ZERO RESOURCE SPEAKING RATE ESTIMATION FROM CHANGE POINT DETECTION OF SYLLABLE-LIKE UNITS

Shekhar Nayak[1]*, Saurabhchand Bhati*[2]*, K. Sri Rama Murty*[1]

[1]Department of Electrical Engineering, Indian Institute of Technology Hyderabad, India
[2]Center for Language and Speech Processing, The Johns Hopkins University, USA

ABSTRACT

Speaking rate is an important attribute of the speech signal which plays a crucial role in the performance of automatic speech processing systems. In this paper, we propose to estimate the speaking rate by segmenting the speech into syllable-like units using end point detection algorithms which do not require any training and fine-tuning. Also, there are no predefined constraints on the expected number of syllabic segments. The acoustic subword units are obtained only from speech signal to estimate the speaking rate without any requirement of transcriptions or phonetic knowledge of the speech data. A recent theta-rate oscillator based syllabification algorithm is also employed for speaking rate estimation. The performance is evaluated on TIMIT corpus and spontaneous speech from Switchboard corpus. The correlation results are comparable to recent algorithms which are trained with specific training set and/or make use of the available transcriptions.

Index Terms— Speaking rate estimation, syllable-like units, unsupervised segmentation, vowel end point detection

1. INTRODUCTION

The impact of speaking rate has been studied broadly on automatic speech recognition (ASR). It has been shown that the accuracy of speech recognition decreases as the speaking rate increases. This has been attributed to the increased variation in pronunciations [1]. Incomplete articulation in fast speech leads to acoustic mismatch [2]. In the case of slow speaking rate, factors which affect ASR performance are hyper articulation and intra-syllabic pauses [3]. Fast speaking rate leads to more substitution and deletion errors whereas slow speaking rate leads to more insertion errors. Therefore, speaking rate dependent decoding and speaker adaptation techniques have been proposed to improve the accuracy of ASR [4]. Speaking rate has also been used as a speaker-specific feature for voice conversion [5]. Several techniques have been proposed in the literature for speaking rate estimation (SRE). The techniques for SRE can be broadly classified into acoustic and linguistic methods.

Acoustic methods: These methods estimate the speaking rate directly from the raw speech waveform. The energy rate or *enrate* was proposed for SRE by calculating the first spectral moment of the energy envelope of speech over short-time windows [6]. *enrate* was combined with two different peak picking estimators from the wideband energy envelope of speech and are averaged to arrive at a multiple rate estimator or *mrate*. A method to detect vowels based on smoothed modified loudness was proposed for SRE [7]. A Gaussian mixture model (GMM) based online SRE approach was proposed in [8]. Wang et. al. have proposed to use subband and temporal correlations to detect syllables for SRE [9]. In the Praat script for SRE, the intensity peaks supported by intensity dips on either side are hypothesized as potential syllable nuclei [10]. In [11], convex cost functions were proposed to estimate temporal density function from time-frequency representation for SRE.

Lexical methods: Lexical methods define speaking rate in terms of phone rate or word rate. Phone rate for an utterance is defined as the ratio of total number of phones to the total duration of phones, essentially, phones per second. The phones are counted after performing phone recognition [6]. A broad class phone recognizer was used for SRE [12]. If accurate phone level transcriptions are not available but correct word level transcriptions are available, then forced alignment can be performed in order to get phone durations [13]. If both accurate phone level or word level transcriptions are not available, then the phonetic segmentation will not be good and will lead to incorrect information about the number of phones and their corresponding durations. Also, if the availability of data with the orthographic transcriptions is limited, then the ASR model training will not be effective, and it, in turn, will affect the SRE. Therefore, in recent times there is an increased interest in the speech community towards zero resource approaches which do not require any labeled data for training or any explicit linguistic knowledge [14].

In this work, we propose a zero resource approach for SRE using a syllabification algorithm based on vowel end point (VEP) detection. A multiple evidence based approach is used to detect the VEPs, and the region between two successive VEPs is considered as a syllable-like unit of C^*V-type, where C^* denotes a non-vowel like region usually consisting of a single or a group of consonants and V denotes a single vowel. The number of detected syllable-like units per second is used quantify the speaking rate. A recent approach based on theta-rate oscillations [15] to detect boundaries of syllable-like units was proposed for unsupervised word discovery [16]. This approach is also used to compare with the proposed approach in the zero-resource settings. The SRE evaluations are done on TIMIT and Switchboard corpus in terms of correlation between the actual and the estimated number of units and the speaking rate.

2. MULTIPLE EVIDENCES FOR VEP DETECTION

The earlier works in SRE defined the speaking rate as the number of syllables per second or the number of phones per second in a given segment of speech. A syllable is a subword linguistic unit consisting usually of a vowel as a nucleus with a preceding onset and a succeeding coda, both are optional and are generally consonants. Jiao et al. [11] exploited this almost certain presence of a vowel in a syllable to reformulate the SRE problem as equivalent to estimating the number of vowels per second in a segment of speech. The authors posed the SRE as a convex optimization problem in which an optimum weighting function has to be determined for the features

derived from the speech segment to estimate the number of vowels per second in that segment. In this work, we use signal processing methods to detect VEPs which are used as anchor points to identify the syllable-like units. Unlike the referred work, our approach does not require any labeled speech data for training.

In this work, we use multiple evidences extracted from the source and spectral characteristics of the speech signal for accurate VEP detection. The evidences from the excitation source information include zero frequency filtered signal and Hilbert envelope of linear prediction (LP) residual of speech signal [17], while the evidences from the spectral characteristics include spectral peaks and modulation spectrum energies [18]. The evidences from the source and spectral features are finally combined with the evidence from the Bessel features to arrive at accurate VEP locations. The significance of these evidences for VEP detection is briefly described in the following subsections.

2.1. Evidence for VEP from source features [17]

Speech production model is considered to be a time-varying system excited with quasi-periodic sequence of impulses or noise for voiced or unvoiced sounds respectively [19]. The change in the nature of excitation from voiced to unvoiced is an important clue to detect the VEPs. This excitation source information extracted from the Hilbert envelope of LP residual and zero-frequency filtered signal has been exploited to detect the VEPs.

Hilbert envelope of LP residual: It is the magnitude of the complex analytic signal formed from the LP residual [20]. It preserves the excitation source characteristics and is given by the following equation -

$$H_e(n) = \sqrt{e^2(n) + \hat{e}^2(n)} \tag{1}$$

where $e(n)$ is the LP residual and $\hat{e}(n)$ is its Hilbert transform. The Hilbert envelope is smoothed by retaining the maximum value for every 5ms with a shift of one sample. The smoothed Hilbert envelope provides evidence for the detection of VEPs.

Zero Frequency Filtered Signal (ZFFS): The ZFFS [21] is obtained by passing the preemphasized speech signal through a cascade of two ideal zero frequency resonators, and subtracting the trend from the resulting signal.

$$\hat{y}(n) = -\sum_{k=1}^{4} c_k y(n-k) + s(n) - s(n-1) \tag{2}$$

$$\hat{y}(n) = y[n] - \frac{1}{2N+1} \sum_{n=-N}^{N} y(n) \tag{3}$$

where the filter coefficients are $c_1 = 4, c_2 = -6, c_3 = 4, c_4 = -1$. The average pitch period is used as the window length $2N + 1$ for the trend removal. Since $\hat{y}(n)$, referred to as ZFFS, is obtained by passing speech signal though a narrowband filter centered around 0 Hz, it predominantly contains the excitation strength information.

To detect the VEP locations, the points at which there is significant change in the excitation information are detected by convolving from right to left, the Hilbert Envelope of LP residual or ZFFS with a 100 ms length first order Gaussian differentiator (FOGD) with standard deviation of one sixth of window length. These evidences are summed up and normalized by the maximum value. The resulting envelope provides excitation source based evidence for the VEP locations.

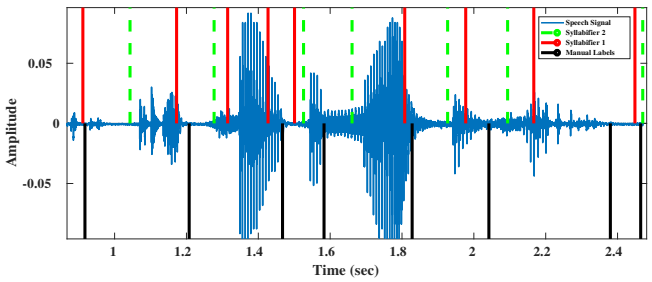

Fig. 1: A Switchboard utterance with manually marked labels (black) and labels from Syllabifier-1 (red) and Syllabifier-2 (green).

2.2. Evidence for VEP from spectral features [18]

The vowels are produced by a relatively open and relatively stationary vocal-tract system compared to the consonants. Hence, the strength of the formants and rate of change of the spectral content provides a strong evidence for the detection of the vowels and their end points.

Spectral Peaks: The shape of the vocal tract which leads to the production of different vowels can be estimated by selecting a few largest spectral peaks. Speech signal is windowed into frames of 20 ms with a frame shift of 10 ms. A 256-point discrete Fourier transform (DFT) is applied to each frame and the sum of ten largest peaks from the first 128 points is computed. This spectral peak sum preserves the evidence for VEP detection.

Modulation spectrum: Change in modulation spectrum energy also corresponds to the vowel end points as it represents change in slowly varying temporal and frequency components of speech signal [22]. The VEP evidence is obtained from the modulation spectrum by passing the speech signal through a band of 18 critical trapezoidal shaped band pass filters in the range of 0-4 kHz. Then, the amplitude envelope of the signal is computed by half wave rectification and low pass filtering at 28 Hz. Thereafter, utterance level normalization is done for the amplitude envelope in each band by dividing by its average value. Further, DFT is computed by using a Hamming window with 250 ms width and 12.5 ms shift to analyze the modulations of the processed amplitude envelope in the range of 4-16 kHz. Finally, the energies from all the bands in this frequency range are summed to get the modulation spectrum energy [23]. The change is further enhanced by computing the slope of the modulation spectrum energy.

Significant changes in the spectral peak and modulation energy envelopes provide evidence for VEP detection. Hence, the individual evidences are convolved with the FOGD operator and added to enhance the VEPs.

2.3. Evidence for VEP from Bessel features [24]

Schroeder argued that any arbitrary signal can be effectively represented by using basis functions which resemble the signal itself [25]. Speech signal can also be considered to be generated by an under-damped time-varying all pole system with a periodic train of impulses or a random noise excitation which produces series of decaying quasi periodic sinusoids resembling voiced speech or narrowband signals resembling whispered/unvoiced speech respectively [26]. Bessel basis functions are damped sinusoids with decaying amplitude and regular zero crossings which makes them suitable representation for speech signals, and for vowel end point

detection as well [27]. The k-th order Bessel function is given by

$$J_k(\lambda) = \sum_{r=0}^{\infty} \frac{(-1)^r}{r!\Gamma(r+k+1)} \left(\frac{\lambda}{2}\right)^{2r+k} \quad (4)$$

A speech signal $s(t)$ can be represented in terms of Bessel functions in the time interval $(0, l)$ as

$$s(t) = \sum_{r=1}^{\infty} C_r J_0(\frac{\lambda_r t}{l}), \quad 0 < t < l \quad (5)$$

where $J_0(.)$ represents 0^{th}-order Bessel funtion, C_r are the coefficients of the Bessel function and $\lambda_r, r = 1, 2, \ldots$ are the positive roots of $J_0(\lambda) = 0$ in the ascending order. Bessel coefficients are given by

$$C_r = \frac{2\int_0^l ts(t)J_0(\frac{\lambda_r t}{l})}{l^2[J_1(\lambda_r)]^2} \quad (6)$$

where $J_1(.)$ represents the Bessel function of first order, $r = 1, 2, \ldots, R$, and the order of Bessel function is R. Bessel coefficients contain both magnitude and phase information and are real [28]. The relation between the index of Bessel coefficient r and the corresponding frequency of the signal f_r at which the maximum peak is achieved can be expressed as

$$f_r = \frac{rf_s}{2N} \quad (7)$$

where f_s is the sampling frequency and N is the number of samples in the duration l.

Representation of speech signal in terms of Bessel functions is effective in enhancing vowel-like regions by considering the appropriate range of Bessel coefficients [24]. The signal $s(t)$ can be bandpass filtered in the discrete range of Bessel coefficients (r_1,r_2) corresponding to the vowel region as computed by (7) using frequency range of the vowel region.

$$\hat{s}(t) = \sum_{r=r_1}^{r_2} C_r J_0(\frac{\lambda_r t}{k}) \quad (8)$$

The discrete version of the bandpass filtered signal $\hat{s}[n]$ is considered an AM-FM signal and its amplitude envelope is extracted using discrete energy separation algorithm. This amplitude envelope is smoothed using a moving average filter of 1 ms duration. A 100 ms size FOGD with 10 ms variance is convolved with the smoothed amplitude envelope from right to left to get the VEP evidence.

Finally, the earlier evidences are combined with the evidences from the Bessel features in the same manner as previous evidences were combined. This provides very reliable and stronger VEPs based on several evidences as described above. The VEPs directly provide the estimate of the number of vowels and in turn the number of syllables in a given segment, the syllable-type being C^*V between consecutive VEPs. This multiple evidence based method for vowel end point detection is further referred to as Syllabifier-1.

3. DETECTION OF SYLLABLE-LIKE UNITS USING THETA OSCILLATOR

Recently, an oscillator based on theta-rate neural oscillations in auditory cortex regions of brain was proposed for unsupervised spoken word discovery [16], which achieved a high word segmentation accuracy on multiple languages. These oscillations coincide well with the syllabic rate according to the speech perception studies [29]. Therefore, Räsänen et al. [16] proposed a damped harmonic oscillator to model the syllabic rate. The input to the oscillator is the amplitude envelope of speech and the minima in the amplitude of the oscillator represent the boundaries of the syllable-like units. The oscillator is modeled as

$$f(t) = e(t) - \frac{1}{f_s}x(t-2)v(t-1) - \frac{2\pi\Delta f}{f_s}v(t-2)f(t-1) \quad (9)$$

where, $e(t)$, $x(t)$, $v(t)$, $f(t)$ denote the amplitude envelope of the speech signal, amplitude, velocity and force of the oscillator, respectively. f_s denotes the sampling frequency and Δf is the bandwidth of the oscillator which is fixed to 8 Hz for critical damping.

This method was originally proposed for unsupervised word discovery from speech. It is computationally efficient, simple and unsupervised. The oscillator is tuned to match the rhythm of syllables. Therefore, in this work we use this method for comparison against the proposed method for SRE. It is referred to as Syllabifier-2 in the rest of the paper. Figure 1 shows a Switchboard utterance segmented into syllable-like units by Syllabifier-1, Syllabifier-2 and the corresponding manual boundaries.

4. SPEAKING RATE ESTIMATION

4.1. TIMIT Evaluations

The TIMIT test set consists of 1680 sentences on which all the results are reported [30]. The results are compared with intensity based Praat script (Praat) [10], the subband and temporal correlation-based method (Sub-band Corr) [9], the GMM based method (GMM) [8], the convex weighting criteria method (Convex OPT) [11]. Sub-band Corr uses TIMIT training set for Monte-Carlo training as in [9]. GMM based model is also trained using the same training set. The Convex OPT method is shown to be dependent on the number of training sentences with speaking rate error reducing almost monotonically with increase in the number of training sentences [9]. Also, the weighting vectors are speaker adapted using a sentence from the test set for each speaker for achieving further improvements.

The Praat script was directly evaluated on TIMIT test set. The syllabifier-1 (proposed) and syllabifier-2 also do not require any training and are evaluated on the test set directly. Further, there is no parameter fine-tuning or cross-validation done using any labeled data in terms of phones/syllables to keep the methods completely zero resource.

Table 1: Speaking rate estimation results for TIMIT test set

Method	Correlation	Mean error	Stddev error	SR error rate%	SR mean error	SR stddev error
Pratt	0.890	1.93	1.38	15.4	0.639	0.49
Sub-band Corr	0.830	1.82	1.48	15.0	0.610	0.40
GMM	0.805	1.61	1.41	14.0	0.528	0.41
Convex-OPT	0.869	1.39	1.24	12.2	0.462	0.36
Syllabifier-1 (Proposed)	0.854	1.60	1.39	13.2	0.537	0.44
Syllabifier-2	0.840	1.98	1.58	15.5	0.662	0.51

The evaluation metrics are the correlation between the actual and the estimated number of vowels, absolute mean error and the corresponding standard deviation (Stddev error), speaking rate (SR) error rate defined by absolute difference between the actual and the predicted vowels normalized by the actual number of vowels, SR mean and stddev error computed from the absolute difference between actual and predicted SR [11]. Table 1 shows the results on TIMIT test set. The correlation for Syllabifier-1 and Syllabifier-2 was found to be comparable to all the methods except Praat which gives relatively higher SR error rate compared to other methods. The mean error, the stddev error and the SR error rate for Syllabifier-1 is better than almost all methods except Convex-OPT which is speaker adapted using test utterances. SR mean error and SR stddev error are also comparable to other methods. This shows that zero resource syllabifiers perform on par with the state-of-the-art on TIMIT without any parameter tuning.

Table 2: Syllable count correlation and statistics for switchboard spontaneous speech

Method	Correlation	Mean error	Stddev error
Convex-OPT	0.971	1.30	1.31
Syllabifier-1 (Proposed)	0.970	1.42	1.59
Syllabifier-2	0.960	1.84	1.82

Table 3: Syllable rate correlation and statistics for switchboard spontaneous speech

Method	Correlation	Mean error	Stddev error
enrate	0.415	0.747	1.405
sub-mrate	0.637	0.530	1.219
mrate	0.671	0.464	1.121
Convex-OPT	0.744	0.600	0.490
Sub-band Corr	0.745	0.339	0.796
Broad Class	0.763	-0.161	0.780
Syllabifier-1 (Proposed)	0.655	0.639	0.668
Syllabifier-2	0.517	0.932	0.830

4.2. Spontaneous speech: Switchboard Evaluations

Spontaneous speech consists of inconsistent number of pauses with varied duration of pauses and spoken phrases. This makes speaking rate estimation a difficult task for spontaneous discourse. The syllabification algorithms are evaluated on ICSI Switchboard corpus subset with 5564 speech utterances which have syllable based manual transcriptions [31]. Acoustically based methods enrate, sub-mrate and mrate are compared which do not require any manual transcriptions for training [6]. The results are also compared with a broad phonetic class recognizer (Broad Class) [12] trained on SCOTUS corpus consisting of large number of tokens for each phonetic class. The methods enrate, sub-mrate, enrate, Sub-band Corr use the pause and noise labels in the manual transcriptions to split the utterances into spurts. Convex-OPT, Sub-band Corr and Broad Class use utterances for training/development set in some form or the other. Syllabifier-1 and Syllabifier-2 are completely zero resource methods and do not use any training/development set. Both these methods also do not use spurts obtained using transcriptions.

Table 2 shows the correlation between the actual and the estimated syllable counts and the mean and standard deviation between the absolute error between the two. The correlation and other statistics for Syllabifier-1 are comparable to Convex-OPT which is a training based method whereas Syllabifier-1 and Syllabifier-2 are directly evaluated on entire Switchboard corpus.

Table 3 shows the correlation between the actual and the estimated syllable based speaking rate and the mean and standard deviation between the corresponding absolute error. The correlation for Syllabifier-1 is comparable to the other zero resource methods and has lesser stddev error compared to the acoustically based methods. The correlations of all the training based methods Convex-OPT, Subband Corr, Broad Class are higher than other methods. But the zero resource methods are more generic and can work on any database or language without any fine tuning under new or unknown settings. This can also be inferred by the fact that the zero resource methods are evaluated as it is on both TIMIT and Switchboard corpus without any specific training/fine-tuning for either of the corpus.

5. CONCLUSIONS

In this paper, zero resource methods for detecting boundaries of syllable-like units are proposed and are evaluated for speaking rate estimation. Multiple evidences from excitation and source information from the speech production model along with the evidences from Bessel feature based representations are used for detecting vowel end points and in turn the boundaries of syllable-like units. This method provides comparable performance with existing methods for speaking rate estimation on TIMIT and Switchboard corpus. A recent zero resource syllabification algorithm based on theta-rate oscillations at the syllabic rate is also evaluated for speaking rate estimation and is shown to perform closer to other methods. Zero resource based methods for speaking rate estimation can be used for any language or database without any training or fine-tuning of parameters using labeled data. Thus, zero resource methods are better suited for improving speech recognition accuracy through speaking rate dependent decoding in low resource settings.

6. ACKNOWLEDGEMENT

The authors acknowledge the Ministry of Human Resource Development (MHRD) and the Ministry of Electronics and Information Technology (MeitY), Government of India, for sponsoring this work under IMPRINT initiative. Special thanks to Dr. Biswajit Dev Sarma and Prof. S. R. Mahadeva Prasanna from IIT Guwahati for their invaluable help regarding vowel end point detection methods. We also thank Okko Räsänen from Tampere University of Technology, Finland for making theta-oscillator code available online.

7. REFERENCES

[1] Eric Fosler-Lussier and Nelson Morgan, "Effects of speaking rate and word frequency on pronunciations in conversational speech," *Speech Communication*, vol. 29, no. 2, pp. 137–158, 1999.

[2] Hiroaki Nanjo and Tatsuya Kawahara, "Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2002*, pp. I–725 – I–728.

[3] Mohamed Benzeghiba, Renato De Mori, Olivier Deroo, Stephane Dupont, Teodora Erbes, Denis Jouvet, Luciano Fissore, Pietro Laface, Alfred Mertins, Christophe Ris, et al., "Automatic speech recognition and speech variability: A review," *Speech communication*, vol. 49, no. 10, pp. 763–786, 2007.

[4] Hiroaki Nanjo and Tatsuya Kawahara, "Language model and speaking rate adaptation for spontaneous presentation speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 4, pp. 391–400, 2004.

[5] Ashish Verma and Arun Kumar, "Modeling speaking rate for voice fonts," in *Proc. Eighth European Conference on Speech Communication and Technology*, 2003, pp. 2917–2920.

[6] Nelson Morgan and Eric Fosler-Lussier, "Combining multiple estimators of speaking rate," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1998.*, vol. 2, pp. 729–732.

[7] Thilo Pfau and Günther Ruske, "Estimating the speaking rate by vowel detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1998*, pp. 945–948.

[8] Robert Faltlhauser, Thilo Pfau, and Günther Ruske, "On-line speaking rate estimation using gaussian mixture models," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2000*, pp. 1355–1358.

[9] Dagen Wang and Shrikanth 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.

[10] Nivja H De Jong and Ton Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior research methods*, vol. 41, no. 2, pp. 385–390, 2009.

[11] Yishan Jiao, Visar Berisha, Ming Tu, and Julie Liss, "Convex weighting criteria for speaking rate estimation," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 23, no. 9, pp. 1421–1430, 2015.

[12] Jiahong Yuan and Mark Liberman, "Robust speaking rate estimation using broad phonetic class recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2010*, pp. 4222–4225.

[13] Nikki Mirghafori, Eric Fosler, and Nelson Morgan, "Towards robustness to fast speech in ASR," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1996*, pp. 335–338.

[14] Ewan Dunbar, Xuan Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, and Emmanuel Dupoux, "The zero resource speech challenge 2017," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2017*, pp. 323–330.

[15] Okko Räsänen, Gabriel Doyle, and Michael C Frank, "Prelinguistic segmentation of speech into syllable-like units," *Cognition*, vol. 171, pp. 130–150, 2018.

[16] Okko Räsänen, Gabriel Doyle, and Michael C Frank, "Unsupervised word discovery from speech using automatic segmentation into syllable-like units," in *Proc. Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[17] Gayadhar Pradhan and SR Mahadeva Prasanna, "Speaker verification by vowel and nonvowel like segmentation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 854–867, 2013.

[18] SR Mahadeva Prasanna, BV Sandeep Reddy, and P Krishnamoorthy, "Vowel onset point detection using source, spectral peaks, and modulation spectrum energies," *IEEE Transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 556–565, 2009.

[19] Lawrence Rabiner and Biing-Hwang Juang, "Fundamentals of speech recognition," 1993.

[20] SR Mahadeva Prasanna and B Yegnanarayana, "Detection of vowel onset point events using excitation information," in *Proc. Ninth European Conference on Speech Communication and Technology*, 2005, pp. 1133–1136.

[21] K Sri Rama Murty and B Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[22] Steven Greenberg and Brian Kingsbury, "The modulation spectrogram: In pursuit of an invariant representation of speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1997*, pp. 1647–1650.

[23] Brian ED Kingsbury, Nelson Morgan, and Steven Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech communication*, vol. 25, no. 1-3, pp. 117–132, 1998.

[24] Biswajit Dev Sarma, S Supreeth Prajwal, and SR Mahadeva Prasanna, "Improved vowel onset and offset points detection using bessel features," in *Proc. International Conference on Signal Processing and Communications (SPCOM), 2014*, pp. 1–6.

[25] Jim Schroeder, "Signal processing via fourier-bessel series expansion.," Tech. Rep., DENVER UNIV CO COLL OF ENGINEERING, 1994.

[26] Shekhar Nayak, Saurabhchand Bhati, and K Sri Rama Murty, "An investigation into instantaneous frequency estimation methods for improved speech recognition features," in *Proc. IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2017, pp. 363–367.

[27] Chetana Prakash, N Dhananjaya, and Suryakanth V Gangashetty, "Bessel features for detection of voice onset time using AM-FM signal," in *Proc. IEEE International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2011, pp. 1–4.

[28] Chetana Prakash, Dhananjaya N Gowda, and Suryakanth V Gangashetty, "Analysis of acoustic events in speech signals using bessel series expansion," *Circuits, Systems, and Signal Processing*, vol. 32, no. 6, pp. 2915–2938, 2013.

[29] Anne-Lise Giraud and David Poeppel, "Cortical oscillations and speech processing: emerging computational principles and operations," *Nature neuroscience*, vol. 15, no. 4, pp. 511, 2012.

[30] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report*, vol. 93, 1993.

[31] John J Godfrey, Edward C Holliman, and Jane McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1992*, pp. 517–520.