Robust Phonetic Segmentation Using Multi-taper Spectral Estimation for Noisy and Clipped Speech

Bhavik Vachhani
TCS Innovation Labs, Mumbai
Email: bhavik.vachhani@tcs.com

Chitralekha Bhat
TCS Innovation Labs, Mumbai
Email:bhat.chitralekha@tcs.com

Sunil Kopparapu
TCS Innovation Labs, Mumbai
Email: sunilkumar.kopparapu@tcs.com

***Abstract*—Robust phonetic segmentation is extremely important for several speech processing tasks such as phone level articulation analysis and error detection, speech synthesis, and annotation. In this paper, we present an unsupervised phonetic segmentation approach and its application to noisy and clipped speech such as mobile phone recordings. We propose a multi-taper-based Perceptual Linear Prediction (PLP) speech processing front-end, together with Spectral Transition Measure (STM) and a novel post-processing technique, to improve over the baseline STM technique. Performance of the proposed technique has been evaluated using precision, recall and F-score measures. Experimental results show an absolute improvement of 11% for TIMIT and 18% for Hindi speech data (clean) over the baseline approach. Significant improvement in phonetic segmentation was observed for noisy speech - simulated as well as mobile phone recordings.**

***Keywords*—Spectral Transition Measure, Multi-taper, Perceptual Linear Prediction, Clipping, Babble Noise**

I. Introduction

Phonetic segmentation is the process of breaking down a given speech utterance into its basic units, namely phones. In [1], the author emphasises the importance of accuracy in determining the phone boundaries, and the importance of Spectral Transition Measure (STM) as a metric for syllable perception. A robust phonetic segmentation technique is essential for speech tasks such as phone level articulation analysis and error detection, speech synthesis, transcription annotation etc. Several supervised and unsupervised phonetic segmentation techniques have been proposed in the literature. In [2]–[4] authors, address the accuracy of phonetic segmentation using a two-step approach; the initial estimate is obtained using an Automatic Speech Recognizer (ASR) and the boundaries are refined using specific boundary level acoustic models [2], using regression tree [3] and using acoustic-phonetic knowledge [4]. A similar approach is stated in [5], wherein the boundaries are improved using powerful statistical models conditioned on phonetic context and duration features. In [6], authors propose phonetic segmentation using a combination of phone posterior features and auditory attention features. Despite excellent results, supervised methods are limited by the availability and quality of speech corpora. Unsupervised phonetic segmentation using Maximum Marginal Clustering (MMC), a kernel method, is shown in [7]. In [8], phoneme boundaries are detected using a two-layered Support Vector Machine (SVM)-based system using frequency synchrony and average signal levels computed using a biomimetic model of the human auditory processing. A time-constrained agglomerative clustering algorithm to find the optimal segmentations is reported in [9]. These techniques have been reported to perform well on clean speech.

Robust phonetic segmentation of speech recordings on smart phones and tablets is valuable to building speech-based applications. Specifically, our objective is to assess articulation errors in the Hindi language, at phone level using mobile applications. Two major concerns when building such applications are (1) Smart phone and tablet recordings are susceptible to environmental noise and clipping. (2) Hindi is a low resource language. In such a scenario, an unsupervised and robust phonetic segmentation approach is desirable. Hence, we adopt Spectral Transition Measure (STM), which is closely correlated with phonetic boundaries [10] and can be exploited to automatically obtain phonetic boundaries in an unsupervised manner. STM based methods have been recently used to analyze the effectiveness of Perceptual Linear Prediction Cepstral Coefficients (PLPCC) [11] based features in speech synthesis [12].

The main contribution of this paper is a robust phonetic segmentation technique for mobile phone and tablet recordings of speech contaminated with environmental noise or clipping. The proposed phonetic segmentation technique uses (a) Multi-taper-PLPCC based speech front-end, together with STM and (b) novel data driven post-processing. Multi-taper spectral analysis introduced by Thomson [13] has been used to determine spectral transition since it gives an enhanced spectral estimation as compared to single taper even under noisy conditions. Multi-taper MFCC features have recently found application in speech related tasks [14]–[17]. The organization of the paper is as follows: Section II describes the STM-based phonetic segmentation approach and its limitations. Section III discusses the proposed unsupervised phonetic segmentation technique. Section IV describes the experimental setup. Section V discusses the evaluation results. We conclude in Section VI.

II. STM-based phonetic segmentation

Our focus is to achieve robust phonetic segmentation in an unsupervised manner for Hindi speech. The STM algorithm for phonetic segmentation was originally proposed using MFCC features [10]. In [12], authors suggest an improved STM-based phonetic segmentation technique using PLPCC features. STM algorithm is elaborated in this section.

Let $s_1, s_2, \ldots, s_m$ be the m frames of a speech signal, such that each frame is of duration $30\,ms$ with an overlap of $20\,ms$ between consecutive frames. Let

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

$f = [\bar{f}_1, \bar{f}_2, \ldots, \bar{f}_m]$ be the spectral features of the speech signal, where $\bar{f}_k$ is the spectral feature vector (dimension D). The rate of change of spectral feature is defined as

$$\bar{a}(m) = \frac{\sum_{k=-n}^{n} \bar{f}_{(k+m)} \cdot k}{\sum_{k=-n}^{n} k^2} \quad \text{where } \bar{a} = [a_1, a_2, \ldots, a_D] \quad (1)$$

STM is defined as the mean-squared value of the rate of change of spectral features as shown in Equation 2.

$$STM(m) = \frac{||\bar{a}(m)^2||}{D} \quad (2)$$

The locations of local maxima obtained in the STM contour indicate a spectral transition and a possible phone boundary.

The limitations of the baseline algorithms for clean speech signal are: (a) Spurious boundaries are introduced in the silence (e.g. non-speech, click) part of the speech signal [10] and (b) Over-segmentation in the long vowel regions and diphthongs [10]. Distortions like clipping and noise further degrade the performance of the baseline algorithm.

III. Proposed technique

Traditional phonetic segmentation techniques have reported good accuracy in terms of precision and recall for a $20\,ms$ tolerance as compared with ground truth, for clean speech. However, over-segmentation is a cause of concern in the presence of clipping and environmental noise. We propose the use of multi-taper based speech processing front-end, together with STM and a novel post-processing mechanism for obtaining optimal phonetic segmentation for noisy speech.

A. Speech processing front-end

Conventionally spectral estimation of speech is done by applying a Hamming-window or a single taper for speech signal processing. A major limitation of the single taper method is that, by using one taper a significant portion of the signal is discarded and the data points at the extremes are down-weighted, giving a high variance for the direct spectral estimate [18]. Hence, a multi-taper method is used so that the statistical information lost by using just one taper is partially recovered by using multiple windows for the same duration. Additionally multi-taper spectrum is more robust to noise due to its low-variance property. Noise robustness of multi-taper spectral estimation has been discussed in [15]. The multi-taper spectrum is a weighted sum of the several tapered periodograms. Spectral estimation of a signal s using multi-taper method is as follows,

$$S_(m,k) = \frac{1}{M} \sum_{p=0}^{M-1} \lambda(p) \sum_{j=0}^{N-1} w_p(j) s(m,j) e^{-i2\pi \frac{k}{N} j} \quad (3)$$

where $w_p(j)$ is the p^{th} data taper function, M is the number of tapers and $\lambda(p)$ is the weight corresponding to the p^{th} taper, N is the speech frame length and k is the FFT points. In practice, weights are designed so as to compensate for increased energy loss at higher order tapers.

For our work, we apply multi-taper spectral estimation to the PLPCC feature extraction and thus have modified speech front-end as shown in Figure 1.

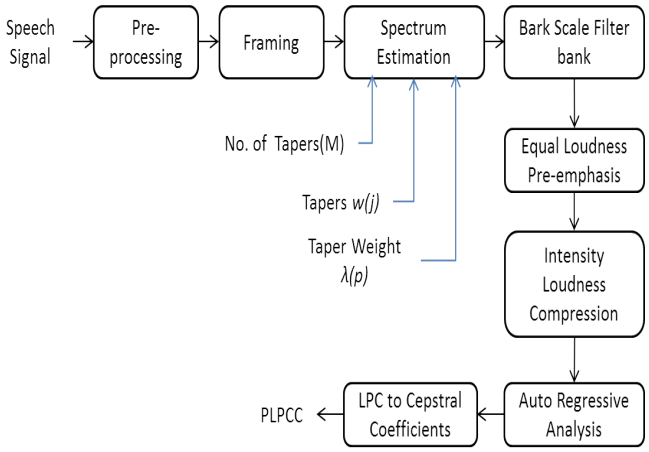

Fig. 1: Modified frond-end processing (Feature Extraction)

Phonetic segmentation using two types of orthonormal multi-tapers for speech front-end have been compared.

1. Sine tapers [19]

$$w_p(j) = \sqrt{\frac{2}{N+1}} \sin\left(\frac{\pi p(j+1)}{N+1}\right), \; j = 0, 1, \ldots, N-1 \quad (4)$$

2. Discrete Prolate Spheroidal sequences (DPSS) or Thomson or Slepian tapers [13]

$$w_p(j) = \frac{\sin[\omega_c T(p-j)]}{(p-j)}, \qquad j = 0, 1, \ldots, N-1 \quad (5)$$

where N denotes the desired window length in samples, ω_c is the desired main-lobe cut-off frequency in radians per second, and T is the sampling period in seconds.

The STM algorithm as described in Section II is applied for phonetic segmentation where the PLPCC features extracted using multi-taper spectral estimation form the feature set $f = [\bar{f}_1, \bar{f}_2, \ldots, \bar{f}_m]$. The STM contour is then subjected to a post-processing technique described in Section III-B, to eliminate spurious peaks and thereby improve the phone boundary location estimation.

B. Dynamic threshold computation

An experimentally determined static threshold has been used for boundary correction in [10]. In this paper we propose, a threshold determined dynamically from the STM itself, i.e. a threshold specific to a given speech utterance. Of the statistical measures such as mean and median, the median of the STM vector (of length *m*) τ_M computed for each speech utterance was found to be the most suitable as threshold for robust phonetic segmentation.

$$STM(m) = \begin{cases} STM(m), & \text{if } STM(m) > \tau_M \\ \tau_M, & \text{otherwise} \end{cases} \quad (6)$$

The median threshold τ_M, also catered to elimination of spurious boundaries inserted within long vowel, diphthong or silence regions.

IV. Experimental setup

A. Data Preparation

To validate language independence of the proposed technique, we experimented on two different language databases i.e., (a) TIMIT and (b) an in-house phonetically balanced Hindi corpus. Robustness of the proposed technique was validated using simulated clipped data and noisy data.

1) Clean data:

- TIMIT American English corpus [20] - contains 2,34,925 between-phone boundaries manually determined by experts.
- Hindi speech corpus [21] - contains 55,104 between-phone boundaries manually marked.

2) Simulated clipped data: We simulate clipping using following transformation for both TIMIT and Hindi speech corpus.

$$x_c(n) = \begin{cases} x(n), & \text{if } |x(n)| < \tau \\ \tau \cdot \text{sgn}(x[n]), & \text{if } |x(n)| \geq \tau \end{cases} \quad (7)$$

where $x(n)$ is the original signal, $x_c(n)$ is clipped signal and τ is the percentage clipping introduced in the speech signal. Clipping percentage was varied from 10 to 50 in steps of 20.

3) Simulated babble-noise data: We chose to add babble noise to clean speech to simulate speech with background noise like characteristics. Clean speech from TIMIT and Hindi speech corpus were combined with babble noise from NOISEX-92 database using the FaNT toolkit [22]. SNR of the noisy speech was varied from 0 dB to 15 dB.

4) Test data: Test data was recorded in Hindi on three different devices simultaneously. 18 sentences from 7 speakers were recorded on (a) a laptop using a close talking microphone, (b) a mobile phone (SAMSUNG Galaxy Ace GT-S5830i) using a hands-free microphone and (c) a tablet (Nexus 7) placed fixed on the table. Speech was recorded at $16\,kHz$ sampling rate. We consider the laptop recording as clean data in this set. This data was annotated manually to establish the ground truth and consists of 3997 phone segments for speech recorded on each device. Speech recorded on these devices was either clipped or contained environmental noise.

Multi-taper spectral estimation was done using 2-256 tapers for sine tapers and 4 to 7 tapers for DPSS. Best results (reported) were obtained with 6 tapers-DPSS.

12 dimension (*D*) PLP cepstral coefficients were extracted for each 30 ms frame with 20 ms overlap. STM computation was done over 5 frames with *n* = 2 in Equation 1. Both the baseline and the proposed techniques were validated on the above data. To measure the performance of phonetic segmentation, we use the standard precision, recall and F-score measures [23].

V. Experimental Results and analysis

A high F-score indicates high precision and high recall, thus ensuring high system accuracy. High precision indicates low over-segmentation. In Table I, we compare the performance of our proposed technique with known unsupervised techniques in the literature, for TIMIT clean speech.

TABLE I: Comparison of unsupervised phonetic segmentation methods for TIMIT clean speech

Method	**Precision** (Pr)	**Recall** (Re)	**F-score** (Fs)
Dusan et al [10]	72.73	75.2	73.94
Qiao et al [9]	78.76	77.5	78.13
Baseline [12]	60.69	80.2	69.1
1.Multi-taper DPSS	64.1	82.46	72.13
2.Multi-taper DPSS + threshold	84.6	75.8	80

TABLE II: Unsupervised phonetic segmentation methods for Hindi clean speech

Method	**Precision** (Pr)	**Recall** (Re)	**F-score** (Fs)
Baseline-Hindi [12]	59.2	82	68.8
1.Multi-taper DPSS	67.98	87.23	76.41
2.Multi-taper DPSS + threshold	93.2	81.8	87.1

Over-segmentation was the key cause of poor F-score in the baseline. Both multi-taper based speech front end and dynamic thresholding have contributed significantly in tackling the over-segmentation problem. Results in Table I show significant improvement in F-score at each step (1.Multi-taper DPSS and 2.Multi-taper DPSS+threshold) of the proposed technique for TIMIT data. Similar improvements were seen for Hindi clean speech as shown in Table II. To the best of our knowledge, no published results exist for this dataset using any other unsupervised technique.

However, using a dynamic threshold as explained in Section III-B impacts the recall adversely, since some of the weak transitions that result in low peaks are lost due to this. An analysis of the results showed that the transitions between broad phone classes such as vowel-glides, vowel-nasals and vowel-liquids are eliminated due to the post-processing. It was found that DPSS (reported in this work) performed better than sine tapers for clean as well as noisy speech.

Tables III and IV show the performance in terms of precision, recall and F-score using the baseline [12] and proposed techniques for noise-induced TIMIT and Hindi speech data. Results comparable to babble noise were seen for additive white Gaussian noise (AWGN) as well. An improvement of 11% for TIMIT and 18% for Hindi speech data (clean) over the baseline approach is seen. An improvement of 7% for 50% clipping and 10% for 10 dB noise was seen for TIMIT speech. Similarly, an improvement of 13% for 50% clipping and 12% for 10 dB noise was seen for Hindi speech. It is evident from the F-scores in Tables III and IV that the proposed technique performs better as compared to the baseline technique for clean data, clipped data and noisy data. Over-segmentation resulting from spectral roughness caused by clipping [24] is taken care of through spectral enhancement using multi-taper windowing and the novel post-processing technique. Higher improvements were seen on Hindi speech corpus; as seen in Table I This can be attributed to distinguishable pauses between words that caused the baseline to perform poorly and

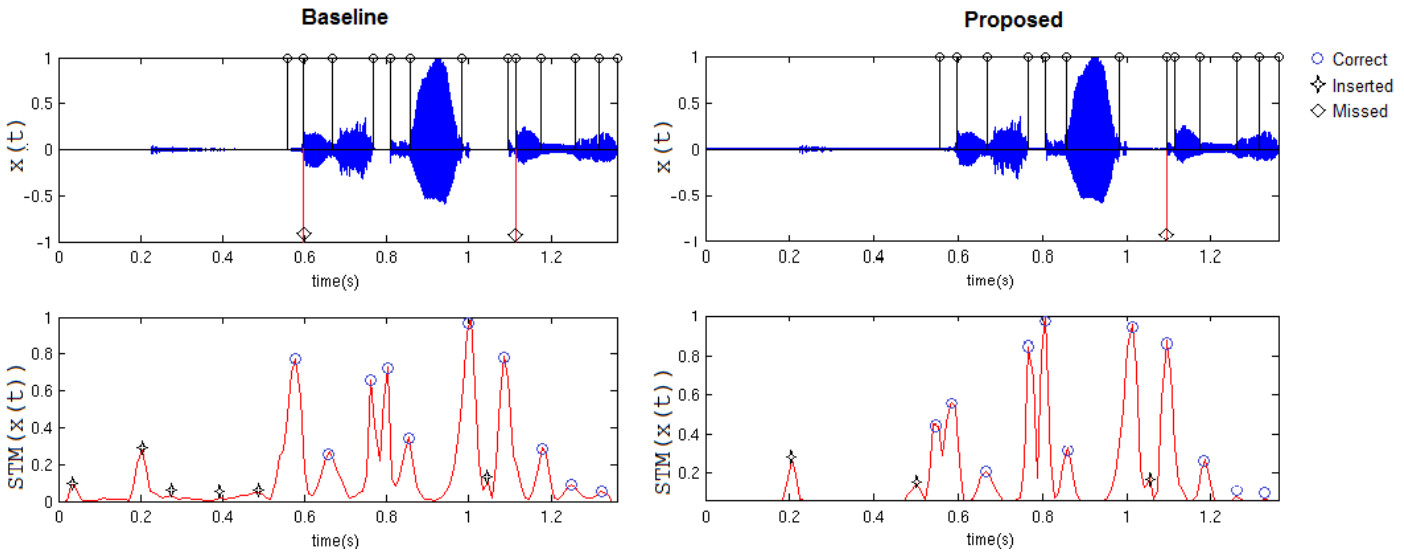

Fig. 2: Phonetic segmentation using baseline and proposed technique (inclusive of Multi-taper and post processing) for TIMIT utterance 'His captain was'; ground truth phonetic segments are marked in black.

TABLE III: Comparison of phonetic segmentation (%) using baseline [12] and proposed technique for clipped speech

Database	TIMIT						Hindi					
Clipping (%)	Baseline			Proposed			Baseline			Proposed		
	Pr	Re	Fs	Pr	Re	Fs	Pr	Re	Fs	Pr	Re	Fs
10	60.3	83.6	70.1	83.1	74.6	78.6	58.6	82.4	68.5	91.8	80.8	85.9
30	58.4	84.5	69.1	80.9	73	76.8	57.2	83.3	67.8	88.9	79.1	83.7
50	56.6	85.7	68.2	78.2	71.9	74.9	55.2	85.4	67.1	83.4	78.2	80.7

TABLE IV: Comparison of phonetic segmentation (%) using baseline [12] and proposed technique for speech with babble noise

Database	TIMIT						Hindi					
Noise (dB)	Baseline			Proposed			Baseline			Proposed		
	Pr	Re	Fs	Pr	Re	Fs	Pr	Re	Fs	Pr	Re	Fs
0	46.5	87.8	60.8	59.6	71.5	65	47.6	88.8	62	63	75.3	68.6
5	49.7	87.7	63.5	68.1	74.3	71.1	51.5	89.2	65.3	73.1	77.9	75.4
10	53.5	87.7	66.4	76.3	76.9	76.6	55.6	89.3	68.5	81.5	80.1	80.8
15	56.5	87.5	68.7	82.1	78.7	80.3	58.6	89.2	70.7	87.4	81.5	84.3

resulted in over-segmentation. The proposed speech processing front-end and post-processing technique handled the over-segmentation caused in the pause and silence regions. Figure 2 shows the phonetic segmentation using the baseline and proposed algorithm for clean speech along with the ground truth boundaries, wherein $x(t)$ is the temporal representation of the signal and $STM(x(t))$ is the STM of $x(t)$.

Similar experiments were carried out on test data mentioned in Section IV. The F-scores for the recordings on the three devices for a tolerance interval of 20 ms are as shown in the Table V.

TABLE V: Comparison of phonetic segmentation (%) for speech recordings on three different devices

Device	Baseline			Proposed		
	Pr	Re	Fs	Pr	Re	Fs
Laptop	63.5	78.6	70.2	83.3	69.4	75.7
Mobile	48.9	80.5	60.9	71.6	72.2	71.9
Tablet	51.4	83	63.5	67.9	68.7	68.3

VI. Conclusion

A robust phonetic segmentation technique is essential for several different types of speech-based applications. With the advent of smart devices, speech-based mobile applications are gaining popularity. However, the performance of such applications is impacted due to the low audio quality of the mobile-device recorded speech. Robust phonetic segmentation techniques become imperative for such a system. In this work, we propose a multi-taper based speech front-end, together with STM and a novel data driven post-processing technique for phone segmentation under noisy conditions such as environmental noise and clipping, commonly present in a mobile phone recording. The proposed approach was validated on TIMIT and a Hindi speech corpus where baseline algorithm provided 68.8% accuracy for Hindi speech corpus and the proposed method provided 87.1% with an improvement of over 18%. Similarly for TIMIT data, the proposed approach gave an

improvement of 11%. Similar results were observed for a set of speech, recorded simultaneously on three devices - a laptop (clean data), a tablet (noisy) and a mobile phone (clipped). We intend to use this technique as the first step for assessment of articulation errors in mobile phone based applications.

REFERENCES

[1] S. Furui, "On the role of spectral transition for speech perception," *Journal of the Acoust. Society of America*, vol. 80, no. 4, pp. 1016–1025, 1986.

[2] J. Yuan, N. Ryant, M. Liberman, A. Stolcke, V. Mitra, and W. Wang, "Automatic phonetic segmentation using boundary models," in *Proc. INTERSPEECH*, August 2013, pp. 2306–2310.

[3] J. Adell and A. Bonafonte, "Towards phone segmentation for concatenative speech synthesis," in *Proc. The* 5^{th} *ISCA Speech Synthesis Workshop*, 2004, pp. 139–144.

[4] V. Patil, S. Joshi, and P. Rao, "Improving the robustness of phonetic segmentation to accent and style variation with a two-staged approach." in *Proc. INTERSPEECH*, September 2009, pp. 2543–2546.

[5] A. Stolcke, N. Ryant, V. Mitra, J. Yuan, W. Wang, and M. Liberman, "Highly accurate phonetic segmentation using boundary correction models and system fusion," in *Proc. ICASSP*, May 2014, pp. 5552–5556.

[6] O. Kalinli, "Combination of auditory attention features with phone posteriors for better automatic phoneme segmentation." in *INTERSPEECH*. ISCA, 2013, pp. 2302–2305.

[7] Y. P. Estevan, V. Wan, and O. Scharenborg, "Finding maximum margin segments in speech," in *Proc. ICASSP*, vol. 4, April 2007, pp. IV–937–IV–940.

[8] S. King and M. Hasegawa-Johnson, "Accurate speech segmentation by mimicking human auditory processing," in *Proc. ICASSP*, May 2013, pp. 8096–8100.

[9] Y. Qiao, N. Shimomura, and N. Minematsu, "Unsupervised optimal phoneme segmentation: Objectives, algorithm and comparisons," in *Proc. ICASSP*, March 2008, pp. 3989–3992.

[10] S. Dusan and L. R. Rabiner, "On the relation between maximum spectral transition positions and phone boundaries," in *Proc. INTERSPEECH*, September 2006, pp. 645–648.

[11] H. Hermansky, "Perceptual linear predictive (plp) analysis speech," *Journal of the Acoust. Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[12] N. J. Shah, B. B. Vachhani, H. B. Sailor, and H. A. Patil, "Effectiveness of PLP-based phonetic segmentation for speech synthesis," in *Proc. ICASSP*, May 2014, pp. 270–274.

[13] D. Thomson, "Spectrum estimation and harmonic analysis," *Proceedings of the IEEE*, vol. 70, no. 9, pp. 1055–1096, September 1982.

[14] J. Alam, T. Kinnunen, P. Kenny, P. Ouellet, and D. O'Shaughnessy, "Multi-taper MFCC features for speaker verification using i-vectors," in *Proc. ASRU*, Dec 2011, pp. 547–552.

[15] T. Kinnunen, R. Saeidi, J. Sandberg, and M. H. Sandsten, "What else is new than the hamming window? Robust MFCCs for speaker recognition via multitapering," in *Proc. INTERSPEECH*, September 2010, pp. 2734–2737.

[16] Y. Hu and P. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE Transactions on Speech and Audio Processing,*, vol. 12, no. 1, pp. 59–67, Jan 2004.

[17] Y. Attabi, M. Alam, P. Dumouchel, P. Kenny, and D. O'Shaughnessy, "Multiple windowed spectral features for emotion recognition," in *Proc. ICASSP*, May 2013, pp. 7527–7531.

[18] G. A. Prieto, R. L. Parker, D. J. Thomson, F. L. Vernon, and R. L. Graham, "Reducing the bias of multitaper spectrum estimates," *Geophysical Journal International*, vol. 171, no. 3, pp. 1269–1281, 2007.

[19] K. S. Riedel and A. Sidorenko, "Minimum bias multiple taper spectral estimation," *IEEE Transactions on Signal Processing,*, vol. 43, no. 1, pp. 188–195, Jan 1995.

[20] J. S. Garofolo, "Getting started with the darpa TIMIT CD-ROM: An acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST)*, 1988.

[21] S. K. Kopparapu, "CSRL Hindi speech corpus," https://sites.google.com/site/awazyp/data/speechcorpus, viewed February 2016.

[22] "FaNT- Filtering and Noise Adding Tool," http://dnt.kr.hs-niederrhein.de/index964b.html, viewed February 2016.

[23] O. J. Räsänen, U. K. Laine, and T. Altosaar, "An improved speech segmentation quality measure: the R-value," in *Proc. INTERSPEECH*, Sep. 2009, pp. 1851–1854.

[24] J. Eaton and P. A. Naylor, "Noise-robust detection of peak-clipping in decoded speech," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 7019–7023.