# Efficient voice activity detection algorithm based on sub-band temporal envelope and sub-band long-term signal variability

*Bin Liu* [1], *Jianhua Tao* [1], *Fuyuan Mo*[2], *Ya Li* [1], *Zhengqi Wen* [1] , *Shanfeng Liu* [1]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190；

[2] Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190）

liubin@nlpr.ia.ac.cn, jhtao@nlpr.ia.ac.cn, mofuyuan@aliyun.com,
yli@nlpr.ia.ac.cn, zqwen@nlpr.ia.ac.cn, sfliu@nlpr.ia.ac.cn

## Abstract

Voice activity detection (VAD) is widely used for various speech-based systems which is an important pre-processing step. This paper proposes a robust voice activity detection algorithm. In the proposed algorithm, the sub-band temporal envelope and the sub-band long-term signal variability are considered to distinguish the speech from all kinds of non-speech which include stationary noise and non-stationary noise. The two features are combined to make a robust VAD decision according to the fusion decision. The proposed algorithm also is an unsupervised low-complexity algorithm and can operate without pre-train models. The experiments results show that the proposed algorithm is prior to the different baseline algorithms and can handle a variety of noise environments over a wide range of signal-to-noise ratios. The proposed algorithm could apply to speech-based systems.

**Index Terms**: voice activity detection, sub-band temporal envelope, sub-band long-term signal variability, fusion decision

## 1. Introduction

Voice activity detection (VAD) is a significant technology for distinguishes speech from non-speech in audio streams automatically. VAD is widely used for various speech-based systems such as speech enhancement, language identification, speaker recognition, speech coding and automatic speech recognition. Accurate VAD greatly reduces error rates and overall computation time for speech recognition [1] and speaker recognition [2]. Noise robust speech detection in the audio signals is an important pre-processing step because it can significantly improve performance. Practical VAD must be able to accurately and robustly detect speech periods and non-speech periods from observed signals in real environments in which complex background noise are existed. However, developing a VAD for noisy environments with low signal-to-noise ratios (SNR) or for any non-stationary noise is still very challenging.

Many methods have been proposed for speech detection. Early VAD approaches were based on simple energy thresholds, pitch and zero-crossing rate rules [3]. More recent approaches consider more advanced parameters such as Mel frequency cepstral coefficients (MFCC), wavelet-based features [4], correlation coefficients [5] and spectrogram entropy [6, 7]. These approaches perform well in where there is little or no background noise. However, the performance will degrade seriously when the SNR decreases [8]. To solve this problem, other VAD have been developed and require noise estimation. The posteriori SNR and the priori SNR are calculated in [9]. Noise estimation and adaptation techniques were considered to improve its robustness under non-stationary noise environments but it has high computational complexity [10]. The most promising approaches are data-driven methods, which a classifier is trained to predict speech and non-speech according to acoustic features [11] such as support vector machines (SVM) [12], Gaussian mixture models (GMM) [13], artificial neural network (ANN) [14] , hidden markov models (HMM) [15] and conditional random fields (CRF) [16]. However, the performance of these techniques may be much lower when the acoustic characteristics of the real environment mismatch the training data. The recent studies show that the long-span features could effectively improves the robustness in complex noise environment because the decision for each frame can be performed in the context of the adjacent frames [17]. In addition, speech signal could be detected according to the temporal envelopes within different frequency bands [18]. In recently, there has been interest in developing the VAD systems which can operate without training data and robust in a variety of noise environments over a wide range of signal-to-noise ratios. Earlier VAD systems, such as G.729B [19] and AMR [20], followed a rule-based approach and required no training data. These speech coding standards specifies a VAD which is often used in VAD performance evaluations.

In this study, we propose a robust and practical VAD system in real environments even where strong stationary or non-stationary noise exists. We analyze the temporal envelope for different frequency bands and compute their statistical characteristics. We also consider the long time information and the long-term signal variability feature will be extracted for the particular frequency bands which reflect the formant characteristic. The VAD's decision is made after multiple observations. Sub-band temporal envelope feature and sub-band long-term signal variability will be combined to detect the speech in different environments. The proposed VAD does not need training data and any pre-trained models.

The remainder of this paper is structured as follows. Section 2 introduces the proposed VAD algorithm. The performance comparisons are demonstrated and evaluated in section 3. The conclusions and future work are shown in section 4.

## 2. Proposed VAD algorithm

In this section, we firstly introduce the framework of the proposed VAD system. Subsequently, the detail is presented.

978-1-4799-4219-0/14/$31.00 ©2014 IEEE

There are total five parts in the proposed algorithm which includes preprocessing, sub-band temporal envelope analysis, sub-band long-term signal variability analysis, fusion decision and post-processing. In the preprocessing stage, different sub-bands signal are extracted. The statistical assessment of the sub-band temporal envelope and the sub-band long-term signal variability analysis for each frame will be implemented after preprocessing respectively. The two parameters mentioned above are fused to make the decision. The length of speech segment is extended in the post-processing stage. The flowchart of proposed algorithm is shown in Fig. 1. The envelope 1 represents the temporal envelope in the second sub-band. The envelope 2 represents the temporal envelope in the fifth sub-band. The entropy includes the second sub-band and the third sub-band.

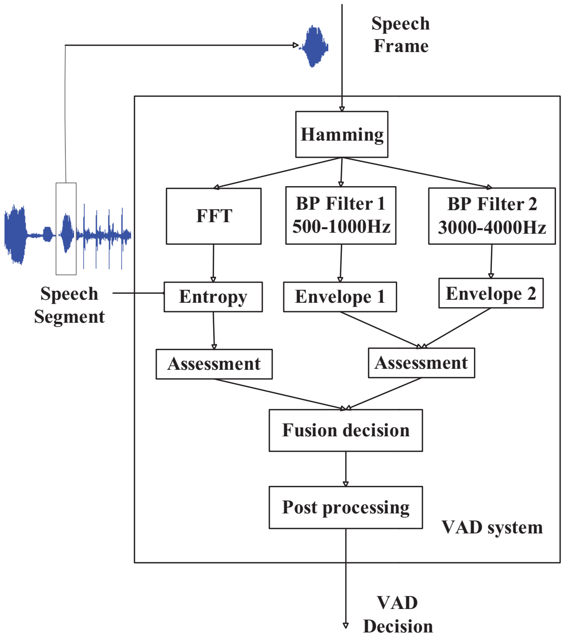

Figure 1: the flowchart of proposed algorithm

## 2.1. Preprocessing

The entire signal is first divided into frames with a Hamming window. The length of each frame is 32 ms where the sampling of the speech signal is 8 kHz.

To extract the sub-band temporal envelope feature, the speech signal analysis begins by filtering the input speech signal into five frequency bands. These filters are 6th order Butterworth, with pass bands of 0-500, 500-1000, 1000-2000, 2000-3000, and 3000-4000 Hz according to the MELP standard [21]. The FFT is implemented and particular frequency bands reflected formant characteristics are selected in order to measure the long-term signal variability.

The sub-band below 2000Hz could represent obvious formant characteristics even in the low SNR environment for the voiced speech which is significant for speech intelligibility. However, the first sub-band possible overlaps some non-stationary noise such as the volvo noise and the gun machine noise. The sub-band above 3000Hz includes abundant noise composition and the energy of speech composition in this frequency band is less than the low frequency bands. Therefore, the proposed algorithm considers three sub-bands, namely the second sub-band (500-1000Hz), the third sub-band (1000-2000Hz) and the fifth sub-band (3000-4000Hz).

## 2.2. Sub-band temporal envelope analysis

The temporal envelope will be analyzed in the second sub-band and the fifth sub-band. For each frame, the temporal envelope of the relevant frequency band is extracted using the Hilbert Transform. Envelope extraction using the Hilbert transform involves the calculation of the analytic signal [22], as illustrated in Eq.1 and Eq.2.

$$E_i(t)=\sqrt{x_i(t)^2+\tilde{x}_i(t)^2} \tag{1}$$

$$\tilde{x}_i(t)=x_i(t)*\frac{1}{\pi t} \tag{2}$$

where $E_i(t)$ is the Hilbert envelope of $x_i(t)$, $i$ represents *i-th* sub-band signal and $\tilde{x}_i(t)$ represents the Hilbert Transform of $x_i(t)$.

The Inter-Quartile Range (IQR) which is shown in Eq. 3 is calculated within the temporal envelopes of the second frequency band by using the difference between the third quartile (the value below which 75% of the values in the distribution, Q3) and the first quartile (the value above which 25% of the values in the distribution, Q1) [18].

$$IQR = Q3 - Q1 \tag{3}$$

The speech signal could be detected according to the sub-band temporal envelope feature. We define $VAD_{envelope}$ as the detection result for each frame and then this parameter could be shown in Eq. 4.

$$VAD_{envelope}=\begin{cases}\log_{10}(IQR-(\mu_2-\mu_5)) & IQR-(\mu_2-\mu_5)>1\\ 0 & IQR-(\mu_2-\mu_5)<=1\end{cases} \tag{4}$$

where $\mu_2$ and $\mu_5$ represent the mean of temporal envelope for the second frequency band and the fifth frequency band respectively.

Fig. 2 illustrates an example of the $VAD_{envelope}$ for one segment speech signal produced by a male speaker. The pink noise and the gun noise are added with 0 dB SNR respectively. As shown in Fig. 2, the $VAD_{envelope}$ of the speech is higher than that of the gun noise which is one of a typical non-stationary noise even in low SNR environment. However, it is not efficient to detect the speech signal from the pink noise.

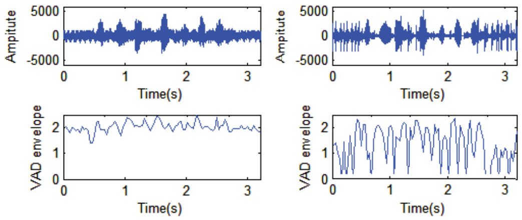

Figure 2: the example of sub-band temporal envelope analysis

## 2.3. Sub-band long-term signal variability analysis

The long-term signal variability measure is calculated to make use of the adjacent frames which may be relevant to the current frame. Both the second sub-band and the third sub-band are considered to analyze the long-term signal variability in the proposed algorithm. We define $S_x(n,\omega_k)$ as the short time amplitude spectrum at $\omega_k$. The entropy on the normalized short-time spectrum is computed at each frequency point. We define the entropy at *k-th* frequency point for the *m-th* frame as [17]

$$\varepsilon_k^x(m) = -\sum_{n=m-R/2}^{m+R/2} \frac{S_x(n,\omega_k)}{\sum\limits_{l=m-R/2}^{m+R/2} S_x(l,\omega_k)} * \log\left(\frac{S_x(n,\omega_k)}{\sum\limits_{l=m-R/2}^{m+R/2} S_x(l,\omega_k)}\right) \quad (5)$$

where $\varepsilon_k^x(m)$ represents the entropy which is calculated over $R$ consecutive frames. The $R$ is set to six because it is difficult to apply for real communication system while the delay is too long.

The mean and variance of $\varepsilon_k^x(m)$ for the frequency points within the second sub-band and the third sub-band are computed.

$$\overline{\varepsilon_k^x(m)} = \frac{1}{f_{end} - f_{begin}} \sum_{k=f_{begin}}^{f_{end}} \varepsilon_k^x(m) \quad (6)$$

$$\zeta_x(m) = \frac{1}{f_{end} - f_{begin}} \sum_{k=f_{begin}}^{f_{end}} (\varepsilon_k^x(m) - \overline{\varepsilon_k^x(m)})^2 \quad (7)$$

where $\overline{\varepsilon_k^x(m)}$ and $\zeta_x(m)$ represent the mean and variance for the frequency points within the second sub-band and the third sub-band; $f_{begin}$ represents the start frequency point in the second sub-band and $f_{end}$ represents the last frequency point in the third sub-band. The speech signal could be detected according to $\zeta_x(m)$.

Fig. 3 illustrates an example of the $\zeta_x(m)$ for the same segment speech signal analyzed in Fig. 2. As shown in Fig. 3, the $\zeta_x(m)$ of the speech is higher than that of the pink noise which is one of a typical stationary noise even in low SNR environment; however, it is not robust to distinguish the speech signal from the gun noise.

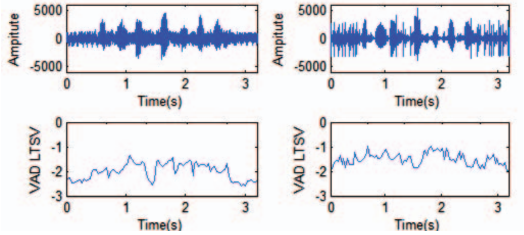

Figure 3: the example of sub-band long-term signal variability analysis

### 2.4. Fusion Decision

The $VAD_{envelope}$ and the $\zeta_x(m)$ are combine to make a more robust VAD decision in a variety of noise environment. The fusion decision is defined as $VAD_{decision}$, then

$$VAD_{decision} = \alpha * VAD_{envelope} + (1-\alpha)\log_{10}(\zeta_x(m)) \quad (8)$$

where $\alpha$ control the trade-off between the sub-band temporal envelope and the sub-band long-term signal variability.

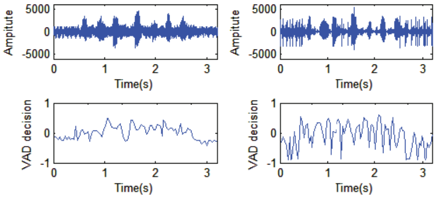

Figure 4: the example of fusion decision

Fig. 4 illustrates an example of the fusion decision for the same segment speech signal analyzed in Fig. 2 (the $\alpha$ is set to 0.5). As is shown in Fig. 4, the $VAD_{decision}$ of speech is higher than that of the noise signal even in low SNR environment for both the gun noise and the pink noise.

### 2.5. Post-processing

The fusion decision in proposed algorithm can effectively detect voiced frame. However, it may perform poorly in detecting short unvoiced frames surrounding a voiced segment because the unvoiced signal is similar to random white noise. To solve this problem, the boundaries of each detected speech segment are extended by 96ms (three frames).

## 3. Experiments and results evaluation

### 3.1. Data and Analysis Methodology

We conducted a series of experiments to evaluate the proposed VAD. The speech data is taken from the TIMIT corpus [23]. We selected 4977 sentences spoken by different speakers from eight different American dialect regions as experiments data. To evaluate the proposed method, we conduct VAD experiments under various noise environments with different SNRs; the noise data is taken from the NOISEX-92 corpus for various noise signals (babble, hf, tank, factory, car, buccaneer, gun, pink and white noise) [24]. We then generated a synthetic data set using speech from held out speakers in the TIMIT database, mixed with a variety of stationary and non-stationary noise samples from the NOISEX-92 database at three SNR levels (10 dB, 5 dB and 0 dB). All the examples are 8 kHz sampling rate and 16 bits PCM quantization.

To optimize the algorithm parameter, the 3000 sentences are selected randomly from the 4977 sentences and mixed with five noise signal (car, buccaneer, gun, pink and white noise) at three SNR levels (10dB, 5dB and 0dB). The rest of data which mixed with four noise signal (babble, hf, tank and factory noise) at three SNR levels (10dB, 5dB and 0dB) is applied to evaluate the proposed algorithm.

We compare the proposed method to two existing methods [17, 18]. Both are natural candidates for comparison to our method because they don't require training data from the user and is an unsupervised algorithm. In [17], the temporal envelope is selected as the features of VAD. In [18], the long-term signal variability is applied to VAD.

The trade-off between True Positive Rate (TPR) and False Positive Rate (FPR) is a key concern in designing robust VAD. For the different threshold, we compare TPR and FPR. To obtain Receiver Operating Characteristics (ROC) curves, we vary the decision threshold for the proposed method and the baseline methods. To plot the ROC curve, we disabled adaptive threshold schemes in the baseline method [18]. Fig. 5 shows the ROC curve between TPR and FPR for the different noisy speech and different SNR levels.

### 3.2. Parameter Determination

In this section, we describe the experiments we performed to choose the optimal parameter for the proposed algorithm. To select the optimal parameter, we will search over a range of parameters. To calculate the $VAD_{envelope}$, we analyzed the assessment of temporal envelope for different sub-bands. The second sub-band is selected after searching from the first sub-band to the third sub-band (their combination is also considered). The fifth sub-band is confirmed after searching the fourth sub-band, the fifth sub-band and their combination. We also optimize the R and R = [2, 4, 6] is considered (the delay is not long due to it is difficult to apply for the real communication system). All the possible sub-band combinations are considered to optimize the $\zeta_x(m)$ and the $\overline{\varepsilon_k^x(m)}$. We optimize the $\alpha$ ranging from 0.3 to 0.7 (the step is 0.05). The $\alpha$ is set to 0.5.

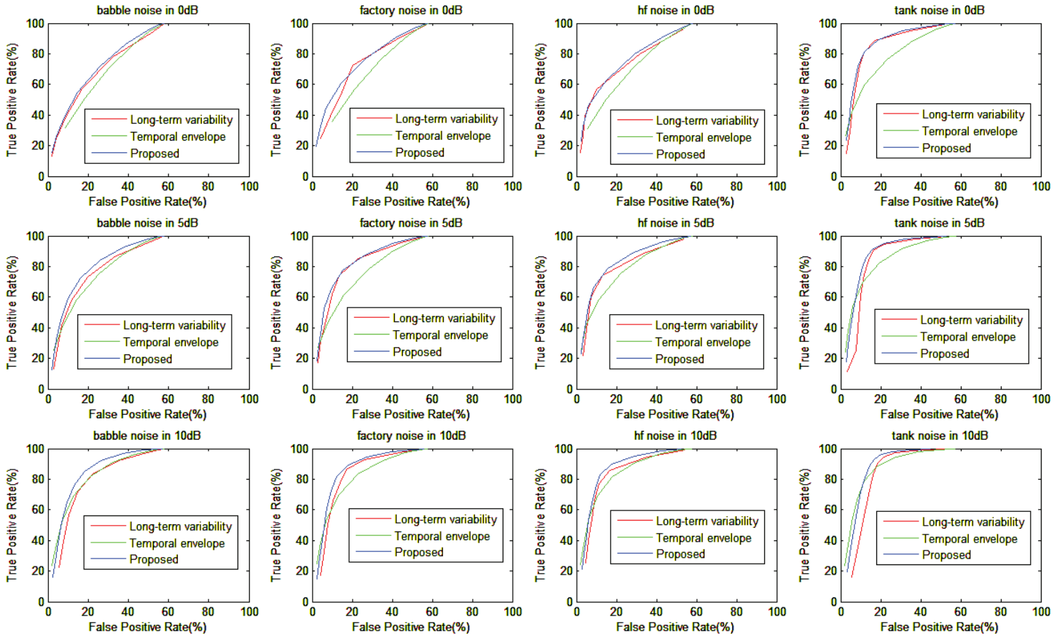

Figure 5: the ROC curve for different VAD algorithms

### 3.3. Results and discussion

We evaluate the TPR firstly. We find that the proposed VAD has higher TPR as compared to both the baseline algorithms. And the proposed algorithm has lower FPR as compared to both the baseline algorithms. The areas of ROC are larger for the proposed algorithm. The proposed algorithm has advantage especially for speech-like sound such as babble noise. The ROC curves indicate that proposed algorithm significantly outperforms the baseline algorithms in a variety of noise environments.

As shown in this subsection, the proposed method outperforms both the baselines. The proposed algorithm analyzes the sub-band which could represent obvious formant characteristics and has the advantages of both the sub-band temporal envelope and the sub-band long-term signal variability through proper fusion decision. Therefore, it is robust to detect the speech signal from all kinds of stationary noise and non-stationary noise even in low SNR. These results confirmed that the VAD we propose was able to operate accurately and could apply to speech-based systems.

## 4. Conclusion and future work

This paper presented a noise robust voice activity detection approach. In the proposed method, the sub-band temporal envelope feature and the sub-band long-term signal variability are combined to distinguish the speech from all kinds of non-speech which include stationary noise and non-stationary noise. The proposed VAD was compared to two existing VAD algorithms using NOISEX-92 signals at various SNR and our experiments results show that the proposed approach significantly outperforms the baseline approaches. The results revealed that the proposed approach could accurately detect speech periods and non-speech periods in the low SNR conditions. In addition, it does not require training data and pre-trained models. It doesn't make assume that the beginning of the signal contains non-speech. The delay for proposed method is below 150ms and could meet the requirement of cellular communications and underwater acoustic communication. It is directly applicable to these systems.

In future, we plan to optimum the fusion decision; the noise environment could be detected automatically and the VAD decision would be adjusted adaptively through different environment. We would improve speech detection performance for unvoiced frame. In addition, we will extend our method into wideband speech signal.

## 5. Acknowledgements

This work is supported by the Major Program for the National Social Science Fund of China (13&ZD189), the National Natural Science Foundation of China (NSFC) (No.61273288, No.61233009, No.61203258, No.61305003, No.61332017 and No.61375027).

## 6. Reference

[1] T. Pfau, D. P. Ellis, and A. Stolcke, "Multispeaker Speech Activity Detection for the ICSI Meeting Recorder," in Proceedings of Automatic Speech Recognition and Understanding, Italy, pp. 107–110, 2001.

[2] D. A. Reynolds, "An overview of automatic speaker recognition technology," In Acoustics, Speech and Signal Processing (ICASSP), USA, pp. 4072– 4075, 2002.

[3] K. Woo, T. Yang, K. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," IET Electronics Letters, vol. 36, no. 2, pp. 180-181, 2000.

[4] Y. C. Lee, "Statistical model-based VAD algorithm with wavelet transform," IEICE Trans. Fundamentals, vol. 89, no. 6, pp. 1594–1600, 2006.

[5] A. Craciun and M. Gabrea, "Correlation coefficient-based voice activity detector algorithm," in Canadian Conference on

Electrical and Computer Engineering, Canada, pp. 1789–1792, 2004.

[6] P. Renevey and A.Drygajlo, “Entropy based voiced activity detection in very noisy conditions,” in Proc. EUROSPEECH, Denmark, pp. 1887–1890, 2001.

[7] R. Prasad, H. Saruwatari and K.Shikano, “Noise estimation using negentropy based voice-activity detector,” in 47th Midwest Symposium on Circuits and Systems, pp. 149–152, 2004.

[8] F. Beritelli, S. Casale, G. Ruggeri, and S. Serrano, “Performance evaluation and comparison of G.729/AMR/fuzzy voice activity detectors,” IEEE Signal Processing Letters, vol. 9, no. 3, pp. 85–88, 2002.

[9] J. Sohn, N. S. Kim, and W. Sung, “A Statistical Model-Based Voice Activity Detection,” IEEE Signal Processing Letters, vol. 6, no. 1, pp. 1998–2000, 1999.

[10] B. Lee and M. Hasegawa-Johnson, “Minimum mean squared error a posteriori estimation of variance vehicular noise,” in Proc. Biennial on DSP for In-Vehicle and Mobile Systems, 2007.

[11] A. Misra, “Speech/nonspeech segmentation in web videos,” in Proc. of INTERSPEECH, USA, pp.1977-1980, 2012.

[12] N. Mesgarani, M. Slaney, and S. A. Shamma, “Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations,” Audio, Speech, and Language Processing, IEEE Transactions on, vol. 14, no. 3, pp. 920–930, 2006.

[13] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, K. Vesely, P. Matejka, X. Zhu, and N. Mesgarani, “Developing a speech activity detection system for the DARPA RATS program,” in Proc. of INTERSPEECH, USA, pp.1969-1972, 2012.

[14] F.Eyben, F.Weninger, S.Squartini, and B.Schuller, “Real-life voice activity detection with LSTM recurrent neural networks and an application to hollywood movies,” In Acoustics, Speech and Signal Processing (ICASSP), Canada, pp. 483-487, 2013.

[15] Y. Liang, X. Liu, Y. Lou, and B. Shan, “An improved noise-robust voice activity detector based on hidden semi-Markov models,” Pattern Recognition Letters, vol.32, no.7, pp. 1044-1053, 2011.

[16] A. Saito, Y. Nankaku, A. Lee, and K. Tokuda, “Voice activity detection based on conditional random fields using multiple features,” in Proceedings of INTERSPEECH, Japan, pp. 2086–2089, 2010.

[17] N. Lezzoum, G. Gagnon and J. Voix, “A low-complexity voice activity detector for smart hearing protection of hyperacusic persons,” in Proc. of INTERSPEECH, France, pp.723-727, 2013.

[18] P. K. Ghosh, A. Tsiartas, and S. Narayanan, “Robust voice activity detection using long-term signal variability,” Audio, Speech, and Language Processing, IEEE Transactions on, vol. 19, no. 3, pp. 600-613, 2011.

[19] A. Benyassine, E. Shlomot, H. Y. Su, D.Massaloux, C. Lamblin, and J. P. Petit, “ITU-T Recommendation G. 729 Annex B: a silence compression scheme for use with G. 729 optimized for V. 70 digital simultaneous voice and data applications,” Communications Magazine, IEEE, vol.35, no.9, pp. 64-73, 1997

[20] E. Ekudden, R. Hagen, I. Johansson and J. Svedberg, “The adaptive multi-rate speech coder,” In Speech Coding Proceedings, pp. 117-119, 1999.

[21] L. M. Supplee, R. P. Cohn, J. S. Collura and A. V. McCree, “MELP: the new federal standard at 2400 bps,” In Acoustics, Speech and Signal Processing (ICASSP), Germany, 1591-1594, 1997.

[22] S. L. Marple, “Computing the Discrete-Time ”Analytic” Signal via FFT,” IEEE Transactions on Signal Processing, vol. 47, no. 9, pp. 2600–2603, 1999.

[23] J.S. Garofolo, “TIMIT: Acoustic-phonetic Continuous Speech Corpus,” Linguistic Data Consortium, 1993.

[24] Rice University, NOISEX-92 Database, [Online] Available: http://spib.rice.edu/spib/select noise.html.