2016 Conference of The Oriental Chapter of International Committee
for Coordination and Standardization of Speech Databases and Assessment Technique (O-COCOSDA)
26-28 October 2016, Bali, Indonesia

Break Index (BI) Annotated Speech Corpus for Urdu TTS

Benazir Mumtaz, Saba Urooj, Sarmad Hussain and Ehsan Ul Haq
Center for Language Engineering
Al-Khawarizmi Institute of Computer Science, University of Engineering and Technology
Lahore, Pakistan
{First name. last name}@kics.edu.pk

***Abstract*—This study presents 10 hours of speech corpus annotated at Break Index (BI) level for Urdu text to speech system (TTS). For the speech corpus annotation, 5-steps scale of BI is presented in this study, which assists us to understand the prosodic relationship between word boundaries and phrase boundaries in Urdu. Moreover, an algorithm for Urdu TTS is designed to detect the 5-steps scale of BI automatically using acoustic cues such as duration of phrasal breaks, intonational patterns, glottalization and phrase lengthening. This algorithm annotates 10 hours of speech automatically and reports 98.4% accuracy compared to perceptual marked BI. This research also discusses the controversial issue of prosody to syntax mapping and syntax to prosody mapping in Urdu. The results indicate: (i) in the context of noun phrase and case phrase (NP-KP), prosodic break occurs between NP-KP to eliminate syntactic ambiguity; (ii) string of KPs does not require explicit prosodic break because case markers act as phrase separators in KP-KP context (iii) verb phrase (VP) in an utterance is preceded by a prosodic break to differentiate it from other phrases in an utterance and (iv) NPs of complex predicates prosodically act as single unit in an utterance.**

Keywords—Break index (BI)/prosodic breaks; BI identification algorithm; syntactic breaks; prosodic and synthectic phrase alignment; Urdu TTS

I. INTRODUCTION

While speaking, humans intend to cluster words together with perceptible pauses. These pauses are termed as phrase breaks or Break Index (BI) and are crucial in grouping speech utterances into meaningful units of information. Text to speech systems require learning of these phrase break models to enhance their naturalness and intelligibility [1], [2].

This paper deals with devising an algorithm for automatic prediction of phrase breaks for Urdu TTS. Phrase breaks are present in the speech signal in the form of pauses [3]. Duration of pauses, phrase lengthening and intonational changes are the acoustic cues to determine the level of the disjuncture i.e. whether the pause is a medium level pause depicting a comma [4] in a sentence or a full pause depicting a full stop in a sentence. There are some other levels of disjuncture as well which show that the pause is not strong in some cases e.g. after conjunction words that join two sentences or phrases or the boundary drop between two words showing them as one.

Traditionally, grammar based parse trees, part-of-speech (POS) information, and decision trees using Ney-Essen clustering algorithm are taken as an input to predict the prosodic phrasal break [5], [6]. However, researches [2], [4] on South Asian languages such as Hindi, Tamil and Bengali claim that although there is some association between syntactic and prosodic phrases, the relationship between them is non-linear. Moreover, to predict prosodic breaks, high quality syntactic parsers are requisites which can automatically generate syntactic information with high level of accuracy. These methods cannot be used in Urdu language context where the syntactic parsers are not readily available and manual marking of data is time consuming and expensive. Therefore, in this research, for the identification and classification of levels of disjuncture in speech, native speakers' intuition is used and this intuition is then mapped on acoustic cues such as pauses' duration, glottalization and pitch changes to devise BI identification algorithm. Moreover, an experiment is conducted to see the extent to which prosodic and syntactic phrases can be aligned in Urdu.

This paper is organized into following sections. Prior studies on Break Index levels are detailed in Section II. The process used to record ten hours of speech, identification criteria of Urdu BI-levels and their corresponding acoustic cues are listed in Section III. In Section IV, annotation of speech corpus at BI-level using BI identification algorithm and annotation of speech corpus at syntax level to observe the correspondence between prosodic and syntactic phrases are discussed. The quality assurance of BI annotation is presented in Section V while results, discussion and conclusion are described in Section VI, VII and VIII respectively.

II. LITERATURE REVIEW

Break index indicates prosodic correlation between two sequential words [7]. Generally, this prosodic correlation can be shown using five levels of disjunction on the scale from '0' to '4'. However, different languages use diverse range of scale to mark BI tier. Japanese TOBI (J-TOBI) system uses 4-step scale of BI ranging from level '0' to level '3' [7] to present prosodic grouping between the words. Level '0' in J-TOBI expresses weakest degree of disjuncture whereas level '3' expresses strongest degree of disjuncture. In addition, all the

978-1-5090-3516-8/16/$31.00 ©2016 IEEE

breaks in J-TOBI are assigned BI-levels aligning exactly with the boundary of the word. Unlike J-TOBI, Finnish TOBI (FIN-TOBI) [8] model uses 5-step scale to interpret the association between words where '3' and '4' coincide with the prosodic constituents i.e. intermediate and full intonational phrases respectively.

In German Tones and Break Indices (GTOBI) model, BI-levels are not explicitly marked unless they deviate from level '0' and '1'. In this case, three levels of break indices are used i.e. '2', '3' and '4' to indicate the strength of phrasal boundaries [9]. In GTOBI, level '2' is further divided into '2r' and '2t' representing a rhythmic break with tonal continuity and a tonal break with rhythmic continuity respectively. Similar to FIN-TOBI, level '3' and level '4' in GTOBI coincide with the intermediate phrase and the full intonational phrase.

In Hindi [10], [11] and Bengali [12], prosodic grouping between words is categorized at three prosodic levels i.e. accentual phrase (AP), intermediate phrase (ip) and intonation phrase (IP). These prosodic levels (AP, ip and IP) roughly correspond to single word, phrase/clause and sentence respectively and are denoted with high boundary tone. Survey of prosodic models of various languages indicates that different languages use different BI-levels to identify and understand the relationship between words and phrases boundaries. However, existence of BI-levels is still an unexplored area in the context of Urdu language. Therefore, this study attempts to find out how many levels of BI exist in Urdu and how these levels can be differentiated from each other using acoustic cues.

Beckman and Ayers [13] claim that the identification of BI-level depends on the subjective interpretation of the listener. Therefore, it is possible that one language offers two diverse scales to mark BI tier. For example, in English language, two prosodic annotation systems exist i.e. English Tones and Break Indices (E-TOBI) [13] and Rhythm and Pitch (RaP) [14]. Both systems offer different levels of BI. E-TOBI claims that degree of perceived disjuncture between words can be expressed on the scale from '0' to '4' whereas RaP opposes five levels of perceived disjuncture between the words. It uses only two levels of break indices to show the major and minor phrase boundaries denoted with '))' and ')' respectively. It also claims that it is not obligatory for tonal labels to coincide with phrasal boundaries. This subjective identification of word or phrase boundaries highlights another motivation of this research i.e. to translate subjective identification criterion of BI-levels into objective identification criterion by developing an algorithm, which will help Urdu TTS to express the strength between word boundaries objectively.

BI-Levels also assist to understand the prosodic hierarchy of a language [15]. Lahiri, Plank [16] and Eisenberg [17] claim that lower prosodic constituents of prosodic hierarchy (i.e. syllable and prosodic word denoted by level '0' and '1') are determined by rhythmic principles whereas higher prosodic constituents (i.e. phonological and intonational phrases denoted by level '3' and '4') are determined by syntactic boundaries. In this research, as an initial step, we will also try to investigate the correspondence between the syntactic phrases and the phonological or intonational phrases of Urdu to see the extent to which syntactic boundaries can be predicated using level '3' and '4' information.

III. METHODOLOGY

For the identification and the classification of BI-levels, ten hours of Urdu corpus is recorded in mono-form at 48 kHz sampling rate. The text for the ten hours of speech is extracted from three different corpora i.e. Urdu news corpus, Urdu digest corpus and 1 million word corpus using Greedy algorithm [18]. For the recording of 10 hours corpus, a professional female speaker has been hired. The recordings are conducted in a soundproof chamber using PRAAT. After the recording, speech is annotated at segment, syllable, word and stress levels [19] using Case Insensitive Speech Assessment Method Phonetic Alphabet (CISAMPA). The multi-tier annotation process used for the development of Urdu TTS is described in [20].

For the automatic identification of break indices levels, one hour of speech containing 1036 sentences is perceptually annotated by two expert linguists. This manually marked data (1036 sentences) is then divided into training and testing data comprising of 854 and 182 sentences respectively. The sentences of the training data are used to understand the prosodic relationship between word boundaries. Based on the analysis of training data, BI identification algorithm for Urdu has been developed whereas the sentences of the testing data are used to test the performance of BI identification algorithm.

Analysis of training data highlights that native speakers use five levels of break indices ranging from level '0' to level '4' to express relationship between word boundaries. Moreover, these levels can be differentiated from each other using acoustic cues i.e. duration of pauses, highness and lowness of pitch track and phrase initial and final glottalization. Details of the break indices levels along with their acoustic cues are described in the subsequent sections.

A. Break Index Level 0

Level '0' is assigned when the word boundary between two words is completely erased. In Urdu, boundaries between words are erased in the following cases:

- A pronoun followed by a case marker as in the words آپ نے (ɑ:p ne:\ You)
- Compound words combined with <ِ> zair or "و" vao izafat as in the words مخلوقِ خدا (məxlu:q e: xʊ̪dɑ:\ creature of God) and غور و فکر (ɣo:r o: fɪkər\ contemplation)
- An aspectual auxiliary followed by a tense auxiliary as in the words ہوں گے (hõ: ge:\ Will be)
- An aspectual auxiliary followed by another aspectual auxiliary as in the words جا رہی (ʤɑ: rəhi: \ is going)

In Urdu, prosodic word boundaries associated with level '1' always carry high boundary tone at the end of last syllable of the word. However, in case of level '0', high boundary tone is

delayed and is realized on the last syllable of the second word indicating two words are uttered as one prosodic word.

B. Break Index Level 1

Level '1' is assigned to the word boundary where there is no disjuncture; phrase initial or final glottalization and vowel lengthening of the last syllable. Acoustically, level '1' correlates with high boundary tone and is donated with H--.

C. Break Index Level 2

Level '2' is assigned to the context where there is:

- Lengthening of the vowel of last syllable but there is no accented syllable within a word or phrase
- A disjuncture/pause but there is no accented syllable within a phrase

D. Break Index Level 3

Level 3 is assigned to the intermediate intonational phrase boundaries (i.e. L-/H-/!H-/^H-). An intermediate intonational phrase has at least one stressed/accented syllable. The acoustic cues used to identify the intermediate phrase are as follows:

1) Weak disjuncture: Intermediate phrase has less strong disjuncture than the full intonational phrase. This disjuncture is visible in the pitch track. In any wave file, if a pause is less than 129ms then it is marked as intermediate intonational phrase boundary by the annotators (see section VII for futher details on how weak disjuncture of level '3' is calculated). Prosodic disassociation found in level '3' lacks the perception of completeness which coincides with the stronger break index level '4'.

2) Pitch reset: Pitch range is also reset for each new intermediate phrase. There are two types of pitch patterns found at the end of a level '3' followed by a weak disjuncture:

- The pitch track rises to the height of the speaker's pitch range suggesting that the utterance has not completed yet. High pitch range followed by a weak disjuncture is an indicator of the end of one intermediate phrase and start of another intermediate phrase.
- The pitch track gradually falls at middle of the speaker's pitch range, not touching the lowest range of the speaker's pitch opposite to level '4' where the pitch falls to the lowest range of the speaker's pitch.

3) Phrase initial and final glottalisation: Perceptually, a native speaker can hear a weak disjuncture before the phrase initial glottalization. Therefore, level '3' is assigned before the start of glottalization even though there is no visible pause between two words. Moreover, phrase final glottalization can also be followed by a small pause or a long pause. If it is followed by small pause of 129ms, it is assigned level '3' but if it is followed by long pause of 456ms, it is assigned level '4'.

E. Break Index Level 4

Level '4' corresponds to full intonational phrase boundaries (i.e. L-L%, L-H%, H-H%, H-L%, H-^H%). The acoustic cues used for identifying full intonational phrase are discussed below:

1) Strong disjuncture: In any wave file, if a pause is 456ms or more than 456ms, it is marked as full intonational phrase boundary by the annotators.

2) Final lowering of f0: Level '4' is assigned to final lowering i.e. a final extra low f0 value at the right edge of an intonational phrase. In level '4', pitch track goes firmly down to the lowest pitch range of the speaker.

IV. ANNOTATION OF SPEECH CORPUS

Annotation process used to mark BI tier on the remaining 9 hours of speech is discussed in Section *'A'* whereas details of syntactic level annotation to find out the correlation between prosodic and syntactic phrases are given in Section *'B'*.

A. Annotation of 9-Hours of Speech Using BI Identification Algorithm

For the automatic annotation of 9 hours of speech, one hour of manually marked data is analyzed. Based on the analysis of training data discussed in previous session, BI identification algorithm has been developed. The BI identification algorithm assigns break indices starting from left to right at the end of each word. The steps of the algorithm are as follows:

1. Find the word boundary and check whether the word boundary is followed by a pause or not.
2. If the word boundary is not followed by a pause, check whether it is followed by a glottalization or not.
3. If the word boundary is not followed by a glottalization, check the phrase lengthening at the end of the syllable using the 'vowel duration analysis table' [20].
4. If there is no phrase lengthening, check the syntactic category of the word. If the word is a pronoun followed by a case marker or aspectual auxiliary followed by a tense auxiliary, or compound word combined with zair or vao izafat, assign level '0'.
5. If the word does not belong to above-mentioned categories, assign level '1'.
6. If the word boundary is followed by a pause, check whether the word has stressed syllable or not. If the word does not have a stressed syllable, assign level '2'.
7. If the word has stressed syllable, check the duration of the pause. If the duration of pause is less than 129ms, assign level '3'. If the duration of the pause is more than 456ms, assign level '4'.
8. If duration of the pause is more than 129ms and less than 456ms, move to the next cue i.e. pitch analysis. If the pitch track rises to the height of the speaker's pitch range and the value is greater than 189 Hz, assign level '3'. If the pitch track goes down to the lowest pitch range of the speaker and the value is less than 139 Hz, assign level '4'.

9. If the pitch is neither in the high range nor in the low range of the speaker, move to the next cue i.e. syntactic structure of the phrase. If the break is preceded by a clause, assign level '4'. If the break is preceded by a syntactic phrase, assign level '3'.

10. Repeat from the step (1) until all the words in an utterance are processed.

Steps from '1' to '8' are incorporated in an 'automatic BI annotation utility'. The remaining word boundaries that require step '9' are considered ambiguous boundaries and are left unmarked. The possible BI-levels for the points that reside inside ambiguous areas could be '3' or '4'. To find out which level from level '3' or '4' is better and brings us closer to native speaker intuition, both '3' and '4' are used to mark the unannotated intervals respectively. The average accuracies of the algorithm with '3' and '4' values for ambiguous area are given in Table 1B. A sample of speech wave file annotated using the BI identification algorithm (marked on first tier) is shown in Fig. 1.

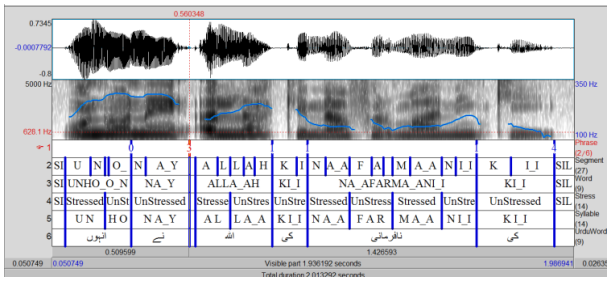

Fig. 1. Break Index Annotated Speech File

B. *Annotation of Speech Corpus at Syntactic Level*

For syntactic level annotation, a data set of 60 sentences is randomly selected from manually marked hour 1. These 60 sentences are manually annotated at syntax level by an expert linguist using Urdu POS tagset [21]**.** After the annotation, the syntactic marked data is automatically compared with break index marked data to investigate the two dimensional analysis:

a) prosodic-syntactic phrase agreement and disagreement
b) syntactic-prosodic phrase agreement and disagreement

Prosodic-syntactic phrase analysis shows the extent to which levels '2', '3', and '4' match with syntactic phrases whereas syntactic-prosodic analysis illustrates the contexts where the prosodic breaks are inserted in the syntactic phrases. The results of prosodic alignment with syntactic phrases and syntactic alignment with prosodic phrases are reported in Section VI.

V. Quality Assurance of Break Index Annotation

Quality assessment of both manually and automatic marked data is done at the BI-levels. Details of quality assessment methods used for manually and automatically marked data are described in sections *'A'* and *'B'* respectively.

A. *Quality Assessment of 1-hour Manually Marked Data*

Quality assessment of 1-hour manually marked data is conducted using following process. Twenty percent of the source data is perceptually annotated by a principal annotator, in addition to the annotation team, as reference annotation. These sentences annotated in duplicate are then automatically compared with the source data to determine the consistency between two annotators in the identification of BI-levels perceptually. In the manually marked data, if the rate of error is more than 5% or the numbers of BI intervals are not same in source and reference files, the data is rejected and sent back for re-annotation. This process is repeated until the rate of error in the training data is less than 5%.

B. *Quality Assessment of Automatically Marked Data Using BI Identification Algorithm*

To explore the stability of BI identification algorithm as compared to the perception of native speaker, 182 sentences have been randomly extracted from the unseen data. These extracted sentences are perceptually annotated by a linguist and then compared with the automatically marked data. The results of alignment between perceptually and automatically marked data are detailed in the Section VI.

VI. Results

Two types of results are reported in this session: BI annotation results and prosody-syntax relationship results. BI annotation results are further subdivided at two levels. Firstly, results of comparative analysis of perceptually marked BI and algorithmic marked BI (Table 1A) are reported. Secondly, annotation of ambiguous region using level '3' and level '4' (Table 1B) are detailed.

TABLE 1A: COMPARATIVE ANALYSIS OF PERCEPTUAL MARKING WITH BI MARKING ALGORITHM

Levels of BI	**BI Annotation Results**			
	Perceptually marked intervals	***Algorithmic marked intervals***	***Difference between perceptually and algorithmic marked intervals***	***Correct intervals Marked***
Level 0	13	12	-1	11 (84.6%)
Level 1	961	963	+2	960 (99.8%)
Level 2	66	53	-13	60 (90.9%)
Level 3	287	218	-69	280 (97.5%)
Level 4	296	259	-37	287 (96.9%)
Total	1623	1505	-118	1598 (98.4%)

Table 1A reports that BI identification algorithm has left 118 intervals (7.27 %) as unmarked. To find out which level from level '3' and '4' brings us closer to native speaker intuition, both '3' and '4' are used to mark the unannotated intervals respectively. After marking the data, the sentences have been compared with perceptually marked data sentences. Table 1B reports the average accuracies of the algorithm with '3' and '4' values for the ambiguous areas.

TABLE 1B: PERFORMANCE STATISTICS OF THE ALGORITHM WITH LEVEL '3' AND LEVEL '4' AS AMBIGUOUS REGION VALUE

BI Levels	Ambiguous Region Values' Results		
	Matched points with perceptual data	*Mismatched points with perceptual data*	*Accuracy*
Level 3	76	42	64%
Level 4	31	87	26%

Table 1B shows that using the value of '3' as a BI-level value for gray areas has improved the accuracy of the algorithm. Therefore, all the unmarked intervals are replaced with level '3'.

Study of prosody-syntax relationship reports two types of alignments: prosodic phrases (i.e. level '2', '3', and '4') alignment with syntactic phrases (i.e. NP, VP etc, see Table 2A) and syntactic phrases alignment with prosodic phrases (Table 2B).

TABLE 2A: PROSODIC PHRASES ALIGNMEMT WITH SYNTACTIC PHRASES

BI levels	Prosody-syntactic alignment	
	Total No. of prosodic phrases	*Alignment between prosodic and syntactic phrases*
Level 2	11	10 (90.9%)
Level 3	77	61 (79.2%)
Level 4	66	66 (100%)
Total	154	137 (88.9%)

TABLE 2B: SYNTACTIC PHRASES ALIGNMENT WITH PROSODIC PHRASES

Syntactic Categories	Syntax-prosodic alignment	
	Total No. of Syntactic Phrases	*Prosodic break between syntactic phrases*
KP-VP	19	12 (63%)
NP-VP	15	6 (40%)
NP-KP	26	20 (76%)
KP-KP	16	2 (12.5%)
Total	76	40 (52.6%)

VII. DISCUSSION

This section presents two types of analyses. Type one analysis gives insight into BI-level annotation whereas type two discusses prosodic-syntactic and syntactic-prosodic alignments.

Data analysis of BI tier indicates that prosodic association between word boundaries in Urdu can be expressed using five level scales ranging from '0' to '4'. Moreover, similar to Hindi and Bengali [11], [12] prosodic words in Urdu end with high boundary tone. However, in the case of level '0' high boundary tone realizes itself differently in different contexts. In the context of 'string of aspectual auxiliaries', the high boundary tone of first word suppresses itself to get connected with the second word and realizes itself at the boundary of second word. This elimination of high boundary tone between two words indicates that two words are uttered as one prosodic word and share '0' level relationship. In the context of 'pronoun with case markers', the high tone does not eliminate itself at the end of the word. It starts from the end of the first word and sustains itself till the end of the second word indicating both words are joined together by sharing the same high boundary tone. In 'compound words' context, the high tone realizes itself completely on the izafat diacritic and the word after izafat diacritic takes the pitch pattern of L* pitch accent and high word boundary. Interestingly, in all the prosodic words combined with level '0', it is the first word that carries stressed syllable. The syllables in second word are always unstressed grouping with the stressed syllables to create a rhythm as shown in the Fig. 2.

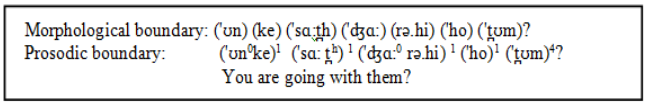

Fig. 2. Prosodic Grouping of Words across Morphological Boundaries

This finding also supports Lahiri's finding [16] that in prosodic constituency, boundary of morphological word dissolves to form the trochaic pattern.

Level '1' in Urdu is a default boundary that appears in the absence of breaks, glottalization and phrase lengthening. Results highlight that most of word boundaries that are considered ambiguous fall within the region of level '3' and level '4'. These two levels are most perplexing levels to differentiate as both share the same acoustic cue i.e. disjuncture. However, duration of disjunction varies for both levels. To find out disjuncture variation, the durations of pauses with BI values of '3' and '4' have been extracted from one hour speech data. The mean values of the pause durations have been calculated to find the two extreme values for marking BI value of '3' or '4'. It has been observed that the mean duration for pause with break index level '3' is 129ms whereas for level '4' it is 456ms. In addition, it is observed that level '3' disjuncture is preceded by a rising pitch contour indicating perception of completeness is lacking which corresponds with level '4'. Moreover, level '2' in Urdu coincides with unaccented coordinate conjunction and subordinate conjunctions (such as اور/ɔːr/and, یا/jɑː/or; لیکن/lekɪn/but and تاکہ/ʈɑːkeː/so that respectively) that join two full or intermediate intonational phrases respectively.

Prosodic-syntactic phrases analysis highlights that the most misaligned prosodic phrase is level '3'. Level '3' misalignment occurs in the contexts of foreign fragments and degree adjectives. Data analysis underlines that foreign fragments comprising of honorifics contain multiple prosodic phrases within one syntactic phrase (as in the phrase □ آپ/ ɑːp səlləllɑːhoːʔəlæːheːvəɑːlehiːvəsəlləm/the holy Prophet peace be upon him). Similarly, degree adjective takes prosodic break if it comes in an adverbial phrase (بڑی احتیاط سے/bəɽi æhʈiɑʈ se/with great care) and acts as a focus particle in the sentence.

Syntactic to prosodic phrase analysis suggests that the maximum alignment occurs in case of NP-KP pair i.e. in 76% of the data NP is followed by a prosodic break. The motivation is if noun is not followed by any case marker, there

is a chance that the noun gets merged with the following noun/pronoun. Therefore, the speaker intends to insert break to differentiate NP from KP and defuse syntactic ambiguity e.g.

1) لوگ اس کو شک کی نگاہ سے دیکھ رہے تھے۔

log/NP/Pause(focus)/ʊs ko/KP/ ʃək ki/KP/ nɪga se/KP/Pause/ ʤekh rəhe ʧhe/VP/[People were suspecting her/him.]

In opposition to this, minimum prosodic alignment (12%) occurs in case of KP-KP pair as the case markers itself act as phrase separator and do not require explicit break e.g. in the example (1) there is no prosodic break observed in both the KP-KP pairs (i.e. ʊs ko/KP/ ʃək ki/KP/ and ʃək ki/KP/ nɪga se/KP). The situation is different when there is an intensifier in the sentence as the intensifier pulls the focus and hence the pause in most of the cases irrespective of intensifier position in the sentence e.g.

2) اسی طرح دنیا میں بھی ایک دوسرے سے بالکل ممتاز تھے۔

isi ʧəhɑ/AdvP/Pause/ ʤʊnya me bʰi/KP/Pause(focus)/ æk ʤusre se/KKP/ bɪlkʊl mʊmʧɑz ʧhe/VP/. [Similarly, they were distinct from each other in the world also.]

The relationship of NP and VP pair suggests that there is 40% alignment and 60% misalignment. The analysis of the data showed that the aligned phrases were all those nouns which were standalone NPs and were not the nouns of complex predicates. The misaligned phrases were nouns of complex predicates and were not followed by prosodic breaks. This can be explained on the basis of the fact that the NPs of complex predicates act as single unit with the verb (or light verb) they are attached with and hence are not followed by any prosodic break. Moreover, the relationship of KP and VP shows 63% alignment. The finding suggests that if a VP is preceded by a KP there would be a preverbal prosodic break to differentiate VP from KP.

VIII. CONCLUSION AND FUTURE WORK

From the above discussion, it can be concluded that Urdu uses 0-4 range scale to express prosodic relationship between words. This prosodic relationship is perceptually predicted and then used to design BI identification algorithm to automate the BI annotation process using acoustic cues (i.e. phrasal breaks, glottalization and pitch patterns). This study also highlights that there is 52.6% correspondence from syntactic to prosodic phases and 88.9% correspondence from prosodic to syntactic phrases. This analysis would act as a seed for developing pause model of speech-prosody interface for Urdu. Currently, simple sentences are used to explore the syntax-prosody mapping. In future, compound sentences and complex predicates would be investigated on larger data to understand the correlation between prosodic and syntactic phrases. Moreover, the role of syntactic and prosodic information in the high quality synthesize speech will also be investigated in the future.

REFERENCES

[1] K. Prahallad, E.V. Raghavendra & A.W. Black, "Semi-supervised learning of acoustic driven prosodic phrase breaks for Text-to-Speech systems," in *Proceedings of 5th International Conference on Speech Prosody,* Chicago, 2010.

[2] A.Vadapalli, K. Prahallad, and P. Bhaskararao, "Significance of word-terminal syllables for prediction of phrase breaks in Text to Speech system for Indian languages," in *Proceedings of 8th speech synthesis workshop*, 2013.

[3] S. Nemala and A.M. Hema, "A new prosodic phrasing model for Indian language Telugu," *INTERSPEECH*, 2004.

[4] L.P. Sivaram and T. Nagarajan, "Estimation of phrase boundaries for Tamil speech synthesizer," no. 3 , vol. III , 2014.

[5] A. Parlikar and A.W. Black, "Data-driven phrasing for speech synthesis in low-resource languages," in *IEEE International conference on acoustics, speech and signal processing*, Japan, 2012.

[6] J. Hirschberg, and O. Rambow, "Learning prosodic features using a tree representation," in *INTERSPEECH*, 2001, pp. 1175-1178.

[7] J.J. Venditti. "The J_ToBI Model of Japanese," in *Prosodic typology: The phonology of intonation and phrasing*. Oxford University Press, 2006, pp.1:172.

[8] J. Valikangas, "FIN-ToBI tones and break indices for Finnish," Department of Phonetics, University of Helsinki, 2002.

[9] S. Baumann, M. Grice, R. Benzmüller. "GToBI-a phonological system for the transcription of German intonation," in *Prosody*, 2000, pp. 21-28.

[10] A. Sengar and R. Mannell, "A preliminary study of Hindi intonation," in *Proceedings of SST*, 2012.

[11] U. Patil, et al., "Focus, word order and intonation in Hindi," *Journal of South Asian Linguistics*, vol. 1.1, 2008.

[12] S. D. Khan, "Intonational phonology and focus prosody of Bengali (Ph. D. thesis)," 2008.

[13] M. E. Beckman and G.A Elam, "Guidelines for ToBI labeling, version 3," in *Ohio State University*, 1997.

[14] M. Breen, L.C. Dilley, and E. Gibson, "Inter-transcriber reliability for two systems of prosodic annotation: ToBI (tones and break indices) and RaP (Rhythm and Pitch)," 2012, pp. 277-312.

[15] W.Y.P. Wong, M. K. Chan, and M. E. Beckman, "An autosegmental-metrical analysis and prosodic annotation conventions for Cantonese," in *Prosodic typology: The phonology of intonation and phrasing 1,* 2005.

[16] A. Lahiri and F. Plank, "Phonological phrasing in Germanic: the judgement of history, confirmed through experiment," in *Transactions of the Philological Society*, 2010, pp. 370-398.

[17] P. Eisenberg, "Floor plan of German grammar : the word/Grundriss der deutschen Grammatik: das wort," Stuttgart; Weimar: Metzler, 3rd ed, 2006.

[18] W. Habib, W. Basit, S. Hussain, and F. Abeeda, "Design of speech corpus for open domain Urdu Text-to-Speech System using Greedy algorithm," in *Proceedings of Conference on Language and Technology 2014 (CLT14)*, Karachi, 2014.

[19] B. Mumtaz, S. Urooj, S. Hussain, and W. Habib, "Stress annotated Urdu speech corpus to build female voice for TTS," in *the Proceedings of 18th Oriental COCOSDA/CASLRE Conference*, Shanghai, China, 2015.

[20] B. Mumtaz, et al., "Multitier annotation of Urdu speech corpus," in *Conference on Language and Technology*, Karachi, Pakistan, 2014.

[21] T. Ahmed, et al. "The CLE Urdu POS tagset," in *poster presentation in Language Resources and Evaluation Conference (LERC 14)* , Reykjavik, Iceland, 2014.