Syllable Nuclei Detection Using Perceptually Significant Features

A. Apoorv Reddy[1], *Nivedita Chennupati*[2], *B. Yegnanarayana*[3]

Speech and Vision Lab, IIIT Hyderabad, A.P, India

{[1]apoorv.reddy, [2]nivedita}@research.iiit.ac.in, [3]yegna@iiit.ac.in

Abstract

Speech can be segmented into syllables by identifying the syllable nuclei, which are points of high sonority. The excitation peaks in the linear prediction (LP) residual and the formant peaks can be interpreted as perceptually significant point features which contribute to the loudness of speech. In this paper, the use of these two point features is described for the use of detecting syllable nuclei. Each of these evidences contain information about different aspects of speech production, namely the glottal vibrations and the time varying vocal tract system. Thus it is possible that they contain complementary information about the syllable nuclei. Performance of the proposed syllable nuclei detection algorithm is evaluated for the TIMIT, Switchboard and the NTIMIT corpus. The proposed method performs comparably against two other state of the art syllable nuclei detection methods, and is shown to perform better for conversational speech. It is very fast and requires no training.

Index Terms: sonority, syllable nuclei detection, glottal closure instant, group delay function, LP residual

10.21437/Interspeech.2013-169

1. Introduction

At the perceptual level, the syllable nuclei are attributed to high energy sonorants or resonant sounds, which are relatively loud and carry a clear pitch. These attributes lead us to infer that the acoustic correlates of syllable nuclei are energy and periodicity properties [1]. An energy-based syllable detection method was proposed in [2], where energy peaks in the range 250 to 2500 Hz are shown to be well correlated with the syllable nuclei. A smoothed modified loudness contour is used to detect vowels for the purpose of estimating speaking rate in [3]. Speech rate estimation methods have mainly used the durations between the syllable nuclei as a method to find an estimate of the speaking rate [4, 5]. Syllable detection in [4] uses spectral correlation envelopes using selected sub-bands with temporal correlation and smoothing. Monte-carlo simulations are performed to find optimal settings for subband selection and the thresholds used for peak picking. A rhythm guided syllable detection algorithm is proposed in [5] where the rhythmic feature of the sequence of syllables in continuous speech is exploited. The parameters of an optimal sinusoid are calculated on the basis of peaks detected apriori in the energy envelope. A least squares fitting criterion is used to calculate the frequency and phase offset of the sinusoid based on the detected peaks. Then the next peak is detected in the energy envelope in a range around the next sinusoid peak. This range is dictated by the frequency of the sinusoid. The sinusoid is updated after calculation of each peak. A hierarchical hidden Markov model (HMM) based method is proposed in [6], which automatically syllabifies the input speech by generating *syl* and *garbage* tags for the input frame. A multilayer perceptron based automatic syllable boundary detection method is described in [7]. Here, the neural network tries to estimate the posterior probabilities of a phoneme being in syllable nuclear position in the context of neighbouring phonemes. Then possible errors are corrected automatically by parsing the decision output string which was obtained from the posterior probabilities of each phoneme. Wu et. al proposed the use of a multilayer perceptron based classifier to detect syllabic onsets which were subsequently shown to improve speech recognition [8]. In [9], a bidirectional long short-term memory neural network model is used to identify potential syllable nuclei in spontaneous and read speech. The neural network uses a 79 dimensional input vector including a 20 sub-band modulation spectrum, their first differences, 12 PLP coefficients, log energy and their first and second differences. The output was specified by Gaussian curves spanning the duration of the syllable nuclei, and setting the rest of the output as zero. The neural network was trained using the gradient descent algorithm.

The acoustic correlates used for detection of syllable nuclei are based on our limited understanding of the production features relating to the perception of the syllable, such as a high energy sonorant or a relatively loud sound carrying clear pitch. The concepts of perception of energy and pitch are due to some features derived from a finite duration segment of speech. For example, energy is generally computed over 20-30 ms, assuming stationarity of the vocal tract system during that interval. Likewise, pitch periodicity can be perceived only if the signal is processed over a few cycles of the glottal vibration. In fact the least periodic sounds like whispers, fricatives, affricates and stops do not correpsond to the syllable nuclei. Most voiceless sounds do not possess the characteristics of syllable nuclei.

In this paper, we examine a set of new acoustic correlates that can contribute to the perception of high energy sonorants and relatively loud sound for detecting syllable nuclei. The new acoustic correlates are based on the fact that in voiced speech, the primary source of excitation of the vocal tract system is by the impulse-like characteristics due to the sharp closure of the vocal folds in each glottal cycle. It is well known that sharper the closure, the louder is the speech sound [10]. This can happen without any relation to the periodicity or energy of the signal. Also, the vowels or sonorants are perceived louder due to the resonances of the obstruction free vocal tract. The sharper the resonances, i.e., lower the bandwidths, the louder is the corresponding sound. These two properties, the impulse-like excitation and low bandwidth resonances, can be interpreted as some kind of point properties, in the sense that impulse-like behaviour is confined to a very short ($<$ 1ms) region in the time domain and the sharp low bandwidth formants have the spectral energy concentration around the formant frequencies.

We hypothesize that the perception of high energy (loudness) and sonority could be due to the impulse-like excitation in time domain and sharp resonances in the frequency domain. Both these are point properties (as opposed to spread) in their respective domains. Subjective experiments on speech, synthe-

sized by suppressing and enhancing these two point features, confirm this hypothesis. Note that both these features are robust in the sense that they are local high SNR regions in their respective domains. We develop acoustic correlates that reflect these features, and show that they can help in identifying syllable nuclei.

In Section 2 we conduct subjective experiments to justify the hypothesis by modifying the excitation and resonance features in the signal. Section 3 gives an algorithm to detect the syllable nuclei. Section 4 gives the performance of the syllable nuclei detection method on TIMIT and Switchboard data. Robustness of the method is examined by evaluating its performance on the NTIMIT corpus. Section 5 gives a summary of the work reported in this paper.

2. Subjective experiments

In the speech signal the characteristics of the time varying vocal tract system can be represented approximately by the linear prediction coefficients (LPCs) derived for each frame (about 10-30 ms) of data using LP analysis. The LP residual represents some of the features of the time varying excitation. In particular, in voiced segments the impulse-like excitation is reflected as large energy of the residual signal around the glottal closure instants (GCIs). The impulse-like behaviour can be seen better in the Hilbert envelope (HE) of the LP residual. The sharpness of these peaks around the GCIs gives a perception of loudness [10]. The sharpness can be increased by multiplying the residual using a sequence of Gaussian-shaped pulses located around the GCIs. The modified LP residual is used to excite the time varying all-pole model represented by the LPCs for each frame.

On the other hand, to decrease the sharpness of the excitation peaks, each sample of the LP residual is divided by the square root of the corresponding sample in the Hilbert envelope of the LP residual.

The LPCs for each frame represent the shape of the vocal tract for that frame, and hence contain the information of the resonances or formants of the vocal tract system. The sharpness of the resonances in the LP spectrum may add to the perception of loudness caused by increasing sharpness of the peaks in the HE of the LP residual around the GCIs. To increase the sharpness of the peaks around the formant peaks, the speech signal (sampled at 8KHz) is passed through an all-pole filter represented by LPCs, which are derived from the LP residual signal of the speech signal obtained using a 1st order LP analysis. The first order LP analysis reduces the slope of the spectrum in the LP residual signal. A 7th order LP analysis of the 1st order LP residual gives LPCs which has peaks at the formant locations, however with a nearly flat overall slope. Hence passing the original speech signal through an all-pole filter represented by the LP7 coefficients emphasize the formants without changing the overall spectral slope.

On the other hand, the sharpness of the formant peaks can be decreased by passing the original signal through an inverse filter represented by the LP7 coefficients.

The original signal is thus modified in four ways, where the formant peaks are enhanced and de-emphasized, and the excitation peaks in the LP residual are sharpened and de-emphasized.

DR (de-emphasized residual) and EG refer (emphasized GCI) to the modified signals generated from the de-emphasized residual and the enhanced residual respectively. FS (formant suppressed)and FE (formant enhanced) represent the modified signal with suppressed formant peaks and emphasized formant peaks respectively.

	DR	*EG*	*FS*	*FE*
Sentence 1	-0.5	0.125	-0.625	0.75
Sentence 2	-0.5	0.125	-0.75	0.5
Sentence 3	-0.75	0.25	-0.75	0.75
AOS	**-0.583**	0.167	**-0.708**	**0.67**

Table 1: Average opinion scores for the modified signals with respect to the original signal.

Listeners were asked to listen to the original speech and the corresponding modified signals to mark the loudness level compared to the original signal. There were overall 3 sentences of 2-3 seconds duration each. Eight subjects were asked to give a score of +1 or -1 and 0 for the modified signals if they perceived the modified signals to be louder, muffled or the same as the original signal, respectively. The sentences were presented in the following manner: original, DR, EG, original, FS and FE. Table 1 gives the average opinion scores for the various modified signals on a scale of -1 to +1.

The the average opinion score (AOS) for the modifications for all three sentences calculated by averaging the opinion scores across all 8 listeners. The AOS indicates that there is a loss in perception of loudness if any one of the two, excitation source information or formant peaks are suppressed.

The subjective listening tests of the signal and modified speech signals indeed confirm that perception of loudness changes if any one of the two, excitation source information or the vocal tract system information is changed. The absence of the sharpness of the excitation peaks in the LP residual and the high bandwidth of the formant peaks in the modified signals give a perception of less loudness as compared to the original signal.

This can be seen in the spectrogram plots of the original signal as compared to the modified signals in Fig.1. The spectrograms in Fig.1 have been computed with a frame length of 20ms and a shift of 10ms. In Fig.1(b) and 1(c), the formant peaks have been suppressed and enhanced, respectively. The formant structure can not be easily observed in Fig.1(b), while it can be clearly seen in Fig.1(c). Fig. 1(d) and 1(e) correspond to the modified signals with a de-emphasized residual and enhanced residual, respectively. We can clearly observe from Fig. 1(d) and 1(e) that though the formant information is preserved, the formant magnitude is de-emphasized in 1(d) and emphasized in 1(e).

This observation gives us motivation to look at the excitation peaks and the formant peaks as acoustic correlates of loudness of speech, and thus to derive an envelope-based syllable nuclei detection method.

3. Envelope-based syllable nuclei detection

In this section we will describe the two evidences which are used to derive envelopes for syllable nuclei detection. They are based on the two point properties discussed before, namely, the excitation peaks in the LP residual and the formant peaks in the group delay spectrum.

3.1. Short time energy of the Hilbert envelope of LP residual (EHE)

A 14th order LP analysis is performed for each frame of 20ms with an overlap of 10ms at a sampling rate of F_s = 8kHz.

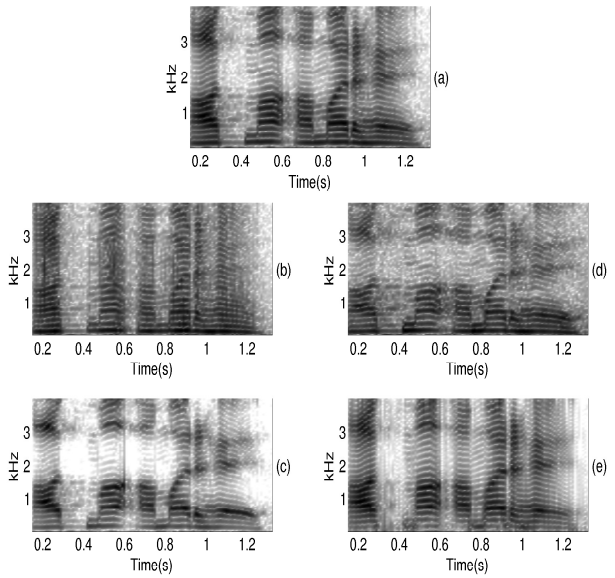

Figure 1: Spectrograms for the Hindi sentence *'mujhe nirantar ki ja:nka:ri acchi lagi'* corresponding to (a) Original speech signal (b) Formant suppressed (FS) (c) Formant Enhanced (FE) (d) De-emphasized LP residual (DR) (e) Emphasized LP residual (EG) .

The excitation peaks can be enhanced by computing the Hilbert envelope of the LP residual, as it serves to remove the phase information present in the excitation source [11]. The Hilbert envelope is the magnitude of the analytic signal $s_a(n)$ of the LP residual $e(n)$. The analytic signal $s_a(n)$ is,

$$s_a(n) = e(n) + je_h(n) \tag{1}$$

where $e_h(n)$ is the Hilbert transform of the LP residual $e(n)$, where $e_h(n)$ is calculated as follows:

$$e_h(n) = \begin{cases} \mathcal{F}^{-1}\{-j\mathcal{F}\{e(n)\}\}, \text{if } f \geq 0 \\ \mathcal{F}^{-1}\{j\mathcal{F}\{e(n)\}\}, \text{if } f < 0 \end{cases} \tag{2}$$

where f is the frequency and $\mathcal{F}$ denotes the Fourier transform.

The Hilbert envelope $h_e(n)$ of the LP residual $e(n)$ is computed as follows,

$$h_e(n) = \sqrt{e^2(n) + e_h^2(n)} \tag{3}$$

The GCIs of the speech signal are extracted using the zero frequency filtering (ZFF) method [12]. Energy of the HE of the LP residual is calculated around the GCIs with a window length of 1 ms and a shift of 1 sample. We take the local maxima of the energies calculated in this region as a measure of the loudness of the speech signal. We will call this energy profile as EHE.

To extract the syllable nuclei we need only observe gross level changes in the EHE. Thus to smear the local variations in the EHE we convolve it with a Hamming window of length 50ms. However a small peak corresponding to a syllable nucleus and lying in the neighbourhood of a relatively large peak will tend to get de-emphasized by this smoothing operation. Thus the square root of the EHE profile is taken before convolving it with the Hamming window.

3.2. Maximum formant magnitude envelope (MFME)

It is known that the group delay spectrum (GD) of a signal is proportional to the squared magnitude spectrum around the formant frequencies [13]. That is,

$$\tau_g(\omega) \propto |X(\omega)|^2 \tag{4}$$

where $X(\omega)$ is the Fourier transform of the signal $x(n)$.

The group delay function (GD) can be represented as a function of the real and imaginary parts of the signal spectrum in the following way [14],

$$\tau_g(\omega) = \frac{X_i(\omega)X_r'(\omega) - X_r(\omega)X_i'(\omega)}{X_r^2(\omega) + X_i^2(\omega)} \tag{5}$$

where,
$X(\omega) = X_r(\omega) + jX_i(\omega)$ and
$X'(\omega) = X_r'(\omega) + jX_i'(\omega)$ is the Fourier transform of the signal $nx(n)$.

The GD function is computed using the method described in [15] for each frame of 20 ms with a shift of 10 ms at a sampling rate of F_s = 8kHz. The formant peak with the highest magnitude in the GD spectrum will carry the most energy for that segment of speech. A contour is constructed by taking the formant peak in the GD spectrum with the maximum amplitude for each frame. The square root of this contour is taken as the MFME contour. The MFME contour is also smoothed by convolving it with a Hamming window of 50ms.

3.3. Combined evidence

Before combining the two evidences, we enhance each evidence. Each evidence is enhanced in the following manner. The first order difference (FOD) of the evidence is calculated. Spurious peaks in the individual evidences are eliminated using a simple slope counting method. The evidence is then amplitude normalized between two consecutive negative to positive zero crossings of the differenced evidence signal. The evidences EHE and MFME are then combined by taking their samplewise mean. This combined evidence is then normalized. A simple peak picking algorithm is used to find the local peaks. These peaks correspond to potential syllable nuclei.

Each peak in the individual evidences will have an amplitude of 1. So an amplitude threshold of 0.5 is used to remove spurious peaks in the combined evidence which fall below this threshold. A minimum spacing threshold of 75ms is used to remove a smaller peak if it lies in the neighbourhood of a larger peak. An adaptive thresholding technique described in [3] is used to further validate the detected peaks. For a peak, the combined evidence must fall below a threshold t, which is a fraction of the local maxima within a range D around the peak. We have taken $t = 0.8$ and $D = 75$ms. Peaks lying in unvoiced regions are removed by performing voice/unvoiced segmentation on the speech signal. Voiced/unvoiced segmentation of speech is done on the basis of the strength of excitation of the voiced epochs [16]. These spurious peaks may correspond to fricatives which have high energy.

Fig.2 illustrates the working of the syllable nuclei detection algorithm. The shaded regions correspond to the vowel regions in the sentence. The peaks in Fig.2(f) marked green are hits.

4. Evaluation

To evaluate the proposed method for syllable nuclei detection, the phonetically transcribed TIMIT and Switchboard corpora

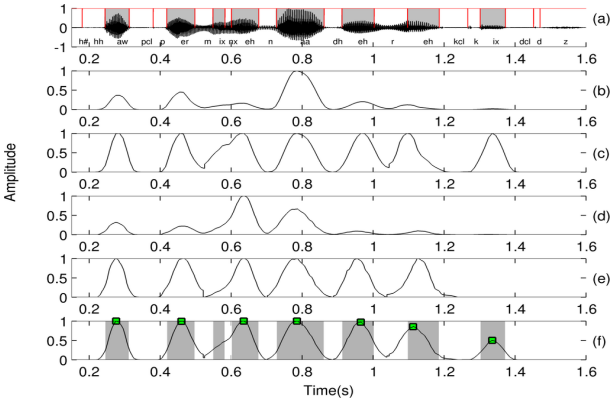

Figure 2: Syllable nuclei detection (a) Speech signal for the TIMIT sentence *'How permanent are their records?'*. (b) Maximum Formant Magnitude Envelope (MFME). (c) Enhanced MFME. (d) Energy of Hilbert Envelope of LP residual (EHE) (e) Enhanced EHE. (f) Combined Evidence. The shaded area correpsonds to the ground truth for syllable nuclei durations.

are used as the ground truths for the location of the vowel phones. If a detected syllable nuclei lies in the region of the vowel phone, it is marked as a hit. First, the performance of peak picking in the individual evidences on the TIMIT database is evaluated, and then compared with the performance of the combined evidence (COMB1).

In addition, the traditional energy contour is used to set a baseline for detecting syllable nuclei. The energy contour is smoothed by convolving it with a 50ms Hamming window and then normalized. The differenced energy contour is calculated. The energy contour is then normalized between the consecutive negative to positive zero crossings of the differenced energy contour.

For the purpose of validating the EHE evidence with the MFME evidence, another system is designed where the two evidences are combined by taking their product instead of their mean. We shall call this system as COMB2.

[%]	Energy	EHE	MFME	COMB1	COMB2
Recall	68.75	87.84	82.34	**92.67**	74.37
Precision	**98.53**	89.95	90.19	91.06	**96.02**
F-measure	80.99	88.88	86.09	**91.86**	83.82

Table 2: Comparison of syllable nuclei detection using the energy envelope, EHE envelope, MFME envelope and the combined envelopes COMB1 and COMB2.

The performance of the individual evidences and the combined evidence are tabled in Table 2. Recall is defined as the ratio of the number of hits to the number of syllable nuclei in the ground truth. Precision is defined as the ratio of the hits to the number of detected nuclei, while the F-measure is the harmonic mean of the recall and precision. The COMB1 system has a higher hit rate and a better F-measure than the COMB2 system, although its precision is lower. The energy contour guided method has the best precision, however its recall rate is very low. From Table 2, we can infer that in most cases, the individual evidences EHE and MFME reinforce each other, though they also provide complementary information to each other when one of the evidences is missing in a syllable nuclear position. This can also be seen in Fig. 2(c). The strength of

(a) **TIMIT**

[%]	RG	BLSTM	COMB1	COMB1 + Energy
Recall	86.59	92.22	**92.67**	**91.24**
Precision	98.86	95.82	**91.06**	**95.6**
F-measure	92.07	93.98	**91.86**	**93.37**

(b) **STP**

[%]	BLSTM	COMB1	COMB1+Energy
Recall	84.44	**87.98**	**85.7**
Precision	83.11	**83.31**	**85.47**
F-measure	83.74	**85.58**	**85.61**

Table 3: Comparison of rhythm guided syllable nuclei detection (RG), BLSTM syllabification method and proposed method for the TIMIT and Switchboard corpora.

the impulse like events in the error residual for the phone /ix/ in Fig. 2 is very low, which results in a small EHE evidence and is thus not considered for evidence normalization. For purposes of comparing with other syllable nuclei detection methods, we will consider the COMB1 system.

We compare our results against the state of the art speech rhythm guided syllable nuclei detection (RG) algorithm described in [5] and the bidirectional long-short-term memory neural network (BLSTM) syllabification method proposed in [9]. Table 3(a) and 3(b) compare the performance of the proposed method with RG and BLSTM for the phonetically transcribed TIMIT and Switchboard (STP) corpora respectively. We have not compared our method with RG for conversational speech as we couldn't find enough time to implement it on our own. The precision of the energy contour based syllable nuclei detection method has been exploited to subsequently reduce the false alarms. The individual evidences EHE and MFME are set to zero if the amplitude of the corresponding sample in the energy contour lies below a certain threshold. The proposed method performs comparably for the TIMIT database, but outperforms the BLSTM and RG methods for the Switchboard corpus. To test the robustness of the proposed syllable nuclei detection method, we have tested it on the NTIMIT database and obtained a recall rate of **91.46%**, precision of **79.11%** and an F-measure of **84.83%**.

5. Summary

The syllable nuclei positions are usually occupied by sonorants which are perceptually louder than other speech sounds. In this paper, we have explored two perceptually significant acoustic features which may be helpful for syllable nuclei detection. The short time energy of the HE of the LP residual and the formant peak information, both are features which are point properties in their respective domains, i.e., time and frequency. We have conducted subjective listening tests which validate that these point features are important for the perception of loudness in speech. These two features are used to generate a profile whose peaks may correpsond to potential syllable nuclei. We then evaluate the performance of our syllable nuclei detection method against RG [5] and BLSTM [9] and find that the results are on par with the current state of the art methods in case of the TIMIT corpus and significantly better for the Switchboard corpus. The advantage of our proposed method is that no training is required, and it is computationally fast.

6. References

[1] Z. Xie and P. Niyogi, "Robust acoustic-based syllable detection," in *Proc. ICSLP*, 2006.

[2] H. Pfitzinger, S. Burger, and S. Heid, "Syllable detection in read and spontaneous speech," in *Proc. ICSLP*, vol. 2, 1996, pp. 1261–1264.

[3] T. Pfau and G. Ruske, "Estimating the speaking rate by vowel detection," in *Proc. ICASSP*, vol. 2, 1998, pp. 945–948 vol.2.

[4] D. Wang and S. Narayanan, "Robust speech rate estimation for spontaneous speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2190–2201, 2007.

[5] Y. Zhang and J. Glass, "Speech rhythm guided syllable nuclei detection," in *Proc. ICASSP*, 2009, pp. 3797–3800.

[6] P. Nel and J. du Preez, "Automatic syllabification using hierarchical hidden markov models," in *Proc. ICASSP*, vol. 1, 2003, pp. 768–771.

[7] J. Tian, "Data-driven approaches for automatic detection of syllable boundaries," in *Proc. ICSLP*. Citeseer, 2004.

[8] S.-L. Wu, M. L. Shire, S. Greenberg, and N. Morgan, "Integrating syllable boundary information into speech recognition," in *Proc. ICASSP*, vol. 2, 1997, pp. 987–990.

[9] C. Landsiedel, J. Edlund, F. Eyben, D. Neiberg, and B. Schuller, "Syllabification of conversational speech using bidirectional long-short-term memory neural networks," in *Proc. ICASSP*, 2011, pp. 5256–5259.

[10] G. Seshadri and B. Yegnanarayana, "Perceived loudness of speech based on the characteristics of glottal excitation source," *The Journal of the Acoustical Society of America*, vol. 126, pp. 2061–2071, 2009.

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

[12] B. Yegnanarayana and K. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 614–624, 2009.

[13] B. Yegnanarayana, "Formant extraction from linear-prediction phase spectra," *The Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1638–1640, 1978.

[14] A. Oppenheim and R. Schafer, "Digital signal processing," *Prentice-Hall, Englewood Cliffs, NJ, 1975*.

[15] M. Anand Joseph, S. Guruprasad, and B. Yegnanarayana, "Extracting formants from short segments of speech using group delay functions," in *Proc. INTERSPEECH*, 2006, pp. 1009–1012.

[16] N. Dhananjaya and B. Yegnanarayana, "Voiced/nonvoiced detection based on robustness of voiced epochs," *IEEE Signal Processing Letters*, vol. 17, no. 3, pp. 273–276, 2010.