A robust frontend for VAD: exploiting contextual, discriminative and spectral cues of human voice

Maarten Van Segbroeck, Andreas Tsiartas and Shrikanth S. Narayanan

Signal Analysis and Interpretation Lab,
University of Southern California, Los Angeles, USA
{maarten, shri}@sipi.usc.edu, tsiartas@usc.edu

Abstract

Reliable automatic detection of speech/non-speech activity in degraded, noisy audio signals is a fundamental and challenging task in robust signal processing. As various speech technology applications rely on the accuracy of a Voice Activity Detection (VAD) system for their effectiveness and robustness, the problem has gained considerable research interest over the years. It has been shown that in highly distorted conditions, an accurate segmentation of the target speech can be achieved by combining multiple feature streams. In this paper, we extract four one-dimensional streams each attempting to separate speech from the disturbing background by exploiting a different speech-related characteristic, i.e. (i) the spectral shape, (ii) spectro-temporal modulations, (iii) the periodicity structure due to the presence of pitch harmonics, and (iv) the long-term spectral variability profile. The information from these streams is then expanded over long duration context windows and applied to the input layer of a standard Multilayer Perceptron classifier. The proposed VAD was evaluated on the DARPA RATS corpora and shows to be very competitive to current state-of-the art systems.

Index Terms: noise robust features, speech activity detection

1. Introduction

Voice activity detection (VAD) is defined as the problem of separating a target speech sound from interfering sound sources that are present in the noisy acoustical environment. In various speech processing modules and applications, such as noise reduction algorithms, language identification, speaker recognition, speech coding and automatic speech recognition, a noise robust detection of speech in the audio signals is an important and fundamental pre-processing step as it can significantly improve performance.

A VAD typically consists of a feature extraction module and a decision mechanism, which can vary from a simple set of rules to advanced classification mechanisms that require intensive training on noisy speech. Trained classifiers have proven to be very effective when the mismatch between training and testing conditions is relatively low.

Recent work has proven that the use of combinations of diverse feature streams could significantly improve the robustness of speech activity detection, especially when the noise conditions are severe [1, 2, 3, 4]. This paper contributes to this trend by proposing a VAD system that attempts to measure specific properties of speech that discriminate it from disturbing background sounds.

Human speech is a complex sound signal characterized by spectral patterns composed of spectro-temporal amplitude modulations, rich harmonic structures and fluctuations of spectral energy at both short and long-term time scales. In this work, we attempt to measure the presence or absence of speech by exploiting spectral cues in the auditory spectrogram of the noisy signal. These cues are related to various aspects of the human speech production process and are reflected in the spectral shape, the spectro-temporal modulations, the voicing character and the long-term spectral variability of speech. For each cue, we will extract a one-dimensional probability stream which can robustly discriminate between speech and non-speech sounds under various noise conditions.

The feature streams are contextually expanded over long-span time windows, normalized and merged in order to be applied to a classifier MLP for speech detection. The system is evaluated on the data corpus of the DARPA RATS[1] project which consists of highly degraded speech recordings that were transmitted over noisy radio communication channels [5]. With a low complexity in terms of computational cost and architectural design, the proposed VAD achieves a performance that is competitive to most recently developed state-of-the art speech detection systems.

The remainder of this paper is structured as follows. Section 2 presents the feature extraction module of the proposed VAD. Subsequent processing of the feature streams to obtain a robust frontend for speech detection are described in section 3. Experimental results are given in section 4. Conclusions and future work are discussed in section 5.

2. Feature extraction

In order to make a robust decision on the presence of speech in audio recordings, we will rely on multiple information streams that are extracted from the noisy signal. Each of these streams measures a distinctive property of the target speech that is relatively invariant under various adverse acoustical situations and which can discriminate the speech from concurrent background sources. In this section, we will examine a set of four feature streams which are derived from analysis windows with different sizes, depending on their optimality of representing the underlying speech attribute.

2.1. Spectral shape

In its physical form, human speech is a pressure waveform generated when the exhaled airflow from the lungs passes through the vocal folds and vocal tract. Articulators such as tongue, velum, jaw and lips, are controlling the vocal tract cavity shape and are responsible for the production of specific speech sound segments called phones. The shape of the vocal tract at a

[1] Robust Automatic Transcription of Speech

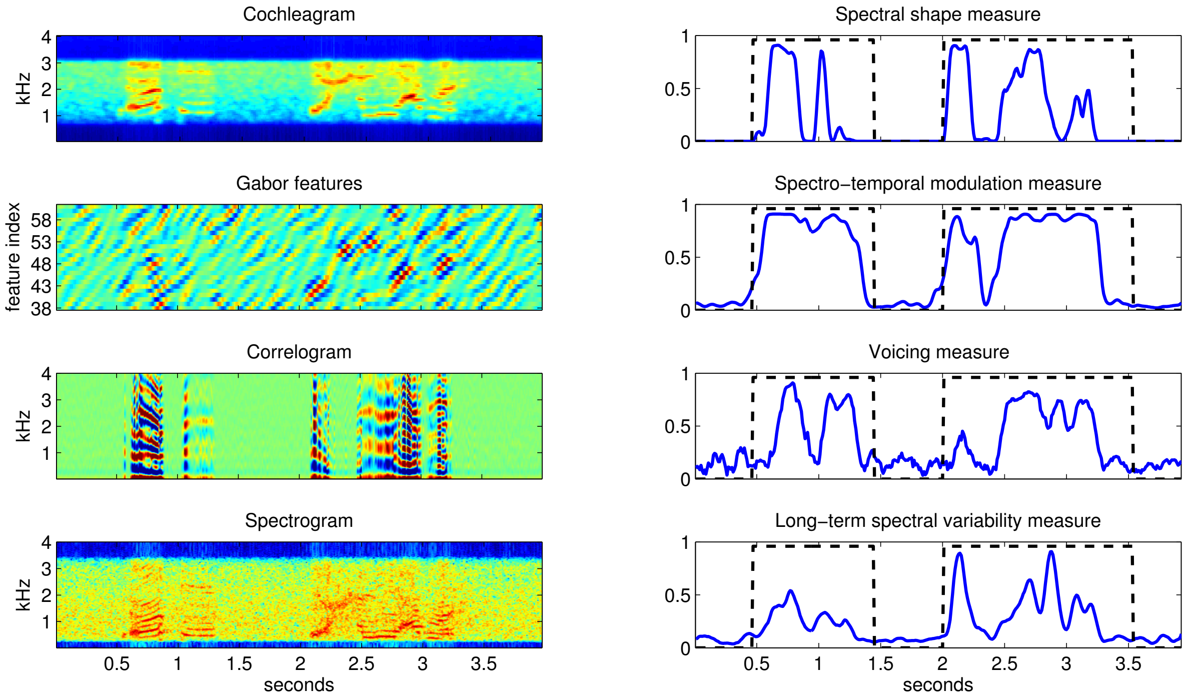


Figure 1: Spectral representations (left) and their corresponding feature measure streams (right) which are used in the proposed VAD-frontend. The dashed line in the right column plots indicates the annotated speech regions in the audio segment.

certain time instance is reflected in the envelope of the short time power spectrum of the signal. The popularity of standard feature representations such as Mel-Frequency Cepstral Coefficients (MFCC) [6], Perceptual Linear Prediction (PLP) coefficients [7], Gammatone Frequency Cepstral Coefficients (GFCC) [8], Power Normalized Coefficient (PNCC) [9], are largely explained by the psycho-acoustically relevant accuracy in which they succeed to represent the spectral envelope. In this paper, we opt the usage of GFCC features for their robustness properties as demonstrated in [10, 11]. GFCC feature vectors are derived by applying a DCT transform to the cochleagram of the audio signal, which is obtained by an auditory filterbank using Gammatone filters.

In this paper, 24-dimensional GFCC features are extracted using a 64 Gammatone filterbank, the outputs of which are temporally integrated by using an Hanning window with a time span of 200 ms and a 10 ms frame shift. The length of the analysis window is a factor 10 times longer than the conventional choice of 20-30 ms. Experiments not reported here, have shown that the use of a longer window for the task of robust speech detection is a better choice in the trade-off between smoothing out the spectral content over successive phonemes in time - which is less critical in VAD as it is for phoneme classification or speech recognition tasks - and the improved frequency resolution in the spectral domain. This yields a more salient representation of important acoustic cues, such as formants tracks.

Next, we derive a one-dimensional time-varying stream which measures the probability that speech is present in the observed spectral content of the audio signal. To this end, a MLP was trained on the speech/non-speech labeled training data of degraded speech and the feature stream was subsequently computed from the output posteriors. This neural network based approach of extracting data-driven features from the acoustic representation of speech was initially proposed by TRAP features [12] and since then many variants of data-driven MLP-based features have been derived. Previous work has already shown the discriminative properties of these features when applied to robust ASR [13, 14] and VAD [4].

In this paper, a three-layer MLP is trained on the GFCC input stream augmented with first-order derivatives. The hidden layer consists of 24 units, chosen as such that each hidden unit corresponds to a Gammatone filter output. After the MLP is trained, its two outputs states generate a posterior probability for the presence or non-presence of speech, respectively. By taking the ratio of both output posteriors, all relevant and redundant content of the spectral shape is then projected onto a single stream that contains confidence information about speech activity over time. This stream is shown at the top right in Figure 1 and the corresponding cochleagram representation from which it is derived is shown at the top left. The audio segment of 3.5 sec that was used here for visualization purposes, belongs to the RATS corpus [5] and was randomly selected from the set of recordings that were retransmitted through channel A.

2.2. Spectro-temporal modulations

The dynamic character of a speech signal originates from the complex interaction of the articulators during speech production. The changes in the vocal tract shape are responsible for the successive articulation of different phonetic sounds and are represented by temporal variations in the spectral energy of speech.

A variety of features have been designed to capture these acoustic modulations patterns of speech by means of wavelet filtering, such as the modulation spectrogram [15], RASTA [16], Gabor features [17] and cortical features [18], to name a few. All these features operate on longer time scale windows, typically of the order of 100 ms, a value that is defined by the wavelet filter size. It has been shown that when the wavelets are specifically tuned to appropriate spectral scales and temporal rates, the extracted features are able to accurately capture the acoustic modulations of speech and could add to the robustness of speech recognition [19, 20, 21] and detection [3, 4].

The movements of the articulators are continuous and restricted by the vocal tract physiology, which imposes constraints to the spectral scales and temporal rates of speech. Hence, one can assume that the temporal and spectral modulations of speech are modelled by a confined subspace in the total parameter space spanned by the dimensions of the spectro-temporal feature vector.

Similar to the previous section, a MLP classifier is trained to discriminate the temporal variations of spectral speech energy from those that originate from other sounds sources. In this paper, Gabor features are extracted and applied to the input layer of the MLP. These features were computed on the 8 kHz downsampled version of the audio signal to make optimal use of the lower narrowband in which speech reveals most prominent cues in spectro-temporal modulations. In this work, we only retained the filter outputs corresponding to the temporal modulation frequencies at 0 Hz and 6.2 Hz as inspired by [22, 23]. After critical subsampling [24], we leads to a 92-dimensional feature vector which is subsequently encoded to a posterior signal by a MLP with empirically chosen hidden layer size of 32 units. The left panel at the second row of Figure 1 shows the Gabor feature representation of the audio segment corresponding to the log-Mel spectrogram that was processed by a Gabor filter with 6.2 Hz temporal modulation frequency and temporal modulation frequency of 0.12 cycle/octave [25]. The corresponding posterior stream of the MLP is given at the right panel of the second row of Figure 1.

2.3. Harmonicity

During voiced speech periods, speech is characterized by strong periodicity due to the quasi periodic vibrations of the vocal folds which are reflected in the spectrum by the presence of a fundamental frequency or pitch and its harmonics. Although this periodicity is readily present in the short-time Fourier transform from which spectral shape features of section 2.1 are derived, it has been integrated out by the auditory filterbank and the dimensionality compression. Features explicitly measuring the voicing state of speech, have the potential to contribute to the noise robustness of ASR [26, 27] and VAD [28, 29]. Therefore, a voicing stream will be derived next to provide an additional cue to the speech/non-speech decision.

As described in [30], a correlogram is a time-frequency representation that succeeds in robustly revealing the presence of periodicity in an audio signal. By finding the peaks in the correlogram that arise from the presence of the pitch frequency, a probability measure for the voicing nature of the speech can be derived as follows. First, we estimate the fundamental frequency by the subharmonic summation method of [31] which suppresses doubling and halving errors and has shown its robustness in previous work [32]. The target pitch value was here confined to the frequency range from 50 Hz to 800 Hz. Next, we compute the correlogram by applying the autocorrelation function on the pre-emphasized, framed and Hamming windowed signal using a frame shift of 10 ms and a frame length of 25 ms. A voicing measure is subsequently derived at each time as the autocorrelation value evaluated at the estimated pitch period normalized by the signal energy, i.e. the autocorrelation value at zero lag. The third row of Figure 1 shows the correlogram at the left and the derived voicing stream is plotted at the right.

2.4. Long-term spectral variability

A fourth stream of information aims to model the variability of speech over a long-term window. Phones in different languages have durations from 10 ms to 200 ms and are uttered during a normal conversation at a rate of three to seven syllables per second [33] with formant characteristics that vary over time. This phoneme switch causes additional variability in the audio signal due to the presence of speech.

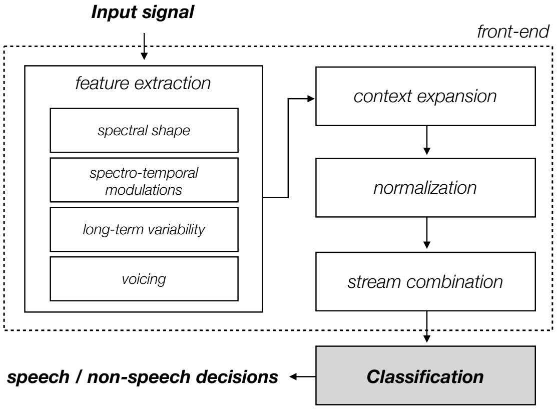


Figure 2: Overview of pre-processing steps in the proposed robust frontend for VAD, represented by the dashed line.

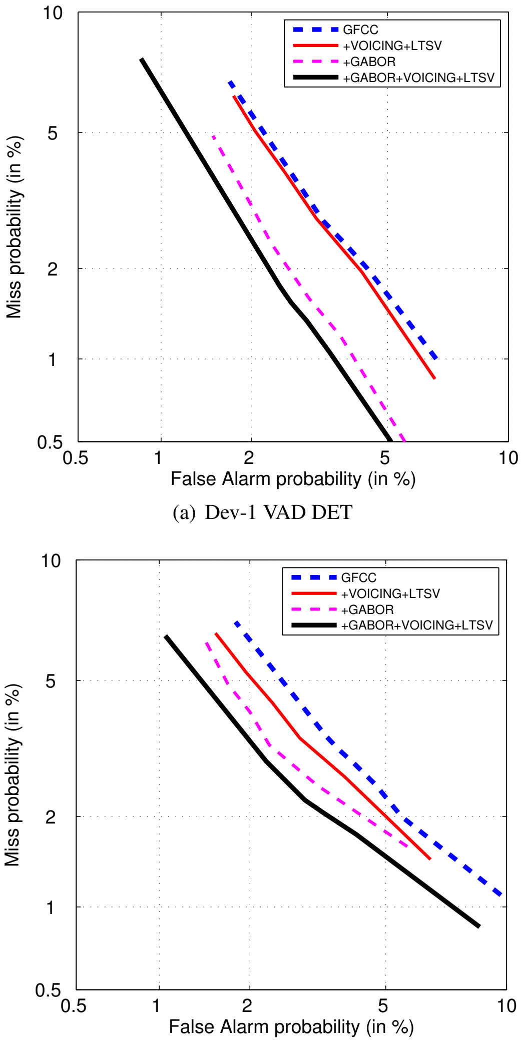
For this purpose, we have used the long-term signal variability (LTSV) measure of [34] to specifically capture the variations of speech caused by the successive generation of phones. The LTSV is a one-dimensional signal defined as the variance of the entropy measured over all frequency bins of the normalized short-time spectrogram. When applied to the problem of voice activity detection, this feature has been shown to have robust performance in stationary noise conditions or slowly varying noise conditions over long-term windows. A more detailed discussion on their implementation can be found in [34].

An important aspect in the derivation of the LTSV is a smoothing of the spectrogram to get better estimates of the stationary noise. We have used spectral smoothing of 100 ms and we estimate the spectral variability over a 500 ms interval from a spectrogram computed at 20 ms window and 10 ms shift. This spectrogram is shown at the bottom left panel of Figure 1, while the LTSV measure is plotted in the bottom right panel.

3. Context expansion, normalization and stream combination

The features streams discussed in the previous section are now recasted in context-probability streams, where *context* refers to the time span over which we integrate the information from the probability measures. This expansion of context information is done for each one-dimensional stream, and is achieved by applying a DCT transform on all samples of a long-term (moving) time window centered around the frame of interest. This way, the successive measures of probability information over the window are compressed to the set of low order DCT coefficients. An optimal window duration lies in the range of 70-120 frames, corresponding to a duration of the order of a second. We found that retaining only the first 5 DCT components is sufficient to extract most relevant context information from those windows.

The resulting 5-dimensional stream vectors are then variance normalized using a global variance vector computed on the training data, which gave us a slightly better performance compared to per-file normalization. Stream combination is then simply obtained by stacking the four context-expanded features stream into a single 20-dimensional frame vector. Finally, a



(a) Dev-1 VAD DET

(b) Dev-2 VAD DET

Figure 3: DET curves on Dev-1 and Dev-2 for different feature combination in the proposed VAD frontend.

MLP classifier was trained on these feature vectors using a hidden layer size equal to the feature dimension. Speech segments are detected by thresholding the ratio of both MLP outputs. Figure 2 shows an overview of the proposed VAD.

4. Experimental results

The performance of the proposed VAD was evaluated on the DARPA RATS corpora [5] that was collected by the Linguistic Data Consortium (LDC). The audio data consists of speech data of multiple languages that was transmitted through eight noisy radio communication channels, labelled from A to H. Two official development sets, denoted by Dev-1 and Dev-2, were provided for "dry run" evaluations[2].

In this paper, a per channel training approach was followed to train the MLPs of sections 2.1 and 2.2, as well as for the MLP that was used for classification. Averaged over all channels, only 30 hours of training data was used from the total provided amount of ~300 hours. The system was tested on both development sets and error computation was performed using the scoring engine software provided by SAIC using the evaluation protocol described in [35].

Set	System	Pfa at PMiss=4%	Pmiss at Pfa=1.5%	EER
Dev-1	GFCC	2.40	5.60	3.15
	+ VOICING	2.20	5.30	3.10
	+ LTSV	2.10	5.10	3.00
	+ GABOR	1.20	3.50	2.40
	all combined	**1.00**	**3.10**	**2.15**
Dev-2	GFCC	3.10	7.10	3.55
	+ VOICING	2.90	6.90	3.40
	+ LTSV	3.00	7.00	3.50
	+ GABOR	2.20	6.20	2.90
	all combined	**1.95**	**5.85**	**2.70**

Table 1: Error results on Dev-1 and Dev-2 for different feature combinations in the proposed VAD frontend.

Figures 3(a) and 3(b) show the DET curves of respectively Dev-1 and Dev-2, where the false alarm probability (Pfa) and miss probability (Pmiss) were computed over all channels A-H, with the exception of channel D that was left out from the official evaluation on Dev-2. The DET-curves were obtained by multiple runs of the VAD system using different values of the threshold that was applied on the outputs of the classifier MLP as discussed in section 3. Baseline numbers were generated by the isolated use of the posteriors streams derived from GFCCs as discussed in section 2.1. The accuracy was improved by stream combinations with the Gabor feature derived posterior stream, the voicing measure and the LTSV. The gain in accuracy obtained by the combination with the latter two streams was more effective on Dev-2 than on Dev-1, but was less significant than combining GFCCs with Gabor features. The combined use of all features yields further relative improvements on Dev-1 and Dev-2 compared to GFCC+GABOR with 10% and 7%, respectively, resulting in a baseline improvement of approximately 32% on both sets.

To conform with the RATS Phase 2 evaluation, Table 1 presents the false alarm rate at 4% Pmiss, the miss rate at 1.5% Pfa and the Equal Error Rate (EER), all expressed in error percentage, for different feature combinations in the proposed VAD frontend.

5. Conclusions

A noise robust frontend for VAD was presented that extract four probability streams, each measuring the presence of a different attribute of speech in the audio signal, i.e. the spectral shape, the spectro-temporal modulations, the harmonicity and the long-term spectral variability. By training a classifier on the combination of the contextually expanded versions of these streams, a high accuracy in robust speech detection was achieved, and this despite the low dimensionality of the features. Future work involves to investigate the robustness of alternative feature representations and the use of more advanced neural nets, e.g. deep neural networks, to either extract more accurate posterior streams in the VAD frontend and to further increase the performance of the speech/non-speech classifier.

6. Acknowledgements

This research was supported by the Defense Advanced Research Projects Agency (DARPA) and the National Science Foundation (NSF)

[2] Annotation Delivery V3 was used in training and testing. 1st and 2nd incremental releases for Dev-1, 3rd incremental release for Dev-2.

7. References

[1] Y. Kida and T. Kawahara, “Voice activity detection based on optimally weighted combination of multiple features,” in *Proc. Interspeech*, 2005, pp. 2621–2624.

[2] M. Fujimoto, K. Ishizuka, and T. Nakatani, “A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme,” in *Proc. ICASSP*, 2008.

[3] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Vesel, and P. Matejka, “Developing a speech activity detection system for the DARPA RATS program.” in *Proc. Interspeech*, 2012.

[4] A. Thomas, S. Mallidi, T. Janu, H. Hermansky, N. Mesgarani, X. Zhou, S. Shamma, T. Ng, B. Zhang, L. Nguyen, and S. Matsoukas, “Acoustic and data-driven features for robust speech activity detection,” in *Proc. Interspeech*, 2012.

[5] K. Walker and S. Strassel, “The rats radio traffic collection system,” in *Odyssey 2012-The Speaker and Language Recognition Workshop*, 2012.

[6] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognitions in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[7] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[8] Y. Shao, S. Srinivasan, and D. Wang, “Incorporating auditory feature uncertainties in robust speaker identification,” in *Proc. ICASSP*, 2002, pp. 277–280.

[9] C. Kim and R. M. R. M. Stern, “Power-normalized coefficients (PNCC) for robust speech recognition,” in *Proc. ICASSP*, 2012.

[10] Y. Shao, Z. Jin, D. L. Wang, and S. Srinivasan, “An auditory-based feature for robust speech recognition,” in *Proc. ICASSP*, 2009.

[11] R. Schluter, I. Bezrukov, H. Wagner, and H. Ney, “Gammatone features and feature combination for large vocabulary speech recognition,” in *Proc. ICASSP*, 2007.

[12] H. Hermansky and S. Sharma, “TRAPS - classifiers of temporal patterns,” in *Proc. Interspeech*, Sydney, Australia, Nov. 1998, pp. 1003–1006.

[13] D. Ellis, R. Singh, and S. Sivadas, “Tandem acoustic modeling in large-vocabulary recognition,” in *Proc. ICASSP*, 2001.

[14] F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky, “Probabilistic and bottle-neck features for LVCSR of meetings,” in *Proc. ICASSP*, 2007.

[15] B. Kingsbury, N. Morgan, and S. Greenberg, “Robust speech recognition using the modulation spectrogram,” *Speech Communication*, vol. 25, pp. 117–132, Aug. 1998.

[16] H. Hermansky and S. Sharma, “Temporal patterns (TRAPs) in ASR of noisy speech,” in *Proc. ICASSP*, vol. 1, Phoenix, Arizona, U.S.A., Mar. 1997, pp. 289–292.

[17] K. M., “Spectro-temporal gabor features as a front end for ASR,” in *Proc. Forum Acusticum Sevilla*, 2002.

[18] N. Mesgarani, M. Slaney, and S. Shamma, “Discrimination of speech from non-speech based on multiscale spectro-temporal modulations,” *IEEE Transactions of Audio, Speech and Language Processing*, 2006.

[19] H. Hermansky and P. Fousek, “Multi-resolution RASTA filtering for TANDEM-based ASR,” in *Proc. Interspeech*, Lisbon, Portugal, Oct. 2005, pp. 361–364.

[20] S. Zhao and N. Morgan, “Multi-stream spectro-temporal features for robust speech recognition,” in *Proc. Interspeech*, 2008, pp. 898–901.

[21] T. S., S. Ganapathy, and H. Hermansky, “Tandem representations of spectral envelope and modulation frequency features for ASR,” in *Proc. Interspeech*, 2009, pp. 2955–2958.

[22] T. Tsai and N. Morgan, “Longer features: They do a speech detector good,” in *Proc. Interspeech*, 2012.

[23] M. Van Segbroeck and S. Narayanan, “A robust frontend for ASR: combining denoising, noise masking and feature normalization,” in *Proc. ICASSP*, 2013.

[24] B. T. Meyer, S. V. Ravuri, M. R. Schadler, , and N. Morgan, “Comparing different flavors of spectro-temporal features for ASR,” in *Proc. Interspeech*, 2011, pp. 1269–1272.

[25] B. Meyer, S. Ravuri, M. Schädler, and N. Morgan, “Comparing different flavors of spectro-temporal features for ASR,” in *Proc. Interspeech*, 2011, pp. 1269–1272.

[26] D. L. Thomson and R. Chengalvarayan, “Use of voicing features in hmm-based speech recognition,” *Speech Communication*, vol. 37, no. 3, pp. 197–211, 2002.

[27] A. Zolnay, R. Schulter, and H. Ney, “Extraction methods of voicing feature for robust speech recognition,” in *Proceedings of EUROSPEECH*, 2003, pp. 497–500.

[28] T. Fukuda, O. Ichikawa, and M. Nishimura, “Long-term spectrotemporal and static harmonic features for voice activity detection,” *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 5, pp. 834–844, 2010.

[29] E. Chuangsuwanich and J. Glass, “Robust voice activity detector for real world applications using harmonicity and modulation frequency,” in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[30] S. Granqvist and B. Hammarberg, “The correlogram: A visual display of periodicity,” *The Journal of the Acoustical Society of America*, vol. 114, 2003.

[31] H. Van hamme, “Robust speech recognition using cepstral domain missing data techniques and noisy masks,” in *Proc. ICASSP*, Montreal, Canada, May 2004, pp. 213–216.

[32] M. Van Segbroeck, “Robust large vocabulary continuous speech recognition using missing data techniques,” Ph.D. dissertation, K.U.Leuven, ESAT, Jan. 2010.

[33] J. M. Pickett, *The sounds of speech communication: A primer of acoustic phonetics and speech perception*. University Park Press Baltimore, 1980.

[34] 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.

[35] H. Goldberg, *RATS Evaluation Plan*, 2011, https://rats.saic.com/index.php/Evaluation_Protocols.