Break Index (BI) Annotated Speech Corpus for Urdu TTS

Saba Urooj, Benazir Mumtaz, Sarmad Hussain & Ehsan Ul Haq

Center for Language Engineering
Al-Khawarizmi Institute of Computer Science
University of Engineering and Technology Lahore, Pakistan

Break Index (BI)

- BI indicates prosodic correlation between two sequential words.
- 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'
  - Finnish TOBI (FIN-TOBI) model uses 5-step scale to interpret the association between words
  - In Hindi and Bengali, prosodic grouping between words is categorized at three prosodic levels i.e. accentual phrase (AP), intermediate phrase (ip) and intonation phrase (IP)

Motivation

- To explore the levels of Break Indices (BI) for Urdu language
- To translate subjective identification criterion of BI-levels into objective identification criteria by developing an algorithm
- To use BI-Levels to understand the prosodic phrasing of Urdu
- To investigate the correspondence between the syntactic phrases and the phonological or intonational phrases

Corpus Description

Recorded Corpus Size	10 Hours (10500 sentences)
Data Extraction for 10 Hours of Speech	35M Corpus, CLE Urdu Digest Corpus and News Corpus
Speaker	A professional female speaker
Recording Environment	Anechoic Chamber
Sampling Rate	48KHz
Levels of Annotation:	Segment, Syllable, Word and Stress levels

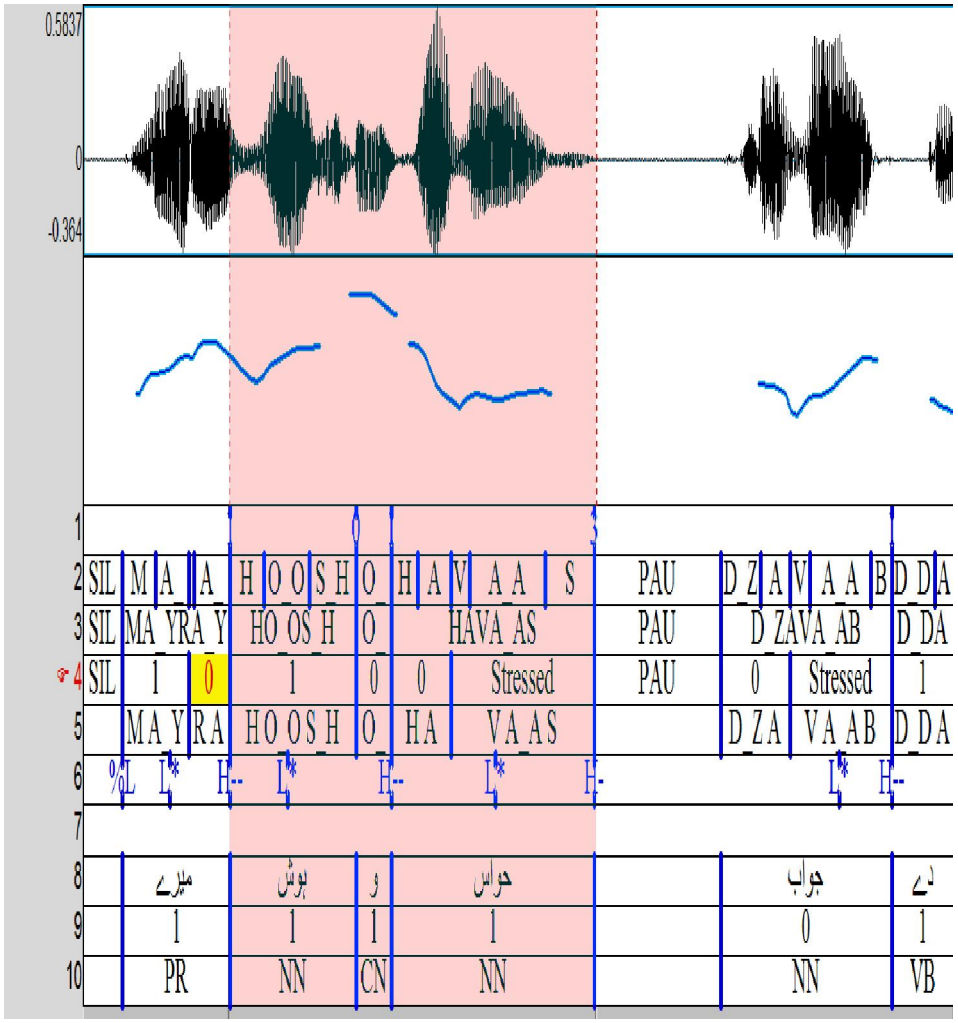
Automatic Identification of BI levels

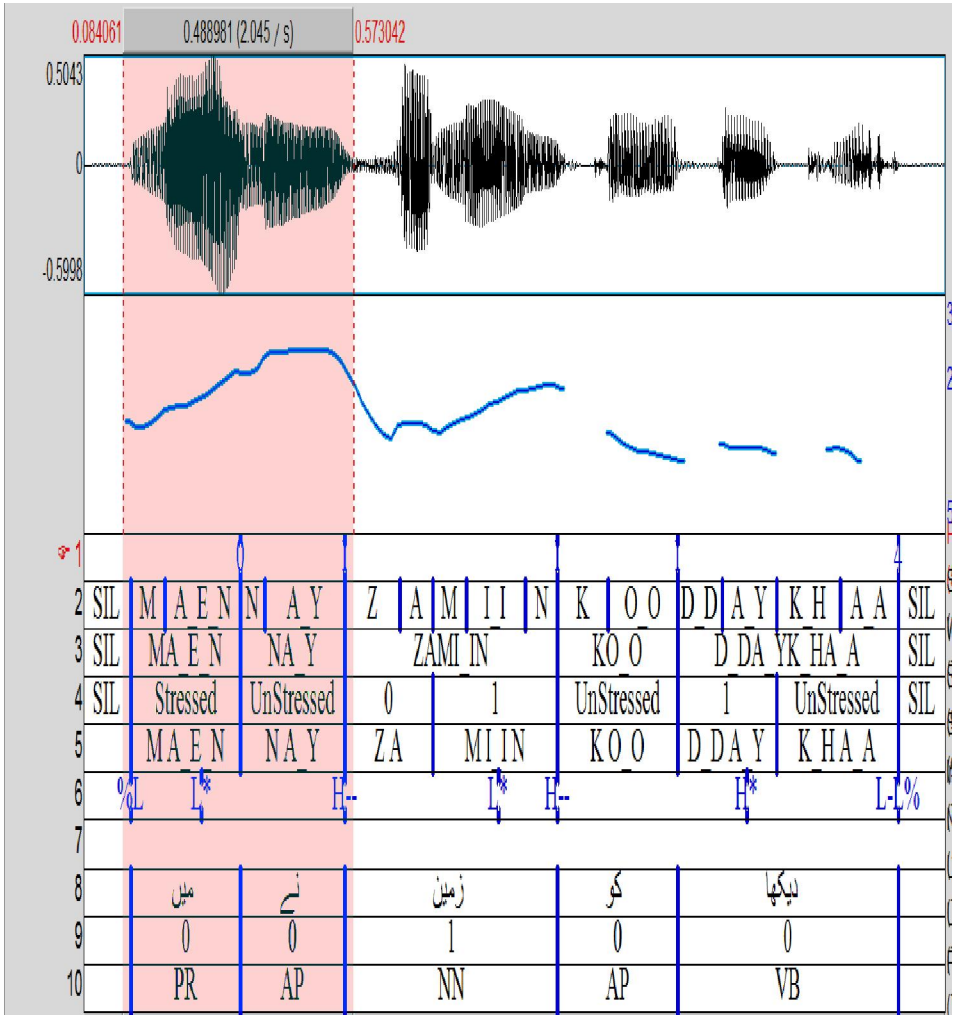
- Data for the Automatic Identification of BI levels: one hour of speech (1036 sentences)
- 1 hour Speech Annotation: perceptual annotation by two linguists
- Division of 1 hours of Speech: training and testing data comprising of 854 and 182 sentences respectively
- Analysis of training data highlights that native speakers uses five levels of break indices ranging from level '0' to level '4' to express relationship between word boundaries

BI Level 0

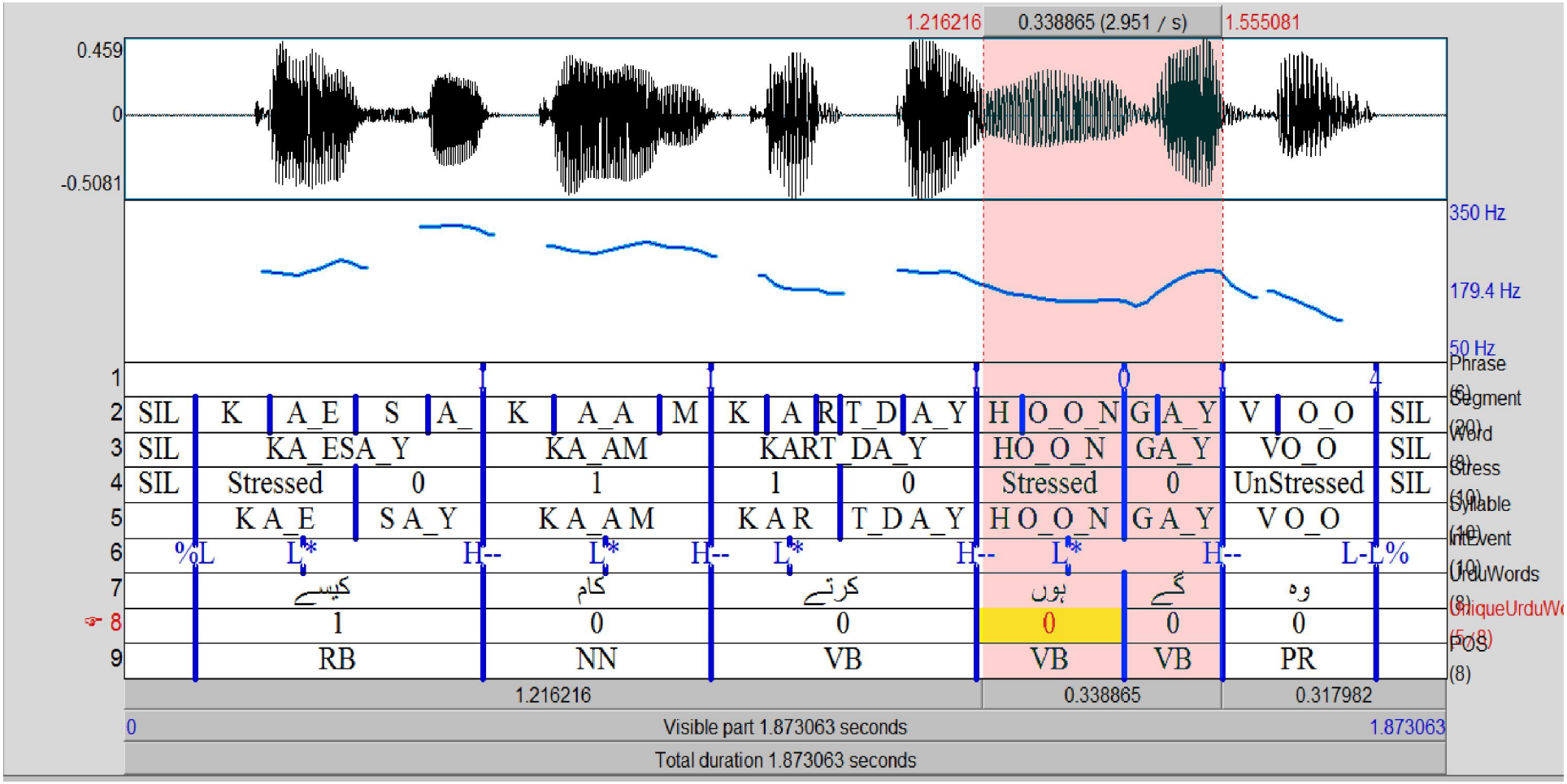
- In Urdu, Level ‘0’ is used in the following contexts:
  - A pronoun followed by a case marker as in the words آپ نے (ɑ:p ne:\ You)
  - Compound words combined with <ِ> zair or "و" vao izafat as in the words مخلوِق خدا (məxlu:q e: xʊd̪ɑ:\ creature of God) and غور و فکر (ɣo:r o: fɪkər\ contemplation)
  - 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
  - بڑھتی جا رہی تھی(ʤɑ: rəhi: \ is going)

BI Level 0: Compound words & Pronoun with Case Marker



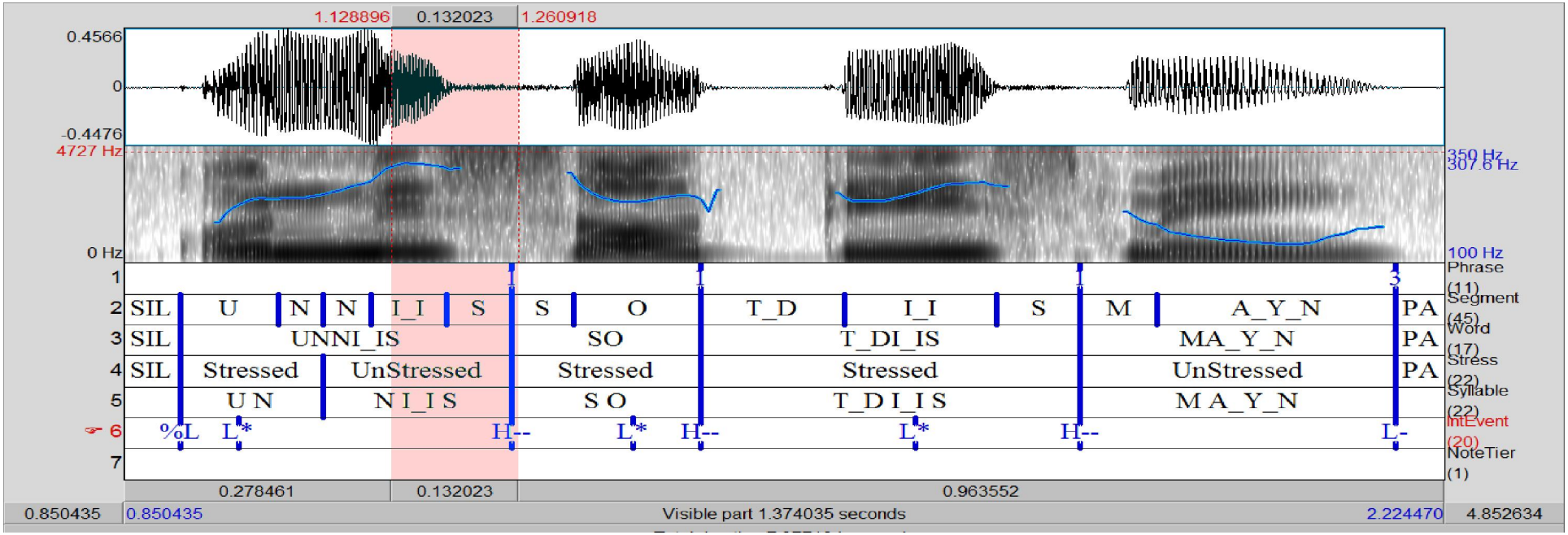


BI Level 0: Aspectual and Tense Auxiliaries



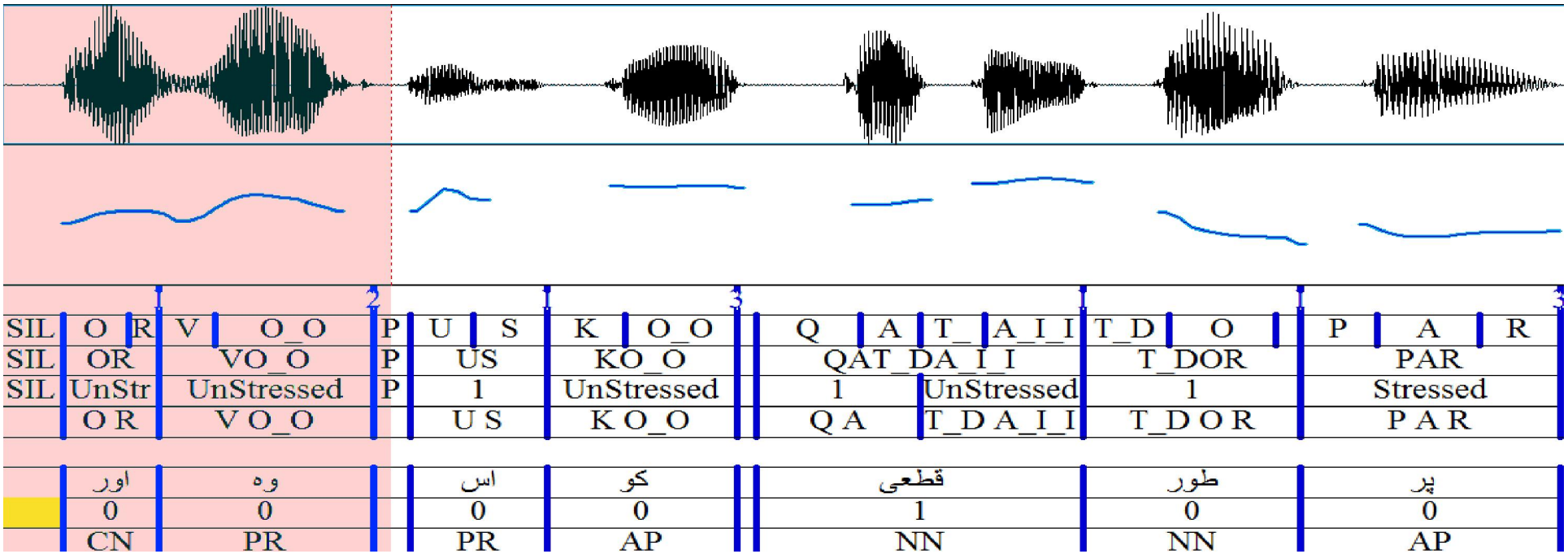
BI Level 1

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



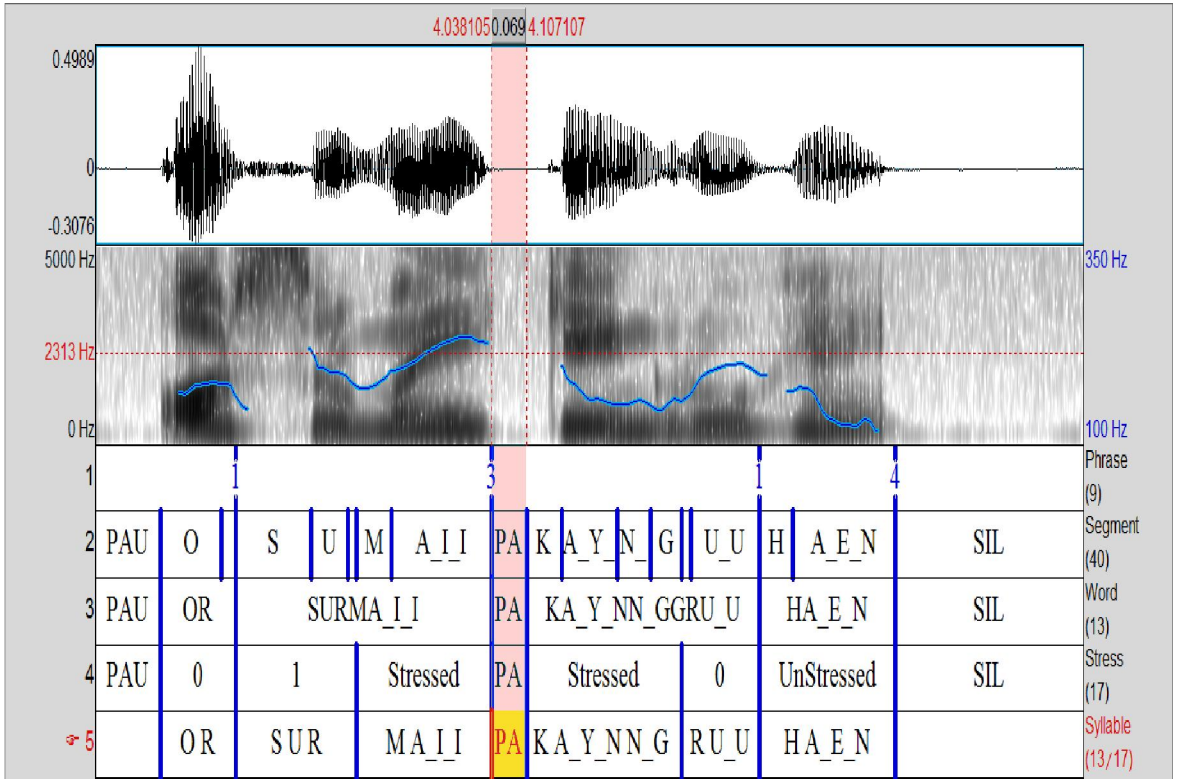
BI Level 2

- 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



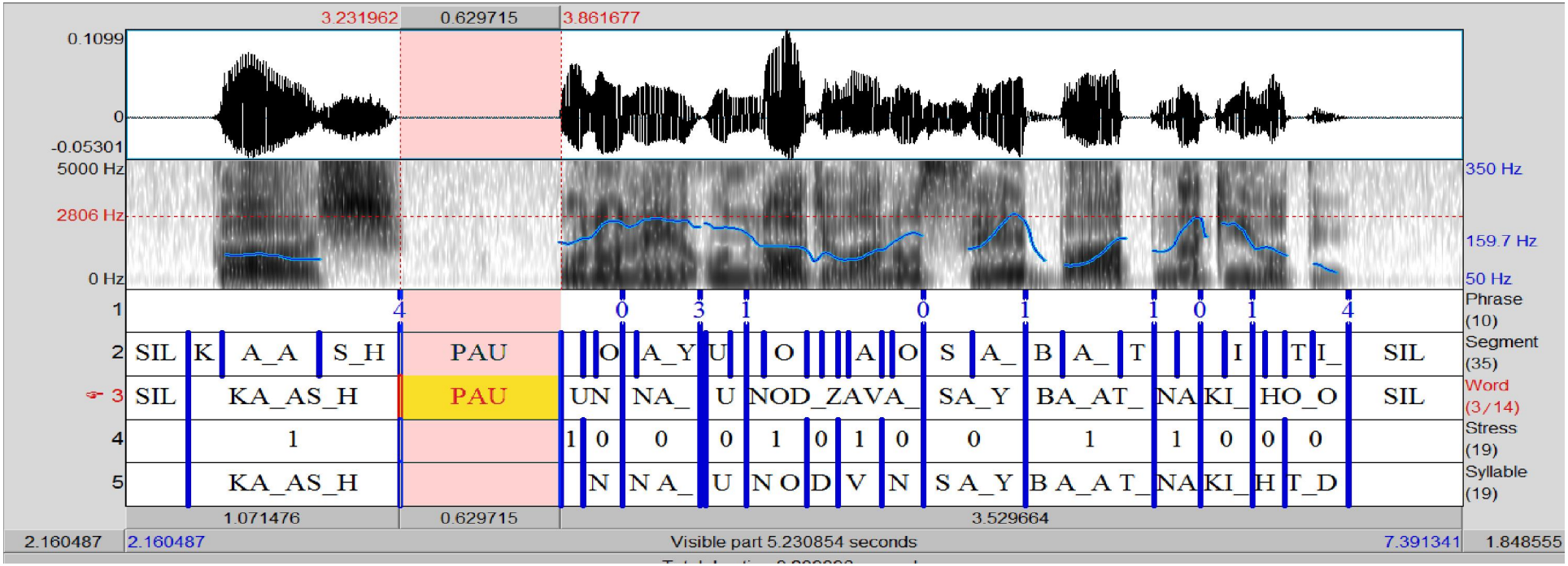
BI Level 3

- 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 intermediate phrase are as follows:
  - Weak Disjuncture
  - Pitch Reset
  - Phrase lengthening
  - Glottalization



BI 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:
  - Strong disjuncture
  - Final lowering of f0



BI Identification Algorithm

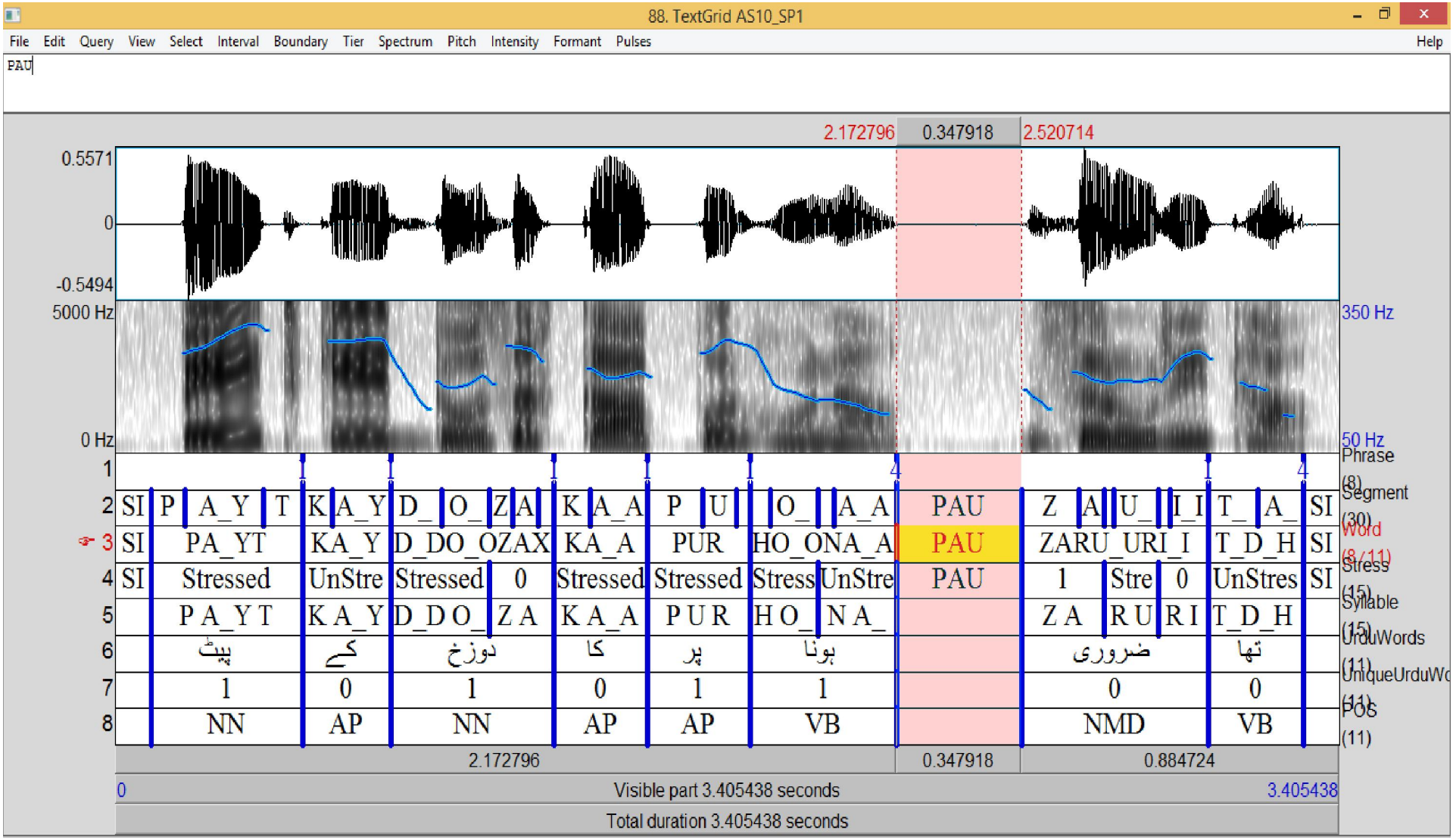
1. Find the word boundary and check whether the word boundary is followed by a pause or not.
2. If the word boundary is not followed by a pause, check whether it is followed by a glottalization or not.
3. If the word boundary is not followed by a glottalization, check the phrase lengthening at the end of the syllable using the vowel duration analysis table at various position of the syllable [1]
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'.

1. Mumtaz, B., Urooj, S., Hussain, S. and Habib, W. "**Stress Annotated Urdu Speech Corpus to Build Female Voice for TTS**", in the Proceedings of 18th Oriental COCOSDA/CASLRE Conference 2015, Shanghai, China.

BI Identification Algorithm

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.
9. 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'.
10. If the pitch is neither in the high range nor in the lower 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'.
11. Repeat from the step (1) until all the words in an utterance are processed.

Syntactic Structure: a Cue to Differentiate level 3 & Level 4



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 (7%)	1598 (98.4%)

Prosodic Phrases Alignment with Syntactic Phrases

- 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.
- After the annotation, the syntactic marked data is automatically compared with break index marked data to find out the two dimensional analysis:
  - prosodic-syntactic phrases agreement and disagreement
  - syntactic-prosodic phrases agreement and disagreement

Prosodic Phrases to Syntactic Phrase Alignment

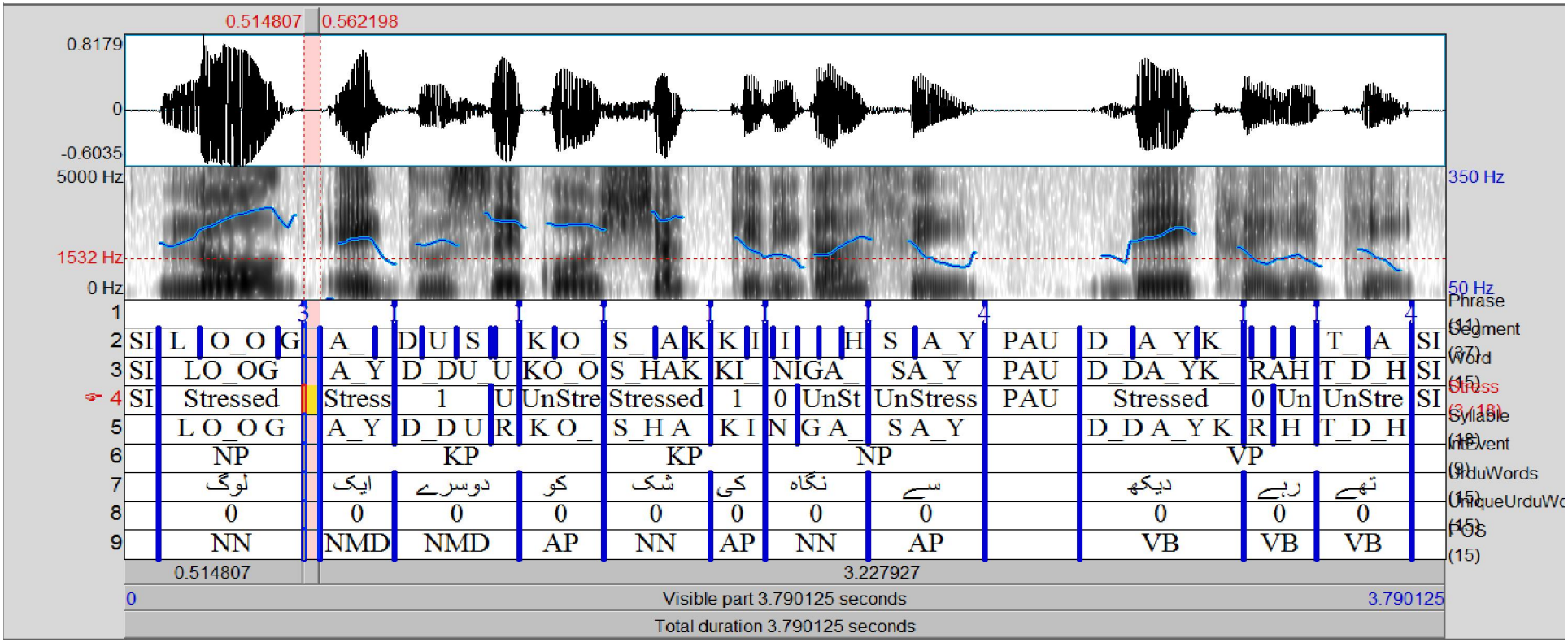
BI levels	**Prosody-syntactic alignment**	
	Total No. of prosodic phrases	***Alignment between prosodic and syntactic phrases***
Level 2	11	10 (90.9%)
Level 3	77	61 (79.2%)
Level 4	66	66 (100%)
Total	154	137 (89%)

Syntactic Phrase to Prosodic Phrases Alignment

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

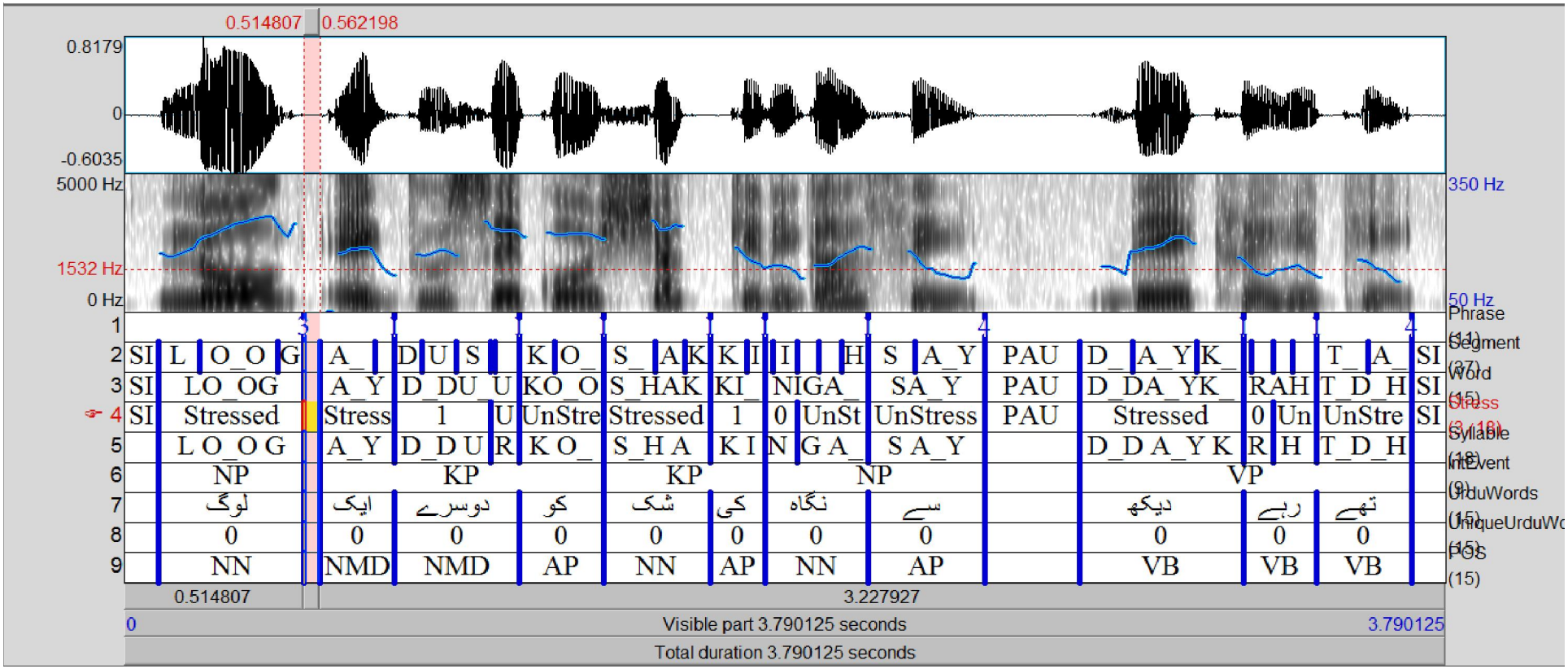
Discussion on Syntax-prosodic Phrases Alignment

- Maximum alignment occurs in case of NP-KP pair (76%). 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 to defuse syntactic ambiguity.



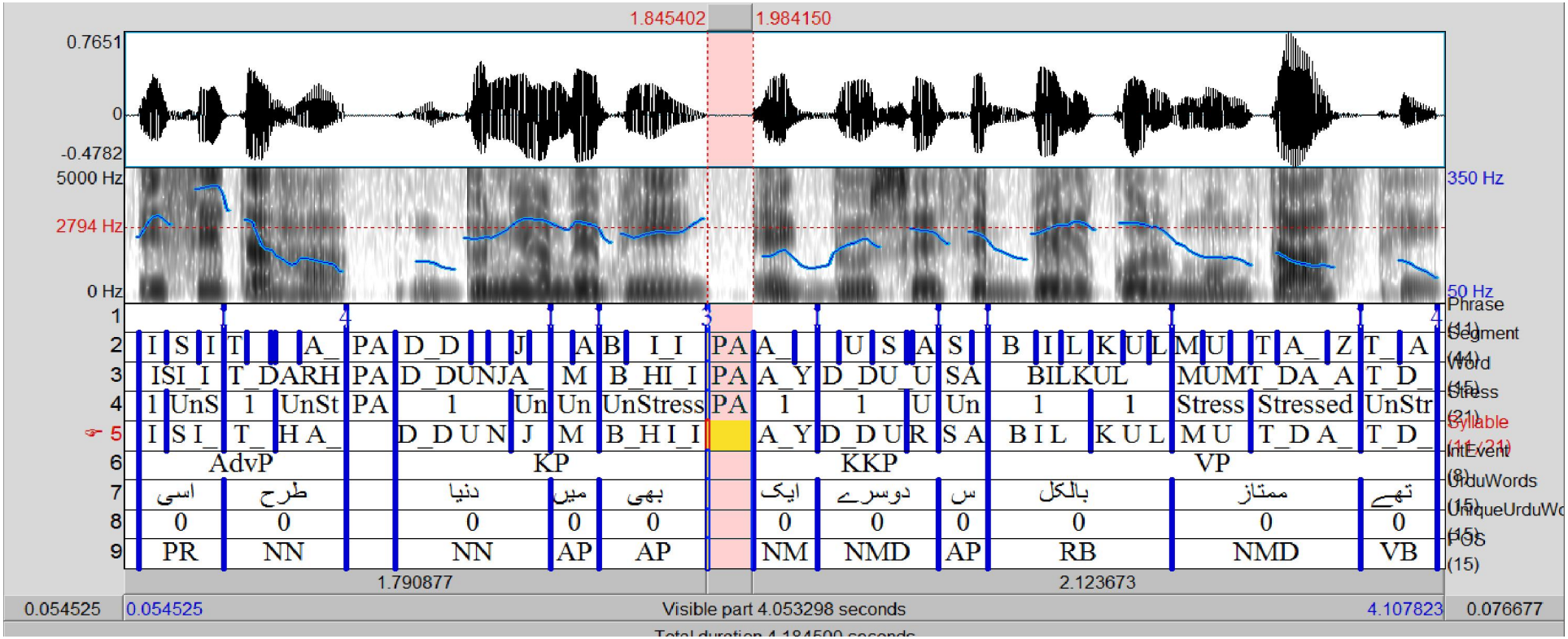
Discussion on Syntax-prosodic Alignment

- Minimum prosodic alignment (12%) occurs in case of KP-KP pair as the case markers themselves act as phrase separator and do not require explicit break.



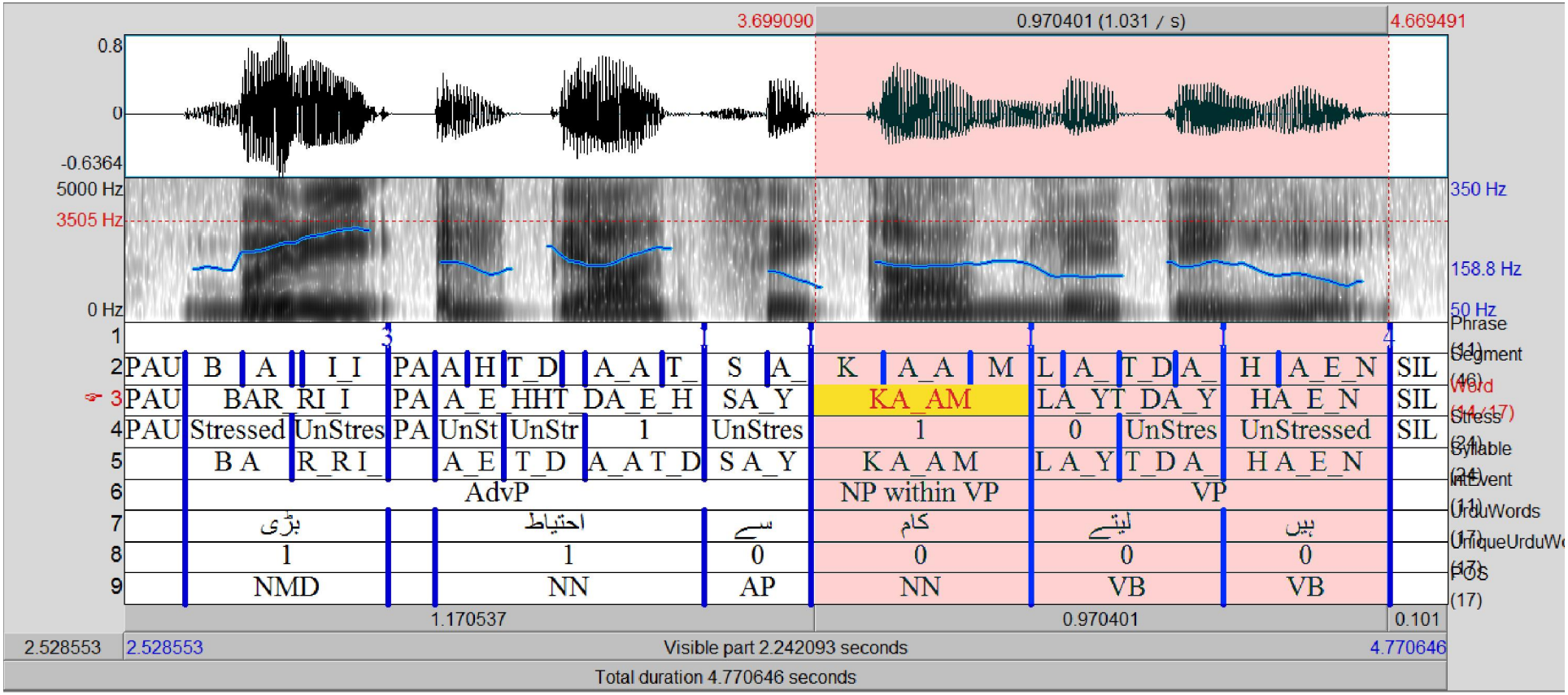
Discussion on Syntax-prosodic Alignment

- 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.



Discussion on Syntax-prosodic Alignment

- NP -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.



Conclusion & Future Work

- Urdu uses 0-4 range scale to express prosodic relationship between words.
- There is 52.6% correspondence from syntactic to prosodic phases and 89% 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.

Thank You