Multi-band long-term signal variability features for robust voice activity detection

Andreas Tsiartas[1], *Theodora Chaspari*[1], *Nassos Katsamanis*[1], *Prasanta Ghosh*[2], *Ming Li*[1], *Maarten Van Segbroeck*[1] , *Alexandros Potamianos*[3], *Shrikanth S. Narayanan*[1]

[1]Signal Analysis and Interpretation Lab, Ming Hsieh Electrical Engineering, University of Southern California, Los Angeles ,USA
[2] IBM Research India, New Delhi, India
[3] ECE Department, Technical University of Crete, Chania, Greece
{tsiartas,chaspari}@usc.edu, nkatsam@sipi.usc.edu, prasantag@gmail.com, mingli@usc.edu, maarten@sipi.usc.edu, potam@telecom.tuc.gr, shri@sipi.usc.edu

Abstract

In this paper, we propose robust features for the problem of voice activity detection (VAD). In particular, we extend the long term signal variability (LTSV) feature to accommodate multiple spectral bands. The motivation of the multi-band approach stems from the non-uniform frequency scale of speech phonemes and noise characteristics. Our analysis shows that the multi-band approach offers advantages over the single band LTSV for voice activity detection. In terms of classification accuracy, we show 0.3%-61.2% relative improvement over the best accuracy of the baselines considered for 7 out 8 different noisy channels. Experimental results, and error analysis, are reported on the DARPA RATS corpora of noisy speech.

Index Terms: noisy speech data, voice activity detection, robust feature extraction

1. Introduction

Voice activity detection (VAD) is the task of classifying an acoustic signal stream into speech and non-speech segments. We define a speech segment as a part of the input signal that contains the speech of interest, regardless of the language that is used, possibly along with some environment or transmission channel noise. Non-speech segments are the signal segments containing noise but where the target speech is not present. Manual or automatic speech segment boundaries are necessary for many speech processing systems. In large-scale or real-time systems, it is neither economical nor feasible to employ human labor (including crowd-sourcing techniques) to obtain the speech boundaries as a key first step. Thus, the fundamental nature of the problem has positioned VAD as a crucial preprocessing tool to a wide range of speech applications, including automatic speech recognition, language identification, spoken dialog systems and emotion recognition.

Due to the critical role of VAD in numerous applications, researchers have focused on the problem since the early days of speech processing. While some VAD approaches have shown robust results using advanced back-end techniques and multiple system fusion [1], the nature of VAD and diversity of environmental sounds suggests the need of robust VAD front-ends. Various signal features have been proposed for separating speech and non-speech segments in the literature. Taking into account short-term information ranging from 10ms to 40ms, various researchers [2, 3, 4] have proposed energy-based features. In addition to energy features, researchers have used zero-crossing rate [5], wavelet-based features [6], correlation coefficients [7] and negentropy [8, 9] which has been shown to perform well in low SNR environments. Other works have used long-term features in the range of 50-100ms [10] and above 150ms [11]. Long-term features have been shown to perform well on noisy speech conditions under a variety of environmental noises. Notably, they offer theoretical advantages for stationary noise [11] and capture information that short-term features lack.

The long-term features proposed in the past focus on extracting information from a two-dimensional (2-D) time-frequency window. Limiting the extracted feature information from 2-D spectro-temporal windows fails to capture some useful auditory spectrum properties of speech. It is well known that the human auditory system utilizes a multi-resolution frequency analysis with non-linear frequency tiling reflected in the Mel-scale [12] representation of audio signals. Mel-scale provides an empirical frequency resolution that approximates the frequency resolution of the human auditory system. Inspired by this property of the human auditory system and the fact that the discrimination of various noise types can be enhanced at certain different frequency levels, we expand the LTSV feature proposed in [11] to use multiple spectral resolution.

We compare the proposed approach with two baselines: the MFCC [13] features and the single-band (1-band) long-term signal variability (LTSV) [11] and show significant performance gains. Unlike [14] where standard MFCC features have been used for this task and experimented with various back-end systems, we use a fixed back-end and focus only on comparing features for the VAD task using a K-Nearest Neighbor (K-NN) [15] classifier. We perform

our experiments on the DARPA RATS data [16] for which an off-line batch processing is required.

2. Proposed VAD Features

In this section, we describe the proposed multi-band extension of the LTSV feature introduced in [11]. LTSV has been shown to have good discriminative properties for the VAD task especially in high SNR noise conditions. We try to exploit this property by capturing dynamic information of various spectral bands. For example, impulsive noise which degrades the performance of LTSV features is often limited to certain band regions in the spectrum. The aim of this work is to investigate the use of a multi-band approach to capture speech variability across different bands. Also, speech variability might be exemplified in different regions for different phonemes. Thus, a multi-band approach could have advantages over the 1-band LTSV.

2.1. Frequency smoothing

The low pass filtering process is important for the LTSV family of features because it removes the high frequency noise on the spectrogram. Also, it was shown that it improves robustness in stationary noise [11], such as white noise.

Let $S(\hat{f}, j)$ represent the spectrogram, where $\hat{f}$ is the frequency bin of interest and j is j^{th} frame. As in [11], we smooth S using a simple moving average of window of size M (assumed to contain even number of samples for our notation) as follows:

$$S_M\left(\hat{f}, j\right) = \frac{1}{M} \sum_{k=j-\frac{M}{2}}^{j+\frac{M}{2}-1} S\left(\hat{f}, k\right) \tag{1}$$

2.2. Multi-Band LTSV

In order to define multiple bands, we need a parameterization to set the warping of the spectral bands. For this purpose, we use the warping function from the warped discrete Fourier transform [17] which is defined as:

$$F_W(f, \alpha) = \frac{1}{\pi} \arctan\left(\frac{1+\alpha}{1-\alpha} \tan(2\pi f)\right) \tag{2}$$

where f represents the frequency to be warped starting from uniform bands and α is the warping factor and takes values in the range $[-1, 1]$. A warping factor of -1 implies a high resolution for high frequencies and, of 1 implies a high resolution for low frequencies. A warping factor of 0 results in uniform bands.

To define the multi-resolution LTSV, we first define the normalized spectrogram across time over an analysis window of R frames as:

$$S_R\left(\hat{f}, j\right) = \frac{S_M\left(\hat{f}, j\right)}{\sum_{k=j-\frac{R}{2}}^{j+\frac{R}{2}-1} S_M\left(\hat{f}, k\right)} \tag{3}$$

Hence, we define the multi-band LTSV feature of window size R and warping factor α at the i^{th} frequency band and j^{th} frame as:

$$L_i(\alpha, R, j) = V_{\hat{f} \in F_i}\left(\sum_{k=j-\frac{R}{2}}^{j+\frac{R}{2}-1} S_R\left(\hat{f}, k\right) \log\left(S_R\left(\hat{f}, k\right)\right)\right) \tag{4}$$

V is the variance function defined as:

$$V_{f \in F}\left(a(f)\right) = \frac{1}{|F|} \sum_{f \in F}\left(a(f) - \frac{1}{|F|} \sum_{f \in F} a(f)\right)^2$$

where $|F|$ is the cardinality of set F. The set F_i includes the frequencies $F_W(f, \alpha)$ for $f \in \left[\frac{N_s*(i-1)}{2N} \cdots \frac{N_s*i}{2N}\right]$, N is the number of bands to be included and N_s denotes the sampling frequency.

3. Experimental setup

To compare across the various features, we used a K-NN classifier for all the experiments. We used 70 hours of data from the RATS[1] corpus (dev1_v2 set) for training and 11 hours for testing for each channel; the RATS data comprises of speech data transmitted through eight different channels (A through H), resulting in varying signal qualities and SNRs. To optimize the parameters, we used a small set of 1 hour for training and a 1 hour development set for each channel. As a post-processing step, we applied a median filter to the output of the classifier to impose continuity on the local detection based output. For each experiment, we searched for the optimal K-NN neighborhood size $K \in [1 \cdots 100]$ and the optimal median filter length for various windows sizes ($[100, 300, 500, 700, 900]$ms). This optimization procedure was performed for each channel separately. We set as baselines the MFCC and 1-band LTSV features and compare against the proposed multi-band LTSV. We experimented with all A-H channels included in the RATS data set.

The test set results have been generated using the DARPA speech activity detection evaluation scheme [18] which computes the error at the frame level and considers the following:

- Does not score 200ms from the start/end speech annotation towards the speech frames.
- Does not score 500ms from the start/end speech annotation towards the non-speech frames.
- Converts to non-speech, speech segments less than 300ms.
- Converts to speech, non-speech segments less than 700ms.

4. EMPIRICAL SELECTION OF ALGORITHM PARAMETERS

In this section, we describe the pilot experiments we performed to choose the optimal parameters for the LTSV-based features. Fig. 1 shows the accuracy for channel A for all the parameters used to fine-tune the optimal LTSV features. To select the set of parameters, we run a grid

[1] www.darpa.mil/Our_Work/I2O/Programs/Robust_Automatic_Transcription_of_Speech_(RATS).aspx

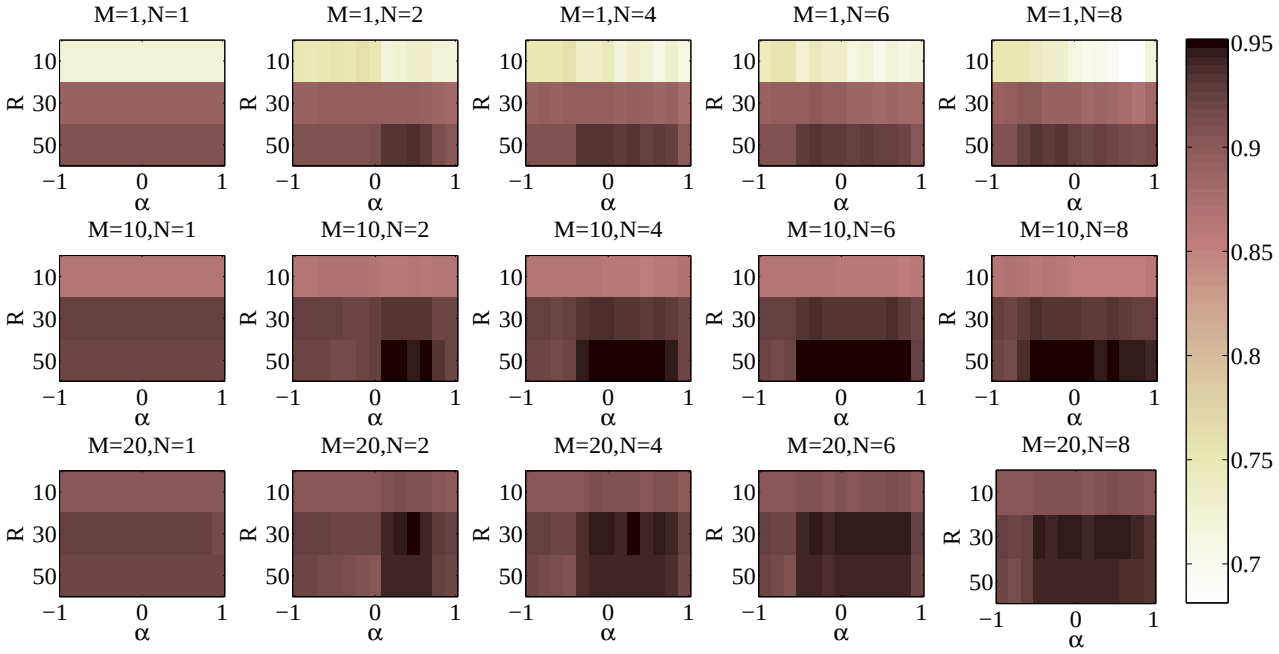

Figure 1: This figure shows the VAD frame accuracy for the development set of channel A for various parameters of the multi-band LTSV. R represents the analysis window length, M the frequency smoothing, α the warping factor and N the number of filters. The bar on the right represents the frame accuracy. This figure indicates that for channel A increasing the number of bands (N) improves the accuracy. Also, indicates that smoothing ($M \geq 100$) and analysis window (R) are crucial parameters for the multi-band LTSV as observed in the original LTSV [11].

search over a range of parameters for each channel separately. In particular, we experimented with 15 different warping factors uniformly in the range $[-0.95 \cdots 0.95]$. We also computed the spectrogram smoothing parameter M as defined in Sec. 2.1. $M = 1$ corresponds to no smoothing whereas $M = [100, 200]$ correspond to smoothing of 100 and 200ms, respectively. In addition, we searched different analysis window sizes $R = [100, 300, 500]$ms. The final parameter we experimented with was the number of bands $N = [1, 2, 4, 6, 8]$. Fig. 1 shows that for channel A the optimal number of filters is 6. The optimal values consist of warping factor $\alpha = 0.3$ with smoothing $M = 200$ms and analysis window $R = 300$ms. Channel A contains band-pass speech in the range 400-4000Hz. This might be one of the reasons a warping factor of 0.3 has been chosen for this channel. Smoothing M and analysis window R depend on how fast the noise varies with time. Very slow varying noise types, i.e. stationary noises can afford to have high values for M and R. However, if impulsive noises are of interest, smaller windows are preferable. The warping factor α depends on which frequency bands have prominent formants. For instance, if strong formants appear in low frequency ranges, values around 0.6 are preferable (i.e. close to Mel-scale).

For all pilot experiments, we have optimized K of K-NN using the Mahalanobis distance [19] and the median filter length. We have observed that a median filter of 700-900ms is best for most of the experiments. This suggests that extracting features with longer window lengths can further improve the accuracy.

5. Results and discussion

Fig. 2 shows the Receiver Operating Characteristics (ROC) curve between false alarm probability (Pfa) and miss probability (Pmiss) for the eight different channels of noisy speech and noise data considered. Channels A-D contain stationary channel noise but non-stationary environmental noise which imposes challenges for the 1-band LTSV. Channels G-H consist of varying channel and environmental noise, causing poor performance for the 1-band LTSV features with equal error rate (EER) exceeding 12%.

Poor classification results due to the non-stationarity of the noise can be improved using multi-band LTSV features. Multi-band LTSV features achieve the best performance compared to both baselines, except for channel C where MFCC has the lowest EER.

In addition, we did an error analysis of individual channels to investigate the cases for which the algorithm fails to classify correctly the two classes. On the miss side at the equal error rate (EER), a common error for all channels was due to the presence of filler words, laughter etc. Also, for channels D and E almost half of the errors contributing to the miss rate were due to background/degraded speech. Filler words have slower varying spectral characteristics than verbal speech. If noise has higher spectral variability than filler words, the LTSV features fail to discriminate them.

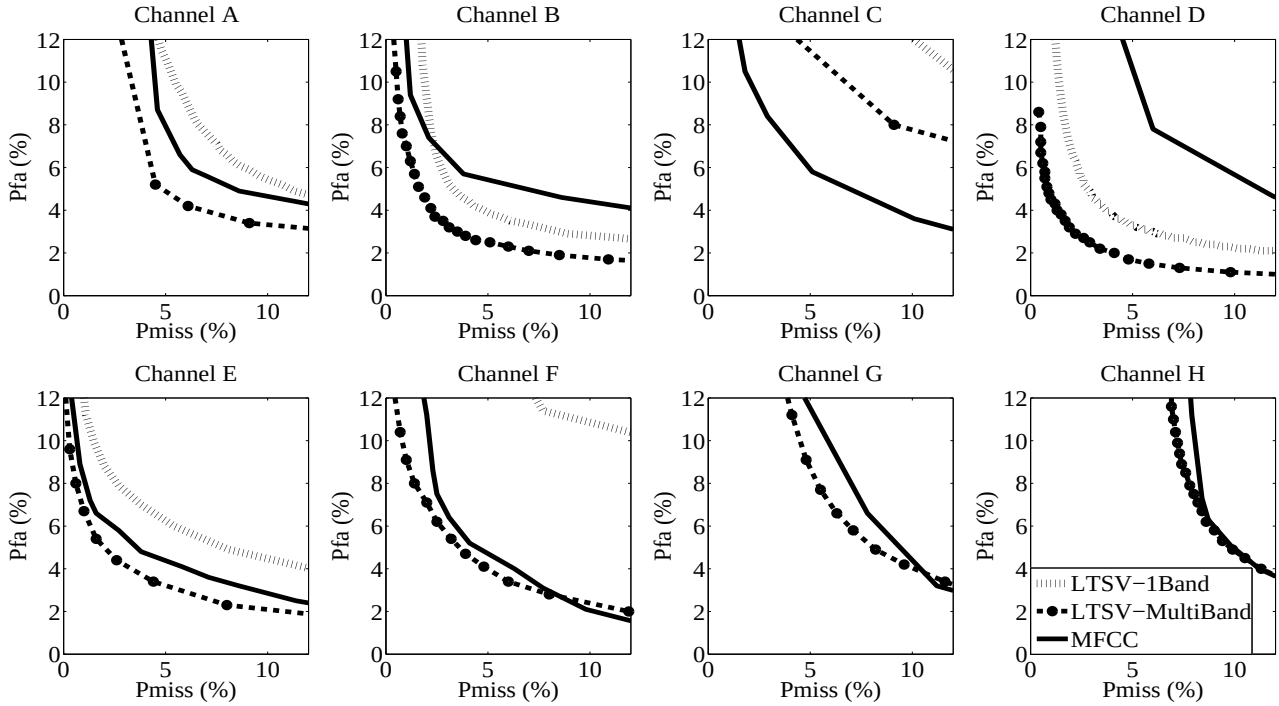

Figure 2: This figure shows the ROC curve of Pfa vs Pmiss for channels A-H of the multi-Band LTSV (LTSV-MultiBand) and the two baselines (1-band LTSV and MFCC). For channels G and H the 1-band LTSV ROCs are out of the boundaries of the plots, hence they do not appear in the figure. The same legend applies to all subfigures.

On the false alarm side, the error analysis at EER reveals that there were a variety of errors including background/robotic speech, filler words and kids background speech/cry. Such errors are expected since background speech shares the spectral variability characteristics of foreground speech; in fact, the classification of background speech by annotators is often based on semantics rather than low-level signal characteristics.

Apart from the speech-like sounds where the multi-band LTSV shows degraded performance, there are non-speech sounds that the multi-band LTSV failed to classify. In particular, false alarms (FA) in channels A,B,D,E and H have been associated with constant tones appearing at different frequencies over time and impulsive noises at varying frequencies. FA in channel C are composed of noise with spectral variability appearing at different frequencies with one strong frequency component up to 2500Hz and bandwidth greater than the speech formants bandwidth. The limited frequency discriminability (although improved in the multi-band version) is an inherent weakness of the LTSV features. Thus, for channel C, LTSVs performed very poorly, even worse than MFCC. FAs of multi-band LTSV in channel G stem from the variability of the channel and not the environmental noise.

Overall, the multi-band LTSV, performs better than the two baselines considered: the 1-band LTSV and MFCC. From the error analysis, we found that the multi-band LTSV not only retains the discrimination of the 1-band LTSV for stationary noises but also improves discrimination in noise environments with variability, even in impulsive noise cases where the 1-band LTSV fails. However, the multi-band LTSV fails to discriminate impulsive noises appearing at different frequencies over time. For speech miss errors, filler words/laughter are challenging for LTSV due to their lower spectral variability over long time relative to the actual speech. Finally, besides channel C where MFCC gives the best performance, the multi-band LTSV gives the best accuracy showing the benefits of capturing additional information using a multi-resolution LTSV approach.

6. Conclusion and future work

In this paper, we extended the LTSV [11] feature to multiple spectral bands for the voice activity detection (VAD) task. We found that the multi-band approach improves the performance in different noise conditions including impulsive noise cases in which the 1-band LTSV suffers. We compare the multi-band approach against two baselines: the 1-band LTSV and MFCC features and we found that we gain significantly in performance for 7 out of the 8 channels tested.

In future work, we plan to include delta features along with additional long-term and short-term features that capture the information the multi-band LTSV fails to capture. One aspect that needs further investigation is how to improve the accuracy at the fine-grained boundaries of the decision due to the long-term nature of the feature set. Also, it would be interesting to explore the potential of these features with various machine learning algorithms including deep belief networks.

7. References

[1] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K Vesely, and Matejka, "Developing a speech activity detection system for the DARPA RATS program," in *Proceedings of Interspeech*. Portland, OR, USA, 2012.

[2] Krishnan P. S. H., Padmanabhan R., and Murthy H. A., "Voice Activity Detection using Group Delay Processing on Buffered Short-term Energy.," in *Proc. of 13th National Conference on Communications*, 2007.

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

[4] Evangelopoulos G. and Maragos P., "Speech event detection using multiband modulation energy.," in *Proc. Interspeech*, Lisbon, Portugal, September 2005, vol. 1, pp. 685–688.

[5] Kotnik B., Kacic Z., and Horvat B., "A multiconditional robust front-end feature extraction with a noise reduction procedure based on improved spectral subtraction algorithm.," in *Proc. 7th EUROSPEECH*, Aalborg, Denmark, 2001, pp. 197–200.

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

[7] Craciun A. and Gabrea M., "Correlation coefficient-based voice activity detector algorithm.," in *Canadian Conference on Electrical and Computer Engineering*, May 2004, vol. 1, pp. 1789–1792.

[8] Renevey P. and Drygajlo A., "Entropy based voiced activity detection in very noisy conditions.," in *Proc. EUROSPEECH*, Aalborg, Denmark, September 2001, pp. 1887–1890.

[9] Prasad R., Saruwatari H., and Shikano K., "Noise estimation using negentropy based voice-activity detector.," in *47th Midwest Symposium on Circuits and Systems*, July 2004, vol. 2, pp. 149–152.

[10] Ramırez J., Segura J.C., Benıtez C., De La Torre A., and Rubio A., "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3, pp. 271–287, 2004.

[11] Ghosh P., Tsiartas A., and Narayanan S., "Robust Voice Activity Detection Using Long-Term Signal Variability," *IEEE Transactions Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 600–613, 2011.

[12] Stevens S.S., Volkmann J., and Newman EB, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

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

[14] Kinnunen T., Chernenko E., Tuononen M., Fränti P., and Li H., "Voice activity detection using MFCC features and support vector machine," in *Int. Conf. on Speech and Computer (SPECOM07), Moscow, Russia*, 2007, vol. 2, pp. 556–561.

[15] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification (2nd Edition)*, Wiley-Interscience, 2000.

[16] K. Walker and S. Strassel, "The RATS Radio Traffic Collection System," in *Odyssey 2012-The Speaker and Language Recognition Workshop*. Singapore, 2012.

[17] Makur A. and Mitra S.K., "Warped discrete-Fourier transform: Theory and applications," *Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on*, vol. 48, no. 9, pp. 1086–1093, 2001.

[18] P. Goldberg, "RATS evaluation plan," in *SAIC, Tech. Rep.*, 2011.

[19] P.C. Mahalanobis, "On the generalized distance in statistics," in *Proceedings of the National Institute of Sciences of India*. New Delhi, 1936, vol. 2, pp. 49–55.