Automatic Phonetic and Prosodic Transcription for Indian Languages : Bengali and Odia

R Ravi Kiran*, **Sunil Kumar. S.B***, **Manjunath K E***, **Biswajit Satapathy***,
Apoorv Chaturvedi*, **Debadatta Pati**+, **K Sreenivasa Rao***
*School of Information Technology
*Indian Institute of Technology Kharagpur, India - 721302
+Balasore College of Engineering and Technology, Sergarh, Balasore - 756060
{r.ravi.kiran.88, sunil220552, ke.manjunath, biswajit2902, chaturvdiap}@gmail.com,
debapati2003@yahoo.com, ksrao@iitkgp.ac.in

Abstract

In this paper, we present an automatic method for transcribing phonetic and prosodic information present in speech. Here, phonetic information refers to sequence of symbols representing sound units present in speech. Prosodic transcription refers to duration patterns of the sequence of sound units, temporal variations of pitch and pause patterns in speech. For deriving automatic phonetic transcription, we have explored Hidden Markov Models (HMMs), FeedForward Neural Networks (FFNNs) and Support Vector Machines (SVM). For deriving prosodic transcription, we use Vowel Onset Point (VOP) to derive duration information of sound units, Zero Frequency Filtering (ZFF) method for pitch contour transcription and short term energy for break indices transcription. For carrying out the present study, we have used speech corpora of two Indian languages namely Bengali and Odia. The same framework could be extended to any Indian languages.

***I*ndex Terms—** Phonetic Transcription, Prosodic Transcription, Bengali, Odia, Pitch, HMM, SVM, FFNN, Segmentation, Zero Frequency Filter, Break Indices

1 Introduction

Transcription is the visual representation of speech using some known notation. Modern speech research activities require advanced speech corpora which contain transcription of message as well as transcription of phonetic and prosodic information. International Phonetic Alphabet (IPA) is one of the most commonly used notational system for getting both prosodic and phonetic transcription.IPA enables us to distinctly represent each sound unit with a unique symbol. Hence, we have transcribed our speech data using IPA symbols (The International Phonetic Association, 2013). The transcription has been organized into multiple layers. One layer is used for phonetic transcription and 3 layers are used for prosodic transcription. Thus PPT results in a multi-tier transcription as shown in Figure 1.

The process of manual transcription involves a human listener, who is capable of understanding the spoken language and transcribing message information into text. The transcriber should also be able to detect variations in prosody of the speech utterance and represent the same using IPA symbols. In languages like Bengali and Odia, the basic spoken units are known as aksharas (consonant + trailing-vowel). These aksharas may be loosely viewed as syllable like units. Thus, prosodic information like tone (extra high, high, mid, low, extra low) and duration (long, half long, extra short) (The International Phonetic Association, 2013) are marked with respect to these syllables only. For pitch contour transcription, the transcriber should represent the temporal variation of pitch at phrase level. Phrase level transcription has been observed to model the human perception of pitch more closely compared to other levels like word level or sentence level. For transcription of break indices, the transcriber should be able to detect variations in pauses made by the speaker and represent the same in the transcription. Transcription is then subjected to objective evaluation, where transcription is exchanged be-

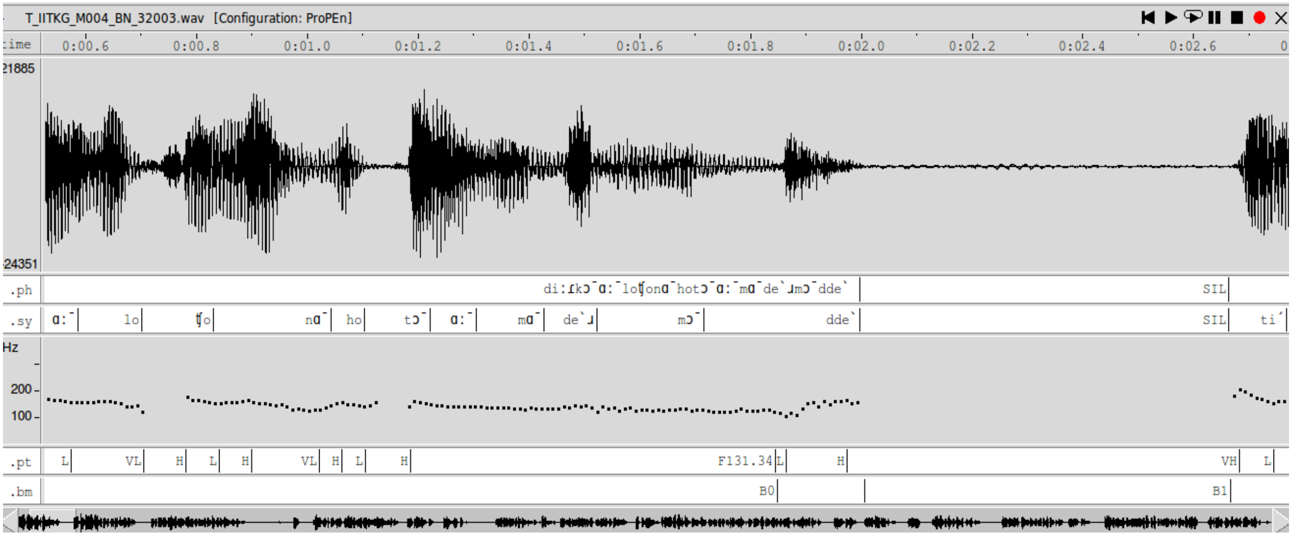

Figure 1: Illustration of Multi-tier Transcription showing speech waveform; Phonetic transcription; Prosodic transcription - Duration pattern transcription, pitch contour transcription, and break indices transcription

tween transcribers and manually verified.

Manual PPT process begins with phonetic transcription, followed by prosodic transcription and completes with evaluation. Phonetic transcription is a one-tier transcription which requires only one layer for representation. Prosodic transcription is a three-tier transcription that requires 3 layers for representation. Different processes followed in prosodic transcription are duration pattern transcription, pitch contour transcription and break indices transcription. In duration pattern transcription, basic sound units like phones or cluster of sound units (like syllables) are analyzed to identify the start and end boundaries of the units. Through identification of the start and end boundaries, we can determine the location and duration of each unit in the speech. Thus, this transcription yields a time aligned labeled data which not only gives us the location and duration of each unit, but also serves as indices for search operation on single query unit (word) or cluster of query units (sentence) in speech database. The pitch contour transcription is a process in which the transcriber listens for variations in the intonational pattern of the speech. These pattern changes are labeled appropriately as very high (VH), high (H), low (L) or very low (VL) depending upon the magnitude of the change. As magnitude difference is relative to one's perception, pitch contour transcription is prone to inconsistency. In break indices transcription, a transcriber is required to label breaks based on the perception of a break's duration. Based upon the size of the duration perceived, the breaks are labeled as B1, B2 or B3. B1 represents word breaks; B2 represents phrase breaks; and B3 represents sentence breaks.

Although manual PPT aims to enrich the transcription with as much phonetic and prosodic information, there are several limitations of this process. As manual transcription is based on human perception, inconsistency is the biggest problem. On a large speech corpus, consistency of transcription is highly critical for research activities and any inconsistency is to be eliminated through evaluation. Manual PPT also requires tremendous manual effort and is known to be tedious, cumbersome and monotonic in nature. In our work, we attempt to overcome some of these limitations via automation. Automated PPT system replaces a human transcriber and produces outputs closely similar to manual transcription.

The rest of the paper is organized as follows: Section 2 describes speech corpus used in our work. Section 3 discusses automatic phonetic transcription. Section 4 discusses automatic prosodic transcription. Lastly, conclusion and our plan of future work is discussed in Section 5.

2 Speech Corpus

The Phonetic and Prosodically Rich Transcribed Speech Corpus (PPRT) developed at IIT Kharagpur (Sunil Kumar, 2013) is the speech corpus used in our study. The PPRT speech corpus contains speech data of two Indian languages - Bengali and Odia. Total duration of the corpus is 20 hours, with 10 hours data for each language. Speech corpus is further classified into three modes based on the nature of prosody: Read speech, Conversation speech and Extempore speech. These modes are common to both languages and amount of data for each mode is tabulated in Tables 1. The number of speakers in each mode is shown in Table 2 .

Table 1: Distribution of Data in Speech Corpus

Mode	Bengali (in hrs)			Odia (in hrs)		
	(F)	(M)	(T)	(F)	(M)	(T)
Read	2.95	2.05	5	2.5	2.5	5
Conver	0.55	1.95	2.5	1.25	1.25	2.5
Extempo	0.75	1.75	2.5	1	1.5	2.5
Total	4.25	5.75	10	4.75	5.25	10

Table 2: Number of Speakers in Speech Corpus

Mode	Bengali			Odia		
	(F)	(M)	(T)	(F)	(M)	(T)
Read	28	13	41	30	30	60
Conver	13	10	23	3	3	6
Extempo	7	7	14	7	9	16

Table 1 displays the amount of data distributed in terms of hours for each mode, gender and language. F denotes the amount of data collected from female speakers, M for male speakers and T denotes the total amount of data collected for a particular mode. Table 2 displays the number of speakers, who amounted to the collected speech data for a given mode and language. The purpose of classifying speech into various modes arises due to the fact that each mode contains a unique variation of prosody.

For each mode, data was collected from speakers of both male and female genders. The age of the speakers considered was in the range 25-40 years. The corpus was recorded at 16 KHz sampling rate and 16 bits precision. Source for data collection for mentioned modes above are : Closed room with controlled acoustics, recordings from digital sources like Internet, Television recording (News channels) and radio.

3 Automatic Phonetic Transcription

Automatic phonetic transcription was carried out at two different levels - at phone level and syllable level. Phone level phonetic transcription provides the sequence of phones present in the speech utterance. Syllable level transcription provides the sequences of syllables present in the speech utterance.

3.1 Automatic Phone Level transcription

In (Manjunath, 2013) , we have developed an automatic Phonetic Transcription System (PTS) to determine the sequence of phones present in the spoken utterance. Two separate PTSs were developed to decode the spoken utterances for Bengali and Odia languages. We have used 35 phones for developing Bengali PTS and 32 phones for developing Odia PTS. We have developed PTS using Hidden Markov Models (HMMs) and FeedForward Neural Networks (FFNNs). Mel-frequency Cepstral Coefficients (MFCCs) are used as features for building the models. Separate models were built for Speaker Dependent (SD) and Speaker Independent (SI) cases. In case of SD system, during training and testing phases, speech utterances from all the speakers are used, but training and test data are completely different. For SI system leave-one-speaker-out approach is used for system building and evaluation. In this approach out of ‘n’ speakers, ‘n-1’ speaker data is used for developing the system and remaining speaker data is used for evaluation. Likewise, in ‘n’ turns, the system is evaluated with respect to all the speakers. The performance of the system is calculated by averaging the performance of all the systems.

The details of PTS developed using HMMs are as follows. A set of context-independent monophone HMMs with one Gaussian per state are flat-initialized. Each HMM has 7 states including 2 non-emitting states (1,7) with transitions from left to right and no skips. Initially flat-start HMMs are created using a prototype model. The flat start HMMs are then re-estimated using the training data with the embedded re-estimation to perform Baum-Welch training. We have carried out re-estimation 8 times to get more accurate models. The viterbi decoding is used for decoding of test speech utterance into data sequence of phones.

The details of PTS developed using FFNNs are as follows. We have used three-layered FFNN with sigmoid nonlinearity at the hidden layer, and softmax nonlinearity at the output layer. The MFCC features will form the feature vectors, which are fed to the input layer of FFNN. The corresponding phone labels will be fed to the output layer of the FFNN. We have used back-propagation algorithm for training the FFNN. During training, multiple passes are made through the entire set of training data. Each pass is called an epoch. The result of training is a weights file. The weights file can then be used as the acoustic model to convert the features of an unseen test utterance into posterior probabilities of each phone class. These posterior probabilities will be decoded to phones using a decoder (QuickNet, 2010).

To evaluate the performance of PTS, the decoded output is compared with the original reference transcription. The comparison is carried out by performing optimal string matching using dynamic programming (HTK, 2009). The performance accuracy of Bengali and Odia PTSs using HMMs and FFNNs are shown in Table 3. From the results, it is observed that the phone recognition performance is better in case of SD system compared to SI system for both languages. The accuracy of phone recognition is found to be superior with FFNN models compared to HMM models. About 12-13% improvement in the accuracy is observed in case of FFNNs for both Odia and Bengali languages. With respect to languages, the phone recognition accuracy for Odia is found to be 5-8.5% more compare to Bengali language.

Table 3: Phone recognition accuracy of Prosodic Transcription System based on HMMs and FFNNs

Mode	Bengali		Odia	
	HMM	FFNN	HMM	FFNN
SD	41.65	53.87	46.18	59.88
SI	38.06	50.16	45.90	58.71

3.2 Automatic syllable level transcription

Automatic transcription at syllable level is carried out using SVM models. We developed two separate systems for Bengali and Odia (Manjunath, 2013). In this work, syllable models are developed using consonant-vowel (CV) units alone. The primary reason for considering CV units is that in Indian context, most of the syllables are of the form CV, CCV and CVC (Anil Kumar, 2012), where CV is common across all forms. The difficulties for other forms (CCV and CVC) are related to lack of enough examples for the syllable units. If we consider all CVs, CCVs and CVCs for recognition, the recognition accuracy is expected to be low due to a large number of classes. Therefore, in this work, we considered only CV units which have enough number of examples for building the models. Based on this assumption, the syllable recognition systems for Bengali and Odia are developed using 67 and 58 CV units, respectively. Syllable recognition systems are built separately for Bengali and Odia.

For each language, two separate systems were developed using SVMs. The systems differed based on the feature vectors used for building the models. For both systems, features are extracted using Vowel Onset Point (VOP) (S.R.M Prasanna, 2009) (Jainath, 2013) as an anchor point. The VOPs are determined using spectral energy present in the glottal closure regions of speech signal. VOP indicates the onset of vowel in speech. For CV, CCV and CVC units, the region before the VOP is constant region and the regions following the VOP can be viewed as transition and steady regions of vowel. The Sequence of steps in the VOP detection method can be seen in (Manjunath, 2013).

For developing system-1, a combination of 5-feature vectors representing MFCCs are used. Among them, the first feature vector is derived from the 20 ms constant speech region adjacent to VOP, and the remaining four feature vectors are derived from the transition region adjacent to VOP. In case of system-2, a combination of 15-feature vectors are derived from constant, transition and steady regions of CV units. Each region contributes 5 feature vectors to form a desired sequence of 15-feature vectors. The constant and transition regions of CV units are marked as 40 ms segments before and after the VOP. Steady vowel region is marked as a segment of 40 ms adjacent to transition region. MFCC feature vectors are extracted using 20 ms frames with 5 ms frame shift. Each feature vector represents 13 MFCC coefficients.

The recognition accuracy of the systems developed in Bengali and Odia is given in Table 4. From the results, it is observed that the recognition accuracy is better for the systems developed using features extracted from constant, transition, and steady region of VOP compared to the system developed using features extracted for constant and transition regions. The performance evaluation is carried out in SD and SI modes. The recognition performance for SD mode observed to be higher compare to SI mode of evaluation. For system-1, the variation between SD and SI modes is about 8% and 28% respectively for Bengali and Odia. Likewise for system-2, the difference between SD and SI modes is about 9% and 28% respectively for Bengali and Odia. The recognition accuracy for both Bengali and Odia is almost same in SI mode whereas for SD mode, the recognition accuracy is very high in case of Odia compared to Bengali.

Table 4: Recognition performance at Syllable level (SD = Speaker Dependent SI = Speaker Independent)

Mode	Bengali		Odia	
	System-1	System-2	System-1	System-2
SD	38.02	49.48	60.00	69.66
SI	30.42	40.26	32.67	41.59

4 Automatic Prosodic Transcription

In this work, prosodic transcription consists of duration pattern for the sequence of syllables, intonation pattern for phrases and pause pattern represented by break indices. Here, duration information for the sequence of syllables is derived using VOP, intonational patterns are determined by using Zero Frequency Filtering (ZFF) method and break indices are determined by short term energy (STE) of the speech.

4.1 Duration pattern transcription

In this work, duration pattern transcription refers to the duration of sequence of syllables. Syllable boundaries are accurately determined using VOP and instants of significant excitation. In general, syllable is defined as C^mVC^n, where m and n are greater than or equal to zero. The syllable constitutes of a vowel and one or more consonants. The vowel is known to be nucleus of a syllable. In Indian languages, aksharas generally correspond to syllables. An akshara is typically in one of the following forms: V, CV, CCV, CCVC and CVCC, where C is consonant and V is vowel. Here, VOPs for the given speech signal are determined by using spectral energy in 500-2500 Hz band in glottal closure regions of the speech signal. The procedure for detecting the VOPs in the speech signals is as can be seen in (Manjunath, 2013).

The syllable boundaries are marked using the VOPs as anchor points. The VOP indicates the onset (beginning) of the vowel in a syllable. Based on previous studies (Jainath, 2013), the beginning of syllables are marked 30 ms before the VOP. This region indicates the consonant segment of a syllable, preceding the vowel. The end of the syllable is marked based on the uniformity of the successive epoch intervals present in the vowel region of the syllable. If the syllable is followed by a pause or unvoiced region, the uniformity in the successive epoch intervals break down at the end of the vowel. In this work, uniformity in successive epoch intervals is determined by analyzing the difference between successive epoch intervals with respect to a threshold. The threshold considered here is 0.5 ms. Based on this threshold, we can determine the continuity of the uniform epoch intervals. If the syllable is followed by a voiced consonant or semivowel of the following syllable, the uniformity in the epoch intervals continue till the following VOP. In this case, the end of the syllable may be marked as 30 ms ahead of the following VOP, which in reality signifies the beginning of the next syllable.

4.1.1 Evaluation

In this work, we have considered 100 sentences to evaluate the accuracy of syllable level segmentation method. To determine the deviation of VOPs, we have used the manual marked VOPs as a reference. The number of VOPs in the 100 sentences is found to be 1454.

Table 5: Syllable-level segmentation accuracy

Avg Deviation	Accuracy (in %)		
	Match	MISS	SPU
25ms	53.63	37.2	9.14
40ms	75.98	16.67	7.35
50ms	85.47	8.34	6.19

Performance of VOP detection method is analyzed using parameters like match rate, missing rate, and spurious rate. VOPs detected within 50-ms deviation to the reference VOPs are considered as genuine VOPs. The ratio (in %) of number of genuine VOPs detected to the total number of reference VOPs are measured for different time resolutions (25-50 ms). The ratio (in %) of undetected VOPs to the total number of reference VOPs is termed as missing rate (MISS). VOPs detected other than genuine VOPs are termed as spurious VOPs. The ratio of spurious VOPs (in %) detected to the number of reference VOPs is termed as spurious rate (SPU). The accuracy of the VOP detection has been tabulated in Table 5. The accuracy of the duration pattern transcription is directly related to the accuracy of the VOP detection. Here, if the VOPs are detected accurately, it has been observed that the transcription is also accurate in terms of boundary alignment of the syllables.

4.2 Pitch contour transcription

Automatic pitch contour transcription was implemented in a two-step process. The first step is the pitch extraction and second step is the transcription process.

4.2.1 Pitch Extraction

Pitch is defined as the rate at which vocal fold vibrates. It is a source feature. Zero-frequency-filtering(ZFF) is a method used to capture source feature. So, an algorithm based on ZFF (KSR Murthy, 2008) (Yegnanarayana, 2009) is used for pitch extraction. In ZFF method, the speech is passed through zero frequency filter and the trend of raising or decaying is removed from the output of the ZFF by using a mean smoothing filter. The Negative to positive zero crossings in the zero frequency filter signal corresponds to location of instants of significant excitation (epochs). The fundamental frequency is then obtained by taking reciprocal of interval between successive epochs as shown in Figure 2.

4.2.2 Pitch transcription

Pitch transcription process works in conjunction with pitch extraction process, where difference between successive pitch values is appropriately labeled. Pitch transcription has also been illustrated in Figure 1.

The sequence of steps used for deriving the transcription has been outlined in Algorithm 1. The $f0\_array$ contains sequence of pitch values for each 10 ms fragment. This sequential arrangement is done to maintain the sequence and timing-details of pitch values. To assign label to each record, we break the $f0\_array$ into segments of smaller sizes like phrases, sentences or in between speaker-breaths. The bandwidth of pitch variation within these smaller segments is then distributed among the four labels of S_0 and labels are assigned to each record of the $f0\_array$. However, not each record goes into transcription but only those records which mark a transition in the label.

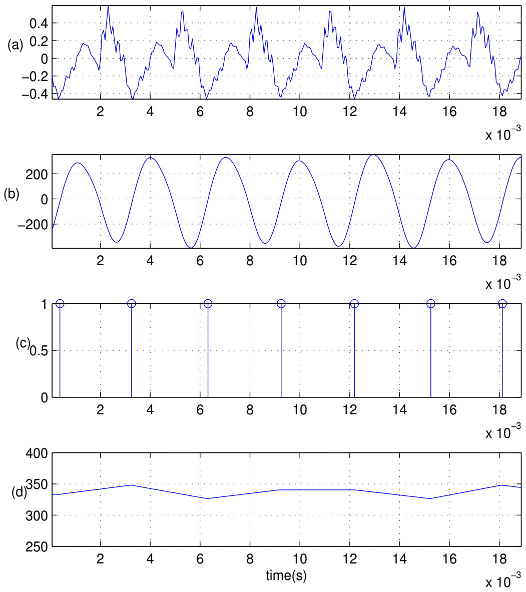

Figure 2: Illustration of pitch extraction using zero frequency filter. A 40ms Segment of (a) speech waveform, (b) zero frequency filter output signal, (c) epoch locations and, (d) estimated pitch contour for the signal (in Hz)

Algorithm 1: Procedure for pitch transcription process

```
Input: Array containing $f0$ values
Output: Transcription

begin
    $S_0 = [VH, H, L, VL]$;
    $S_1 \leftarrow$ scan pitch from $f0\_array$;
    forall elements of S1 do
        assign label $r$ from $S_0$ to element $e$ of $S_1$;
        record$_e \leftarrow [start\_time, end\_time, label]$;
        put record$_e$ in set $S_2$;
    end
    forall elements of S2 do
        if curr label! = next label then
            put recorde in set S3 ;
        else
            skip record$_e$;
        end
    end
    return elements of $S_3$ as $transcription$;
end
```

4.2.3 Evaluation

To evaluate the automatic pitch transcription, a pitch contour is rebuilt from the automatic transcription and manual transcription. For generating the pitch contour from the transcription, we have used linear interpolation technique to generate the sequence of pitch values. The degree of similarity between the reconstructed contours of the manual transcription and automatic transcription against the original pitch-contour of the speech gives us a score of which process (manual or automatic) performs better in representing the original pitch contour. The score is computed using dynamic-time warping technique, by tracing the minimum distance via cost matrix. A graph of the reconstructed contours and original contour is shown in Figure 3.

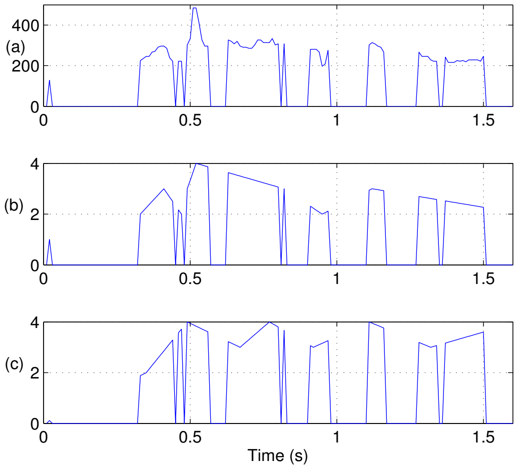

Figure 3: Illustration of the pitch contour of (a) The original contour, (b) Reconstructed contour from the automated labeling process and (c), Reconstructed contour from the manual labeling process

For performance evaluation, 100 phrases were considered. Out of the 100 phrases, it was observed that for 92 phrases, the automatic transcription process fared a higher similarity score to the original pitch contour. Thus automatic pitch transcription helps in speeding up the transcription process. It was also observed that the consistency had significantly improved across the pitch contour transcription.

4.3 Automatic Break indexes transcription

To automate the transcription process, we used short term energy of speech signal computed for the sequence of frames with frame size of 20 ms and frame shift of 10 ms. By nature of production, we know that clean speech signal consists of voiced, unvoiced and silence regions. The energy associated with voiced region is significantly larger when compared to unvoiced region and the energy associated with silence region will be minimum or negligible. Also we know that the duration of unvoiced segments is very small compared to voiced speech segments or silence segments. Thus an analysis of short term energy can be used for primary detection and transcription of break indices.

We observed a drastic fall in the number of breaks, as we move from B1 to B3. This is because a speaker tends to take lesser number of larger breaks during speech production. For automatic transcription, we first detect the location of break and then depending on its duration, the break is transcribed as B1, B2 or B3. The duration thresholds for labels B1, B2 and B3 were derived from the histograms. To evaluate the performance of the automatic transcription system, the transcriptions obtained from the manual process and automatic process was compared manually. The following measures were derived for analyzing the performance of automatic break indices transcription system.

- Total number of Breaks - B1, B2 and B3 in manual transcription.
- Total number of Breaks - B1, B2 and B3 in automatic transcription.
- Match - Number of breaks, whose position and label match in automatic and manual transcription.
- Mismatch (Mis)- Number of breaks in which only the positions matched, but the break was incorrectly labeled.
- Spurious marking (Spu)- Number of breaks which were detected only in automatic transcription, but not present in manual transcription.

A sample evaluation has been shown in Table 6. It displays the total number of break marks de-

Table 6: Comparison of Manual and Automatic Break indices transcription

Label	Manual	Auto	Match	Mis	Spu
B1	54	240	52	2	186
B2	8	4	2	2	0
B3	3	10	3	0	7

tected in manual and automatic transcription process for a 1.5 minute speech file. It also displays the count of matching labels in which the classification of the break matches and the location coincides in both transcriptions. The count of misclassified breaks is also shown. By examining numbers in the table, one can infer that the automatic transcription process detects a large number of spurious breaks which are not generally perceived by human transcribers in manual transcription. This number shows a sharp increase as the size of the break reduces.

5 Conclusion and Future Works

In this paper, we explored different techniques to automate phonetic and prosodic transcription. In case of phonetic transcription, the PTS was designed to work at two levels viz, phone level and syllable level. The phone level PTS was designed using HMMs and FFNNs. The best accuracy obtained in this case is 59.88% for Odia and 53.87% for Bengali. The syllable level PTS has been designed using VOP and SVM based system to recognize CV units in the speech. The accuracy of the PTS obtained in this case is 69.66% for Odia and 49.48% for Bengali. For automatic duration pattern transcription VOPs were explored to mark syllable boundaries. The accuracy obtained in this case is 85.47%. For automatic pitch contour transcription, the transcription system was designed based on ZFF method for pitch extraction and a labeling function for generating the automatic transcription. The performance of the automatic pitch contour transcription system is found to be 92%. Finally, the break indices transcription system was implemented by using short term energy of speech signal. From the analysis of each of these results, it can be concluded that although automatic transcription system is not as accurate as manual transcription process, it still is very helpful, as it simplifies manual work by a large extent. It also increases the consistency of transcription. However so, newer techniques are still being explored to improve the performance and accuracy of the automatic transcription system to model it closer to the manual transcription.

References

The International Phonetic Association. *International Phonetic Alphabet, Available: http://www.langsci.ucl.ac.uk/ipa/index.html*.

Cambridge University Engineering Department. 2009. *The Hidden Markov Model Toolkit, Available :http://htk.eng.cam.ac.uk* .

Speech Group at the International Computer Science Institute. 2010. *QuickNet Software, Available : http://www1.icsi.berkeley.edu/Speech/qn.html*.

Sunil Kumar. S. B, K. Sreenivasa Rao, Debadatta Pati. 2013. *Phonetic and Prosodically Rich Transcribed Speech Corpus in Indian languages : Bengali and Odia*. International Oriental COCOSDA (Accepted).

Manjunath K E, K. Sreenivasa Rao, Debadatta Pati. 2013. *Development of Phonetic Engine for Indian languages : Bengali and Oriya*. International Oriental COCOSDA (Accepted).

Manjunath K E, Sunil Kumar. S. B, Debadatta Pati, Biswajit Satapathy and K. Sreenivasa Rao. 2013. *Development of Consonant-Vowel Recognition Systems for Indian Languages : Bengali and Odia*. INDCON-2013 (Accepted).

K.S.R. Murthy and B Yegnanarayana. 2008. *Epoch Extraction From Speech Signals*. IEEE Transactions on Audio, Speech, and Language Processing, volume-16:1602-1613.

B Yegnanarayana and K.S.R Murty. 2009. *Event-Based Instantaneous Fundamental Frequency Estimation From Speech Signals*. IEEE Transactions on Audio, Speech, and Language Processing, volume-17:614-624.

S. R. Mahadeva Prasanna, B. V. Sandeep Reddy and P. Krishnamoorthy. 2009. *Vowel Onset Point Detection Using Source and Spectral Peaks and and Modulation Spectrum Energies*. IEEE Transactions on Audio, Speech, and Language Processing, volume-17:556-565.

Anil Kumar Vuppala, Jainath Yadav, Saswat Chakrabarti and K. Sreenivasa Rao. 2012. *Vowel Onset Point Detection for Low Bit Rate Coded Speech*. IEEE Transactions on Audio, Speech, and Language Processing, volume-20:1894-1903.

Jainath Yadav and K. Sreenivasa Rao. 2013. *Detection of Vowel Offset Point From Speech Signal*. IEEE Signal Processing Letters, volume-20:299-302.