] O. Tanrikulu, B. Baykal, A. G. Constantinides, and J. A. Chambers, "Critically sampled sub-band acoustic echo cancellers based on IIR and FIR filter banks," *IEEE Trans. Signal Proccessing*, vol. 45, pp. 901–912, Apr. 1997.

[3] Malah, D., R. V. Cox, *et al*., "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," *Proc. IEEE ICASSP*, vol. 2, pp 789-792, 1999.

[4] Enqing D., Heming Z., and Yongli L., "Low bit and variable rate speech coding using local cosine transform," *Proc. IEEE TENCON*, vol 1, pp 423-426, 2002.

[5] Vlaj D., Kotnik B., Horvat B., and Kacic Z., "A Computationally Efficient Mel-Filter Bank VAD Algorithm for Distributed Speech Recognition Systems," *EURASIP Journal on Applied Signal Processing*, issue 4, pp 487-497, 2005.

[6] L. R. Rabiner, and B. H. Juang, *Fundaments of Speech Recognition*. Prentice Hall, 1993.

[7] Craciun A., and Gabrea M., "Correlation coefficient-based voice activity detector algorithm," *Canadian Conference on Electrical and Computer Engineering*, vol. 3, pp 1789-1792, 2-5 May 2004.

[8] J. A. Haigh, and J. S. Mason, "Robust Voice Activity Detection using Cepstral Features," *Proc. IEEE TENCON*, pp. 321–324, 1993.

[9] J. L. Shen, J. W. Hung, and L. S. Lee, "Robust entropy based endpoint detection for voice recognition in noisy environments," *Proc. ICSLP*, vol. 98, pp 232-235, 1996.

[10] R. Tucker, "Voice Activity Detection Using A Periodicity Measure," *Proc. Inst. Electr. Eng.*, vol. 139, pp. 377~380, 1992.

[11] Sohn J., Kim N. S., and Sung W., "A statistical model-based voice activity detection," *IEEE Signal Proc. letters*, vol. 6, no. 1, pp 1-3, Jan 1999.

[12] Cho Y. D., and Kondoz A., "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Proc. Letters*, vol 8, no. 10, pp 276-278, Oct 2001.

[13] Chang J. H., and Kim N. S., "Voice activity detection based on complex Laplacian model," *IEE Electronics letters*, vol. 39, no. 7, pp 632-634, April 2003.

[14] Davis A., Nordholm S., and Togneri R., "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 14, no. 2, pp 412-424, March 2006.

[15] Ramırez, J., J. C. Segura, *et al*., "A new adaptive long-term spectral estimation voice activity detector," *Proc. of EUROSPEECH*, pp 3041-3044, 2003.

[16] Ramirez, J., J. C. Segura, *et al*., "Voice activity detection with noise reduction and long-term spectral divergence estimation," *Proc. IEEE ICASSP*, vol Ii, pp 1093-1096, 2004.

[17] Ramirez, J., J. C. Segura, *et al*., "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, issue 3-4, pp 271-287, 2004.

[18] Fukuda, T., O. Ichikawa, *et al*., "Phone-duration-dependent Long-term Dynamic Features for a Stochastic Model-based Voice Activity Detection," *9th Annual Conference of the International Speech Communication Association*, vol. 1-5, pp 1293-1296, 2008.

[19] Fukuda, T., O. Ichikawa, *et al*., "Long-Term Spectro-Temporal and Static Harmonic Features for Voice Activity Detection," IEEE Journal of Selected Topics in Signal Processing, vol. 4, issue 5, pp 834-844, 2010.

[20] Ghosh, P. K., A. Tsiartas, *et al*., "Robust voice activity detection using long-term signal variability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, issue 3, pp 600-613, 2011.

[21] Ma, Y. and A. Nishihara, "Efficient voice activity detection algorithm using long-term spectral flatness measure," *EURASIP Journal on Audio, Speech, and Music Processing* vol. 2013, issue 1, pp 1-18, 2013.

[22] DARPA-TIMIT, Acoustic-Phonetic Continuous Speech Corpus, NIST Speech Disc 1-1.1,1990.

[23] ITU, "A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70," *ITU-T Recommendation G.729-Annex B*, 1996.

[24] ETSI, "Voice activity detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels," *ETSI EN 301 708 Recommendation*, 1999.

[25] ETSI, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," *ETSI ES 201 108 Recommendation*, 2002.

[26] Boersma, P., "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," *Proceedings of the institute of phonetic sciences*, Amsterdam, vol. 17, pp 97-110, 1993.

[27] Mike Brookes, "VOICEBOX: Speech Processing Toolbox for MATLAB".

[28] 3GPP, "ANSI-C code for the fixed-point distributed speech recognition extended advanced front-end," *3GPP TS 26.243 Recommendation*, 2004

[29] 3GPP, "ANSI-C code for the floating-point Adaptive Multi-Rate (AMR) speech codec," *3GPP TS 26.073 Recommendation*, 1999

[30] ITU, "Coding of Speech at 8 kbit/s using Conjugate Structure Algebraic Code – Excited Linear Prediction. Annex I: Reference Fixed-Point Implementation for Integrating G.729 CS-ACELP Speech Coding Main Body With Annexes B, D and E," *Int. Telecommun. Union*, 2000.

[31] Green D.M. and Swets J.M., *Signal detection theory and psychophysics*, John Wiley and Sons Inc., New York, 1966.