Multitier Annotation of Urdu Speech Corpus

Benazir Mumtaz, Amen Hussain, Sarmad Hussain, Afia Mahmood, Rashida Bhatti, Mahwish Farooq, Sahar Rauf
Centre for Language Engineering,
Al-Khawarizmi Institute of Compute Science,
University of Engineering and Technology,
Lahore
firstname.lastname@kics.edu.pk

Abstract

This paper describes the multi-level annotation process of Urdu speech corpus and its quality assessment using PRAAT. The annotation of speech corpus has been done at phoneme, word, syllable and break index levels. Phoneme, word and break index level annotation has been done manually by trained linguists whereas syllable-tier annotation has been done automatically using template matching algorithm. The mean accuracy achieved at phoneme and break index label and boundary identification is 79.07% and 89.67% respectively. The quality assessment of word and syllable tiers is still under investigation.

1 Introduction

Annotated or tagged speech corpus is an electronic corpus which contains information about the language at phoneme, syllable, stress, word, phrase/ break index and intonation levels. An annotated speech corpus is very significant from computational linguistics perspective as it gives an opportunity to the researchers to observe, optimize, evaluate and re-evaluate the linguistics hypotheses [15]. Moreover, it plays a significant role in the development of a text to speech (TTS) synthesizer.

TTS system needs linguistic input to produce a language, similar to humans. Human child acquires this linguistic information from his environment, stores it in his memory and gradually starts using this information. The TTS similarly takes its linguistic input from annotated speech corpus. Thus for the development of a TTS it is very crucial that speech corpus is annotated very precisely at multiple levels. This paper describes the development, annotation, and quality assessment process for thirty minutes of Urdu speech corpus at phoneme, word, syllable and phrase.

The paper is organized in the following sections. The previous research in the annotated speech corpus development is presented in Section 2. The methodology of Urdu speech corpus annotation at phoneme, word, syllable and break-index level is detailed in Section 3. Quality assessment for each level of annotation is presented in Section 4. The current status of the Speech corpus annotation is given in Section 5 while future work and conclusions are discussed in Section 6.

2 Literature Review

Speech corpus, annotated at phoneme, syllable, word, and phrase level, is a pre-requisite to the development of a robust TTS system [19].

Phoneme level segmentation is a two step process; in the first step the individual phonemes are identified and in the second step their corresponding boundary marks are adjusted. Several methods for automatic phoneme level annotation have been proposed that try to mimic this two step process of annotation. Toledano et al. [4] used the HMM-based models for phoneme identification and proposed the fuzzy logic based post correction rules for the accurate boundary marking. Kuo and Wang [11] proposed a minimum boundary error framework that attempts to minimize the boundary error using manually annotated data. Wang et al. [5] proposed an HMM and SVM based method for automatic phoneme level annotation.

Besides the automatic annotation of speech corpus, manual annotation process has also been used. Sunitha et al. [12] used manual annotation process for the development of TELUGU TTS system. Similarly Chu et al. conducted both manual as well automatic phoneme level annotation of the speech corpus. The obtained results show that manual speech annotation produces good results in the development of a text to speech synthesis system [13]. Although the automatic annotation process is less time consuming, it fails to produce accurate phoneme level annotation.

For the word-tier annotation, both manual and automatic annotation process had been used. Matoušek and Romportl [9] proposed a two phase manual annotation process. In the first phase a skilled annotator annotates the speech at word level and in the second phase the initial annotated speech is revised and corrected by another skilled annotator. Arvaniti [1] had also manually annotated the Greek speech corpus at word level. She had used romanized form of Greek language to annotate the word-tier. In contrast to Arvaniti [1], Goldman [10] had introduced "EasyAlign" tool which automatically aligns continuous speech at three stages: macro segmentation at utterance level, grapheme-to-phoneme conversion and phone segmentation. At phone segmentation level, phone and word are computed using Viterbi-based HVite tool within HTK. Utterances are also verified to its phonetic sequences at this level. The EasyAlign tool uses HMM based models for word identification. Therefore, the proposed method can also not identify accurate word boundary points.

Syllabification of the speech can also be done in two ways; manually or automatically. Sunitha et al. [12] has taken the syllable as the basic unit and used manual syllabification to attain accuracy but manual syllabification is a time consuming process. Therefore, automatic syllabification has been used in different languages. For Telugu TTS system, Sunitha et al. [12] and Tsubaki [7] generated the syllable tier automatically. Hussain [18] has also proposed an algorithm for automatic syllabification of Urdu language words. He has used both Nucleus projection and template matching techniques for Urdu language word syllabification.

For break index/phrase level marking, TOBI system has been used. Japanese language has used J_ToBI tool to annotate the break index-tier manually as well as automatically [8]. J_ToBI is a prosodic labeling tool. Along with the BI (Break Index) 0, 1, 2 and 3, it also assigns; (-) to show uncertainty, (p) to show disfluent disjuncture, and (m) to show mismatch in disjuncture and tone. C-ToBI is used to mark prosodic events in Chinese [20]. A software package SFS (Speech File System, from UCL) is used for C-ToBI transcription but sometimes boundaries need to be modified manually. The software assigns a scale as 2 for the normal break level, 1 for reduced boundary, 3 for more prolonged boundary than the normal, 0 for extremely reduced boundary and 4 for extremely prolonged boundary.

The quality of a TTS system substantially depends on the accuracy of speech segment identification. Therefore, the quality estimation of annotated speech is very essential before providing the annotated data for the training of speech synthesizer. Several methods have been used to ensure the quality of annotated speech. Matouˇsek & Jan [9] used a two step process to produce a better quality of speech annotation. They computed the word error rate and the sentence error rate by comparing the raw text, and the first and second time annotated speech.

Pollák and Cˇernocky [14] proposed a three step process for the assessment of annotated speech. In the first step the annotated speech goes through a syntax test. The syntax test checks the usage of allowed characters and special marks, and ensures that all the annotated fields are non-empty. In the second step the pronunciation of the annotated word is compared with a standard pronunciation. The annotated pronunciation is marked erroneous after the confirmation from a specialized annotator. The final test involves the listening of a random utterance. If the listened utterance is same as the transcription then the annotated speech passes this test. The labeled data will be accepted if all the above mentioned tests are passed. The merits of evaluation of the annotated speech described by [16] used the metric of annotated unit's label as well as the timing boundary of units having identical label to estimate the quality of annotation.

While a lot of research has been conducted in developing annotated speech corpora of various languages, only limited work has been conducted for Urdu language speech corpus development [6]. Thus the current research aims to build on the previous research efforts and develop a speech corpus annotated at the defined four levels. The following section presents the methodology followed for its development.

3 Methodology

To build a speech corpus, thirty minutes of speech has been recorded by a single speaker in the anechoic chamber. This speech is recorded in ‘mono’ form at a sampling rate of 8 kHz. PRAAT software has been used for the recording, annotation and quality assessment of the speech corpus.

The recorded speech corpus is segmented at multiple tiers using Case Insensitive Speech Assessment Method Phonetic (CISAMPA). See appendix 1 for the detailed description of CISAMPA symbols. The methodology for multitier annotation is discussed in the following sections.

3.1 Annotation of Speech Corpus at Phoneme Level

Phoneme tier is annotated manually in this work. At phoneme level, each consonant and vowel is distinctly marked in the Text Grid file after conducting the careful analysis of their properties in the spectrum and time wave form. Following guidelines have been used for the phoneme level annotation:

- Silence is marked in the start and end of the sentence.
- Each segment boundary is marked at the zero crossing point where the sound wave amplitude is going from negative to positive value.
- While splitting a vowel and consonant sound, boundary of the consonant is marked where the personality of the vowel disappears.
- If a few periods of the wave form are creating ambiguity in determining the personality of the vowel then the periods having mixed properties (both of the consonant and the vowel) are included in the vowel.
- While splitting the vowel and vowel junction, the periods with mixed properties of both the vowels are divided into equal halves.
- In case of consonant clusters within or across the words, the wave time periods with mixed properties of both consonants are divided into equal halves and mark as two distinct sounds.
- In case of gemination across the words or within the word, phonemes are divided Sinto equal halves and marked as two distinct sounds but in case of geminated stops and affricates, the closure period is divided into equal halves.
- If a sentence or phrase is starting with the voiceless stop or affricate, the closure duration taken for the onset voiceless stop is 100 milliseconds for the stressed syllable and 87 milliseconds for the unstressed syllable [17].
- If a sentence or phrase is ending with a voiceless stop (there should be silence after the word) and the burst of the stop is not visible, the closure duration taken for the coda voiceless stop is 77 milliseconds for the stressed syllable and 73 milliseconds for the unstressed syllable [17].
- A vowel is labeled as a nasal vowel only if it is contrastively nasalized, if a vowel is contextually nasalized, it is labeled as an oral vowel.

Once the corpus is annotated at phoneme level, it then undergoes the phoneme level quality assessment explained in Section 4.1. The annotated data is passed to word level annotation phase if it is accepted by the phoneme level quality assessment process.

3.2 Annotation of Speech Corpus at Word Level

Annotation at word level is done in two stages. Firstly, the annotator listens and observes the spectrogram of the wave file very carefully to find out that all the words in the file are pronounced properly. In case of mispronunciation/misreading, insertion of extra phoneme in a word or deletion of required phoneme from the word, the wave file is rejected and sent back for the rerecording. In the next stage, the word boundaries of correctly pronounced words are marked manually. These boundaries are completely aligned with the boundaries of the segments. The annotator does not write the word labels between the word boundaries. Symbols are automatically extracted from the phone-tier to fill the word boundaries.

Since the boundaries of words in Urdu language cannot always be identified on the basis of space, it becomes very difficult to determine where the word boundary mark be placed, especially in the case of compound words. For example it is challenging to decide that the word "خوش شکل" (xʊʃ ʃəkəl\good looking) should be marked as one word or two. Therefore, following principles have been used to mark the boundaries between compound words:

- A compound word consisting of two words that are both meaningful is marked as two different words as in the case of the compound word "موم بتی" (mo:m bəʈʈi:\candle).
- A compound word consisting of a meaningless prefix and meaningful word is marked as one word as in the case of compound word "بہ معنی" (bəhməʔni:\ as a meaning of).
- A compound word consisting of meaningful word and meaningless suffix is marked as one word as in the word "خیال آرائی" (xəja:la:ra:i:\imagination).

- A compound word consisting of two meaningless words is marked as one word as in the case of the compound word " ناکردہ " (na:kərɖa:\undone).
- A compound word consisting of a meaningful prefix as well as a meaningful second word is marked as one word as in the case of the compound word " خوب صورت" (xu:bsu:rət̪\beautiful).
- A compound word consisting of meaningful word and meaningful suffix is marked as one word as in the case of the compound word " رنگ ساز" (rəŋg sa:z\dyer).
- A compound word consisting of two meaningful words, combined with a conjunction vao "و " is marked as three different words as in the case of the compound word " غور و فکر" (ɣɔ:r o: fikr\contemplation).
- A compound word combined with <ے> " یائے اضافت" is marked as one word as in the case of the compound word " دریائے راوی " (ɖərja:e:ra:vi:\ Ravi River).
- A compound word combined with <ِ> zair is marked as two different words. The zair phoneme should be the part of the first word while marking word boundary as in the case of " مخلوقِ خدا " (məxlu:qe: xʊɖa:\ creature of God).

Once the corpus is annotated at word level, it then undergoes the word level quality assessment explained in Section 4.2. If the work package is within the acceptance quality threshold it is then shipped to the next level of annotation process.

3.3 Annotation of Speech Corpus at Syllable Level

Syllable tier is automatically generated for Urdu speech corpus by using the algorithm for syllabification presented by Hussain [18]. The algorithm for the syllabification is as follows:

I. Convert the input phoneme string to consonant and vowel string
II. Start from the end of the word (i.e., right to left)
III. Traverse backwards to find the next vowel
IV. If there is a consonant before a vowel than mark a syllable boundary before the consonant
V. Else mark the syllable boundary before this vowel
VI. Repeat from step (iii) until the phonemic string is consumed completely.

3.4 Annotation of Speech Corpus at Break Index/Phrase Level

Annotation at break index level is done manually. Four TOBI levels have been used to annotate the Urdu speech corpus at break index tier. These levels are; 4, 3, 1 and 0. Level 2 has not been used to avoid confusion between Level 1 and Level 3.

The process of assigning break indices starts from left to right. Level 4 is assigned at full intonational phrase boundary. It is assigned at a pause that should be around 100 ms or more than 100 ms.

Level 3 is assigned at intermediate intonational phrase boundary. Three important clues are used in determining the level 3; weak disjuncture, lengthening of the vowel of last syllable and glottalisation. Level 3 has weak disjuncture that is usually visible in the pitch track. This weak disjuncture should be less than 100ms. The duration of the closure period of a voiceless stop and affricates should be carefully separated from the weak disjuncture while assigning level 3. To find out the lengthening of the vowel, the vowel of the last syllable is compared with the same or similar shortest vowel in the file. The lengthened vowel should be 50% long than the shortest. Glottalisation is also a clue of assigning level 3 at two intermediate intonational phrases.

Level 1 is assigned at typical word boundary where there is no lengthening of the vowel, glottalisation and pause. Level 0 is assigned when the boundary between two words is completely removed as in case of clitics.

A sample of annotated speech wave file showing all the layers has been given below:

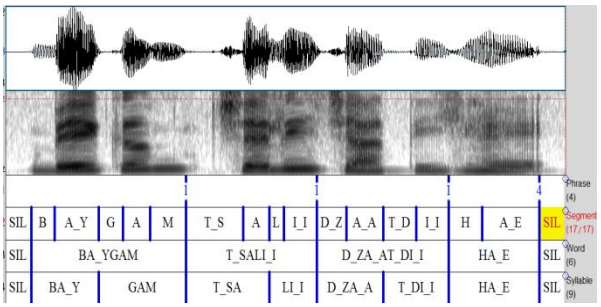

Figure 1: Annotated speech file

4 Speech Annotation Quality Assessment

In this section the quality assessment procedure of annotated layers has been discussed. All the manually labeled files go through different tests at each layer of annotation before they are accepted. Scripts are written in PRAAT [2] which performs the quality estimation tests and produce analysis files. These are explained in the respective annotation layer quality assessment sections below.

The general strategy for quality assessment is that a certain percentage of speech files are manually annotated by an experienced annotator known as the reference files. These files are then compared automatically with the corresponding same speech files annotated by the speech corpus annotation team, called the source files. The mismatches are manually verified by the quality assurance personnel to identify possible errors in the source files. If the error rate is more than 5%, then the source file is rejected and the work package is re-annotated.

4.1 Phoneme Level Assessment

The phoneme level annotation is graded using a two step process. In the first step the phoneme labels are checked whether they are from a defined phone set (as given in appendix 1) and in the second step it is estimated that all the starting boundary of each segment is marked at zero crossing point; amplitude going from negative to positive. The source file is rejected even if a single marked label is not listed in the phone set.

In the second phase the correctness of phoneme label text and boundary is assured. The source files are compared with their respective reference files on the basis of phoneme label and phoneme boundary.

For the phoneme label text comparison maximum string alignment algorithm is used [14]. The alignment algorithm aligns the source and reference phoneme based strings. The output of this alignment algorithm is shown in Figure 2.

U_USKI_I A - MI_I
U SKI_IAMMI_I

Figure 2: Label comparison through maximum string alignment algorithm

In Figure 2 the upper phoneme based string is extracted from a source file and the second one is fetched from the corresponding reference annotated file. After the string alignment, all the reported errors are reviewed manually. This manual review is done by considering the error margin involved in the generation of reference files. After this review only those files will pass this test whose phoneme labels are 100% accurate. For checking the phoneme boundary, the phoneme boundary marked in source file is compared with the corresponding boundary mark in reference file. A time period (T_1) at the surroundings of every boundary point (B_1) is calculated and if in the duration of$(B_1 \pm (T_1 * 1.2))$, there is no boundary point in its counter annotated file then a boundary misalignment is reported. This process is followed for the reference and the source annotated files separately. If the accumulated mismatch with respect to source and reference files is more than 5% then before rejecting the file all the reported mismatches are checked manually to confirm the rejection of the source file.

4.2 Word Level Assessment

An annotated word goes through four types of tests before it gets accepted that are explained below.

In the first test it is assured that a word label should not contain any non speech phoneme label; SIL, PAU as given in phoneset defined in appendix 1. In the second step it is tested that the number of words in text form should be equal to the number of annotated words in the source file. The third test at the word layer is designed to check that all the labeled words can be syllabified according to the Urdu syllabification rules [18]. The words that cannot be syllabified are reported and these rejected words are reviewed by an expert linguist to confirm their incorrectness.

In the final test, the pronunciation of labeled word is compared with the standard Urdu pronunciation available in the pronunciation lexicon and all the erroneous pronunciations are reported after a manual confirmation. In the pronunciation comparison two possible scenarios occur: a word is not found in lexicon or the annotated pronunciation is not found in lexicon. If a word doesn't exist in the pronunciation lexicon, an Urdu linguist is given with the following options:

1. Add the annotated pronunciation in the lexicon
2. Report the annotated pronunciation as an erroneous pronunciation and add the correct pronunciation in the lexicon

In a case that a word exits in the lexicon but the lexicon pronunciation doesn't match the annotated pronunciation then the Urdu linguist is prompted with the following options:

1. Replace the annotated pronunciation with the lexicon's pronunciation
2. Report the annotated pronunciation as an erroneous pronunciation
3. Add the annotated pronunciation as an alternative pronunciation

The pronunciation lookup test will fail if the pronunciation is reported as erroneous by the expert linguist.

The source file will be rejected if even a single word fails any of the above mentioned tests.

4.3 Phrase Level Assessment

Phrase level annotation assessment is a two step process. In the first step, the time of break index in the source file is compared with a reference file. In the second step the level of break index mark are compared. Both these comparisons are done by using the algorithms discussed in section 4.1. In the phoneme level comparison, string alignment algorithm [3] is used where the levels (0-4) are used as a basic unit contrary to the phoneme label. After the analysis the reported errors are reviewed manually. Files that contained even a single error after the manual verification are rejected. The methodology for assessing the syllable tier is under process. Therefore, it has not been discussed in this paper.

5 Current Status of the Urdu Speech Corpus Annotation

Reference annotated files were generated for the complete thirty minutes of speech for the quality assessment of annotated corpus. Results of segment level assessment have been reported in Table 1 to present the overall accuracy of annotation at this level.

Table 1: Phoneme level annotation quality assessment

Annotation Quality Assessment Tests	Total Number of Phones	Total Number of Erroneous Phones	Percentage of Accuracy
Phoneme Label Comparison	19600	2083	89.37%
Phoneme Boundary Comparison	38162	11916	68.77%

The percentage of accuracy achieved after applying the phrase level quality evaluation tests is presented in Table 2 below.

Table 2: Phrase level annotation quality assessment

Annotation Quality Assessment Tests	Total Number of Break Indices	Total Number of Erroneous Break Indices	Percentage of Accuracy
Break Index Level Comparison	5055	978	80.65%
Break Index Time Mark Comparison	9356	122	98.70%

6 Discussion

It is very important for the quality of TTS system that the annotated speech corpus does not contain any errors. Therefore, after the quality assessment results, manual review both at phoneme and break index levels has been carried out by the trained linguists to correct all the errors.

Although this paper present sufficient details about the process of annotating data at phoneme, word and break index levels, there are still issues that need to be resolved. In Urdu language, the existence of diphthongs is still indeterminate. It cannot be precisely stated that how many diphthongs exist in Urdu language. Therefore, at phoneme tier, while segmenting words such as "کیوں" (kɪũ:\Why), " بھائی " ($b^{h}ɑ:i:$\ Brother), "آئے" (ɑ:e:\ Came), " کیا " (kæa:\ What) it is difficult to decide whether the vowels be marked as diphthongs or the boundary should be marked between them to make them two individual phonemes.

Besides diphthongs, co-articulation factor has also created problem in the identification of phonemes. For example, due to co-articulation affect, voiced consonants become voiceless, aspirated consonants become unaspirated, and oral vowels become nasal vowels when they are preceded and followed by the nasal consonants.

Similarly, Level 0 is not used in marking break indices as this level is reserved for clitics. This phenomenon that Urdu language has clitics is still under investigation and needs further research.

Building on this research, development of ten hours of annotated speech corpus is underway. Currently the phoneme, word, syllable and break indices tiers are

annotated but in future intonation tier will also be focused. Automatic annotation methods will also be further investigated in future.

7 Conclusion

In this paper, annotation and testing of 30 minutes of Urdu speech corpus at phoneme, word, syllable, and break index levels has been described. This annotation is done using both manual and automatic methods. On average 79.07% accuracy is achieved at phoneme tier and 89.67% accuracy is achieved at break index tier. After quality assessment results, manual review is also conducted to correct all errors at phoneme and break index levels. This work is in process and the knowledge generated through this process will be used to develop ten hours of annotated speech corpus.

8 Acknowledgement

This work has been conducted through the project, Enabling Information Access for Mobile based Urdu Dialogue Systems and Screen Readers supported through a research grant from ICTRnD Fund, Pakistan. We would also like to thank Wajiha Habib who was the speaker of 30 minutes speech corpus.

9 References

[1] A. Arvaniti, & M. Baltazani, "GREEK ToBI: A System for the Annotation of Greek Speech Corpora", LREC. 2000, (May, Friday, 2014). Available: http://www.kent.ac.uk/secl/ell/staff/amalia-arvaniti/docs/ArvanitiBaltazani_LREC.pdf

[2] B. Paul, W. David, Retrieved September 10, 2013, Available: http://www.fon.hum.uva.nl/praat/

[3] D. Jurafsky, Minimum Edit Distance. 2013. Retrieved June, 6, 2014. Available: http://www.stanford.edu/class/cs124/lec/med.pdf

[4] D. T. Toledano, M. A. Crespo, J. G Sardina, "Trying to Mimic Human Segmentation of Speech Using HMM". Third ESCA/COCOSDA International Workshop on Speech Synthesis, (pp. 1263-1266). Australia. 1998.

[5] H.M. Wang, J.-W. Kuo, H.-Y. Lo, "Towards A Phoneme Labeled Mandarin Chinese Speech Corpus", International Conference on Speech Databases and Assessments (ICSDA). Istanbul. 2008.

[6] H. Sarfraz, S. Hussain, R. Bokhari, A. A. Raza, I. Ullah, Z. Sarfraz, S. Pervez, A. Mustafa, I. Javed, R. Parveen. "Speech Corpus Development for a Speaker Independent Spontaneous Urdu Speech Recognition System." proceeding of OCOCOSDA (2010).

[7] H. Tsubaki, M. Kondo, "Analysis of L2 English Speech Corpus by Automatic Phoneme Alignment", Proceedings from SLaTE 2011, Jan 8, 2012.

[8] J. J. Venditti, "The J_ToBI model of Japanese intonation." Prosodic typology: The phonology of intonation and phrasing, 2005, 172-200. Available at:

[9] J. Matoušek, J. Romportl, "Recording and annotation of speech corpus for Czech unit selection speech synthesis". In Text, Speech and Dialogue. 2007. (pp. 326-333). Springer Berlin Heidelberg.

[10] J. P. Goldman, "EasyAlign: An Automatic Phonetic Alignment Tool Under Praat." INTERSPEECH. 2011.

[11] J.W. Kuo, H. M. Wang, "A minimum boundary error framework for automatic phonetic segmentation". ISCSLP. 2006. (pp. 399-409). Springer-Verlag Berlin Heidelberg.

[12] K.V.N. Sunitha and P. Sunitha Devi, Bhaashika: Telugu TTS System, International Journal of Engineering Science and Technology Vol.2 (11), 2010 , Hyderabad, India , pp.66272.

[13] M. Chu, Y. Chen, Y. Zhao, Y. Li, and F. Soong, "A Study on How Human Annotations Benefit the TTS Voice."Microsoft Research Asia,.2006, Beijing,China,pg4.

[14] P. Pollák, & J. Cˇernocky, "Orthographic and Phonetic Annotation of Very Large Czech", in proc, Language Resources and Evaluation Conference, LREC, (pp. 595-598). Lisabon, (2004)

[15] R. Mitkov, C.Orasan, and R Evans. "The importance of annotated corpora for NLP: the cases of anaphora resolution and clause splitting." Proceedings of Corpora and NLP: Reflecting on Methodology Workshop. 1999.

[16] S. Cox, R. Brady, & P. Jackson, "Techniques for Accurate Automatic Annotation of Speech Waveforms", in proc. International Conference on Spoken Language Processing, (pp. 1947–1950). Sydney, 1998.

[17] S. Hussain, "Phonetic Correlates of Lexical Stress in Urdu", PhD, Northwestern University, Illinois, 1997.

[18] S. Hussain, "Phonological Processing for Urdu Text to Speech System", Lahore: Center for Research in Urdu Language Processing, National University of Computer and Emerging Sciences, B Block, Faisal Town, Lahore, Pakistan. (2007).

[19] S. Kiruthiga, and K. Krishnamoorthy. "Annotating Speech Corpus for Prosody Modeling in Indian Language Text to Speech Systems." International Journal of Computer Science Issues (IJCSI) 9.1 (2012).

[20] Z. Weibin, S. Liqin, and N. Xiaochuan, "Duration Modeling For Chinese Systhesis from C-ToBI Labeled Corpus." ICSLP, 2000.

Appendix 1:

Sr.#	Urdu Letter	IPA	CISAMPA	Sr.#	Urdu Letter	IPA	CISAMPA
1	پ	P	P	35	ر	R	R
2	پھ	p^h	P_H	36	رھ	r^h	R_H
3	ب	B	B	37	ڑ	ɽ	R_R
4	بھ	b^h	B_H	38	ڑھ	$ɽ^h$	R_R_H
5	م	M	M	39	ی	J	J
6	مھ	m^h	M_H	40	یھ	j^h	J_H
7	ت،ط	t̪	T_D	41	چ	ʧ	T_S
8	تھ	$t̪^h$	T_D_H	42	چھ	$ʧ^h$	T_S_H
9	د	d̪	D_D	43	ج	ʤ	D_Z
10	دھ	$d̪^h$	D_D_H	44	جھ	ʤh	D_Z_H
11	ٹ	T	T	45	ُو	u:	U_U
12	ٹھ	t^h	T_H	46	ُوں	ũ:	U_U_N
13	ڈ	D	D	47	و	o:	O_O
14	ڈھ	d^h	D_H	48	وں	õ:	O_O_N
15	ن	N	N	49	وَ	ɔ:	O
16	نھ	n^h	N_H	50	وَں	ɔ̃:	O_N
17	ک	K	K	51	ا،آ	ɑ:	A_A
18	کھ	k^h	K_H	52	ان،آں	ɑ̃:	A_A_N
19	گ	g	G	53	ی	i:	I_I
20	گھ	g^h	G_H	54	یں	ĩ:	I_I_N
21	ن in نگ, نک, نکھ, نگھ	ŋ	N_G	55	ے	e:	A_Y
22	ق	Q	Q	56	یں	ẽ:	A_Y_N
23	ع	ʔ	Y	57	ہ ِ	E	A_Y_H
24	ف	F	F	58		e	A_E_H
25	و	V	V	59	ہ ُ	O	O_O_H
26	س،ص،ث	S	S	60	ے	æ:	A_E
27	ذ،ز،ظ،ض	Z	Z	61	یَں	æ̃:	A_E_N
28	ش	ʃ	S_H	62	ِ	ɪ	I
29	ژ	ʒ	Z_Z	63	ُ	ʊ	U
30	خ	X	X	64	ء،َ	ə	A
31	غ	ɣ	G_G	65	ں ِ	ĩ	I_N
32	ہ،ح	H	H	66	ں ُ	ʊ̃	U_N
33	ل	L	L	67	ں َ	ə̃	A_N
34	لھ	l^h	L_H				

The grey highlighted sounds are used rarely.