INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Multi-level Adaptive Speech Activity Detector for Speech in Naturalistic Environments

Bidisha Sharma, Rohan Kumar Das, Haizhou Li

Department of Electrical and Computer Engineering,
National University of Singapore, Singapore
{s.bidisha, rohankd, haizhou.li}@nus.edu.sg

Abstract

Speech activity detection (SAD) is a part of many speech processing applications. The traditional SAD approaches use signal energy as the evidence to identify the speech regions. However, such methods perform poorly under uncontrolled environments. In this work, we propose a novel SAD approach using a multi-level decision with signal knowledge in an adaptive manner. The multi-level evidence considered are modulation spectrum and smoothed Hilbert envelope of linear prediction (LP) residual. Modulation spectrum has compelling parallels to the dynamics of speech production and captures information only for the speech component. Contrarily, Hilbert envelope of LP residual captures excitation source aspect of speech. Under uncontrolled scenario, these evidence are found to be robust towards the signal distortions and thus expected to work well. In view of different levels of interference present in the signal, we propose to use a quality factor to control the speech/non-speech decision in an adaptive manner. We refer this method as multi-level adaptive SAD and evaluate on Fearless Steps corpus that is collected during Apollo-11 Mission in naturalistic environments. We achieve a detection cost function of 7.35% with the proposed multi-level adaptive SAD on the evaluation set of Fearless Steps 2019 challenge corpus.

Index Terms: speech activity detection, Fearless Steps Challenge, naturalistic environments

1. Introduction

The progress in speech and audio processing has shown scope towards wide range of applications. Speech activity detection (SAD) is one of the important modules in applications like speaker verification and speaker diarization [1–4]. The importance of SAD is more evident in real-world scenario as the correct detection of speech regions is necessary for feature extraction and modeling of required information [5,6]. It is found that the performance of many working systems degrades when there is ambient noise present in the surrounding environment [7, 8]. Therefore, there is always a need to have a robust SAD approach, which is susceptible to unseen conditions.

The traditional approaches for SAD consider signal energy as the primary evidence for detecting the speech regions [9]. Such methods use threshold over average energy of the signal that works well in controlled environments. However, these methods cannot work that effectively in noisy scenario. To address such issues, there have been different attempts for proposal of alternate SAD approaches. The statistical approaches are one such direction. A statistical SAD proposed in [10] showed robustness to low signal-to-noise (SNR) and vehicular environments. Further, improvements over statistical SAD approach are investigated in [11]. The multiple statistical models and combination of multiple statistical models have been studied for SAD in [12] and [13], respectively. However, the statistical approaches either require large training data or fails for short segments.

Another way of SAD resorts to signal processing techniques. The periodic nature of speech signal is used to derive an alternate approach for periodicity based SAD [14, 15]. Similarly, entropy is considered as an evidence to detect the speech in noisy conditions [16]. The vowel-like regions belong to high SNR portion of speech signals and are less affected by noise [17–19]. Similarly, glottal activity detection and sonorant region detection are performed to identify the speech regions in a noisy scenario [20, 21]. The study in [22] suggests a self-adaptive method for SAD that has been useful. Further, the fusion of multiple evidence as well as SAD approaches also helps and reflects in improved speaker verification performance [23–25]. Some other novel approaches include semi supervised [26], supervised/unsupervised [27] and parametric distribution based SAD [28]. The robustness of SAD techniques are mostly investigated for speaker verification and related areas.

The Fearless Steps Challenge[1] 2019 is organized to evaluate the state-of-the-art methods for different applications with naturalistic audio signals in challenging environments [29, 30]. A corpus from the data captured during Apollo-11 Mission is released for the challenge. There are five different tracks in challenge, SAD being one among them. The microphones used during the Apollo-11 Mission were mostly far-field that captured most of the environmental noise, some of which have complex harmonic structure [31–33]. This makes the recorded data very challenging for accurately detecting the speech regions. The combo SAD has been proposed by the challenge organizers in their previous investigations [33]. The speech recorded in the corpus are unprompted, and hence subject to significant variations in speech characteristics for every speaker [29].

In this work, we focus on finding an effective SAD for detection of speech regions on naturalistic audio from Fearless Steps Challenge 2019 corpus. Based on the variable noise level and sparse speech regions in the database, we use noise robust features under controlled thresholds. Hilbert envelop of linear prediction (LP) residual has been previously used as an evidence for vowel-like region detection [17]. Modulation spectrum of a signal represents the rate of articulation of a vocal signal. We consider these two evidence for designing a novel approach for SAD. Further, we propose a quality factor (Q-factor) of a signal that is derived from the denoised signal to decide an adaptive threshold. The two evidence are then used in a multi-level order to detect the speech regions. We refer this

[1] http://fearlesssteps.exploreapollo.org/

http://dx.doi.org/10.21437/Interspeech.2019-1928

proposed method as multi-level adaptive speech activity detector. The studies in this work are evaluated on the Fearless Steps 2019 challenge corpus.

The rest of the paper is organized as as follows. Section 2 describes the proposed multi-level adaptive SAD. In Section 4, the details of the experiments is presented. Section 5 reports the results and discussion. The work is concluded in Section 6.

2. Challenges in Naturalistic Environments

The Fearless steps (Apollo-11 Mission) audio data is collected from 30 individual analog communication channels with multiple speakers in different locations, working real-time to accomplish NASA's Apollo missions [34]. Three primary phases of the Apollo-11 mission are selected: (i) lift off, (ii) lunar landing, and (iii) lunar walk. For the SAD task, 20 hours and 10 minutes (39 files) of human verified ground truth labels and transcripts are provided as development set and we are encouraged develop unsupervised method of SAD with limited ground truth available. It is important to note that the dataset consists of different levels of noise, amount of speech content, and amount of silence over different files. The development set comprises of about 60% audio from clean channels and the other 40% is from degraded channels, while the channel information is not provided. This makes the problem of SAD even more challenging. Due to very long silence duration, the speech activity density of the corpus varies.

Here we bring out a discussion on the distribution of the energy and duration of speech/non-speech segments over the development set to highlight the challenges associated in naturalistic environments. Figure 1 (a) shows the continuous histogram plot corresponding to short-term energy (with 20 ms frame-size, 10 ms frame-shift) of all the speech and non-speech frames of the development set data, which shows a significant overlap between the two histograms. This gives us an idea that traditional energy-based SAD approaches may not be useful in this scenario. Further, we have also shown the same histogram plot for duration of all speech segments and non-speech segments in Figure 1 (b). It can be observed that the non-speech segments have a mean duration of 8.60 seconds, while speech segments have a mean duration of 2.49 seconds. This introduces difficulty in making a decision based on duration threshold to classify a short segment into speech or non-speech.

The discrete histogram illustrated in Figure 2 shows that the speech to non-speech duration ratio is lower for majority of the files provided in the development set. This indicates there are speech sparse, speech dense and speech/non-speech balanced examples in the dataset. Looking into these issues, we propose a novel framework for SAD, which works for variable types of channels and noises.

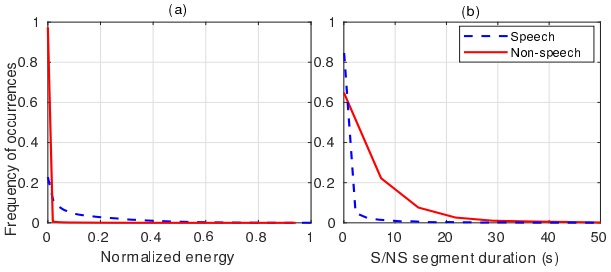

Figure 1: *The histograms of (a) energy, (b) duration of all the speech and non-speech segments on the development dataset.*

[2]http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

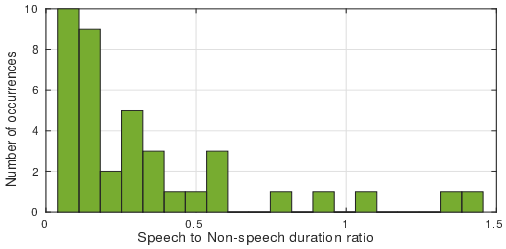

Figure 2: *The histogram of ratio of speech to non-speech duration for each file on the development dataset.*

3. Multi-level Adaptive SAD

Looking into the complexity involved with the Fearless Steps data (Apollo-11 Mission), we propose a method, which uses robust features in a controlled manner based on the amount of noise present in the signal. We selected two different evidence, which are modulation spectrum energy and smoothed Hilbert envelope energy (SHE) from analysis of different noise robust features. It is important to note that these two evidence are extracted after applying a basic denoising method over the noisy speech signal.

The modulation spectrum represents the evolution of the amplitude content of various frequency bands in short time Fourier transform (STFT) [35] spectrum over time. It reflects dynamics of speech production, in which the articulators move at rates of 2-12 Hz [36], and to the sensitivity of auditory cortical neurons to amplitude modulations at rates below 20 Hz. Capturing modulation spectrum energy over specific bands of the noisy signal can be helpful to extract information only for speech component present. In this work, the method of extracting modulation spectrum is followed from [35, 37].

The audio signal is analyzed into approximately 18 critical band filters between 0 and 8 kHz. These filters are trapezoidal in shape, and there is minimal overlap between adjacent bands. In each band, an amplitude envelope signal is computed by half-wave rectification and low pass filtering with cutoff frequency of 28 Hz. Each amplitude envelope signal is then downsampled to 80 samples/s and normalized by the average envelope level in that channel, measured over the entire utterance. The modulations of the normalized envelope signals are analyzed by computing the discrete Fourier transform (DFT) over 250-ms Hamming windows with shift of 12.5 ms to capture the dynamic properties of the signal. Finally, the 416 Hz components are added together, across all the critical bands.

The Hilbert envelope of LP residual is a good approximation of the excitation source of a speech signal, which is found to be useful in various speech processing tasks [38]. In this case, the LP residual is derived by performing LP analysis on overlapped segments of an audio signal (size of frame-size 25 ms, frame-shift 5 ms, LP order 12 and sampling rate 8 kHz). To consider only gross level changes in excitation characteristics of the signal, we have smoothed this evidence over a period of 5 ms, which we termed as SHE. The SHE is significant only in the regions with human voice and not in the noise segments.

In the proposed method, instead of directly combining the two evidence, we applied them in two-levels. As shown in Figure 3, we first applied a basic denoising method over the noisy signal, which is spectral subtraction based enhancement [10, 22]. The MATLAB based implementation (*specsub*) available in Voicebox[2] is used. From the denoised audio signal, we extract the modulation spectrum energy. As the modulation spectrum shows very low values for most of

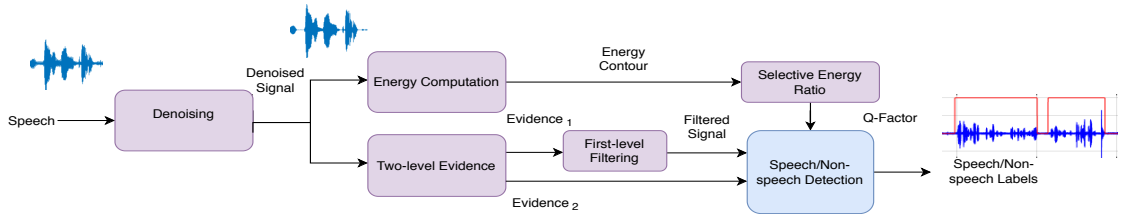

Figure 3: *Block diagram of proposed multi-level adaptive SAD.*

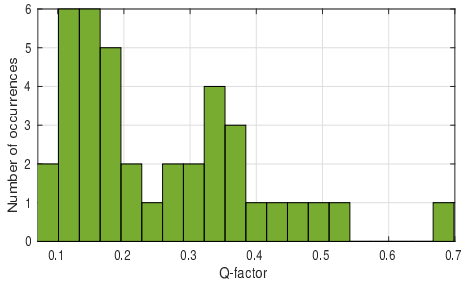

Figure 4: *Histogram of Q-factor obtained on development set.*

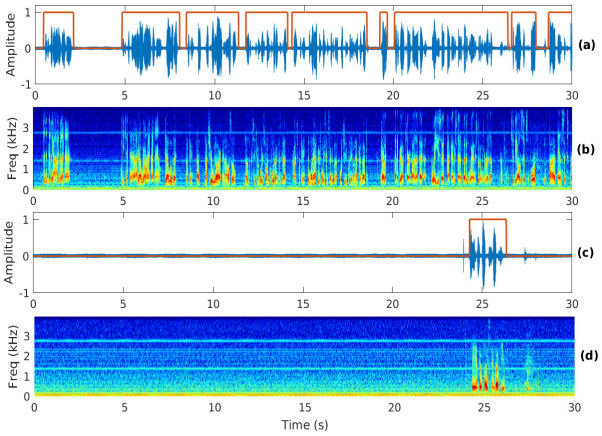

Figure 5: *Illustration of examples with different Q-factors; (a) signal corresponding with Q-factor=0.64, (b) spectrogram for signal in (a), (c) signal corresponding with Q-factor = 0.06, (d) spectrogram for signal in (c).*

the noisy segments, we applied a very low threshold ($0.1 \times median$(modulation spectrum energy)) over the normalized modulation spectrum energy. The regions detected as non-speech in this level are considered with high confidence level and are suppressed in the audio signal by replacing them with silences. In this stage, some noise segments are confused and are not classified as non-speech. In the second level, we used noise suppressed signal and SHE evidence to further filter the audio signal. We apply a threshold over the normalized SHE feature in both time and amplitude to filter the final non-speech segments. We note that different thresholds on SHE and duration are set by observation on the development dataset, for different Q-factors.

The Q-factor represents the ratio of speech and non-speech components present in the signal. To calculate this Q-factor, we derive the short-term energy contour (with frame-size 20 ms, frame-shift 10 ms and sampling rate 8 kHz) from the denoised signal and sort the energy frames in an ascending order. The mean values of energy corresponding to the lowest 20% frames and same for the highest 20% frames are obtained. The ratio of high to low energy ratio, which we termed as selective energy ratio, is considered as the Q-factor. The histogram corresponding to Q-factors obtained for all 39 examples of the development set data is shown in Figure 4. We observe that the speech samples with higher values of Q-factor are speech dense and those with lower values of Q-factor are speech sparse. The waveform and corresponding spectrogram (20 ms frame-size and 10 ms, frame-shift with sampling rate 8 kHz) for a speech dense example, with Q-factor 1.02 are shown in Figure 5 (a) and (b). The same for speech sparse example with Q-factor 0.12 are shown in Figure 5 (c) and (d). This type of variability over the database, makes it difficult to set a threshold. In this work depending on the Q-factor we categorize the audio files and accordingly set different parameters to detect speech/non-speech segments. Therefore, the proposal of Q-factor is found to be an effective solution to the issues. An implementation of this proposed SAD is made available[3] for use.

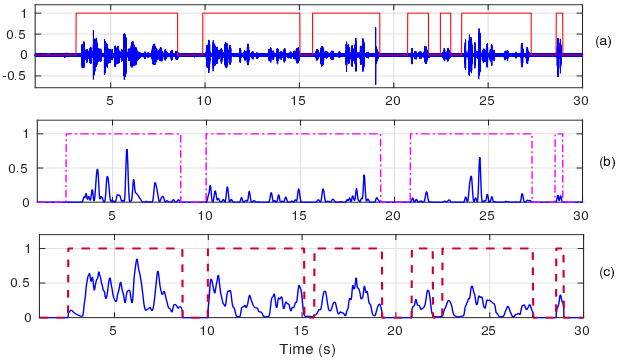

Figure 6: *SAD with the proposed method, (a) signal corresponding to input speech along with reference labels, (b) modulation spectrum evidence along with decision from first-level filtering, (c) SHE evidence with final SAD decision.*

Figure 6 (a) shows a denoised audio sample along with the ground-truth speech/non-speech labels. Corresponding modulation spectrum energy contour and with segmentation obtained in first-level filtering is shown in Figure 6 (b); In Figure 6 (c) we show the SHE contour with final speech/non-speech segments.

4. Experiments

4.1. Database

The Fearless Steps Challenge database contains two parts, namely, development and evaluation for the SAD task. There are 39 examples, each of around 30 minutes duration on the development set. On the contrary, the evaluation set consists of 40 examples of similar duration. The SAD labels are provided for the development set, on which the comparative studies are performed and then to be applied on evaluation set.

4.2. Experimental Setup

As explained in Section 2, we first compute the Q-factor for the denoised test file. In view of the development set data, based

[3]https://github.com/bidishasharma/MultiSAD

Table 1: *Thresholds on smoothed Hilbert envelope energy (SHE), modulation spectrum energy (MSE) and duration of segment based on Q-factor.*

Signal Type	Q-factor	Thresholds		
		Th_{SHE}	Th_{MSE}	Th_{Dur} (sec)
Speech Sparse	< 0.3	0.03	0.10×median(MSE)	1
Speech Balanced	> 0.3 & < 0.5	0.02	0.10×median(MSE)	1
Speech Dense	> 0.5	0.01	0.10×median(MSE)	0.5

on the Q-factor we divide the speech files into three categories; speech sparse, speech balanced and speech dense as shown in Table 1. This decision of Q-factor threshold is made based on the observation on development set data. Using the modulation spectrum evidence extracted from the denoised audio, we perform the first level filtering and suppress the non-speech regions. Here, the threshold is 10% of the median of modulation spectrum energy, which doesnot vary with Q-factor. Here we consider median instead of mean, because of distortion in some audio signals, which results in very high value of modulation spectrum energy for few frames. However, the modulation spectrum energy is very low for the noisy segments.

Further, we extract the SHE feature from the same denoised audio signal, on which we suppress the non-speech frames obtained from the first-level filtering. This SHE feature along with the previously calculated Q-factor is applied to final speech/non-speech decision module. In this case, we consider two types of thresholds on normalized SHE value (Th_{SHE}) and minimum duration of a non-speech segment (Th_{Dur}) that varies for speech sparse, speech balanced and speech dense examples, while the threshold on modulation spectrum energy (Th_{MSE}) is constant over all the examples.

All the thresholds are depicted in Table 1, which are set by analyzing the nature of SHE and duration in both the cases. We observe that the duration of individual non-speech segments are always more in speech sparse examples, accordingly Th_{Dur}. As there are less speech segments in speech sparse case, Th_{SHE} is set higher compared to speech dense and speech balanced case. The Fearless Steps Challenge evaluation plan considers the two distinct evaluation metrics, False positive (FP) and False negative (FN), along with their combination to find efficacy of the SAD methods. FP is incorrect detection of speech in a segment where the reference identifies as non-speech, whereas FN is missed the detection of speech in a segment where the reference identifies as speech.

In the SAD task of Fearless Steps challenge, missing, or failing to detect, actual speech is considered a more serious error than misidentifying its start and end times. The primary evaluation metric for the challenge is detection cost function (DCF), which is calculated as follows.

$$\text{D}CF(\theta) = 0.75 \times \text{P}_{\text{FN}}(\theta) + 0.25 \times \text{P}_{\text{FP}}(\theta),$$

where, P_{FP} and P_{FN} are probabilities of FP and FN, respectively. $\text{DCF}(\theta)$ is the DCF value for a system at a given system decision-threshold setting. Further, we show the computation of P_{FP} and P_{FN},

$$\text{P}_{\text{FP}} = \frac{\text{T}_{\text{FP}}}{\text{T}_{\text{non-speech}}}; \quad \text{P}_{\text{FN}} = \frac{\text{T}_{\text{FN}}}{\text{T}_{\text{speech}}},$$

where, T_{speech} and $\text{T}_{\text{non-speech}}$ are total speech and non-speech durations in an example, respectively.

5. Results and Analysis

The proposed multi-level adaptive SAD is evaluated on Fearless Steps Challenge 2019 corpus. We also compared the performance with the existing methods. We note that the existing methods consider their default parameter settings and they are used here without any tuning based on the development set. We also attempted to denoise the signal using spectral subtraction based enhancement [10, 22], and apply energy based VAD with suitable threshold to detect speech/non-speech segments (Specsub+Energy). Further, the results are reported for two different collars of 0 and 5 seconds to observe the trend. However, the challenge considers 5 second collars to benchmark the results.

Table 2: *Comparison of the proposed multi-level adaptive SAD with different existing methods and given baseline of Fearless Steps Challenge in terms of DCF (%), FN (%) and FP (%).*

Development Set						
Collar (sec)→	**0**			**5**		
SAD Method↓	**DCF(%)**	**FN(%)**	**FP(%)**	**DCF(%)**	**FN(%)**	**FP(%)**
Energy [9]	13.87	17.5	2.96	13.37	17.50	0.98
Statistical [10]	23.84	29.67	6.35	23.69	29.67	5.78
Self Adaptive [22]	23.84	29.67	6.35	23.69	29.67	5.78
Specsub+Energy	15.80	20.29	2.30	15.32	20.29	0.36
Proposed	6.68	7.05	5.59	**5.75**	7.04	1.85
Given Baseline [34]	–	–	–	8.60	–	–
Evaluation Set (DCF (%) for 5 sec collar)						
Proposed	**7.35**			Given Baseline [34]	11.70	

Table 2 shows performance comparison of different methods for SAD on the development set. We can observe that for existing methods, the FN is certainly higher than FP, which results in high DCF value. However, for the proposed method the difference between FN and FP is lower, which depicts that the method is not biased for either speech or non-speech regions. For the proposed method FP decreases as we increase the collar from 0 to 5 sec, whereas FN remains the same.

The proposed multi-level SAD outperforms the existing methods as well as the given baseline. The given baseline from the challenge organizers is cited from their SAD report [34]. This method learns a speech model from another corpus and then use that knowledge in Combo SAD [33]. The existing approaches could not perform that well due to the challenges in the corpus as discussed previously in Section 2. In contrast, the proposed SAD considers an adaptive threshold followed by multi-level evidence for SAD that could handle such speech in naturalistic environments. We then apply the proposed multi-level adaptive SAD on the evaluation set and obtain a DCF of **7.35%**, which is ranked ninth out of 27 system submissions in the Fearless Steps SAD Challenge, with baseline system at DCF 11.7% and first ranked system at DCF 3.31%.

6. Conclusions

This work focuses on a novel multi-level adaptive SAD method. It is derived using two evidence based on modulation spectrum and HE of LP residual in a multi-level combination. In addition, we introduce a Q-factor that demonstrate the quality of the signal in terms of noise levels present. This Q-factor is then used to classify the signal into speech dense, balanced and sparse regions, respectively. Thus, the proposed SAD involves an adaptive strategy and uses the stated two evidence in a multi-level way. We study the method on Fearless Steps Challenge 2019 database that shows the proposed SAD performs effectively compared to the existing methods as well as the given baseline for the challenge. The future work will focus to apply this SAD for different applications in naturalistic environments.

7. Acknowledgement

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project No. A18A2b0046).

8. References

[1] J. Ramirez, J. M. Gorriz, and J. C. Segura, "Voice activity detection. fundamentals and speech recognition system robustness," in *Robust Speech*, M. Grimm and K. Kroschel, Eds. Rijeka: IntechOpen, 2007, ch. 1.

[2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, pp. 12 – 40, 2010.

[3] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, Nov 2015.

[4] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, Sept 2006.

[5] J. Ramrez, J. C. Segura, C. Bentez, ngel de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3, pp. 271 – 287, 2004.

[6] M.-W. Mak and H.-B. Yu, "A study of voice activity detection techniques for nist speaker recognition evaluations," *Computer Speech & Language*, vol. 28, no. 1, pp. 295–313, 2014.

[7] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The 2016 speakers in the wild speaker recognition evaluation," in *Interspeech*, pp. 823–827.

[8] R. K. Das, S. Jelil, and S. R. M. Prasanna, "Multi-style speaker recognition database in practical conditions," *International Journal of Speech Technology*, vol. 21, no. 3, pp. 409–419, Sep 2018.

[9] 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," *IEEE Communications Magazine*, vol. 35, no. 9, pp. 64–73, Sep 1997.

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

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

[12] J.-H. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 1965–1976, June 2006.

[13] T. Petsatodis, C. Boukis, F. Talantzis, Z. Tan, and R. Prasad, "Convex combination of multiple statistical models with application to vad," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2314–2327, Nov 2011.

[14] R. Tucker, "Voice activity detection using a periodicity measure," *IEE Proceedings I - Communications, Speech and Vision*, vol. 139, no. 4, pp. 377–380, Aug 1992.

[15] V. Hautamaki, M. Tuononen, T. Niemi-Laitinen, and P. Franti, "Improving speaker verification by periodicity based voice activity detection," in *In Proc. International Conference on Speech and Computer (SPECOM).*, Moscow, Russia, Oct. 2007, pp. 645–650.

[16] P. Renevey and A. Drygajlo, "Entropy based voice activity detection in very noisy conditions," in *EUROSPEECH*, 2001, pp. 1887–1890.

[17] S. R. M. Prasanna and G. Pradhan, "Significance of vowel-like regions for speaker verification under degraded conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2552–2565, Nov 2011.

[18] B. Sharma and S. R. M. Prasanna, "Vowel onset point detection using sonority information." in *INTERSPEECH*, 2017, pp. 444–448.

[19] A. Paul, D. Mahanta, R. K. Das, R. K. Bhukya, and S. R. M. Prasanna, "Presence of speech region detection using vowel-like regions and spectral slope information," in *INDICON*, Dec 2017, pp. 1–5.

[20] A. Pandey, R. K. Das, N. Adiga, N. Gupta, and S. R. M. Prasanna, "Significance of glottal activity detection for speaker verification in degraded and limited data condition," in *TENCON*, 2015, pp. 1–6.

[21] B. Sharma and S. R. M. Prasanna, "Sonority measurement using system, source, and suprasegmental information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 505–518, 2017.

[22] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data," in *ICASSP*, May 2013, pp. 7229–7233.

[23] T. Kinnunen, A. Sholokhov, E. Khoury, D. A. L. Thomsen, M. Sahidullah, and Z.-H. Tan, "Happy team entry to nist opensad challenge: A fusion of short-term unsupervised and segment i-vector based speech activity detectors," in *Interspeech*, pp. 2992–2996.

[24] M. Graciarena, L. Ferrer, and V. Mitra, "The SRI system for the NIST OpenSAD 2015 speech activity detection evaluation," in *Interspeech*, 2016, pp. 3673–3677.

[25] S. Jelil, R. K. Das, S. R. M. Prasanna, and R. Sinha, "Role of voice activity detection methods for the speakers in the wild challenge," in *Twenty-third National Conference on Communications (NCC)*, March 2017, pp. 1–6.

[26] A. Sholokhov, M. Sahidullah, and T. Kinnunen, "Semi-supervised speech activity detection with an application to automatic speaker verification," *Computer Speech & Language*, vol. 47, pp. 132–156, 2018.

[27] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "Supervised/unsupervised voice activity detectors for text-dependent speaker recognition on the RSR2015 corpus," in *Odyssey*, 2014, pp. 123–130.

[28] J. W. Shin, J.-H. Chang, and N. S. Kim, "Voice activity detection based on a family of parametric distributions," *Pattern Recognition Letters*, vol. 28, no. 11, pp. 1295 – 1299, 2007.

[29] J. H. L. Hansen, A. Sangwan, A. Joglekar, A. E. Bulut, L. Kaushik, and C. Yu, "Fearless steps: Apollo-11 corpus advancements for speech technologies from earth to the moon," pp. 2758–2762.

[30] "Fearless steps challenge (FS-I) 2019 evaluation plan," 2019.

[31] A. Sangwan, L. Kaushik, C. Yu, J. H. L. Hansen, and D. W. Oard, "'houston, we have a solution': using nasa apollo program to advance speech and language processing technology," in *Interspeech*, 2013, pp. 1135–1139.

[32] C. Yu, J. H. L. Hansen, and D. W. Oard, "'Houston, we have a solution': a case study of the analysis of astronaut speech during NASA apollo 11 for long-term speaker modeling," in *Interspeech*, 2014, pp. 945–948.

[33] A. Ziaei, L. Kaushik, A. Sangwan, J. H. L. Hansen, and D. W. Oard, "Speech activity detection for nasa apollo space missions: challenges and solutions," in *Interspeech*, 2014, pp. 1544–1548.

[34] J. H. Hansen, A. Joglekar, M. Chandra Shekhar, V. Kothapally, C. Yu, L. Kaushik, and A. Sangwan, "The 2019 inaugural fearless steps challenge: A giant leap for naturalistic audio," in *proc. Interspeech*, 2019.

[35] S. Greenberg and B. Kingsbury, "The modulation spectrogram: In pursuit of an invariant representation of speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings*, 1997, pp. 1647 – 1650.

[36] C. L. Smith, C. P. Browman, R. S. McGowan, and B. Kay, "Extracting dynamic parameters from speech movement data," *The Journal of the Acoustical Society of America*, vol. 93, no. 3, pp. 1580–1588, 1993.

[37] H. Dudley, "Remaking speech," *The Journal of the Acoustical Society of America*, vol. 11, no. 2, pp. 169–177, 1939.

[38] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Transactions on Acoustics, Speech and Signal Processing,*, vol. 27, no. 4, pp. 309–319, 1979.