Automatic Pitch Accent Contour Transcription for Indian Languages

Gurunath Reddy M, Procheta Sen, Manjunath K E, Arup Dutta,
Arijul Haque, Parakrant Sarkar, K Sreenivasa Rao
School of Information Technology
Indian Institute of Technology Kharagpur, India - 721302
{mgurunathreddy, senprocheta, ke.manjunath, arupdutta1990}@gmail.com
{rjlhq05, parakrantsarkar}@gmail.com, ksrao@iitkgp.ac.in

***Abstract*—In this paper, an automatic method to transcribe the pitch accent contour from the speech signal is presented. Pitch contour transcription refers to the labeling of temporal variations of the pitch contour of the speech signal with finite number of discrete labels. Pitch contour is derived from the zero frequency filtered (ZFF) speech signal. A non-linear smoothing technique is used to remove the spurious pitch values in the pitch contour. An intonation like contour is obtained by removing trend in the pitch contour. The location of the tonal variations in the intonation phrases is identified and assigned with appropriate tone label. Pitch contour transcription is derived using tonal labels and the corresponding timing information from the pitch contour. The Automatic pitch contour transcription system is evaluated using read, extempore and conversation modes of speech from 11 Indian languages. For each mode of speech, the speaker-wise subjective evaluation is carried out for 11 Indian languages to validate the correctness of the proposed automatic pitch contour transcription method.**

Keywords*—Automatic pitch contour transcription; Intonation Contour; Zero frequency filter; Saddle points; Intonation phrase; Pitch contour; Read speech; Extempore speech; Conversation speech

I. Introduction

Prosodic transcription consists of transcribing temporal variation of pitch, speaking rate and pause breaks in the speech signal using finite set of discrete labels. The temporal variation of pitch in a spoken utterance can be viewed as intonation. All vocal languages use pitch pragmatically in intonation for emphasis, to convey surprise or irony or to pose a question. The intonation contour has wide range of applications such as identifying speaker and language characteristics. The speaker characteristics represent the emotions and attitudes of the speaker. The language characteristics can be identified using the grammatical structure generated using intonation contour [1] [2] [3]. Intonation contour is used to extract the information out of important segments in the spoken message. The transcribed pitch contour is very useful in analysis and synthesis of the speech signal.

The work presented in this paper is an extension of our previous work [4]. In our previous work [4], a method to transcribe the pitch contour into four discrete labels is proposed. The four labels used are VL(very low), L (low), H(high) and VH(very high). The range of the pitch contour is divided into four regions and each sub-range is assigned with corresponding pitch label. During transcription, the frame belonging to the sub-range is assigned with the corresponding pitch label. The pitch contour is regenerated from the transcribed labels using linear interpolation technique to validate the transcription process. The limitation in our previous work is that the regenerated pitch contour failed to capture the variations present in the original pitch contour due to discrete nature of pitch labels assigned to each frame. But, many real-world applications require to capture the pitch changes within a pitch label. The pitch changes within a pitch label provide information related to speaker's emotional state, attitude and language information. Hence, it is very essential to capture the variations within a pitch label. To overcome this drawback, we have proposed a new method to label pitch contour using combination of four different labels. A pitch label marked as $L-H$ indicates that pitch starts from low pitch accent and ends in high pitch accent. Likewise, a pitch label marked as $L-L$ indicates that both starting and ending of pitch accent are in low intonation-range. By using the proposed method, it is possible to represent the way in which the pitch varies from start to end of the utterance. In our previous work, we have considered only two Indian languages namely: Bengali and Odia. In this work, we have considered a total of 11 Indian languages including Bengali and Odia.

In general, speech may be broadly classified into three modes namely *read speech, extempore speech and conversation speech*. **Read speech** involves reading out from the notes such as reading the news. It is more structured, planned and prepared well in advance. Read speech is delivered using more formal language and it is one-sided. In **Extempore mode** of speech, the speech is delivered without the aid of notes. Delivering a lecture to students in a class could be an example of extempore speech. It is more vigorous, flexible and spontaneous. The Extempore mode of speech is also called lecture mode of speech or public speaking. The **Conversation mode** of speech is a form of interactive, spontaneous communication between two or more people who are following rules of etiquette. Conversation speech is spontaneous because a conversation proceeds unpredictably. It is informal, unstructured and unorganized. Conversation speech involves free speaking style with no constraints. Hence, it is essential to analyse the characteristics of pitch contour transcription across all the three modes of speech.

From the existing literature, it is observed that there are some works related to automatic pitch contour transcription. Silverman et al. (1992) developed tones and break indices

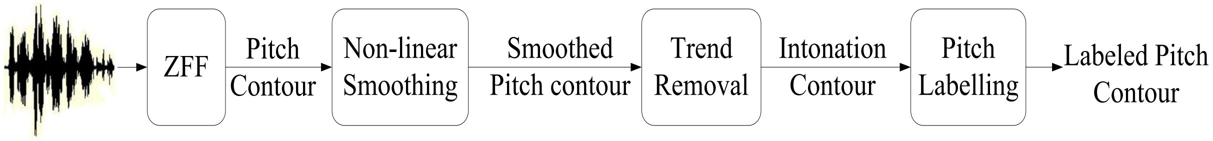

Fig. 1. *Illustration of the block diagram of proposed Automatic Pitch Contour Transcription System.*

(ToBI) annotation standard to represent the prosodic events in spoken language. In ToBI the prosodic events in the speech signal are captured using four tiers of labeling. Pitch accents are defined based on the value and shape of the fundamental frequency contour [5]. Wightman et al. (1994) developed a method to derive accent and boundary labels based on the posterior probabilities of the bi-gram probabilistic model [6]. Ostendorf et al. (1996) developed automatic pitch accent and boundary tone labeling system for predicting pitch accent labels and boundary tone types using decision tree [7]. Chen et al. (2004) used Gaussian mixture model based acoustic-prosodic model and an artificial neural network based syntactic prosodic model to achieve binary pitch accent detection and intonation boundary detection at the word level [8]. Ananthakrishnan et al. (2005) modelled dynamic prosodic features using Hidden Markov models (HMM). An n-gram-based syntactic-prosodic model is used to obtain boundary tones at the syllable level [9]. Most of the existing approaches have used machine learning techniques to derive the automatic pitch contour transcription from the speech signal. Hence, in this work, a method to derive the automatic pitch accent transcription without application of machine learning algorithms is proposed due to nature of the languages considered. There are no existing works on automatic pitch contour transcription in the context of Indian languages are reported. Hence, in this work, we have studied the automatic pitch contour transcription in the context of 11 Indian languages.

The rest of the paper is organized as follows: Section II describes speech corpus used in our work. Section III discusses development and evaluation of automatic pitch contour transcription system. Summary and inferred conclusions are discussed in Section IV.

II. Speech Corpus

In the present work, speech data in 11 Indian languages is considered namely: Assamese, Bengali, Telugu, Kannada, Urdu, Manipuri, Gujarati, Marathi, Odia, Malayalam and Punjabi. The speech data from different languages, speaking modes and genders is considered to derive the general trend in the pitch variations specific to language, speaking mode in a language and the speaker speaking the language.

The speech corpus contains speech data collected in three different modes namely: *read mode, extempore mode and conversational mode*. Speech data in all languages is recorded from the speakers in the age range of 25-40 years. The data is collected from the closed room with controlled acoustics, recorded from digital sources like Internet, Television and Radio. Speech corpus contains 16 bit precision, 16 KHz speech wave files along-with their International Phonetic Alphabet (IPA) transcription for all the three modes. Four speakers are chosen from each mode of speech. Among four speakers two are male and two are female speakers. This accounted to (3 modes X 4 speakers) = 12 speakers per language. And total number of speakers resulted to (11 languages X 12 speakers) = 132 speakers for 11 languages. For each speaker, 10 sentences are considered for analysis. For 11 languages, a total of (132 speakers X 10 sentences) = 1320 sentences are used for analysis. The conversation mode of speech contained both genders together in the same conversation. Hence, for conversation mode of speech four conversations per language is considered.

III. Automatic Pitch Accent Contour Transcription System

The Automatic Pitch Contour Transcription System (APCTS) has 2 steps. In the first step, pitch contour is extracted from the speech signal using zero-frequency filtering (ZFF) based method. In the second step, pitch contour transcription is derived from the pitch contour using a set of pitch labels. Figure 1 shows the block diagram of the proposed APCTS. Pitch contour is extracted by passing the speech signal through the Zero frequency filter. A non-linear smoothing is applied on the pitch contour to smooth the pitch contour. Trend removal technique is applied on smoothed pitch contour to get intonation like contour. Finally, pitch labels are assigned based on the range of intonation contour to get pitch contour transcription.

A. Pitch Contour Extraction

Instantaneous fundamental frequency (F0) at epochs are obtained as described in [10]. Speech signal is passed through a Zero frequency resonator to obtain the epoch locations. An epoch location indicates the instants of significant excitation of the vocal tract system. The Pitch contour extraction steps are shown in Figure 2. A segment of speech signal, the corresponding ZFF signal is shown in (a), (b). Epoch locations and the corresponding pitch contour is shown in (c) and (d) respectively.

The steps involved in deriving pitch contour from ZFF signal is as follows:

1) The speech signal s[n] is differenced to de-emphasise low frequency components

$$x[n] = s[n] - s[n-1] \tag{1}$$

2) The differenced speech signal is passed through a cascade of zero-frequency resonators given by

$$y_0[n] = -\sum_{k=1}^{4} a_k y_0[n-k] + x[n] \tag{2}$$

where $a1 = -4$, $a2 = 6$, $a3 = -4$, and $a4 = 1$

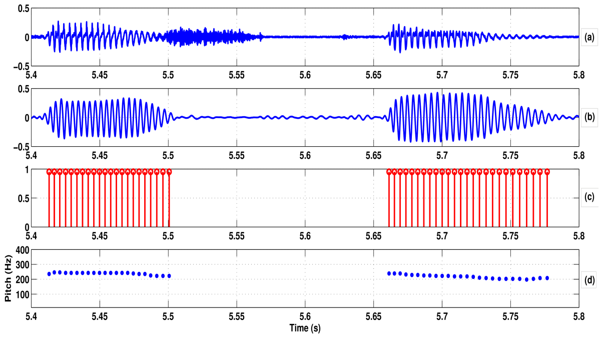

Fig. 2. *Illustration of pitch contour extracted from zero frequency filter signal. (a) A Segment of speech waveform, (b) zero frequency filter output signal, (c) epoch locations and, (d) estimated pitch contour (in Hz)*

3) The trend in $y_0[n]$ is removed by subtracting the mean computed over a window of length equal to average pitch period of long segment of speech signal at each sample. The resulting signal $y[n]$ is the ZFF signal, given by

$$y[n] = y_0[n] - \frac{1}{2N+1}\sum_{m=-N}^{N} y_0[n+m] \qquad (3)$$

4) The epoch locations are obtained from the instants of positive zero crossings of the ZFF signal.
5) Pitch contour is obtained by the reciprocal of the difference between successive epoch locations.

The steps involved in the extraction of pitch contour from ZFF signal is illustration in Figure 2.

B. Automatic Pitch Contour Transcription

The pitch contour obtained from the ZFF signal in III-A is not free from the local roughness and sharp discontinuities. Hence, a non-linear smoothing technique as in [11] is used to smooth the pitch contour. A non-linear smoothing technique consists of combination of non-linear filter followed by a linear filter. A linear filter such as low pass filter is capable of removing noise like variations in the signal but does not preserve sharp discontinuities, while a non-linear filter such as running median preserves the sharp discontinuities in the signal. A 20 point Hanning window is used as a linear filter. Hanning window has a lowpass filtering characteristics, which is capable of removing high frequency noise like component superimposed on the pitch contour. A median filter with 3 point followed by 5 point is used as a non-linear filter to remove the isolated peaks due to measuring errors and to preserve the sharp discontinuities in the pitch contour. The smoothed pitch contour is segmented into pitch and non-pitch regions. The pitch region refers to voiced segments in speech signal with time continuous pitch values. The pitch contour obtained from the ZFF signal and pitch and non-pitch region marked smoothed pitch contour is shown in Figures 3 and 4 respectively.

The steps involved in deriving the pitch contour transcription is given as follows:

1) Input speech signal is Zero frequency filtered to obtain pitch contour as discussed in subsection III-A.

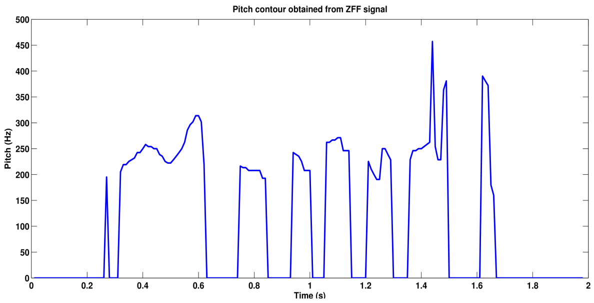

Fig. 3. Illustration of pitch contour obtained from ZFF filtering with pitch outliers

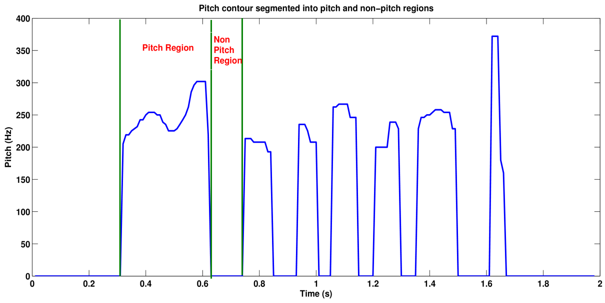

Fig. 4. Illustration of smoothed pitch contour with pitch and non-pitch regions

2) Pitch contour obtained from the ZFF signal is smoothed by non-linear smoothing technique.
3) The histogram of the smoothed pitch contour is obtained and a probability distribution function (PDF) is fitted to fix the pitch range for transcription. The histogram distribution of the pitch contour and the fitted PDF is shown in 6.
4) The pitch transcription range is obtained by thresholding the fitted PDF with two standard deviations from the mean. The pitch range is set to alienate extreme pitch values.
5) The pitch contour obtained in step 2 is divided into pitch and non-pitch regions. For each pitch region, the trend in the pitch contour is removed by subtracting each pitch value by mean calculated for that region. The trend removed pitch contour represents a intonation like contour.
6) The saddle points in the each intonation contour is obtained by computing local maxima's and minima's, which represents the pitch accent changes in the pitch contour.
7) The pitch transcription range as obtained in step 4 is divided into four equal sub-bands from very low to very high. The labels considered for each sub-band are very-low (VL), low (L), high (H) and very-high (VH).
8) The location (in time) of each saddle point obtained from step 6 is marked as either VL, L, H or VH based on the the pitch value in Hz at that location belonging to any of the sub-band range as mentioned in step 7.
9) A label no-pitch (NP) is assigned to the silence and unvoiced regions.

The steps involved in the automatic pitch contour transcription

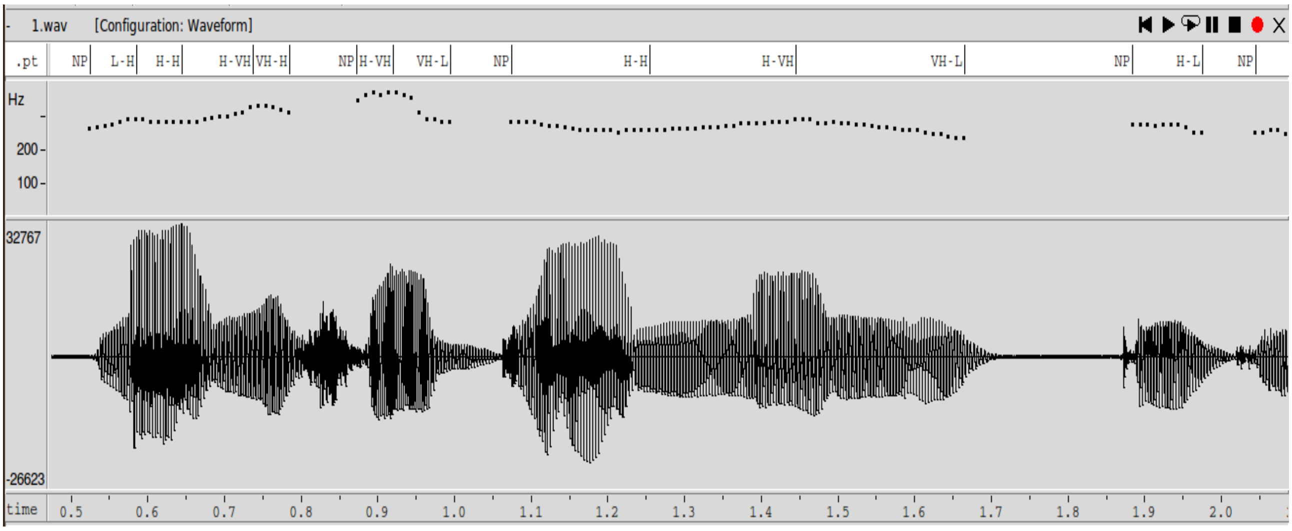

Fig. 5. Illustration of Pitch Contour Transcription obtained from proposed APCTS.

TABLE I. Average Misclassification Error Rate for 11 Indian languages. (M1 = Male1, M2 = Male2, F1=Female1, F2=Female2)

Language	Read				Extempore				Conversation				Mean MER
	M1	M2	F1	F2	M1	M2	F1	F2	C1	C2	C3	C4	
Assamese	6.22	5.63	7.33	3.36	4.45	5.35	6.48	4.91	17.23	20.41	23.27	20.26	10.40
Bengali	4.67	5.00	5.62	4.11	2.29	3.82	3.96	2.15	14.76	10.65	16.21	17.19	7.53
Gujrati	7.37	8.73	2.09	6.47	7.52	8.36	4.93	7.06	7.08	9.12	10.76	19.97	8.28
Kannada	4.49	6.11	5.86	4.11	7.22	6.17	5.33	8.21	21.93	10.91	14.48	12.98	8.98
Malayalam	6.23	3.67	6.16	4.69	6.31	7.82	12.61	5.62	17.44	19.60	5.88	14.66	9.22
Manipuri	7.39	8.54	5.60	7.53	7.33	6.76	5.75	5.59	23.30	15.93	11.64	12.11	9.78
Marati	9.74	7.78	9.74	5.05	5.65	6.74	6.47	5.99	9.55	15.08	18.55	9.55	9.15
Odia	5.64	5.65	4.54	5.62	4.79	5.52	6.97	4.65	7.39	15.72	12.04	14.02	7.71
Punjabi	9.02	7.11	6.17	10.07	9.61	7.15	5.32	6.77	26.00	17.89	15.65	13.50	12.12
Telugu	7.43	8.80	9.75	7.86	10.97	14.18	10.07	11.90	16.67	9.01	8.70	9.74	10.42
Urdu	6.74	6.67	5.59	8.55	7.50	7.23	4.69	5.33	22.23	21.08	13.46	12.99	10.17

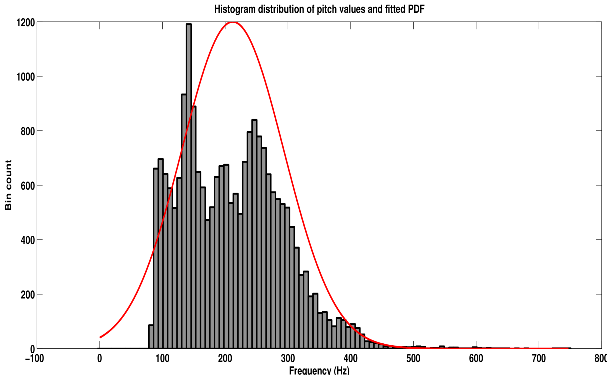

Fig. 6. Illustration of histogram distribution of pitch values and the fitted PDF

is illustrated in Figure 7.

C. Evaluation of Automatic Pitch Contour Transcription System

Subjective evaluation is carried out for evaluating the proposed APCTS. Subjects are made to visualise the variations in the intonation contour and asked to check whether the variation is reflected in automatically derived pitch contour transcription. Wavesurfer tool is used to display the pitch contour obtained from ZFF signal in one pane, pitch contour transcription in the another pane and speech waveform in yet another pane. A screen shot of the multi layer pitch contour transcription

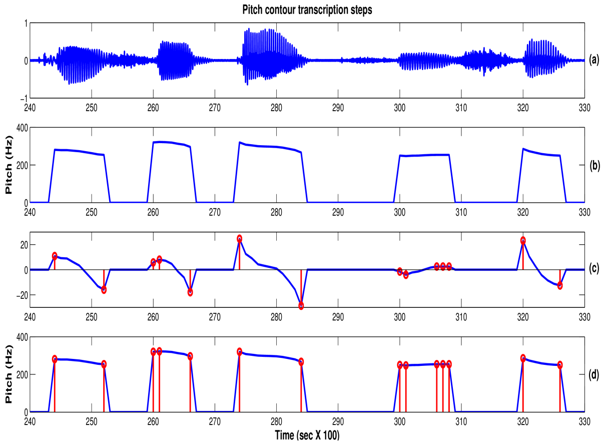

Fig. 7. Illustration of steps involved in pitch contour transcription. Segment of speech signal, smoothed pitch contour, saddle points marked on intonation contour and the location of saddle points with the corresponding pitch values are shown in (a), (b), (c) and (d) respectively.

is shown in Figure 5. Subjects are instructed to note down the false labelling, no labelling and any other error if present in the automatic transcription. Based on the metrics obtained, confusion matrices are created for all the speakers. The average misclassification rate is determined for each speaker.

Table I shows the language-wise average misclassification error rates (MER) for 11 Indian languages. The results are tabulated separately for read, extempore and conversation modes of speech. For read and extempore modes of speech, average misclassification error rates of 2 male and 2 female speakers is tabulated, whereas for conversation mode of speech MERs for 4 conversations are shown. First column shows the list of Indian languages. Second to fifth columns show the details of read speech. The extempore mode of speech is tabulated in sixth to ninth. Next four columns lists the MERs for conversation mode of speech. Last column shows Mean MER, which is the average of MER of each row. From the table it is observed that, Bengali has least Mean MERs among 11 Indian languages, while the Punjabi has highest Mean MERs. In all the languages considered, read speech has lower MERs compared to extempore and conversation modes of speech. Among extempore and conversation modes of speech extempore has better MER than conversation mode of speech. The MERs of Odia are more consistent compared to other languages. Gujarati has least MER of 2.06% for a female speaker, while the Punjabi has highest MER of 26.00% for a conversation. It can be observed that the highest Mean MER is 12.12% for Punjabi. This means Mean MER has never crossed 12.12% for any language. Hence, we conclude that performance of proposed APCTS is more than 87.88% for any language. The read speech has higher MERs for male speakers compared to female speakers, whereas the extempore speech has higher MERs for female speakers than male speakers.

IV. Summary and Conclusion

In this paper, a method to automatically transcribe the pitch contour from the speech signal is presented. The pitch contour is generated using ZFF-based method. The saddle points on the intonation phrase contour are identified and labelled with appropriate tonal labels. The proposed automatic pitch contour transcription is evaluated on 11 Indian languages. Read speech has shown lower average misclassification error rates compared to extempore and conversation modes of speech. Among extempore and conversation modes of speech, extempore has better average misclassification error than conversation mode of speech. Bengali has least Mean MERs, while Punjabi has highest Mean MERs. The performance of proposed APCTS is observered to be more than 87.88% for all 11 Indian languages.

Acknowledgment

We would like to thank Department of Information Technology, Ministry of Communication and Information Technology (DIT), Govt. of India for providing their sponsorship to carry out this work.

References

[1] Lee W.R, *English Intonation: A New Approach*. North-Holland, 1958.

[2] Wells J.C., *English Intonation: An Introduction*. Cambridge University Press, 2006.

[3] Cooper-Kuhlen Elizabeth, *Introduction to English Prosody*. Edward Arnold, 1986.

[4] R Ravi Kiran, Sunil Kumar. S.B, Manjunath K E, Biswajit Satapathy, Apoorv Chaturvedi, Debadatta Pati and K Sreenivasa Rao, “Automatic Phonetic and Prosodic Transcription for Indian Languages : Bengali and Odia,” in *Tenth International Conference on Natural Language Processing*, 2013.

[5] Silverman K, Beckman M, Pitrelli J, Ostendorf M, Wightman C, Price P, Pierrehumbert J, Hirschberg J, “Tobi: A standard scheme for labeling prosody,” in *Proc. Int. Conf. Spoken Lang.*, 1992.

[6] Wightman C and Ostendorf M, “Automatic labeling of prosodic patterns,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 2, p. 469481, 1994.

[7] Ross K and Ostendorf M, “Prediction of abstract prosodic labels for speech synthesis,” in *Computer Speech and Language*, 1996.

[8] Chen K Hasegawa-Johnson M Cohen A, “An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model,” in *International Conference on Acoustic Speech Signal Processing*, 2004.

[9] Ananthakrishnan S and Narayanan S, “An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model,” in *International Conference on Acoustic Speech Signal Processing*, 2005.

[10] K. S. R. Murty and B. Yegnanarayana, “Epoch extraction from speech signals,” *IEEE Transactions on Audio, Speech, Languange Processing*, pp. 1602–1613, 2008.

[11] Lawrence R. Rabiner, Marvin R. Sambur and Carolyne E. Schmid, “Applications of a nonlinear smoothing algorithm to speech processing,” *IEEE Transactions on Acoustics, Speech and Signal Processing* , vol. ASSP.23, pp. 552–557, 1975.