Development and Automation of Phrase Model for Urdu Speech Corpus

[1]Aneeta Niazi, [1]Saba Urooj, [1]Benazir Mumtaz, [1,2]Tania Habib
[1]Center for Language Engineering (CLE),
Al-Khwarizmi Institute of Computer Science (KICS),
[2]Computer Science and Engineering Department
University of Engineering and Technology (UET),
Lahore, Pakistan.
aneeta.niazi@gmail.com, {first name.last name}@kics.edu.pk

Abstract

A phrase model for the annotation of Break Indices (BI) in Urdu speech corpus has been presented. A detailed acoustic analysis has been carried out to understand the patterns of phrase breaks in 1 hour of recorded Urdu speech. A four level phrase model has been proposed, including BI levels 0, 1, 2 and 4. From the outcomes of this analysis, rules have been formulated for automating the process of BI tagging in Urdu speech corpus. For this purpose, the annotated information of word boundaries, Part Of Speech (POS), intonation, stress and pauses from the Urdu speech corpus has been utilized. As the features indicating the prosodic behavior of the pitch contour, including stress and intonation, have already been accurately tagged in the speech corpus, the results obtained from the automatic BI tagging are quite promising. The automatic BI labeling system has provided coverage of 97.8%, and accuracy of 98.3% for the tagging of unseen data.

Keywords: *Break Index, BI, Phrase model, Urdu speech, Automatic annotation, prosodic modeling*

1. Introduction

In a language, a word or a group of words co-existing as a single conceptual unit is known as a phrase [1]. Human speech contains words clustered together to form phrases. These phrases are separated by pauses, or a change in the speaker's tone.

In written text, punctuation is commonly used to indicate phrase boundaries e.g. comma, full stop etc. However, while speaking, humans insert phrase breaks, even in the absence of punctuation. These breaks usually occur in speech while moving from one word to another, mostly for expressing emotions and intentions [2]. A phrase break is also referred to as Break Index (BI).

For building Text-to-Speech (TTS) systems, an adequately large, well annotated speech corpus is one of the basic requirements. The corpus should contain accurate annotations for prosodic features, such as BI, stress and intonation, to make the TTS sound as human-like as possible.

Several efforts have been made to develop an international standard for annotating prosody in speech. One of the earliest and most popular prosody tagging standards is ToBI (Tones and Break Indices) [3]. In ToBI, the different types of pauses in speech are represented by numbers from 0 to 4. A typical break between two adjacent words is represented by BI level 1. A break in place of a comma is indicated by BI level 3, whereas a distinctive pause in between speech segments is represented by BI level 4. This tagging convention is used by many languages. However, this could not be generalized for all the languages. Researchers developed variations of ToBI to suit the requirements of speech annotation for their particular languages e.g. J-ToBI for Japanese, G-ToBI for German, ToDI for Dutch, B-ToBI for Bengali etc. [4].

The existing Urdu speech corpus [4] has been annotated for all the necessary prosodic features of speech. A 5 level scale (from 0 to 4) has been used for marking BI, based on, duration of breaks, and lengthening of phrases, pitch contour and glottalization. During an acoustic analysis for understanding the tonal patterns of Urdu speech, it has been observed that the structure of Urdu phrases is very different from English phrases in the spoken Urdu sentences. In Urdu, word and phrase boundaries are not marked in accordance with the rules followed by English; i.e. the behavior of Urdu BI is very different from the conventions followed by English. Due to the presence of accentual phrase, Urdu phrase structure resembles with the phrase structure of South Asian languages. Therefore, there is a need to re-design the phrase model, and develop new standards for marking BI for Urdu speech.

The manual labeling of phrase boundaries in speech corpus is a time consuming and laborious activity. Machine learning based systems can be used for making the annotation process efficient, but such

systems require large amount of annotated data, which is not readily available for under-resourced languages, such as Urdu.

In this paper, we present a detailed acoustic analysis that has been carried out for developing a phrase model for Urdu speech. An automatic BI labeling system has been proposed to annotate the BI in the Urdu speech corpus. Section 2 presents a survey of the existing work for the topics of phrase modeling and break index annotation. The results and discussions are covered in section 4, whereas the findings of this research are concluded in section 5. In section 6, we propose the future directions which can be pursued to further investigate the process of phrase modeling.

2. Literature Review

For long-form reading, phrase model serves as one of the most important components for improving the naturalness of a TTS system [2]. The phrase model has been constructed by analyzing textual features such as dependency tree features, Part Of Speech (POS) and word embeddings. For improving the prediction of phrase boundaries, these features have been given as input to train Bidirectional Long Short Term Memory (BiLSTM) and Classification And Regression Trees (CART) based systems. Both subjective and objective testing has been carried out to compare the performance of BiLSTM and CART systems. The evaluation results have shown that better performance has been obtained by using word embeddings and BiLSTM.

A language independent BiLSTM-CRF (Conditional Random Fields) model has been proposed for prosodic boundary prediction [5]. The architecture consists of three layers, i.e. word embeddings, BiLSTM and CRF. These three layers learn from task-specific embeddings, past and future features and sentence level information respectively. The system has been evaluated for Mandarin and English speech. The results show that using the proposed model, the intonational phrase prediction has been significantly improved as compared to the traditional BiLSTM method.

BI labels have been automatically annotated for Japanese and English speech, using only the information extracted from the speech signal [6]. The automatic labeling is carried out without using any other prior information, such as transcriptions or word boundaries. For this purpose, spontaneous Japanese speech has been used to train BiLSTMs. The trained system is used to annotate Japanese and English speech, and a cross-lingual comparison is made with the monolingual English labeling system. The evaluation results have shown that the system trained with Japanese speech performed better for the BI labels 1 and 2, while the system trained with English speech performed better for the Break Index label 3. The less frequent labels in the data have not been accurately detected. The proposed cross-lingual model can be applied when sufficient amount of data is not available for training a monolingual break index labeling system.

An analysis is carried out to observe the impact of the size of focus constituents on phrase boundaries in French [7]. The experimental results have shown that an accentual phrase boundary gets converted into an intermediate phrase boundary, if it forms the right edge of a narrow focus constituent. However, an intermediate phrase boundary remains unaffected in the presence of a narrow focus constituent in its surrounding context.

Intonational phrase break prediction models have been developed to automatically predict phrase breaks in American English [8]. Binary classifiers, based on logistic regression from the LLAMA machine learning toolkit are used. 50 hours of recorded speech are used for building the system. The prediction models are data driven, based on features including lemmatized words, POS, punctuation, distance from punctuation, as well as dependency-relation features. An overall prediction accuracy of 84.7% has been obtained.

A model for detecting prosodic boundaries in Russian speech, using syntactic as well as acoustic information, has been presented [9]. It is based on a two level architecture, where the possible phrase boundaries are marked by using syntactic information, with the help of a dependency tree parser in the first step. In the second step, a Random Forest (RF) classifier uses a small set of acoustic features, such as tempo, pitch range and amplitude etc., to mark the actual prosodic boundaries. The duration of pauses has been reported to be the best amongst all acoustic features used for predicting prosodic boundaries.

For Indian languages, the analysis of phrases becomes very difficult if there is no punctuation in the text [10]. In read sentences, the units in between the pauses are considered as phrases for analysis. It has been observed that the length of inter-pausal units follows a Gamma distribution. An analysis of shape and scale parameters of speech has shown that these parameters have dependence on the location of inter-pausal units. This information is utilized to improve the prosody modeling of TTS system for four Indian languages. The results have shown considerable improvement in the naturalness of synthesized speech.

An automatic prosodic transcription system has been reported for Bengali and Odia languages [11]. 3 levels of breaks are annotated, i.e. word breaks are represented as B1, phrase breaks as B2 and sentence breaks as B3. For labeling BI automatically, short term energy (STE) of speech signal is considered. The

energy associated with silence is negligible as compared to unvoiced and voiced regions. Also, unvoiced segments have very small duration as compared to silence and voiced segments. The duration thresholds for B1, B2 and B3 have been determined by histograms. From the results, it is observed that the automatic BI tagging system detected many spurious breaks, which were not perceived during manual tagging.

For Urdu speech, a 5 level scale (from 0 to 4) has been presented for annotating break indices [4]. Acoustic features including pitch contour, duration of pauses and glottalization have been considered for analysis. 1036 files from CLE Urdu speech corpus are used; comprising of simple sentences. An automatic BI labeling system is developed to annotate 10 hours of speech. The reported analysis does not include complex predicates and compound sentences.

3. Proposed Methodology

This section includes the details of the data collection, analysis and rules developed to formulate phrase model for Urdu.

3.1. Data Acquisition

From CLE Urdu Speech Corpus, 1403 files have been acquired for carrying out break index analysis. Out of these files, 983 files are used as training data to develop the automatic BI labeling utility. The remaining 420 files have been kept as unseen data for testing the performance of the automatic BI labeling system.

Table 1 shows the counts of the BI tags found in the training data.

TABLE 1 Training data counts

Tag	Total Count
0	194
1	1320
2	3948
4	1410
Total	**6872**

Table 2 shows the counts of the BI tags found in the testing data.

TABLE 2 Testing data counts

Tag	Total Count
0	86
1	1097
2	2663
4	1012
Total	**4877**

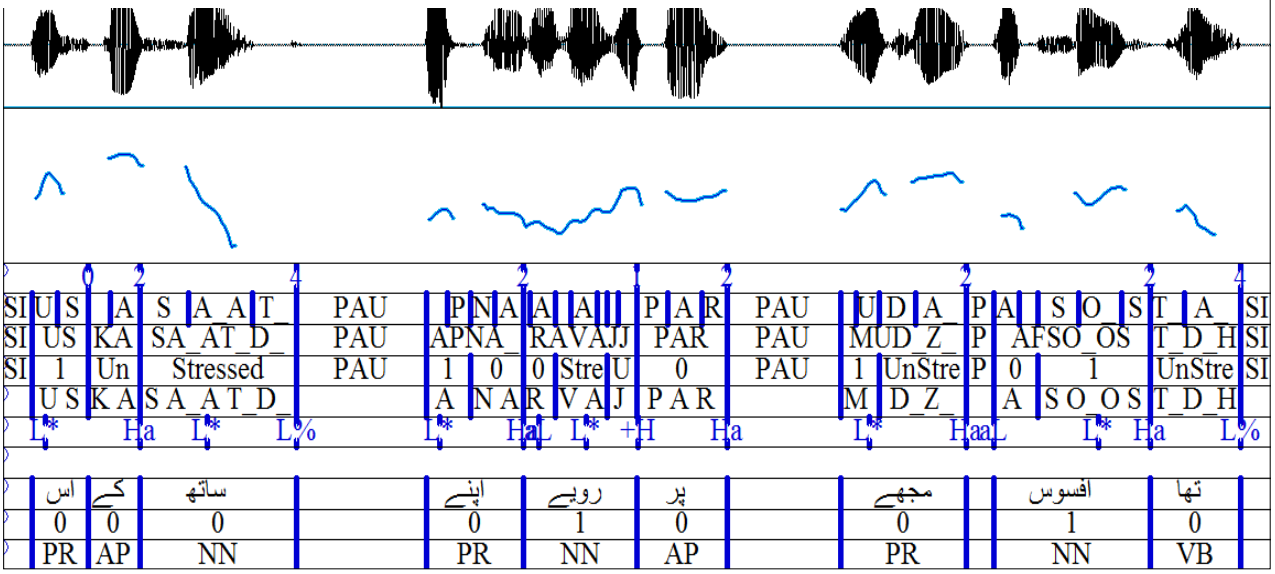
From tables 1 and 2, it can be observed that in Urdu speech corpus, BI level “2” tag has the highest frequency, whereas BI level “0” tag has the lowest frequency. This shows that accentual phrase boundary i.e. level '2' BI occurs most frequently as accentual phrase is the basic unit of Urdu prosody. BI level '0' i.e. the words using zair-e-izafat, vao izafat and the pronoun case marker combinations occur less frequently in the data selected for the phrase model analysis of Urdu.

3.2. Rules for Automating BI Tagging

As BI is marked between words and from words to silence, so the onset of each word of a sentence would not be assigned any BI level. The stress [12], intonation [13] and POS [14] tiers have already been accurately marked in the Urdu speech corpus. The information provided by these tiers is used to formulate rules for automatic BI marking.

Figure 1 shows an example of a speech file containing the Urdu sentence, “US KA_Y SA_AT_D_H APNA_Y RAVAJJA_Y PAR MUD_Z_HA_Y AFSO_OS T_D_HA_A”. (Translation: *I was sorry for my behavior with him*),

FIGURE 2 An example of a speech file that has been automatically annotated for all BI levels (0, 1, 2 and 4).



in which all the 4 levels of break indices (0, 1, 2 and 4) have been automatically labeled, according to the automatic BI labeling rules.

The rules formulated for automating the process of BI tagging are given as follows.

3.2.1. Break Index 4

1. In the first run, mark "4" if the following tag is "SIL".
2. In the second run, mark "4" aligning with every "%" symbol on the intonation tier i.e. "LH%", "H%" and "L%" as % symbol denotes a full intonation phrase boundary and is used at the end of clauses and sentences only.

In Figure 1, it can be observed that BI level '4' is marked at the end of the word "T_D_HA_A", in accordance with the first rule, as it is the last word of the sentence, followed by "SIL". At the end of the word "SA_AT_D_H", BI level '4' is marked in accordance with second rule, as it aligns with the intonation tag "L%".

3.2.2. Break Index 2

Mark 2 aligning with every "Ha" or "La" tag on intonation tier. "Ha" and "La" tags show an accentual phrase boundary. An accentual phrase usually comprises of a pitch accent and a boundary tone and it is the smallest unit of Urdu prosody instead of a word as multiple words are joined to form one accentual phrase in Urdu. In other languages e.g. English, BI level '2' is used to mark strong juncture with no tonal markings. But we have used BI level '2' for accentual phrases as accentual phrase is found in South Asian languages only.

In Figure 1, it can be observed that BI level '2' is marked at the end of words, "KA_Y", "APNA_Y", "PAR", "MUD_Z_HA_Y" and "AFSO_OS", aligning with the "Ha" intonation tag.

3.2.3. Break Index 0

Mark "0" in the following contexts:

1. At the onset boundary ofzair-e-izafat, A_Y with the POS tag CN
2. At the onset boundary of vao-izafat, O_O with the POS tag CN
3. Between the following pronouns and case markers:
   a. Between US/اس with the pos tag PR and KA_Y/ے with the pos tag AP
   b. Between UN/ان with the pos tag PR and KA_Y/ے with the pos tag AP

Personal pronouns in Urdu completely lose their word boundary when followed by certain case markers taking BI level '0' between them and behaving as one prosodic word. See Appendix A for the list of such pronouns and case markers.

In Figure 1, it can be observed that BI level '0' is marked at the end of the word "US", as its POS tag is "PR" and it is followed by the word "KA_Y" with the POS tag "AP", in accordance with the rule 3 (a).

3.2.4. Break Index 1

Mark 1 at all the remaining word boundaries as BI level '1' is used for default word boundary, when two words are not merged as in BI level '0', or there is no

accentual phrase as in BI level '2', or there is no silence or full intonation phrase as in BI level '4'.

In Figure 1, it can be observed that BI level '1' is marked at the end of the word "RAVAJJA_Y", as it does not follow the rules mentioned for BI levels '0', '2' and '4'.

4. Results and Discussion

Table 3 shows the results obtained after automatically labeling the unseen testing data, and comparing it with the manually labeled gold standard corpus.

TABLE 3 Automatic BI labeling results obtained with unseen testing data.

Tag	Total Count	Marked Count	Coverage (%)	Accuracy (%)
0	86	86	100	100
1	1097	1084	96	97
2	2663	2657	99	97
4	1012	983	96	99
Total	**4887**	**4810**	**Avg=97.8**	**Avg=98.3**

From the above table, it can be observed that the level "0" tag has been automatically marked with 100% accuracy, whereas its coverage is also 100%. A very high percentage of coverage has been obtained for all of the four BI tags, with an overall coverage of 97.8%. The results obtained for accuracy are also quite promising, as an average accuracy of 98.3% is obtained.

The reason for obtaining such high quality performance is the fact that the automatic BI labeling system only utilizes the information from already accurately annotated tiers i.e. stress, intonation and part of speech (POS) tags from the speech corpus, and does not rely on extracting information from the pitch contour at run time for BI tagging.

The stress tier contains information about the stressed and unstressed syllables. The intonation tier indicates the high and low tones of the pitch at accentual phrase boundaries and pitch accents. This annotated information has been used to get an idea of the pitch contour, for marking BI in the speech corpus at any point of time.

5. Conclusion

A detailed acoustic analysis has been carried out for understanding the behavior of phrase breaks, to develop a phrase model for Urdu speech. The features considered for this purpose include POS, annotated intonation and stress information.

It has been observed that Urdu speech contains four levels of break indices i.e. 0, 1, 2 and 4, for establishing prosodic relationships between words.

The outcomes of this analysis have been used to develop an automatic BI labeling system. The developed system has provided coverage of 97.8%, and an accuracy of 98.3% with unseen testing data, which is quite promising.

6. Future Work

In future, the automatic BI labeling utility developed during this research will be used to annotate phrase breaks in the remaining 9 hours of CLE Urdu speech corpus. This annotated corpus will be used as input to train the speech synthesis module of Urdu TTS system, to improve the naturalness of synthesized Urdu voice.

Analysis for phrase modeling of long-form Urdu speech can be carried out in order to observe the patterns of phrase breaks during the reading of long Urdu paragraphs. The outcomes of such analysis can be utilized to improve the naturalness of Urdu TTS for audio books and screen readers.

References

[1] (2019) Lexico powered by Oxford. [Online]. https://www.lexico.com/en/definition/phrase

[2] Viacheslav Klimkov et al., "Phrase break prediction for long-form reading TTS: exploiting text structure information," in Interspeech, Stockholm, Sweden, 2017.

[3] Joe Crumpton and Cindy L. Bethel, "A Survey of Using Vocal Prosody to Convey Emotion in Robot Speech," International Journal of Social Robotics, vol. 8, no. 2, pp. 271-285, April 2015.

[4] Benazir Mumtaz, Saba Urooj, Sarmad Hussain, and Ehsan Ul Haq, "Break Index (BI) Annotated Speech Corpus for Urdu TTS," in Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Technique (O-COCOSDA) , Bali, Indonesia, 2016.

[5] Yibin Zheng, Jianhua Tao, Zhengqi Wen, and Ya Li, "BLSTM-CRF Based End-to-End Prosodic Boundary Prediction with Context Sensitive Embeddings in A Text-to-Speech Front-End," in Interspeech, Hyderabad, India, 2018.

[6] Marco Vetter, Sakriani Sakti, and Satoshi Nakamura, "Cross-lingual Speech-based Tobi Label Generation using Bidirectional LSTMs," in International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019.

[7] Amandine Michelas and James S. German, "Focus Marking and Prosodic Boundary Strength in French," International Journal of Phonetic Science (Phonetica), 2018.
[8] Taniya Mishra, Yeon-jun Kim, and Srinivas Bangalore, "Intonational phrase break prediction for text-to-speech synthesis using dependency relations," in International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, QLD, Australia, 2015, pp. 4919-4923.
[9] Daniil Kocharov, Tatiana Kachkovskaia, and Pavel Skrelin, "Prosodic boundary detection using syntactic and acoustic information," Computer Speech and Language, vol. 53, pp. 231-241, 2019.
[10] Jeena J. Prakash and Hema A. Murthy, "Analysis of Inter-Pausal Units in Indian Languages and Its Applications to Text-to-Speech Synthesis," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 10, pp. 1616-1628, June 2019.
[11] R. Ravi Kiran et al., "Automatic Phonetic and Prosodic Transcription for Indian Languages : Bengali and Odia," in 10th International Corference on Natural Language Processing, Noida, India, 2013.
[12] Benazir Mumtaz, Saba Urooj, Sarmad Hussain, and Wajeeha Habib, "Stress Annotated Urdu Speech Corpus to Build Female Voice for TTS," in 18th Oriental COCOSDA/CASLRE Conference, Shanghai, China, 2015.
[13] Benazir Mumtaz, Saba Urooj, and Sarmad Hussain, "Urdu Intonation," Journal of South Asian Linguistics, vol. 10, October 2019.
[14] Tafseer Ahmad et al., "The CLE Urdu POS Tagset," in Language Resources and Evaluation Conference (LERC 14), Reykjavik, Iceland, 2014.

Appendix A

Lists of Urdu pronouns and case markers to be considered for Break Index 0 rule.

Pronouns
MA_E_N
MUD_Z
T_DUD_Z
A_AP
T_DU_U
T_DUM
HAM
IS
US
UN
SAB
VO_O
IN

Case Markers
SA_Y
NA_Y
KO_O
KA_A
KI_I
KA_Y
MA_Y_N