STRESS ANNOTATED URDU SPEECH CORPUS TO BUILD FEMALE VOICE FOR TTS

Benazir Mumtaz, Saba Urooj, Sarmad Hussain, Wajiha Habib

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

ABSTRACT

This research describes the stress annotation process for the two hours of Urdu speech corpus containing 18,640 words and 28,866 syllables to build a natural voice for Text-to-speech (TTS) system. For the stress annotation of speech corpus, two algorithms i.e. phonological and acoustic stress marking algorithms have been tested in comparison to perceptual stress marking. Urdu phonological stress markings algorithm [1] reports 70% accuracy whereas Urdu acoustic stress marking algorithm developed through this research reports 81.2% accuracy. This acoustic stress marking algorithm is then used to annotate two hours of Urdu speech corpus. It is a semi-automatic acoustic stress marking algorithm, which annotates 54% data automatically using duration cue whereas 46% data is marked manually using the acoustic cues of pitch, glottalization and intensity.

Index Terms— Text-to-speech system, stress/prominence, phonological stress marking algorithm, perceptual stress annotation, acoustic stress marking algorithm

1. INTRODUCTION

The quality, intelligibility and naturalness of TTS system depends on the prosodic labeling of the speech corpus [2,3,4] i.e. the segmentation of corpus at stress, tone and break index levels. The prosodic annotated speech corpus also assists TTS to determine the focus, meaning and purpose of the discourse. This paper deals with annotation of speech corpus at stress level. Multiple methods can be used to annotate speech corpus at stress tier such as phonological stress marking [1], perceptual annotation [2] and stress identification using acoustic cues [4].

Phonological analysis of Urdu stress pattern has indicated that stress in Urdu is predicable depending on the weight of the syllable [1]. However, annotation of speech corpus using phonological stress marking highlights few constraints in phonological algorithm indicating rules defined for lexical stress marking cannot entirely be applied to mark the stress on speech. Therefore, this current study focuses to build on the earlier research efforts and present a new algorithm for determining stressed syllables in speech using perceptual intuition of native speakers and acoustic cues to build a natural female voice for Urdu TTS.

This paper is structured in the following sections. Prior studies on the stress annotation using phonological, perceptual and acoustic approaches are presented in Section 2. The methodology used to record two hours of Urdu speech corpus, comparison of phonological stress marking algorithm with perceptual stress marking, development of acoustic stress marking algorithm and comparison of acoustic stress marking algorithm with perceptual stress marking are detailed in Section 3. Annotation of two hours of speech corpus and quality assessment of annotated speech corpus is detailed in Section 4. The results are given in Section 5 while discussion and conclusions are discussed in Section 6 and 7 respectively.

2. LITERATURE REVIEW

Stress is described as the display of prominence on a certain syllable [5]. To investigate phonological aspects of stress, several studies have been conducted in various languages. As far as phonological analysis of stress in Urdu is concerned, it depends on the weight of the syllable and is explored by Hussain [1] who has proposed an Urdu phonological stress marking algorithm. This algorithm classifies Urdu syllables as monomoraic, bimoraic, and trimoraic. Given these definitions, it states that starting from the end of the word, the first heavy syllable is always stressed in Urdu, and if all syllables are light, the penultimate syllable is stressed.

Buhmann et al. [6] conducted the perceptual experiments for spoken Dutch corpus and claimed that high quality prosodic annotations (including prominence) can be achieved by non-expert native speakers. Streefkert et al. [7] investigated prominence marking in read aloud Dutch sentences and found out a reliable inter-labeler agreement. Wagner [8] conducted empirical studies for native and non-native speakers of German speech read at fast and normal speed. They found out that native speakers rely more on their prediction or native language intuition especially if acoustic cues are difficult to interpret, as it is the case in very fast speech while non-native speakers rely more on acoustic cues.

Studies also focus that stress has acoustic implication and they try to acoustically predict prominence using

978-1-4673-8279-3/15/$31.00 ©2015 IEEE

perceptual intuition of native speaker. For example, Portele [9] conducted experiment for American English and revealed that there is some strong relation between prominence and acoustic-prosodic parameters (duration and F0 mainly). These parameters can serve to automate American English prosody in a content-to-speech system. Moreover, Cutler [10] is of the view that to find out a stressed or unstressed syllable, most of the researches have mainly focused on the acoustic characteristics of stressed versus unstressed syllables. He claims that the three acoustic dimensions are involved in the realization of stress i.e. duration, fundamental frequency and intensity. These acoustic properties correspond to the perceptual phenomena of length, pitch and loudness respectively.

In addition, some phoneticians make more specific claim as to which parameters play a larger role in the realization of stress. Ladefoged [11] states it is likely to be some combination of pitch, length, and loudness, with the first two playing the greatest role. Thus, some languages base the distinction between their stressed and unstressed syllable more on F0 differences while other languages more on duration differences and amplitude differences. Moreover, acoustic cues are further analyzed in relation to the position of syllable in a word. The distinction made for the syllable position is of two types i.e. penultimate and final syllable position [12].

For Urdu language, stress marking algorithm marks lexical stress based on syllable weight [1] but the potential of secondary stress and emphatic stress needs to be explored as a word can have more than one stressed syllables in spontaneous speech. Therefore, this paper explores how predictive and stable is the result of phonological marking as compared to native speaker intuition and how this intuition can be automated using acoustic cues for prosodic learning of Urdu TTS.

3. METHODOLOGY

Two hours of speech corpus is recorded in 'mono' form at a sampling rate of 48 kHz to build a female voice for Urdu TTS. These two hours of speech have been extracted from three large corpora using Greedy algorithm [13] to include maximum coverage of Urdu words. The recording of this corpus is obtained from a female professional speaker in an anechoic chamber using PRAAT software. During the recording, speaker is instructed to maintain the same range of f0, rate of speaking and level of intensity within a recording session and across the recording sessions.

After the recording, speech is segmented at sentence level and sent forward for the multiple levels of annotation using CISAMPA phonetic character set. The process used for the annotation of Urdu speech corpora at multiple levels is described in [14].

For the stress level annotation of speech corpus, two experiments have been conducted. The data used for these two experiments is classified into two data sets: Data Set A and Data Set B. Data set A consists of 544 words. This data is the testing data and is used for two types of testing: to test the stability of phonological stress marking algorithm and to test the accuracy of acoustic stress marking algorithm. Data Set B consists of 2255 words and is used for the training of acoustic stress marking algorithm. Acoustic stress marking algorithm is then used to mark stress for another unseen Data Set C which consists of two hours of speech (18,640 words). Description of Data Set A, B and C is given below in Table 1:

Table 1: Description of Data sets

	Data Set A	Data Set B	Data Set C
Total no. of sentences	60	290	2,008
Total no. of words	544	2255	18,640
Total no. of syllables	848	3990	28,866
Agreed words for stress	386	1165	NA

3.1 Experiment 1

Data Set A is marked perceptually from 2 hours of speech corpus to assess the reliability of phonological stress marking of Urdu. At perceptual level, stress is assigned on this data set by two expert linguists independently. The strategy used for the annotation is that annotators listen to the wave file using sub phrases ending in pauses or glottalization. They assign stress to the syllables within the selected sub phrase after analyzing the vocalic properties of the syllable in the spectrogram and the time wave form.

At phonological level, the sentences of Data Set A are automatically annotated using Urdu stress marking algorithm[1]. After the annotation, the phonological marked data is compared with mutually agreed perceptually marked words in Data Set A to find out the three dimensional analysis:

a) stress annotation agreement between the phonological and perceptual annotation
b) stress annotation disagreement between phonological and perceptual annotation
c) stress annotation additive agreement (i.e. the syllables marked as stressed in addition to primarily stressed syllable) found in the perceptual annotation

Comparative analysis of the Data Set A with phonological annotation reports 30% mismatch indicating phonological stress marking is different from the perception of native speakers due to some constraints in phonological algorithm, which are discussed in section 6. This study aims to map the native speaker intuition of stress identification and tries to automatically predict this by developing a semi-automatic algorithm using acoustic cues. Detail of this process has been reported in Experiment 2.

3.2 Experiment 2

Data Sets B is used for the experiment 2. This data set is also marked perceptually by two linguists. For semi-automatic stress annotation, a detailed analysis of Data Set B has been conducted and average duration of each stressed and unstressed vowel at three positions in a syllable i.e. penultimate, final and final with pause has been automatically calculated (See Appendix 1). The strategy used for automatic stress annotation is that the duration of targeted vowel is compared with the average duration of same vowel at same position. If the targeted vowel's duration is less than the average duration of same unstressed vowel, the syllable is marked as unstressed. However, if the duration of targeted vowel exceeds the average duration of same stressed vowel, the syllable is marked as stressed. Durations of the vowels which fall neither on the unstressed regions nor on stressed regions are left unmarked. The intersecting area of the green and red lines in Figure 1 shows the number of syllables which are left unmarked.

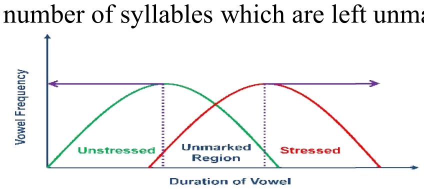

Figure 1: Strategy used for automatic stress annotation

Before applying the stress annotation system on two hours of speech, rate of speaking within a recording session and across the recording sessions has also been evaluated to ensure the consistency in the duration of the stressed and unstressed vowels throughout the speech corpus (see Appendix 3). After that, automatic stress annotation system has been used which assigns stress to 15,483 syllables and the rest of the 13,383 syllables are left unmarked for manual tagging. Unmarked syllables are assigned stress manually using a stepwise process.

The stepwise process has been formulated to map the native speaker intuition at acoustic level. In this stepwise process, acoustic cues of stress marking have been prioritized i.e. duration, fundamental frequency, glottalization and intensity of the vowel respectively. This priority order has been set after using the theoretical knowledge acquired from the literature survey and then practically applying that knowledge on training data (Data Set B). The process developed for stress tier annotation using acoustic cues is discussed in the following sections.

3.2.1 Stress annotation of Urdu speech corpus using duration cue

For annotating the stress tier, number '1' is assigned to a stressed syllable, number '0' to an unstressed syllable and '?' to an ambiguous syllable. The first cue used to annotate stress tier manually is the duration of a vowel. Data analysis of Data Set B reports that vowel of a stressed syllable in Urdu has more duration than the duration of the same vowel in an unstressed syllable. Following guidelines have been developed to mark stress tier using duration of a vowel as a cue.

- For durational analysis, vowels are categorized into five categories i.e. short vowels, high long vowels, low long vowels, medial vowels and diphthongs.
- After categorizing the vowel, position of a vowel in a syllable is analyzed. A vowel can occur at three positions i.e. penultimate position, final position and final position of a syllable with pause.
- After categorizing the vowel and analyzing its position, the duration of the targeted vowel is compared with the duration of the same shortest vowel in a wave file. Two points are considered while selecting a shortest vowel:
  - Do not select a vowel, which comes at the "final syllable with pause" position.
  - The duration of the shortest same vowel for the short, medial and long vowels must be equal to or less than 57ms, 63ms and 100ms (as the durational analysis results indicated) respectively.
- If the duration of targeted vowel is more than its stressed duration (see Appendix 1), number '1' is assigned to the targeted syllable but if the duration of targeted vowel is less than its stressed duration, number '0' is assigned to the targeted syllable.
- If the same vowel is not present in the file, the duration of the target vowel is compared with the duration of the similar shortest vowel.
- Duration is not used as a cue if there is no same or similar vowel in a file for comparison.

3.2.2 Stress annotation of Urdu speech corpus using stylize pitch cue

The second cue used to annotate stress tier manually is the stylized pitch track of the vowel. This acoustic cue of stress marking is used when there are no instances of same or similar vowels in a wave file for durational analysis. While analyzing the pitch track, following steps are used:

- Stylize the wave file of a sentence using PRAAT software.
- Consider only the pitch points of the stylized file that come within the middle of a vowel. A middle pitch point must make a pitch contour. Two types

of pitch contours can be found while stress annotation: low/L pitch contour and high/H pitch contour.

- Number '1' is assigned, if the falling or rising slope between L* and H* is abrupt and steep. Number '0' is assigned, if the falling or rising slope between L* and H* is gradual and flat.
- Pitch track is not used as a cue if the pitch track of a vowel has no pitch point or more than two pitch points.

3.2.3 Stress annotation of Urdu speech corpus using phrase initial glottalization cue

The third cue used to annotate stress tier is phrase initial glottalization. Glottalization can occur at two positions: phrase initial position and phrase final position. Glottalization at phrase initial position is an indicator of stress in Urdu whereas at phrase final position, glottalization indicates that the vowel is tapering off. Data analysis shows that a stressed syllable is strongly glottalised at phrase initial position. Number '1' is assigned to the context where the word initial syllable has strong glottalization. If the syllable has a weak glottalization or no glottalization, then an annotator moves towards next cue, which is intensity.

3.2.4 Stress annotation of Urdu speech corpus using vowel intensity cue

The fourth and last acoustic cue used to segment speech corpus at stress level is the intensity of a vowel. Intensity of a speaker can vary from one recording session to another. Therefore, mean intensity values of each stressed and unstressed vowel are automatically extracted for every recording session separately. Analysis shows that the vowel of a stressed syllable must have at least 3 db more intensity than its unstressed version. While comparing the vowel intensity, following points are considered:

- Select only three middle periods of a vowel from the time wave form to find out the intensity of a targeted vowel.
- Ignore the tapering off from the total duration of a vowel at the final position while finding out the intensity of a vowel that is followed by a pause or silence.
- Assign number '1', if the intensity is more than the stressed value (see Appendix 2). Assign number '0', if the intensity is less than the stressed value. Assign '?', if the intensity value is in between the unstressed and stressed value of a targeted vowel.

Using these guidelines, a semi-automatic acoustic stress marking algorithm is developed.

3.3 Comparative analysis between perceptual and acoustic stress marking annotation

Data Set A is also annotated using semi-automatic acoustic algorithm for comparing the accuracy of this algorithm with the perceptually marked data. The result indicates 22.8% mismatch. This algorithm is further refined after analyzing the 20% mismatched syllables of Data Set A. Based on the analysis, acoustic stress marking algorithm is revised and re-tested on 80% remaining data. The result of the re-test reports 18.7% mismatch. This revised algorithm is now used to annotate two hours of speech corpus/Data set C.

4. ANNOTATION OF TWO HOURS OF SPEECH CORPUS

Two hours of speech corpus containing 18,640 words and 28,866 syllables is annotated using semi-automatic acoustic stress marking algorithm. A sample of annotated Urdu speech corpus is given below in Figure 2.

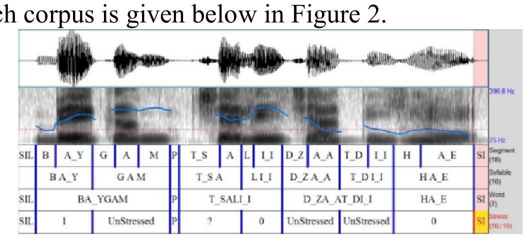

Figure 2: Annotated Urdu Speech Corpus

Once the speech corpus is segmented at stress level using acoustic cues, it is sent forward for the quality assurance process described below.

4.1 Quality assurance of two hours of speech corpus

This section describes the assessment methods used to ensure and improve the quality of acoustically stress annotated two hours of speech corpus. This quality assessment is conducted to check the inter-annotator accuracy for manual stress marking.

4.1.1 Inter-annotator quality assurance of manual stress annotation

The strategy used to assess the quality of manually tagged corpus is that 25% of speech files termed as the reference files are tagged manually by an annotator and are compared with the corresponding same speech files annotated by the speech corpus annotation team, known as the source files. The mismatches between the source and reference files are manually verified and analyzed by an expert linguist. The data is re-annotated until the error rate is less than 5%.

A testing framework consisting of multiple assistive tools is also developed to ensure the quality of stress tier annotation. This framework ensures that all the stress tier labels are from a defined numbering scheme (0, 1, ?), no interval/syllable is left unmarked and no change has been conducted at the automatically marked syllabification tier while annotating the stressed tier. The source package is re-annotated even if a single interval is left unmarked or marked using the undefined numbering scheme.

5. RESULTS

This section reports two levels of results. At first level, results of comparative analysis of perceptual with phonological (Table 2a) and perceptual with acoustic stress marking algorithm (Table 2b) are given. The second level describes the results of two hours of speech corpus annotated using semi-automatic acoustic stress marking algorithm (Table 3).

Table 2a: Comparative analysis of phonological and perceptual stress annotation

	Data Set A
Total no of agreed syllables	617
Matched syllables	434 (70%)
Mismatched syllables	183 (30%)

Table 2b: Comparative analysis of acoustically and perceptual marked stress annotation

	Result 1	Result 2
Total no of syllables	675	537
Total no. of marked syllable	617	438
Matched syllables	476 (77%)	356(81.27%)
Mismatched syllables	141(22.8%)	82 (18.72%)

Table 2b shows the comparative analysis of acoustically and perceptually marked Data Set A. Result 1 reports the percentage of accuracy achieved before revising the acoustic stress marking algorithm whereas Result 2 reports the accuracy achieved after analyzing the 20% mismatch between acoustically and perceptually marked syllables.

The annotation of two hours of speech corpus that is marked using semi-automatic acoustic stress marking algorithm is given in Table 3.

Table 3: Annotation of two hours of speech corpus

Sentences	2,008
Words	18,640
Syllables	28,866
Automatically assigned stress syllables	6,604
Automatically assigned unstressed syllables	8,879
Automatically left unmarked syllables	13,383
Manual marked stressed syllables	5,107
Manual marked unstressed syllables	5,805
Ambiguous syllables marked with '?'	2,471

6. DISCUSSION

This section provides three levels of analyses. Level 1 analyzes the mismatches between the phonological and perceptual marked data (Data Set A) which highlights the constraints of phonological stress marking algorithm. Level 2 reviews the process used for the development of semi-automatic acoustic algorithm and presents the analysis of the mismatches between acoustic and perceptual stress marked data. Level 3 discusses the ambiguous data marked with the symbol of '?'.

The comparison between the phonological and perceptual data shows 29.6% mismatch. Analysis of mismatch data highlights three major constraints of phonological algorithm i.e. evasion of secondary stress marking, moraic dependent marking of monosyllabic content and functional words (e.g. کو/ko:/case marker, نے/ne:/case marker, (دل /d̪ɪl/heart, غم /ɣəm/ grief) and restriction to mark at least one syllable stressed in a bi-syllabic and tri-syllabic words. Analysis shows that phonological stress marking algorithm neglects the possibility of the secondary stress whereas additive agreement is found in the perceptual annotation indicating secondary stress does exist in Urdu. Similarly, monosyllabic content and functional words are always marked unstressed based on their moraic weight by phonological algorithm. Analysis of mismatches suggests that these words should not be treated alike as monosyllabic content words have the probability of becoming stressed. Moreover, bi and tri syllabic words mismatch indicates that phonological marking algorithm must assign stress to a syllable in a bi or tri syllabic word. If both syllables are light, it will assign stress to the penultimate syllable whereas in speech it is not necessary that one syllable in a word must carry stress.

The acoustic analysis of semi-automatic acoustic algorithm highlights that unlike other studies two types of distinction in syllable positions (i.e. penultimate syllable and word final syllable position) cannot be used in Urdu context. Acoustic analysis of data reports that duration of a same vowel in Urdu behaves differently at three positions of syllables: penultimate syllable, word final syllable and word final with pause. Analysis shows that the duration of penultimate syllable vowel is less than final syllable vowel and the duration of final syllable vowel is less than the duration of final syllable with pause vowel. However, nasalized vowels need to be further explored for this generalization.

The data analysis of intensity cue for one recording session (see Appendix 2) highlights lot of variation. It reports that few vowels (/ɑ̃/, /e/, /ũ:/) increase their the intensity with stress whereas intensity of few vowels (/ẽ:/, /æ:/, /i:/) decrease with stress and few (/e:/, /o:/) do not show any change with stress. Due to the unpredictable behavior of intensity cue, it is considered a least reliable cue. Furthermore, a rule has been incorporated in the acoustic stress marking algorithm that heavy coda syllable would always be marked as stressed. This rule has been generalized based on perceptually stressed marked data analysis, which reports 84% alignment of stress with heavy coda syllable as in the word وقت/Time/vəkt̪ and درخت/Tree/d̪ə.rəxt̪.

The analysis of the mismatch between acoustic and perceptual stress marked data underlines that:

- The vowels of case markers were marked stressed using acoustic cue of duration but perceptually they were marked unstressed. This finding clues that the word class has also got some role to play in determining prominence. In this particular context, it can be generalized that independent of vowel duration of case markers, they should be marked unstressed.
- During the selection of shortest vowels for duration analysis, diphthongs, compound words and epenthetic vowels should be ignored as the vowels do not show regular duration in these contexts.
- When the combination of (C)VV comes at phrase final syllable with pause position, it was perceptually marked unstressed irrespective of its long duration. This analysis also supports the findings of Hussain [1] that final mora is invisible to stress making algorithm. Therefore, this combination would be marked unstressed.

The analysis of two hours of speech shows that there is still 9% data, which is left unmarked. This 9% data is reviewed and tagged with '?'. The review of this 9% data indicates that annotators found confusion in assigning stress to three contexts i.e. consonant lengthening, increase in intensity of the vowel and data scarcity. It was analyzed that the syllables tagged with '?' contain consonants with increased duration of more than 100 ms at onset and coda positions irrespective of the fact that there is no increase in vowel's duration. Similarly, the syllables marked '?' because of intensity showed more than 3db increase in intensity than its unstressed version although there is no increase in the duration of vowel. Likewise the syllables are also assigned '?' when an annotator runs out of all acoustic cues.

Currently, only duration cue for automatic stress annotation is being used. In future, stylized pitch track and phrase initial glottalization cues will also be further investigated for the automatic annotation of stress tier.

7. CONCLUSION

In this paper, segmentation of two hours of Urdu speech corpus at stress level has been described. Different comparative analyses such as phonological verses perceptual and perceptual verses acoustic have been conducted to annotate the corpus at stress level. Result reports that 70% accuracy can be achieved using phonological stress marking algorithm and 81.2% accuracy can be achieved using acoustic stress marking algorithm. This work is in process and the acoustic stress marking algorithm developed via this study will be used to annotate eight hours of speech corpus at stress level in future.

8. REFERENCES

[1] Hussain. S, Phonetic Correlates of Lexical Stress in Urdu, 1997, Northwestern University. Unpublished Dissertation.

[2] Z. Yiqing and H. Zheng, "Effect of prosodic structure on segmental variants ," in *Speech Prosody 2002, International Conference*, 2002.

[3] Janez, Erdem, C. Sttergar, "Adapting Prosody in a Text-to-Speech System," *INTECH Open Access Publisher*, 2010.

[4] Ananthakrishnan, Sankaranarayanan, and Shrikanth, S. Narayanan, "Automatic prosodic event detection using acoustic, lexical and syntactic evidence ," *Audio, Speech and Language Processing, IEEE Transactions*, vol. 16, no. 1, pp. 216-228, 2008.

[5] J. Laver, *Principles of Phonetics*.: Cambridge University Press, 1994.

[6] J. Buhmann et al., "Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the spoken Dutch Corpus," in *Proceedings of LREC 2002*, Las Palmas, 2002, pp. 779-785.

[7] B. M. Streefkert, L. C. Pols, and L. F. M. ten Bosch, "Prominence in read aloud sentences as marked by listeners and classified automatically," in *Proceedings of IFA*, Amsterdam, 1997, pp. 101-116.

[8] P. Wagner, "Great expectations-Introspective vs. perceptual prominence ratings and their acoustic correlates," in *Proceedings of Interspeech 2005*, Lisbon, 2005, pp. 2381-2384.

[9] T. Portele, "Perceived Prominence and Acoustic Parameters in American English," in *Proceedings of ICSLP98*, Sydney, 1998, pp. 667-670.

[10] A. Cutler, *11 Lexical Stress: The handbook of speech perception*.: Blackwell Publishers, 2008.

[11] P. Ladefoged, *Phonetic data analysis: An introduction to fieldwork and instrumental techniques*.: Wiley-Blackwell, 2003.

[12] Merkx, M. Marjolein, and P. Monaghan, "A Hierarchy of Prosodic cues in speech processing ," in *Proceedings of the 28th Annual Conference of the Cognitive Science Society* , 2006.

[13] W. Habib, R. Basit, H. Hussain, and F. Adeeba, "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.

[14] B., Hussain, A., Hussain, S., Mahmood. A., Bhatti, R., Farooq, M. and Rauf, S. Mumtaz, "Multitier Annotation of Urdu Speech Corpus," in *Proceedings of Conference on Language and Technology 2014 (CLT14)*, Karachi, 2014.

Appendix1: Mean duration of Vowels at three position of syllables

Vowels	Penultimate syllable 0	Penultimate syllable 1	Final syllable 0	Final syllable 1	Final with pause syllable 0	Final with pause syllable 1	Increased Duration at Non-final	Increased Duration at final	Increased Duration at final with pause
Short Vowels									
A	57	81	61	86	75	107	24	25	32
A_N	NA	80	NA	NA	NA	NA	NA	NA	NA
U	57	85	60	82	89	99	28	22	10
U_N	NA	NA	NA	NA	NA	NA	NA	NA	NA
I	54	85	56	79	77	88	31	23	11
I_N	NA	NA	NA	NA	NA	NA	NA	NA	NA
Low Long Vowels									
A_A	104	132	98	149	133	186	28	51	53
A_A_N	101	155	78	152	148	211	54	74	63
A_E	88	115	NA	175	159	189	27	NA	30
A_E_N	NA	143	NA	145	167	219	NA	NA	52
O	114	143	85	128	141	176	29	43	35
O_N	NA	99	NA	NA	NA	NA	NA	NA	NA
High Long Vowels									
A_Y	70	116	81	140	135	188	46	59	53
A_Y_N	NA	NA	93	149	154	213	NA	56	59
O_O	94	119	87	138	141	210	25	51	69
O_O_N	NA	148	89	112	225	192	NA	23	-33
U_U	80	107	97	134	152	185	27	37	33
U_U_N	NA	176	104	178	158	231	NA	74	73
I_I	78	126	86	131	144	204	48	45	60
I_I_N	NA	NA	91	135	140	192	NA	44	52
Medial Vowels									
A_E_H	67	87	57	78	NA	99	20	21	NA
A_Y_H	57	83	60	96	87	99	26	36	12
O_O_H	65	114	NA	NA	NA	NA	49	NA	NA
Diphthongs									
A_I_I	NA	134	113	195	201	245	NA	82	44
A_A_Y	NA	NA	NA	263	176	242	NA	NA	66
A_A_A_Y	NA	NA	189	209	204	275	NA	20	71
A_A_I_I	NA	177	NA	210	204	264	NA	NA	60
I_U_U_N	NA	147	110	117	NA	323	NA	7	NA
A_E_H_A_A	NA	149	110	187	170	231	NA	77	61
U_U_I_I	NA	NA	NA	NA	NA	NA	NA	NA	NA

Appendix 2: Intensity values of stressed and unstressed vowels

Vowel	Intensity of unstressed vowel	Intensity of stressed vowel
A_A_A_Y	74	80
A_A_I_I	70	79
A_A_N	77	80
A_Y_H	78	81
A_A_Y	71	Not found
A_E	73	74
A_E_H	80	79
A_E_H_A_A	72	80
A_E_N	79	74
A_I_I	70	Not found
A_N	Not found	Not found
A_Y	79	79
A_Y_N	77	78
I	79	79
I_I	77	76
I_I_N	77	76
I_N	Not found	Not found
I_U_U_N	79	Not found
O	81	81
O_N	Not found	Not found
O_O	78	78
O_O_H	81	Not found
O_O_N	80	63
U	80	77
U_N	72	79
U_U	76	79
U_U_I_I	Not found	Not found
U_U_N	70	74

Appendix 3: Rate of speaking within a recording session and across the recording sessions

Rate of speaking within a recording session/RS			
Vowel	**Recording session 1_Batch 1**	**Recording session 1_Batch 10**	**Recording session 1_Batch 11**
Duration of non-final unstressed ə	.057	.060	.054
Duration of non-final stressed ə	.081	.084	.087
Duration of final unstressed ɑ:	.110	.102	.112
Duration of final stressed ɑ:	.169	.165	.161
Rate of speaking across recording sessions			
Vowel	**Recording session 1**	**Recording session 6**	**Recording session 14**
Average duration of non final unstressed ə	.056	.059	.057
Average duration of final unstressed i:	.089	.081	.088