Voice Activity Detection for Best Signal Selection in Air Traffic Management and Control Systems

Radu-Sebastian Marinescu and Corneliu Burileanu

Abstract–**The Best Signal Selection (BSS) in air traffic management and control systems has to decide among several signal instances of the same source which one offers the highest speech intelligibility. In these systems, the source signal is not available, thus, objective speech quality tests could not be used. However, information with regards to the speech quality could be obtained from the score of voice activity detection (VAD) algorithms. In this paper the correlation between speech quality and the score of VAD algorithms is analyzed. The results showed that the VAD score-based methods do not saturate for higher SNR, as the Perceptual Evaluation of Speech Quality (PESQ) does. A new VAD algorithm as a solution for the best signal selection problem is also proposed.**

Keywords— **Speech Intelligibility, Voice Activity Detection.**

I. Introduction

IN the last decades, the air traffic spread more and more in the world, connecting more and more places. At the same time, the need to manage all the flights correctly and securely increased. Air traffic authorities imposed and updated several standards for the air traffic management (ATM) system, keeping in pace with the growing traffic flow. To achieve this, special voice communication systems (VCS) were developed. They ensure the communication between the pilots and the operators from the ground control centers. When a communication is initiated between the aircraft's pilot and the ground air traffic control operator, various systems are used. The pilot speaks through the aircraft's radio station and the signal is received by several ground radio stations. Then, the signal from each ground radio station arrives on different paths to the control center. Here one of the received signals is played to the operator. Ideally, this should be the one which is the clearest and offers the highest intelligibility. This is the equivalent of the BSS for the received signals, which could be achieved through special signal processing algorithms.

The solution for BSS has to take into account various problematic aspects such as: a) the speech sequence lasts in average, for less than a few seconds; b) in general, two consecutive communication are initiated by different aircrafts; c) the selection has to be done relatively quick, in less than 300 ms, imposed by specific standard [1]; d) in the first part of 300ms of the received signal, the noise level undergoes significant variations due to the automatic gain control activation; e) for each reception the signals could be received with different delays in VoIP environments; f) on the same channel, two consecutively received signals may be significantly different in terms of speech intelligibility.

In order to activate properly the communication between the pilot and the operator, these systems must have a previous voice activity detection (VAD) block. Despite the fact that the main goal of VAD algorithms is to spot the speech segments, some of them could offer some information regarding the voice quality. Based on this observation and assuming that the signals are perfectly aligned by a previous processing block, it could be useful finding an adequate VAD algorithm for speech detection and quality estimation, for BSS issue. Thus, recent VAD algorithms were reviewed and selected for a possible BSS solution. Usually, they look after speech features of the signal and then assign a VAD score. Based on specifics of the algorithm and VAD score, a speech/non-speech decision is taken. Over the time, one or more different features of voice signals and techniques were included in VAD algorithms, such as energy or/and subband energy [2]-[4], entropy [5], correlation coefficients [6], wavelet transform [7][8], Walsh basis function representation [9], long-term speech information [10]-[12], periodic to aperiodic component ratio [13][14]. Also, in [15] it is shown that the image processing technique called "local binary pattern" could be used for VAD algorithms. Besides these, some statistical methods train voice and noise mixtures in labeling them in supervised or semi-supervised mode like in [16]-[19].

However, not all the above mentioned VAD algorithms could offer information with respect to the voice quality. Thus, only some of them could be used for BSS. Among these options, a new VAD algorithm is proposed in this paper. In Section II, the correlation between the score of voice activity detection algorithms is analyzed. The new VAD algorithm proper to solve the BSS is proposed in Section III. Further, in Section IV experimental values and discussion regarding the BSS solution are presented. Finally, conclusion and further work stand for Section V.

II. Correlation Between Speech Intelligibility and VAD Scores

The use of a VAD algorithm for speech detection does not imply a solution for BSS, because VAD is not focusing on speech quality and intelligibility. However, the score of some VAD algorithms could offer information for BSS. A quick verification of the correlation between the speech quality and VAD score for any algorithm could be made by analyzing the VAD score characteristics at different SNR. In Fig. 1 we

Manuscript received February 04, 2014.

R.S. Marinescu is with the Department of Research and Development, Rohde & Schwarz Topex, Bucharest, RO 014186 Romania, and with the Speech and Dialogue Research Laboratory, University Politehnica of Bucharest, Bucharest, RO 060042 Romania, (corresponding author, e-mail: radu-sebastian.marinescu@rohde-schwarz.com).

C. Burileanu is with Speech and Dialogue Research Laboratory, University Politehnica of Bucharest, Bucharest, RO 060042 Romania (e-mail: corneliu.burileanu@upb.ro).

present 5 noisy instances. These were created by corrupting the clean signal with Gaussian white noise at different SNR levels, from 5 to 25 dB. Because some of the above presented VAD algorithms could not offer a usable VAD score for our needs and make their decision based on several test conditions, we found that a good correlation between the SNR and VAD score was provided by two VAD algorithms. These are based on the wavelet transform [7] and on the long-term spectral flatness measure (LSFM) [12]. Their characteristics are shown in Fig. 2 and 3, with the following SNR relationship: green – 25 dB, magenta – 20 dB, cyan – 15 dB, red – 10 dB, blue – 5dB.

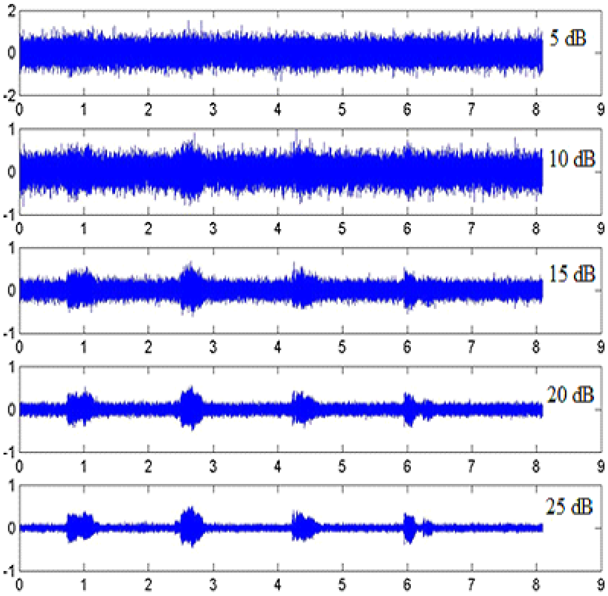

Figure 1 – Noisy signal waveforms at different SNR

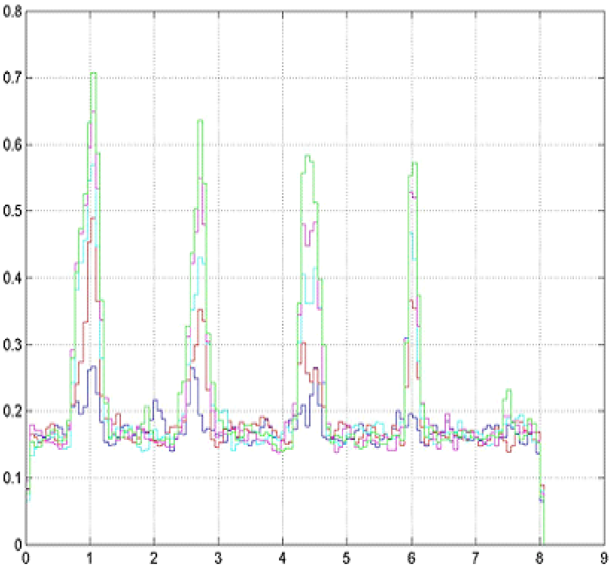

Figure 2 – Wavelet based VAD characteristics: green – 25 dB, magenta – 20 dB, cyan – 15 dB, red – 10 dB, blue – 5dB

In Fig. 2 we notice that signals corrupted with higher SNR have higher VAD scores. Thus, for wavelet-based VAD, the higher the VAD score, the higher speech intelligibility is assumed.

We can notice in Fig. 3, that, after the initialization process is finished, the LSFM features of each corrupted signal have similar values during the non-speech segments. On the other hand, on the speech periods, the LSFM values are smaller for the signals which were less corrupted by the additive noise. Thus, the fifth corrupted signal, with the highest SNR (25 dB), leads to the smallest LSFM values during the speech periods, while the first signal, which is the heaviest noise corrupted signal (5 dB), does not lead to a obvious decrease of the LSFM values. Therefore the LSFM feature could be associated with SNR.

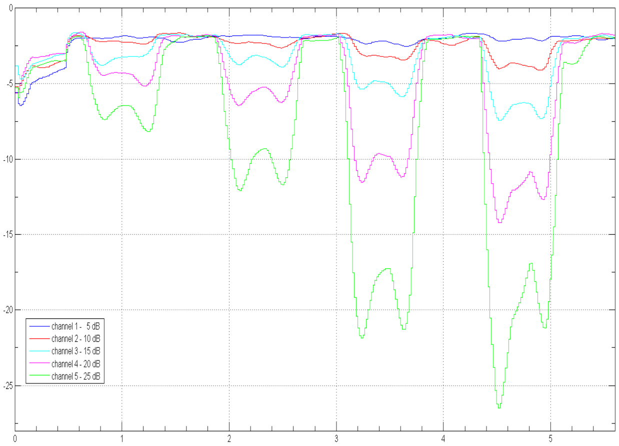

Figure 3 – LSFM based VAD characteristics - green – 25 dB, magenta – 20 dB, cyan – 15 dB, red – 10 dB, blue – 5dB

III. Proposed VAD and BSS Algorithm

In order to find a proper solution for the BSS problem, in this paper we proposed a new VAD algorithm, called smoothed sub-band spectral flatness measure (3SFM). It combines the lower complexity from maximum values sub-band SNR (MVSS) method [4] and the effectiveness of the spectral flatness measure (SFM) [3][11][12]. The idea is to benefit from the strong points of the SFM, but with less memory and time processing than LSFM [12] because it does not need to store the previous power spectrum values.

The input signal is processed frame-by-frame with an overlap factor of 75%. Then, for each k frame the power spectrum $\boldsymbol{S(k)}$ is computed with the Discrete Fourier Transform. At this point, the power spectrum is then divided in nine sub-bands, similar to the ETSI-AMR1 [20]. Next, for each sub-band j the SFM is computed with the following formula:

$$SFM_j(k) = \log_{10}\frac{GM(j,k)}{AM(j,k)} \tag{1}$$

where $\boldsymbol{GM(j,k)}$ and $\boldsymbol{AM(j,k)}$ represent the geometric and arithmetic means. The above means are computed as:

$$GM(j,k) = \sqrt[L_j]{\prod_{n=1}^{L_j} S_j(k,\omega_n)} \tag{2}$$

$$AM(j,k) = \frac{1}{L_j}\sum_{n=1}^{L_j} S_j(k,\omega_n) \quad (3)$$

where $\boldsymbol{S_j(k)}$ and $\boldsymbol{L_j}$ stand for the power spectrum and for the number of frequency bins of the j^{th} sub-band, respectively.

Going further, a robust feature is achieved by computing the distance gain for the current frame with the formula from [4], as below:

$$DSFM(k) = \sum_{i=1}^{9} SFM_i(k) + \sum_{i=1}^{9}\left[SFM_i(k) - \overline{SFM_\iota(k)}\right]^2 \quad (4)$$

where $\boldsymbol{\overline{SFM_\iota(k)}}$ represents the average of all sub-bands spectral flatness measure:

$$\overline{SFM_\iota(k)} = \frac{1}{9}\sum_{i=1}^{9} SFM_i(k) \quad (5)$$

Finally, the smoothed sub-band spectral flatness measure is computed as:

$$3SFM(k) = \alpha \cdot DSFM(k) + (1-\alpha)\cdot 3SFM(k-1) \quad (6)$$

where $\boldsymbol{\alpha}$ is the exponential smoothing factor.

To obtain a BSS solution, a fixed threshold is set to detect the beginning of the first utterance. Then, for each frame of each channel the *3SFM* is computed. If the 3SFM of *k* consecutive frames of a channel is higher than the fixed threshold, then we assume that the first utterance appeared on the respective channel. Depending on the length of the frame size, the overlap factor and the maximum response time imposed for the BSS, the *k* parameter may vary. Further, if voice activity was spotted on channel *j*, then the accumulated *3SFM* of channel *j* is computed as:

$$A3SFM_j = \sum_{i=0}^{i=k-1} 3SFM_j(m-i) \quad (7)$$

where $\boldsymbol{3SFM_j(m)}$ is the speech activity envelope of the m^{th} speech frame.

IV. Experimental Results and Discussions

For a brief check of the correlation between the speech intelligibility and the values of the proposed 3SFM speech feature we can perform a test in the same way as in Section 2 for the signals in Fig 1. Thus, for this implementation we used 32 ms frames with 24 ms overlap and the smoothing parameter was set as $\boldsymbol{\alpha = 0.9}$. The 3SFM characteristics for different SNR, using this configuration are shown in Fig. 4.

We notice that the 3SFM feature is able to indicate the channels with the highest and lowest SNR, hence the feature can be used to decide the BSS. Moreover, we can easily differentiate between speech and non-speech segments using a threshold with a value around –2.2. As the 3SFM levels have similar values during non-speech periods, the feature is robust to noise and can be used in VAD algorithms.

Using the proposed accumulating score scheme for wavelet and LSFM based VAD we obtained two other BSS solutions. In the following experiment we analyzed how well these methods discriminate the best speech quality signal among several noisy signals. For this, all three methods were configured to suit the ATM-VCS demands, analyzing around 200 ms from the first detected utterance. The wavelet-based approach used 128 ms frames with an overlap factor of 75%, a fixed threshold $\boldsymbol{th = 1.1}$ and $\boldsymbol{k = 4}$. The LSFM-based approach used 32 ms frames, overlapped 50%, a fixed threshold $\boldsymbol{th = -1.3}$ and $\boldsymbol{k = 12}$. In order to ensure a real-time processing for ATM-VCS, the default values of LFSM in [12] were modified as $\boldsymbol{M = 5}$ and $\boldsymbol{R = 15}$. For the proposed 3SFM-based BSS solution, we used 32 ms frames, which overlap 75%, $\boldsymbol{\alpha = 0.9}$ and $\boldsymbol{th = -2.2}$. Besides these three accumulated VAD scores based methods, the speech quality of the noisy signals was also evaluated with the PESQ standard, Mean Opinion Score – Listening Quality Objective [21]. The results are presented in Table I.

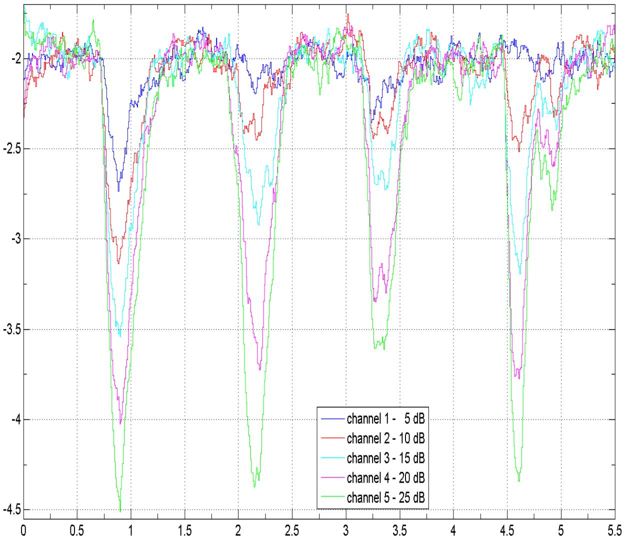

Figure 4 – 3SFM based VAD characteristics

We notice that for wavelet-based solution the accumulated score has positive values and for the LSFM- and 3SFM -based solutions the accumulated score has negative values. This is explained by the way in which these algorithms compute their score; the wavelet-based solution accumulates positive speech features, while others accumulate negative speech features. However, the sign difference does not affect the discriminative capacity of these methods.

TABLE I
EVALUATION OF THE BSS SOLUTIONS

SNR level (dB)	Wavelet [7]	LSFM [12]	**3SFM** *proposal*	PESQ MOS-LQO
5	5.56	0	-60.87	1.95
10	6.90	-17.79	-71.34	2.07
15	7.70	-23.27	-80.12	2.16
20	8.39	-33.57	-89.80	2.17
25	9.05	-50.43	-100.02	2.17

While the original LSFM feature leads to negative values, for 5 dB SNR level, the LSFM-based solution yields the accumulated score equal to 0. This is explained by the fact that for this SNR level the feature was not able to detect the first utterance. Therefore, no LSFM scores were added to characterize the speech intelligibility for this noisy signal.

From Table I we notice that any approach based on accumulated VAD score could indicate the signal with the best speech quality. On the other hand, the PESQ standard saturates above 15dB SNR. Therefore, for several signals with higher SNR, the PESQ could hardly choose the most intelligible one. Moreover, this standard needs the source signal to compute the MOS-LQO, which in ATM-VCS is not available. We can conclude that, in these conditions, the standard score is not useful and the accumulated VAD score methods represent an effective BSS solution.

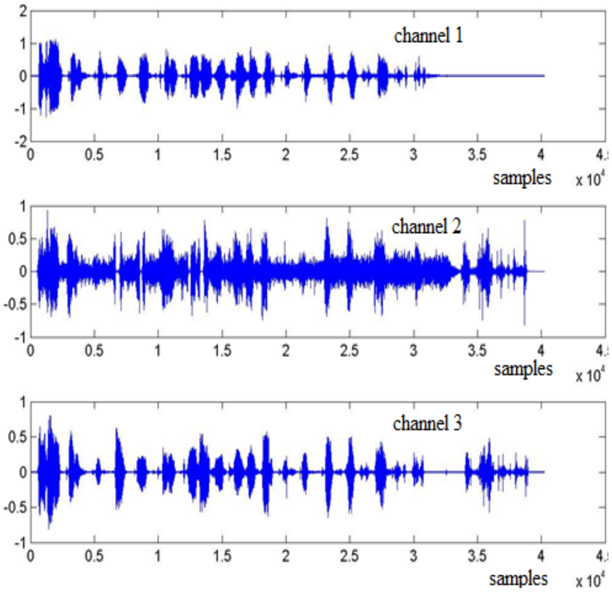

Figure 5 – Typical signals in ATM systems

In order to obtain more information regarding the proposed 3SFM feature, real signals from the ATM system were also used. A typical pair signal from ATM systems is represented in Fig. 5. As we can see, the second channel is the noisiest one. Further, the result of the 3SFM feature for these signals is shown in Fig. 6, where a very high smoother value, $\boldsymbol{\alpha = 0.99}$ was used.

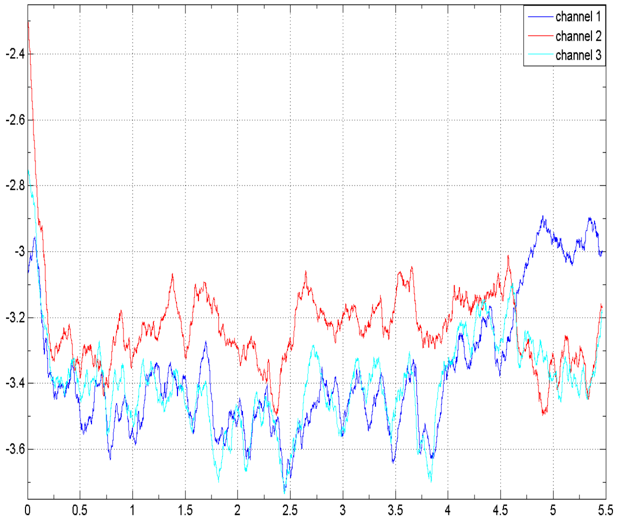

Figure 6 – 3SFM feature for typical signals from ATM systems

In Fig 6 we notice that the 3SFM feature of the second channel has in generally the lowest values. This means that the 3SFM feature will indicate channel 2 as having the worst speech quality. This could also be detected even for the first utterance. The first channel is indicated as offering the best speech quality after computing the accumulated 3SFM score for this utterance for all channels. Because the third channel yields similar 3SFM values, this indicates comparable speech quality for the first and the third channel. This can be verified by looking again over Fig 5, where we can notice similar waveforms. This confirms again the correlation between speech quality and the 3SFM feature. In Fig. 6, at the end part of the signals we can notice that the first channel has the lower score. This is explained by the packet-loss, visible in Fig 5, which leads to a poor VAD score, suggesting no speech activity.

The heavy smoothing operation ($\alpha = 0.99$) is used for signals from ATM systems, to reduce the high variations of the 3SFM feature, which could affect the first utterance detection. Because of the heavy smoothing, for non-speech periods, the 3SFM features in Fig. 6 do not have similar values. This is explained by the fact that the features need more time to rich the non-speech level. However, this does not affect the first utterance detection, maintaining enough discriminative power for an effective BSS solution.

V. Conclusion and Future Work

In this paper we presented possible solutions for the BSS problem in the air traffic management systems. They are based on the correlation between speech intelligibility and scores of VAD algorithms. We also showed that the VAD score-based methods do not saturate for higher SNR, as the PESQ does.

Besides the proposed solutions, which are based on previous VAD algorithms, we proposed here a new VAD algorithm for the BSS problem, the smoothed sub-band spectral flatness measure (3SFM). Moreover, it is shown that this speech feature is robust to noise and could be used in speech detection applications.

Because the 3SFM was proposed as a solution for BSS, future work has to evaluate and compare this feature in the VAD context.

References

[1] ED-136, "Voice over Internet Protocol (VoIP) Air Traffic Management (ATM) System Operational and Technical Requirements", *European Organization for Civil Aviation Equipment*, Mar 2009.

[2] S.A. Soleimani and S.M. Ahadi, "Voice Activity Detection based on Combination of Multiple Features using Linear/Kernel Discriminant Analyses", *International Conference on Information and Communication Technologies: From Theory to Applications*, pp. 1-5, Apr 2008.

[3] M.H. Moattar and M.M. Homayounpour, "A Simple But Efficient Real-Time Voice Activity Detection Algorithm", *EUSIPCO 2009*, Scotland, pp. 2549-2553, Aug. 2009.

[4] W. Jiang, W.K. Lo, and H. Meng, "A new voice activity detection method using maximized Sub-band SNR", *International Conference on Audio Language and Image Processing*, pp. 80–84, 2010.

[5] A. Ouzounov, "Robust Feature for Speech Detection", *Cybernetics and Information Technologies*, vol.4, No.2, pp.3-14, 2004.

[6] Z. Shuyin, G. Ying, and W. Buhong, "Auto-correlation property of speech and its application in voice activity detection", *First International Workshop on Education Technology and Computer Science* (*ETCS '09*), vol. 3 (IEEE, Piscataway), pp. 265–268, 2009.

[7] B.F. Wu and K.C. Wang, "Voice Activity Detection Based on Auto-Correlation Function Using Wavelet Transform and Teager Energy Operator", *Computational Linguistics and Chinese Language Processing*, vol. 11, No. 1, pp. 87-100, Mar 2006.

[8] Y.C. Lee and S.S. Ahn, “Statistical model-based VAD algorithm with wavelet transform”, *IEICE Trans. Fundamentals*, vol. E89-A, no. 6, pp 1594-1600, Jun 2006.

[9] M. Pwint and F. Sattar, “A new speech/non-speech classification method using minimal Walsh basis functions”, *IEEE International Symposium on Circuits and Systems*, vol. 3, pp. 2863-2866, May 2005.

[10] J. Ramirez, J.C. Segura, C. Benitez, A. Torre, and A. Rubio, “Efficient voice activity detection algorithms using long-term speech information”, *Speech Communication*, vol. 42, issues 3-4, pp 271-287, 2004.

[11] P.K. Ghosh, A. Tsiartas, and S. Narayanan, “Robust Voice Activity Detection Using Long-Term Signal Variability”, *IEEE Transactions on Audio*, *Speech, and Language Processing* , ISSN :1558-7916, pp. 600-613, Mar. 2011.

[12] Y. Ma and A. Nishihara, “Efficient voice activity detection algorithm using long-term spectral flatness measure”, *EURASIP Journal on Audio, Speech, and Music Processing*, ISSN 1687-4722, 2013:21, Jan. 2013.

[13] N. Seo, “Individual Voice Activity Detection Using Periodic to Aperiodic Component Ratio Based Activity Detection (Parade) and Gaussian Mixture Speaker Models”, *University of Maryland, Final Project*, 2007.

[14] K. Ishizuka, T. Nakatani, M. Fujimoto, and N. Miyazaki, “Noise robust voice activity detection based on periodic to aperiodic component ratio”, *Speech Communication*, vol. 52, pp. 41-60, Jan. 2010.

[15] N. Chatlani and J.J. Soraghan, “Local Binary Patterns For 1-D Signal Processing”, *EUSIPCO-2010*, Denmark, pp. 95-99, Aug. 2010.

[16] D. Ying, Y. Yan, J. Dang, and F.K. Soong, “Voice Activity Detection Based on an Unsupervised Learning Framework” *IEEE Transactions on Audio, Speech, and Language Processing* 19(8), pp. 2624 – 2633, 2011.

[17] P. Harding and B. Milner, “On the use of Machine Learning Methods for Speech and Voicing Classification”, *In Proceedings of Interspeech*, 2012.

[18] M.K. Omar, “Speech Activity Detection for Noisy Data using Adaptation Techniques”, *In Proceedings of Interspeech*, 2012.

[19] F.G. Germain, D.L. Sun, and G.J. Mysore, “Speaker and Noise Independent Voice Activity Detection”, *Interspeech*, France, pp. 732 – 736, Aug. 2013.

[20] “Adaptive Multi-Rate (AMR) speech codec; Voice Activity Detector (VAD)”, *European Telecommunications Standards Institute,* 3GPP TS 26.094 version 10.0.0 Release 10, 2011.

[21] “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs”, *International Telecommunication Union – Telecommunication Standardization Sector (ITU-T)*, 2001.