ROBUST FRONT-END PROCESSING FOR SPEAKER IDENTIFICATION OVER EXTREMELY DEGRADED COMMUNICATION CHANNELS

Seyed Omid Sadjadi and John H.L. Hansen

Center for Robust Speech Systems (CRSS),
The University of Texas at Dallas, Richardson, TX 75080–3021, USA
{sadjadi, john.hansen}@utdallas.edu

ABSTRACT

Effective front-end processing, which often involves feature extraction and speech activity detection (SAD), is essential for robustness in speech systems. In this study, we propose an unsupervised SAD scheme based on four different speech voicing measures which are combined with a perceptual spectral flux feature. Effectiveness of the proposed scheme is evaluated and compared against several commonly adopted unsupervised SAD methods under actual harsh acoustic conditions. As an example application, we also evaluate performance of the proposed SAD in the context of an i-vector based speaker identification (SID) system, where the recently introduced mean Hilbert envelope coefficients (MHEC) are benchmarked against conventional MFCCs. Long and spontaneous conversational audio recordings from DARPA program RATS (Phase-I) are used in our evaluations. Experimental results indicate that the proposed SAD solution is highly effective and provides superior performance compared to other unsupervised SAD techniques considered. In addition, it is shown that MHECs are effective alternatives to MFCCs for SID tasks under severe degraded channel conditions.

Index Terms— Mean Hilbert Envelope Coefficients (MHEC), speaker identification (SID), spectral flux, speech activity detection (SAD), voicing measures

1. INTRODUCTION

Effective front-end processing, in which the raw acoustic waveform is converted into a more useful and compact representation called feature vectors, is a necessary first step for robust speech applications. Specifically, speech activity detection (SAD) and feature extraction are two primary components of almost all front-ends. SAD has applications in a variety of contexts such as speech coding [1], automatic speech recognition (ASR) [2, 3, 4], speaker and language identification [5, 6, 7, 8], and speech enhancement [9]. In addition, for surveillance and monitoring applications that involve listening to long conversational audio recordings with small *a priori* speech presence probability, SAD can help mitigate the excessive cognitive load on human listeners by removing long and often noisy non-speech intervals. This can in turn increase the efficiency of listeners and reduce overall *listener fatigue*. In this study, our goal is to develop a robust and unsupervised framework for speech detection in data recoded over extremely degraded communication channels [10].

State-of-the-art SAD techniques include both supervised and unsupervised approaches. Supervised methods, which are often based on either Gaussian mixture models (GMM) [7, 11], hidden Markov models (HMM) [3, 4], or multi-layer perceptrons (MLP) [5], work well given that pre-trained models for both speech and non-speech classes broadly match the acoustic characteristics of the test environment. Hence, they are limited to applications where a large amount of training data is available and the acoustic properties of the test environment are consistent and known a priori as well.

On the other hand, unsupervised SAD techniques assume no a priori knowledge about the acoustic characteristics of the test environment, and can be categorized as feature-based [1, 12, 13], or statistical model-based [14, 15, 16, 17]. Feature-based methods perform well under stationary noise conditions at relatively high signal-to-noise ratio (SNR) levels, however their performance degrades rapidly as the noise level increases. Statistical model-based techniques employ a likelihood ratio test (LRT) of speech presence and absence hypotheses in the short-time Fourier transform (STFT) domain, assuming that an estimate of the noise power spectrum is available. The LRT based techniques are generally robust and effective, however their performance is dependent on the accuracy of the noise spectrum estimate, which is assumed to be uncorrelated and additive, making them vulnerable to the presence of non-stationary and rapid changing noise. Moreover, there is no meaningful measure from speech production physiology in the algorithm as it works entirely based on the criteria derived from statistics. A remedy to this vulnerability issue was proposed in [17], where the harmonicity information was integrated into the LRT framework which led to significant gains in SAD performance, particularly at extremely low SNR conditions.

In this study, we propose a robust and unsupervised SAD system solely based on features that convey fundamental traits of speech, which are governed by the speech production process. In particular, four different voicing measures as well as a perceptual spectral flux (SF) feature are linearly combined through the principal component analysis (PCA) to form a 1-dimensional soft-decision vector. For hard-decision making, a 2-mixture GMM is fit on the soft-decision vector, which exhibits a bimodal distribution, and a threshold is estimated based on the weighted average of the GMM means. The effectiveness of the proposed technique is evaluated under actual noisy channel conditions using dry-run (part-1) speech material from Phase-I of the DARPA program Robust Automatic Transcription of Speech (RATS) [10, 18] (distributed by the Linguistic Data Consortium - LDC). Performance of the proposed SAD method is benchmarked against that of commonly adopted feature-based (e.g., ITU G729 Annex B [1]) and statistical model-based (e.g., single and multiple observation LRT [14, 15, 17]) approaches. In addition, as an example application, we also evaluate performance of the proposed SAD in the context of an i-vector based speaker identification (SID) system [19], where the recently introduced mean Hilbert envelope coefficients (MHEC) [20, 21] are benchmarked against conventional MFCCs using extremely degraded speech recordings from Phase-I of the RATS program for the SID task. Developing robust alternative features for SID tasks has recently attracted significant research effort (e.g., see [22, 23, 24, 25]), it is therefore useful to explore the effectiveness of the MHEC versus baseline MFCCs for SID over extremely degraded channels.

2. UNSUPERVISED SAD ALGORITHM

In this section, we describe the procedure for extraction of a 1-dimensional feature vector that is used in our system as soft-decision for the speech/non-speech discrimination task. This "combo" feature is *efficiently* obtained from a linear combination of four different voicing measures as well as a perceptual spectral flux feature. The voicing measures include harmonicity, clarity, prediction gain, and periodicity. The perceptual spectral flux and periodicity are extracted in the frequency domain, while the harmonicty, clarity, and prediction gain are all time domain features. For feature extraction, the audio signal is blocked into 32 ms frames with a 10 ms skip rate. In order to extract the periodicity, harmonicity, and clarity, an approximate knowledge of the plausible pitch range in human speech is required. Here, we choose a pitch period duration within the interval of [2, 16] ms (or equivalently [62.5, 500] Hz in the frequency domain), where the lower limit is imposed by the analysis frame length, and the fact that each frame should at least cover two pitch periods for a reliable voicing estimate.

2.1. Time Domain Features

All time domain voicing measures in this section, directly or indirectly, use the normalized autocorrelation proposed in [26] for noise robust pitch estimation. The deterministic autocorrelation of a short-time windowed segment $x(n)$ is computed as,

$$r_{xx}(t,k) = \frac{\sum_{j=0}^{N-1} x(j)x(j+k)w(j)w(j+k)}{\sum_{j=0}^{N-1} w(j)w(j+k)}, \qquad (1)$$

where $w(j)$ is a Hanning window, and t and k are frame and autocorrelation lag indices, respectively. It has been shown that normalization by autocorrelation of the window function in (1) effectively mitigates the impact of strong formants on the maximum autocorrelation peak in the pitch range, and obviates the need for low-pass filtering and/or center-clipping [27]. In addition, it also compensates for the windowing effect which tapers the autocorrelation function towards zero for larger lags.

2.1.1. Harmonicity

Harmonicity (a.k.a. harmonics-to-noise ratio) is defined as the relative height of the maximum autocorrelation peak in the plausible pitch range. Mathematically, it can be expressed as,

$$h(t) = \frac{r_{xx}(t,k_{\max})}{r_{xx}(t,0) - r_{xx}(t,k_{\max})},$$

$$k_{\max} = \underset{2\text{ ms} \le k \le 16\text{ ms}}{\arg\max}\ r_{xx}(t,k). \qquad (2)$$

Note that the autocorrelation of a periodic signal is also periodic with the same period, and its maximum takes on values close to the autocorrelation at zero lag. Accordingly, for voiced segments which have periodic structure, the harmonicity shows sharp peaks.

2.1.2. Clarity

We define clarity as the relative depth of the minimum average magnitude difference function (AMDF) valley in the plausible pitch range. Computing the AMDF from its exact definition is costly; however, it has been shown that the AMDF can be derived (analytically) from the autocorrelation as [27],

$$D(t,k) \approx \beta(k) \cdot \sqrt{2\Big[r_{xx}(t,0) - r_{xx}(t,k)\Big]}, \qquad (3)$$

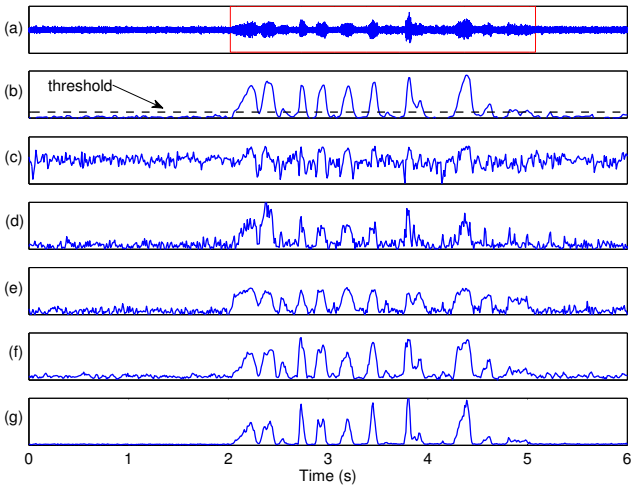

Fig. 1. Individual features as well as their combination for a sample waveform with an average segmental SNR of 2.59 dB. (a) noisy speech signal, (b) the combo feature along with the decision threshold (dashed), (c) the perceptual SF, (d) the periodicity, (e) the clarity, (f) the prediction gain, and (g) the harmonicity.

where $\beta(k)$ is a scale factor that can vary between 0.6 and 1.0. We have found that the clarity feature is not very sensitive with respect to the value of this parameter; therefore, we set $\beta(k)$ to 0.8 in our experiments. The clarity is then extracted as,

$$c(t) = 1 - \frac{D(t,k_{\min})}{D(t,k_{\max})},$$

$$k_{\min} = \underset{2\text{ ms} \le k \le 16\text{ ms}}{\arg\min}\ D(t,k). \qquad (4)$$

Subtracting the term from 1 in (4) simply converts the minima to maxima which are more desirable for our application. In this manner, the clarity exhibits large values for voiced and speech-like segments, while maintaining a minimum for background sounds.

2.1.3. Prediction gain

The prediction gain is defined as the ratio of the signal energy to the linear prediction (LP) residual signal energy. The signal energy can be obtained from the autocorrelation at zero lag. In order to calculate the LP residual signal energy, however, the Levinson-Durbin recursion [27] is applied and the error from the last step yields the energy of the residual signal. The prediction gain is then computed as,

$$G_p(t) = \log\left(\frac{r_{xx}(t,0)}{\epsilon^p}\right), \qquad (5)$$

where ϵ^p is the error in the last step of the recursion, and p is the order of LP analysis which is set to 10 in this study (assuming a sampling rate of 8 kHz). In short-time frames, there is a high correlation among speech samples, making it easier to predict, or in other words the denominator in (5) becomes smaller. Therefore, the average predication gain reaches its highest value for voiced and speech-like frames.

2.2. Frequency Domain Features

Both features from this section are extracted in the STFT domain which is formed by taking a 2048-point DFT from Hamming windowed frames after zero padding. Magnitude information is only used and the phase response is discarded.

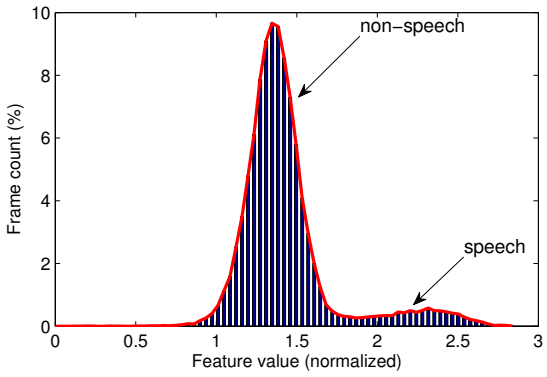

Fig. 2. Distribution of the combo feature values for a sample waveform.

2.2.1. Periodicity

In the STFT domain, the harmonics of the pitch frequency are apparent in the magnitude spectrum of speech during voiced and speech-like segments. This observation serves as the basis for the harmonic product spectrum (HPS) technique [27] which has been widely applied for pitch detection in noisy environments. The HPS in the log-spectral domain is defined as, $P(t,\omega) = \sum_{l=1}^{R} \log \left|X(t,l\omega)\right|$, where R is the number of frequency-compressed copies of the original spectrum, which is fixed to 8 in this study. The frequency-compressed copies coincide at the fundamental frequency and reinforce the amplitude, while other harmonics are cancelled or attenuated in the final product. The periodicity is computed as the maximum peak of (6) in the plausible pitch range,

$$P_{hps}(t) = P(t,\omega_{\max}),$$

$$\omega_{\max} = \underset{62.5\,\text{Hz} \leq \omega \leq 500\,\text{Hz}}{\arg\max} P(t,\omega). \tag{6}$$

The periodicity is especially impervious to noise and other background sounds, since their spectral harmonics cannot combine coherently in the HPS. The periodicity can thus be used to effectively discriminate speech from non-speech sounds.

2.2.2. Perceptual spectral flux

Over short-time frames, speech is a quasi-stationary and slowly varying signal, meaning that its spectrum does not change rapidly from one frame to another. Hence, one can effectively exploit this quality to develop a feature capable of discriminating speech from other more rapidly varying sounds. An example of such a feature is the SF [28], which measures the degree of variation in the spectrum across time. Given the benefits of incorporating perceptual models into speech processing frameworks (e.g., MFCC), in this study we define the perceptual SF as, $SF_p(t) = \left|\left|X_m(t,\omega) - X_m(t-1,\omega)\right|\right|_1$, where $||\cdot||_1$ denotes the L^1-norm, and $X_m(t,\omega)$ is the energy normalized mel-spectrum at frame t which is calculated using an 80-channel mel-filterbank spanning the frequency range from 0 to the Nyquist frequency. The perceptual SF exhibits relatively deep valleys for speech segments, while maintaining a maximum value for background sounds/silence. Accordingly, we employ the negative of this parameter as a feature for speech/non-speech discrimination.

After extracting the above noted features, a 5-dimensional vector is formed by concatenating the voicing measures along with the perceptual SF. Each feature dimension f_i is then normalized according to, $f_i' = \frac{f_i - \mu_i}{\sigma_i}$, where the mean μ_i and standard deviation σ_i are computed over the entire waveform. The normalized 5-dimensional feature vectors are linearly mapped into a 1-dimensional feature space represented by the most significant eigenvector of the feature covariance matrix. This is realized through PCA and retaining the dimension that corresponds to the largest eigenvalue. The 1-dimensional "combo" feature is smoothed via a 3-point median filter to serve as soft-decisions for the SAD task. Fig. 1 shows sample time domain plots of individual features as well as their combination for part of a noisy speech waveform with an average segmental SNR of 2.59 dB. It is seen that, although the individual features might not be as discriminative for the SAD task, their combination exhibits a great potential for noise robust speech/non-speech detection. Distribution of the combo feature values for a 366-second noisy speech signal is depicted in Fig. 2 (note that this is an example for illustration; the distribution of the combo feature can vary across different waveforms based on the distortion level and also the proportion of speech versus non-speech frames, although it still remains bimodal). It is evident that the combo feature has a bimodal distribution in which speech and non-speech classes are well separated. We exploit this property for hard-decision making by fitting a 2-mixture GMM to the feature and estimating a detection threshold (Th) from a weighted average of the mixture means as, $Th = \alpha\,\mu_{sp} + (1-\alpha)\,\mu_{sil}$, where μ_{sp} and μ_{sil} are the speech and non-speech mixture means, respectively. The weight parameter α can be tuned to achieve the desired false accept/reject rate (P_{fa} and P_{miss}).

3. EXPERIMENTS AND RESULTS

3.1. Speech Activity Detection

The proposed unsupervised SAD system is evaluated using speech material from the RATS Phase-I dry-run data (only part-1) [10]. RATS dry-run data consists of a total of 111 conversational telephone speech (CTS) waveforms that were retransmitted (through LDC's Multi Radio-Link Channel Collection System) and recorded over 8 extremely degraded communication channels, labeled A–H, with distinct noise characteristics (the distortion type is nonlinear and the noise is to some extent correlated with speech). Each CTS file is 900 seconds long with sparse speech activity. As previously noted in Section I, our goal here is to develop a robust unsupervised SAD system for long audio recordings with small *a priori* speech presence probability, in order to assist human listeners avoid auditing long noisy non-speech intervals. This justifies our choice of the aforementioned dataset for evaluations. However, as we shall see in the next section, our system has the potential to be adapted for automatic speech applications such as speaker and language identification, with no or minimal modifications.

Given that majority of the features described in Section II can primarily detect voicing, the proposed SAD may perform poorly in detecting short unvoiced frames surrounding a voiced segment. To alleviate this issue, similar to [7] and [12], the boundaries of each detected speech segment are extended by 0.1 s (i.e., 10 frames). The same post-processing is applied for other techniques used in our evaluations.

Fig. 3 shows receiver operating characteristic (ROC) curves obtained from evaluating the proposed unsupervised SAD scheme as well as five commonly adopted SAD solutions, namely ITU G729B [1], single-observation LRT (SOLRT) [14], SOLRT paired with an HMM-based hangover smoothing scheme [14], multiple-observation LRT (MOLRT) [15], and a recently introduced modification to MOLRT that incorporates harmonicity information into the LRT framework [17] (for individual voicing feature performances see [29]). It is evident from the ROC curves that our unsupervised system performs well on extremely degraded data from RATS dry-run

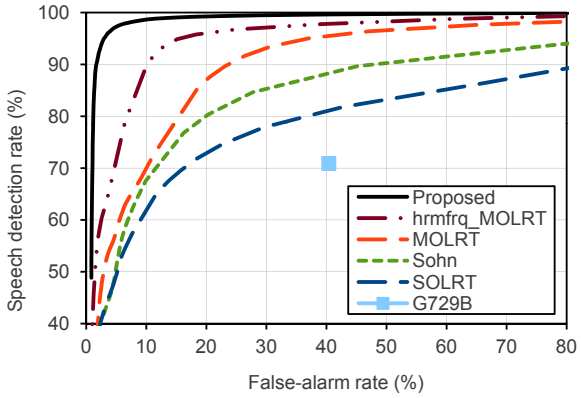

Fig. 3. Comparison of ROC curves for the proposed SAD method versus other feature-based and LRT-based techniques considered in the evaluations using RATS dry-run (part-1) data for Phase-I.

(part-1), which points to its robustness against nonlinear channel distortions. The basic assumption in formulating different flavors of statistical LRT based SAD is that the noise is additive and uncorrelated, however, it should be noted that the type of distortion seen in RATS data is nonlinear (i.e., noise is correlated with speech and not additive). This is a major reason for the poor performance of this class of SAD methods on RATS data.

3.2. Speaker Identification

As an example application, we evaluate the proposed unsupervised SAD in the context of an i-vector based SID system. A total of 18366 speech recordings, including clean source files along with their degraded (retransmitted) versions (A through H), from material distributed within the RATS program for the SID task (LDC2012E40 and LDC2012E49 [18]) are partitioned into development, enrollment, and test sets. The development set contains 11634 audio files from 641 speakers with a minimum of 8 sessions/channels per speaker. For enrollment, we predefine 6 source files for each speaker, thus providing 6×8 retransmitted segments that we could choose from. Accordingly, four different test conditions can be identified as: (i) matched, (ii) mismatched, (iii) seen, and (iv) unseen. Speaker models for matched and mismatched conditions are chosen from the same retransmission channel (e.g., speaker "1A" would come from the 6 source files retransmitted over channel A). A test is categorized as matched if it comes from the same channel on which the speaker model is enrolled (in our example a test segment from channel A is a matched test because the channels match). On the other hand, a test is labeled as mismatched if it comes from a channel other than the one used in enrollment (in our example any test segment from channels B–H). Speaker models for the seen and unseen conditions are enrolled using numerous channels selected randomly for each of the 6 enrollment segments. A test is labeled as seen trial if the model against which it is scored has observed the test channel during enrollment. If the speaker model has not observed the test channel during enrollment, the associated trial is part of the unseen trials. The total number of trials for each of the 4 test conditions are: matched (26,428), mismatched (177,197), seen (554,988), and unseen (195,448).

We consider two acoustic features in our SID experiments; 12-dimensional MFCC (HTK [30]) and MHEC [20, 21] feature vectors are extracted using 32-channel mel and Gammatone filterbanks spanning the telephone bandwidth (i.e., 300–3400 Hz), respectively. For both acoustic features, frame log-energies are appended and the first and second temporal cepstral derivatives are computed to from 39-dimensional vectors, and finally cepstral mean and variance normalization (CMVN) is applied. In order to perform non-speech frame dropping, we use (i) time labels generated by LDC on clean source data (before retransmission) and re-aligned to the retransmitted versions, and (ii) time labels obtained using the proposed unsupervised SAD on clean as well as retransmitted waveforms. In this experiment, the SAD threshold parameter, α, is set to 0.55.

A 1024-mixture gender independent UBM is constructed per feature-SAD combination (4 in total) and used to generate zeroth and first order Baum-Welch statistics for training the i-vector extractor [19]. We extract 400-dimensional i-vectors for each file and use the average for 6-sided enrollment. The dimension of i-vectors is then reduced to 250 using LDA, and Gaussian PLDA models [31, 32] with 250 columns in the Eigenvoice matrix are trained on the development data to score the trials. Results of the SID experiments using LDC and the proposed SAD labels are given in Tables 1 and 2, respectively, in terms of EER and P_{fa} at $P_{miss} = 10\%$ (FA10m, which is the RATS program target). In summary, the SID performance is consistently superior with time labels produced by the proposed SAD which further confirms its effectiveness for speech detection over extremely degraded audio recordings. In addition, irrespective of the time label source, the MHEC based subsystems outperform the systems trained with MFCCs on all 4 test conditions considered. Furthermore, a simple additive fusion of the MHEC and MFCC based subsystems provides consistent gain in performance. This indicates that the two acoustic features are complimentary for the SID task.

Table 1. Performance of the MFCC and MHEC based subsystems as well as their fusion with LDC speech/non-speech labels.

Condition	EER (%)			FA10m (%)		
	MFCC	MHEC	Fusion	MFCC	MHEC	Fusion
seen	5.45	4.84	4.79	2.58	2.12	2.00
unseen	6.08	5.76	5.48	3.42	2.76	2.64
matched	4.05	3.74	3.72	1.69	1.34	1.34
mismatched	6.66	6.27	6.00	4.33	3.66	3.41

Table 2. Performance of the MFCC and MHEC subsystems and their fusion with speech/non-speech labels from the proposed SAD.

Condition	EER (%)			FA10m (%)		
	MFCC	MHEC	Fusion	MFCC	MHEC	Fusion
seen	4.46	4.27	3.97	1.66	1.35	1.37
unseen	5.05	4.86	4.56	1.99	1.80	1.63
matched	3.62	3.36	3.30	1.23	1.04	1.05
mismatched	5.67	5.37	5.11	3.02	2.61	2.36

4. ACKNOWLEDGMENT

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. D10PC20024. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the DARPA or its Contracting Agent, the U.S. Department of the Interior, National Business Center, Acquisition and Property Management Division, Southwest Branch. The authors would like to thank Mitchell McLaren of SRI for providing development lists for the SID experiments.

5. REFERENCES

[1] 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 Commun. Mag.*, vol. 35, pp. 64–73, Sept. 1997.

[2] B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 24, no. 3, pp. 201–2012, Jun. 1976.

[3] R. Sarikaya and J. H. L. Hansen, "Robust detection of speech activity in the presence of noise," in *Proc. ICSLP*, Dec. 1998, pp. 1455–1458.

[4] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan, and R. Sarikaya, "Robust speech recognition in noisy environments: The 2001 IBM SPINE evaluation system," in *Proc. IEEE ICASSP*, May 2002, pp. 53–56.

[5] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical structures of neural networks for phoneme recognition," in *Proc. IEEE ICASSP*, May 2006, p. I.

[6] I. McCowan, D. Dean, M. McLaren, R. Vogt, and S. Sridharan, "The delta-phase spectrum with application to voice activity detection and speaker recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 7, pp. 2026–2038, Sept. 2011.

[7] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Vesely, and P. Matejka, "Developing a speech activity detection system for the DARPA RATS program," in *Proc. INTERSPEECH*, Sept. 2012.

[8] P. Matějka, O. Plchot, M. Soufifar, O. Glembek, L. D'Haro, K. Veselý, F. Grézl, J. Ma, S. Matsoukas, and N. Dehak, "Patrol team language identication system for DARPA RATS P1 evaluation," in *Proc. INTERSPEECH*, Sept. 2012.

[9] A. Davis, S. Nordholm, S. Y. Low, and R. Togneri, "A multi-decision sub-band voice activity detector," in *Proc. EUSIPCO*, Sept. 2006.

[10] K. Walker and S. Strassel, "The RATS radio traffic collection system," in *Proc. ISCA Odyssey*, Jun. 2012.

[11] M. K. Omar, "Speech activity detection for noisy data using adaptation techniques," in *Proc. INTERSPEECH*, Sept. 2012.

[12] J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Commun.*, vol. 42, pp. 271–287, Apr. 2004.

[13] M. C. Huggins, B. Y. Smolenski, and A. D. Lawson, "Adaptive high accuracy approaches to speech activity detection in noisy and hostile audio environments," in *Proc. INTERSPEECH*, Sept. 2010, pp. 3094–3097.

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

[15] J. Ramirez, J. Segura, C. Benitez, L. Garcia, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Process. Lett.*, vol. 12, no. 10, pp. 689–692, Oct. 2005.

[16] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, pp. 412–424, Mar. 2006.

[17] L. N. Tan, B. J. Borgstrom, and A. Alwan, "Voice activity detection using harmonic frequency components in likelihood ratio test," in *Proc. IEEE ICASSP*, Mar. 2010, pp. 4466–4469.

[18] DARPA Robust Automatic Transcription of Speech (RATS). [Online]. Available: http://projects.ldc.upenn.edu/RATS/

[19] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[20] S. O. Sadjadi and J. H. L. Hansen, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions," in *Proc. IEEE ICASSP*, May 2011, pp. 5448–5451.

[21] S. O. Sadjadi, T. Hasan, and J. H. L. Hansen, "Mean Hilbert envelope coefficients (MHEC) for robust speaker recognition," in *Proc. INTERSPEECH*, Sept. 2012.

[22] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, and S. Shamma, "Linear versus mel frequency cepstral coefficients for speaker recognition," in *Proc. IEEE ASRU*, Dec. 2011, pp. 559–564.

[23] C. Hanilci, T. Kinnunen, F. Ertas, R. Saeidi, J. Pohjalainen, and P. Alku, "Regularized all-pole models for speaker verification under noisy environments," *IEEE Signal Process. Lett.*, vol. 19, pp. 163–166, Mar. 2012.

[24] T. Kinnunen, R. Saeidi, F. Sedlak, K. A. Lee, J. Sandberg, M. Hansson-Sandsten, and H. Li, "Low-variance multitaper MFCC features: A case study in robust speaker verification," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 7, pp. 1990 –2001, Sept. 2012.

[25] M. J. Alam, T. Kinnunen, P. Kenny, P. Ouellet, and D. O'Shaughnessy, "Multitaper MFCC and PLP features for speaker verification using i-vectors," *Speech Commun.*, vol. 55, no. 2, pp. 237 – 251, Feb. 2013.

[26] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of the sampled sound," in *Proc. Institute of Phonetic Sciences*, vol. 17, 1993, pp. 97–110.

[27] L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech Processing*, 1st ed. Upper Saddle River, NJ: Prentice Hall Press, 2010.

[28] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. IEEE ICASSP*, Apr. 1997, pp. 1331–1334.

[29] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Process. Lett.*, vol. 20, pp. 197–200, Mar. 2013.

[30] HTK - Hidden Markov Model Toolkit v3.4.1. [Online]. Available: http://htk.eng.cam.ac.uk/

[31] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE Int. Conf. Computer Vision, ICCV 2007*, Oct. 2007, pp. 1–8.

[32] D. Garcia-Romero and C. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. INTERSPEECH*, Sept. 2011, pp. 249–252.