Subjective Testing of Urdu Text-to-Speech (TTS) System

Kh.Shahzada Shahid[1], Tania Habib[2], Benazir Mumtaz[2], Farah Adeeba[2], Ehsan ul Haq[3]
Centre for Language Engineering
Al-Khawarizmi Institute of Computer Science
University of Engineering and Technology, Lahore
[1]{khawaja.shahzada}, [2]{firstname.lastname}, [3]{ehsan.ulhaq}@kics.edu.pk

Abstract

Text-to-speech (TTS) systems for many widely spoken languages have been developed and evolved over the last few decades. Such systems are being used in many different fields. Since these TTS systems have differences in the perceived sound quality, many speech quality test methods have been proposed to compare and evaluate their performance. Test materials for these tests, however, are language specific and hence cannot be used for TTS systems developed for other languages such as Urdu. In this work, we have presented a speech quality test material specially designed for Urdu TTS systems. The proposed test is conducted using the perception of both blind and non-blind native speakers to evaluate naturalness as well as phoneme, word and sentence-level intelligibility of recently developed Urdu TTS system. Furthermore, a qualitative comparison is performed between two most popular methods for building TTS systems.

1. Introduction

Text-to-speech systems (TTS) are commonly used in everyday life, e.g., in navigation devices, public announcement systems [1] and entertainment productions [2] . It also plays a crucial role in the field of telecommunication, industrial and educational applications. TTS systems for foreign languages such as English, German and Japanese, have been developed long ago and are well established today [3]–[5]. However, research on the development of TTS system for the Urdu Language, which is a national language of Pakistan and is spoken by more than 162 million people worldwide [6], is still in its earlier stages [7]. This paper is an attempt to assess the speech quality of recently developed Urdu TTS system [8]. This effort will enhance man to machine interaction possibilities and overcome the literacy barrier for the semi-urban and rural population of Pakistan.

Speech quality is a multi-dimensional term and its evaluation contains several problems [9][10]. Speech quality of a synthesizer is determined by its similarity to the human voice (i.e., *naturalness*), its ability to be easily understood (i.e., *ineligibility*) [11] and its suitability for certain applications [10][12]. Moreover, it is reported that different applications prefer different features' evaluation. For instance, the high speaking rate with speech intelligibility features is usually preferred over naturalness in reading machines for the blind. On the other hand, in multimedia applications or electronic mail readers, prosodic features and naturalness are considered as essential features [13].

Subjective evaluation of speech synthesis is usually done by listening tests according to standards described by ITU-T Rec. P.85 [14]. Several methods have been developed during last decades for assessment of synthetic speech. However, no single evaluation provides a foolproof assessment method that focuses on both naturalness and intelligibility aspects of speech at different levels (phoneme, word, sentence or comprehension) and can provide useful and reliable information about the quality of TTS system. In addition, prior studies indicate that test materials developed for subjective evaluation of TTS need to be language specific [15]. Moreover test material should be large enough to represent a variety of language features (*representativeness*), while at the same time short enough not to distract listeners' attention (*compactness*).

In this study, we have designed both compact and representative subjective testing material for the evaluation of Urdu TTS systems. The proposed tests have been conducted on blind and non-blind Urdu native speakers and results have been reported about speech quality of Urdu TTS system. These results not only evaluate TTS speech quality but also help to

figure out areas that need to be considered for further improvements in TTS. Furthermore, this work also compares the two widely recognized approaches to build speech synthesizers, i.e., unit selection [16] and Hidden Markov Models (HMMs) [17] , with the aim to identify which one is better choice for generating Urdu synthetic speech in terms of both naturalness and intelligibility.

The remainder of this paper is divided into following sections: Section 2 briefly describes the architecture of Urdu TTS system. Section 3 explains the design of subjective quality test and testing materials selected for this purpose. The procedure and comparative results of two voice synthesis approaches are reported in Sections 4 and 5 respectively. Finally, Section 6 concludes the findings of this research.

2. Urdu TTS System Architecture

Urdu TTS system converts Urdu text into synthetic speech waveform as shown in Figure 1. TTS system generally consists of two main modules, Natural Language Processor (NLP) and Speech Synthesizer. NLP pre-processes the input text including abbreviations, dates, and numbers; and converts into its appropriate phonetic description annotated with prosodic and context dependent information. Speech Synthesizer then generates corresponding speech signal using the description provided by NLP. Overall speech quality of TTS system relies on both of these modules.

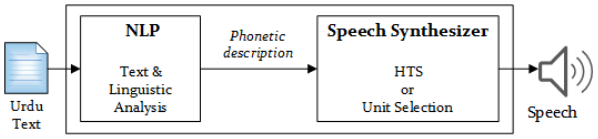

Figure 1 Architecture of TTS system.

Two different types of concatenative synthesis approaches have been used in Urdu TTS system. First, the classical unit selection (US) method that synthesizes speech by concatenating pre-recorded human speech waveforms and hence requires a large amount of speech database [4]. Second is Hidden Markov Model-based synthesis (HTS) that uses statistical models instead of actual speech units [18] and for this reason its footprint is very small (less than 10MB), compared to unit selection approach. More architectural details of Urdu TTS system are available in [18] and [19].

3. Design of Subjective Test

The design of subjective test highly depends on the application domain where TTS system is to be deployed. For example a TTS destined to provide traffic information asks for a more specific type of test materials than TTS to be used as news/screen-reader for the blind, where test materials should cover vocabulary from a wide range of topics (e.g., religion, sports, literature, health etc.) and multiple sentence structures [20]. Urdu TTS system belongs to the second type of category, and hence quality test is designed comprehensively. The test contains a total of 1010 words out of which 496 are unique. These words are taken from news, literature, and daily life conversational vocabulary. Total speaking time of the test is approximately 9 minutes and response time is around 20 minutes.

The theme of this subjective test revolves around four questions: (a) Is Urdu TTS system mature enough to deliver any type of speech content with the acceptable clarity of voice and the underlying message? (b) Is Urdu TTS' voice as pleasant as that of human beings? (c) Is Urdu TTS system suitable for both the blind and non-blind communities? (d) Which one of the two speech synthesis approaches (HTS or US) is a better choice for Urdu TTS based on the criteria set by above questions? To answer these questions, a group of subjective tests is conducted categorized under intelligibility and naturalness tests that are briefly explained below.

3.1. Intelligibility Tests

Intelligibility tests focus on the ability to identify what is spoken regardless of whether it sounds robotic or human-like, noisy or clear. Good quality in intelligibility includes an understanding of spoken utterances with correct perception at each level of speech units from phonemes to sentences [21]. Intelligibility tests designed at segmental, sentence and comprehension levels for Urdu TTS systems are discussed below.

3.1.1. Segmental Test With segmental evaluation methods intelligibility is tested at smallest speech units, like phonemes. In contrast to vowels, consonants are difficult to recognize in synthetic speech, because of sudden spectral transitions and multiple excitation signals [20] and hence test materials usually focus on consonants [13]. Moreover, syllable-initial and syllable-final consonants are perceived differently by listeners [22]. For this reason, it makes sense to break

down the segmental-quality evaluation of TTS for consonants in both initial and final positions within monosyllabic words. For this purpose, a test set is designed containing 64 pairs of confusable rhyme words. Words in a pair differ in their initial or final consonants. The consonants are equally distributed among 4 phonemic distinctive features (8 word-pairs per feature per position). Few examples are shown in Table 1, for complete dataset please refer to Tables A-1 and A-2 in Appendix A.

Table 1 Examples of segmental evaluation test.

Phonemic features	Description	Pairs with different initial consonants	Pairs with different final consonants
Voicing	voiced - unvoiced	پات، /pɑːʈ/ bɑːʈ	باپ، /bɑːp bɑːb/
Nasality	nasal - oral	مول، /mol bol/	تام، /ʈɑːm ʈɑːb/ تاب
Aspiration	Aspirated – Non-Aspirated	بال، /bɑːl bʰɑːl/ بھال	باپ، /bɑːp bɑːpʰ/ باپھ
Sibilation	sibilated - unsibilated	چال، /ʧʰɑːl kɑːl/ کال	ساز، /sɑːz sɑː tʰ/ ساتھ

These rhyme words are tested through following carrier sentence:

(1) کیا آپ اردو لغت سے لفظ ---- کا مطلب بتا سکتے ہیں؟

kæɑː ɑːp ʊrɖu lʊɣəʈ se ləfz ----- kɑː məʈləb bəʈɑː səkʈe hæː

What- kæɑ: you- ɑːp Urdu- ʊrɖu dictionary-lʊɣəʈ case marker-se word- ləfz ---- case marker- kɑ: meaning- məʈləb tell- bəʈɑ: can-səkʈe tense aux- hæ:

"Can you inform me the meaning of --- word from the dictionary?"

First, a pair of rhymed words is visually presented to the subject. Then one word of the pair embedded in the carrier sentence is aurally presented and the listener's task is to indicate which of the two words was spoken as part of the sentence. The carrier sentence is sensitive to segmental errors in the word to be tested, as there is a lack of contextual information that can assist listeners to predict the segment not heard. Furthermore, all the cognitive information that is required for this recognition task is provided to the listener before the auditory presentation. Hence, an error in identifying the word can be regarded as a direct measure of TTS systems' inaccuracy.

This diagnostic test can highlight the misidentified phonemes and help to localize the problem points for improvements. The obtained measure of segmental intelligibility is simply the percentage of correctly identified words distributed among 4 phonemic distinctive features.

Table 2 SUS test sets.

Sr. No.	Sentences-Set1	Sentences-Set2
1	میز تیز رفتاری سے بیٹھ گیا۔ meːz ʈeːz rəfʈaːri se bæTʰ gæaː	جہاز کے پتے رونے لگے dʒəhaːz keː pəʈʈeː roːneː ləgeː
2	باغ میں کاغذ بجنے لگا۔ baːɣ mẽː kaːɣəz bæhneː ləgaː	کتاب کی آواز چبھنے لگی kiʈaːb ki aːvaːz ʧəməkneː ləgiː
3	جہاز اپنے پیروں پر لیٹ گیا۔ dʒəhaz əpne pærõː pər leːT gæaː	درخت لوکی طرح اڑنے لگے ɖərəxʈ ləhuː ki ʈərha uɳe ləgeː
4	کتاب کے پتے ٹوٹنے لگے۔ kiʈab ke pəʈʈeː TuːTneː ləgeː	ریت اپنے پیروں پر لیٹ گئی reːʈ apneː pærõː pər leːT gəiː
5	درخت پر سے سڑک ٹوٹنے لگی۔ ɖərəxʈ pər se sərək TuːTneː ləgiː	باغ کا میز پاگل ہو کر بیٹھ گیا baːɣ ka meːz paːgəl ho kər bæTʰ gæɑː
6	ہوا لوٹنے کی آواز چبھنے لگی həva ʈʊTne ki avaz ʧəməkne ləgi	ہوا بیٹھ کر مر جھا گئی həvaː bæTʰ kər mʊrdʒʰaː gəiː
7	درخت کر سیوں پر ناچنے لگے ɖərəxʈ kʊrsĩjõː pər naʧne ləgeː	کچے انڈے درخت پر ناچنے لگے kəʧʧeː ənDeː ɖərəxʈ pər naːʧneː ləgeː
8	ریت کی آواز مر جھا گئی reːʈ ki aːvaːz mʊrdʒʰaː gəiː	خشک جوتیاں میز پر ٹوٹنے لگیں xʊʃk dʒuːʈiːjãː meːz pər TuːTneː ləgĩː
9	خشک جوتیاں لو کی طرح اڑنے لگیں xʊʃk dʒuːʈiːjãː ləhuː ki ʈərha uɳe ləgẽː	کچے انڈے سے سڑک بجنے لگی kəʧʧe ənDeː se sərək bæhneː ləgiː
10	کچے انڈے پاگل ہو کر رونے لگے kəʧʧeː ənDeː pagəl ho kər roːneː ləgeː	آواز کے بجنے سے میز لیٹ گیا aːvaːz ke bæhneː se meːz leːT gæɑː

3.1.2. Sentence level Test Segmental intelligibility at sentence level is usually evaluated through transcription task of semantically unpredictable sentences (SUS) [23][24]. SUS sentences have grammatically correct syntax, however, they are unpredictable semantically. They have no inherent meaning, therefore minimize the possibility of deriving phonetic information from textual context but the speech signal itself, e.g.,

(2) میز تیز رفتاری سے بیٹھ گیا-

mez tezrɑːftɑːri: se bæt gəɑː

Table- mez speedily- tezrɑːftɑːri case marker-se sat: bæt tense-gəɑ

"Table sat down speedily"

SUS sentences are constructed using high-frequency words from language specific lexica. Instead of forced-choice, subjects are asked to transcribe the sentence as they listen. This helps to avoid ceiling effect in listeners' responses. An overall percentage of correct recognition is calculated based on the percentage of correctly transcribed words per sentence. Higher the percentage more intelligible is the synthesized voice.

One inherent problem with sentence level tests is that each sentence can be presented to a subject only once during the test [21]. This fact becomes a major concern when the purpose of the test is to compare two different TTS technologies. In order to avoid learning effect, separate SUS test sets have been designed for both HTS and US voice synthesis and are shown in Table 2. For a fair comparison, the same set of vocabulary is used for both test sets.

Table 3 MOS rating scales [14]

Naturalness (Quality)	**How do you rate the quality of the sound that you just heard**? 1. Bad 2. Poor 3. Fair 4. Good 5. Excellent
Speaking rate	**What was the average speed of delivery?** 1. Much slower 2. Slower 3. Normal 4. Faster 5. Much faster
Pronunciation	**Did you notice any anomalies in pronunciation?** 1. Yes, very annoying 2. Yes, annoying Poor 3. Yes, slightly annoying 4. Yes, but not annoying 5. No.

3.1.2. Comprehension Test Intelligibility test methods discussed so far focus on the accuracy of individual sounds or words, rather than correct reception of the underlying message. For some TTS applications, such as news readers, it is not required to recognize every single phoneme, as long as the meaning of whatever is being spoken is understood [25]. In comprehension tests, synthesized speech sample containing few sentences or paragraph is presented to the subject, followed by a questionnaire about the content of the passage. Hundred percent segmental intelligibility is not needed to answer the questionnaire. Two news paragraphs from BBC Urdu website were selected for testing Urdu TTS. Topic selection was made from the category that is less likely to be familiar to most of the listeners such as latest research reports from health sciences domain.

3.2. Naturalness Test

The goal of an ideal TTS system is to mimic human speech style, so it should also be evaluated against overall speech quality parameters, such as *speaking rate*, *pronunciation*, and *naturalness*, in addition to intelligibility. Naturalness and overall quality of synthetic speech are difficult to quantify as they are abstract subjective attributes and subjects' may have different preferences for these attributes [21]. Mean opinion scoring (MOS), recommended in ITU-T Rec. P.85 [14], is a most widely used method for speech quality evaluations.

Table 4 MOS test set

Sr.	Sentences
1	اس دوران ترکی اور ایران کے مابین مجموعی تجارت کا حجم ۸۔۲۱ بلین ڈالر رہا Is ɖo:ra:n ʈʊrki: ɔr æra:n ke ma:bæn mədʒmu:i: ʈədʒarəʈ ka hʊdʒəm a:Th se Ikki:s blljən Dɔlər rəha:
2	جہاں خوکے بعد رش یا د واقع ہووہ بھی واو معدولہ ہو سکتی ہے dʒəhɑ̃: xʊke: ba:ɖ rəʃ ja:ɖ vaQəja: ho: vo: bhi va:ve dʒ məro:la: ho: səkʈi: hæ.
3	اس کی تاریخ پیدائش ہے ۶ / ۹ / ۱۹۸۰ Is ki ʈari:x pe:ɖa:lʃ hæ tʃe: no ʊnni:s sɔ əssi:
4	دونوں کھلاڑیوں نے: ۲،۲ وکٹیں لیں۔ ɖo:nõ: khllaɾljõ: ne: do: do: vikTẽ: lĩ:
5	ماہانہ تنخواہ ۲۰ ہزار روپے علاوہ ۱۶٪ کمیشن دی جائے گی فوری رابطہ کریں ۰۳۰۰ ۸۲۳۹۴۲۳۰ məha:na: ʈənxa: bi:s həza:r rʊpe: əla:va: ba:ra: fi:səɖ kəmi:ʃən ɖi: dʒa:e: gi: fɔ:ri: ra:bəʈa kərẽ: do: tʃɑ:r e:k a:Th do: tʃɑ:r no tʃɑ:r do: ʈi:n slfər
6	علینا نے کل ۱۹:۰۰ بجے مارکیٹ جانا ہے۔ əli:na: ne kəl ʊnni:s bəje: ma:rkIT dʒa:na: hæ
7	علم صرف میں ف، ع، ل کو حرف کی ہجائے کلمہ کہا جاتا ہے۔ Ilme sərf mẽ: fe: æn la:m ko: hərf ki bədʒa:e kəlma: kəha: dʒa:ʈa hæ
8	لاہور میں فروری ۱۴۰۶ء کو خوب طوفان آیا la:hɔr mẽ: fərvəri: tʃo:ɖa sɔ ɖo: ko: xu:b ʈʊfa:n a:ja:
9	آخری وقت اشاعت: اتوار ۹ فروری ۱۷:۳۵ جی ایم ٹی ۲۲:۳۵ پی ایس ٹی ، ۲۰۱۴ a:xəri: vəQʈ əʃa:ʈ lʈva:r no: fərvəri: səʈa:ra: sɔ pæɳʈi:s dʒi: æm Ti: ba:i:s sɔ pæɳʈi:s pi: æs Ti: do: həza:r tʃɔɖa
10	آجکل لوگ بہت سی لوک داستانوں سے دور ہیں a:dʒ kəl bohəʈ si: lo:k ɖa:sʈa:nõ: se ɖu:r hãe:

3.2.1. MOS Test This method is a grading-based procedure, where subjects are asked to rate given speech samples by asking questions such as "How do you rate the quality of the sound that you just heard?" and responses are collected on a 5-point scale, where

high score means better perceived quality. Values from 1 to 5 are presented with descriptions from "bad" to "Excellent", or similar depending on what is asked. Complete range of scales and their descriptions for the subjective attributes are presented in Table 3. The arithmetic average of scores given by all respondents represents mean opinion score (MOS) and TTS technologies are ranked accordingly. Meaningful sentences covering a wide variety of sentence structures, e.g., sentences with definitions, date, time, contact numbers, and facts & figures are selected (Table 4).

4. Experimental Setup

Total of 23 naïve subjects (3 female, 20 male) aged between 18 and 22 participated in the testing process. Out of 23 subjects 5 were blind males. Blind's subjects' response collected and interpreted separately. All of them were native Urdu speakers. None of them suffered from any hearing problems or dyslexia. All subjects participated as volunteers. Experiments were conducted under control environment where each subject was listening synthesized voices using headphones. Urdu TTS was manually optimized for pronouns and other mispronounced technical terms. The optimization included an adjustment of wrong articulated words and an improvement of pauses between sentences and paragraphs.

4.1. Procedure

The test was composed of four major sections each corresponding to one of the four tests discussed in Sec. 3. In MOS section, each subject's screen displays one sentence at a time synthesized in both HTS and unit selection voices. Voices' identity was kept hidden from the subjects in order to avoid biases. Voices were displayed with names like voice A and voice B. Subjects were asked to listen to a sentence in both voices and rate them according to their naturalness, speaking rate, and pronunciation. Each subject was given a proper explanation of these terms and meaning of rating scale used for voices' quality.

In the comprehension section, one paragraph synthesized in each voice played one by one. After listening paragraphs, respondents were asked to answer three questions taken from the paragraph. Subjects were allowed to listen to the paragraphs again if they need. The third section contains the transcription task for SUS sentences. Fourth section consists of segmental evolution using DRT and MRT test sets, where each respondent has to pick one of the two possible options against the played voice.

5. Results and Discussion

5.1. Intelligibility

5.1.1. Segmental Test This sub-section provides summarized results of segment level tests for both blind and non-blind groups. Table 5 and 6 show that at the segmental level, all features (i.e., voicing, Nasality, Aspiration and Sibilation) are understood better at word-initial place as compared to word-final place for both voices (HTS and US). Moreover, US voice performs better than HTS voice across most of the features except voicing and aspiration. Note: The metric reported in Table 5 and 6 is average percentage of correctly identified words from the pool of word pairs discussed under the heading ***Segmental Test*** in Sec 3.

Table 5 Segmental test results (in percentage) for non-blind group

	Non-Blind			
	HTS		US	
	Word Initial	Word Final	Word Initial	Word Final
Voicing	89.6	65.3	73.5	64.6
Nasality	97.2	95.1	97.9	95.1
Aspiration	95.8	51.4	84.5	62.5
Sibilation	97.9	97.9	100	99.3

Table 6 Segmental test results (in percentage) for blind group

	Blind			
	HTS		US	
	Word Initial	Word Final	Word Initial	Word Final
Voicing	72.5	67.5	75	63.75
Nasality	90	97.5	95	95
Aspiration	77.5	42.5	82.5	52.5
Sibilation	100	85	97.5	95

5.1.2. Sentence level Test Participants were allowed to listen SUS sentences maximum of two times. However, most of them played each sentence for

once only. The obtained measure of intelligibility was based on a percentage of correctly recognized words. Results for both voices (HTS and US) are summarized in the graph shown in Figure 2. According to results, intelligibility at word level is better for HTS voice as compared to US and this result is consistent among both subject groups (blind and non-blind).

5.1.3. Comprehension Test Total of three questions were asked per paragraph. Answers to the open-ended questions were scored according to a 3-point scale (0, 0.5, and 1) where 0 points are given to incorrect or unanswered responses; partially correct or too general, yet not wrong answers are given 0.5 points; and only correct and specific answers are marked with 1 point. Results are summarized in the graph shown in Figure 3. Again in this intelligibility test HTS voice's performance is slightly better than US voice.

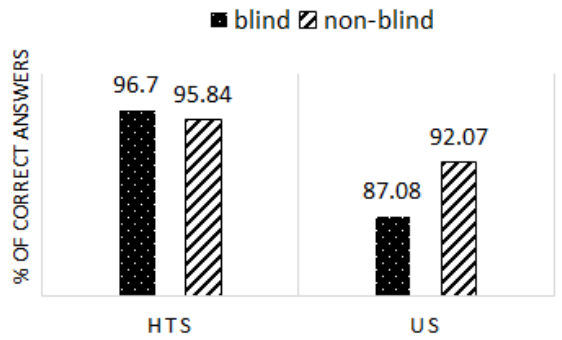

Figure 2 SUS test results.

5.2. Naturalness

For the overall quality rating, subjects were allowed to repeat sentences. Mean rating of both voices w.r.t naturalness, speaking rate and pronunciation are reported in tabular format as shown in Table 7. Entries of this table can be interpreted according to MOS rating scales described in Table 3. According to both (blind and non-blind) groups, US voice is closer to human voice as compared to HTS; US voice speaking rate is almost normal while HTS's is slightly faster than normal; and pronunciation of US is also better than that of HTS voice.

Table 7 MOS test scores.

	Naturalness		Voice Rate		Pronunciation	
	HTS	US	HTS	US	HTS	US
Non-Blind	2.89	3.11	3.28	2.81	2.94	3.32
Blind	2.78	3.22	3.49	3.08	2.94	3.54

6. Conclusion

From the results it is clear that both synthesized voices (HTS and US) are reasonably intelligible for humans and most of the respondents easily understood the synthesized sentences. Moreover, this work also pinpoints the shortcomings of Urdu TTS, e.g., from Table 5 and 6 we can see that these voices are weak in modeling aspiration feature as compared to nasality feature. Improvements of these aspects will be done in future work. When it comes to overall intelligibility, i.e., how accurately message is understood, HTS synthesis approach performs better than US, the reason is when pre-recorded speech units are concatenated in US approach they get affected by sudden changes in pitch values that create distractions for listeners.

From the naturalness point of view, however, US is preferable among both types of subjects (blind and non-blind). The reason is that in US approach speech waveform is synthesized by concatenating actual human voice units while in the case of HTS it is generated through statistically trained models. Currently the speech corpus used for training is annotated at phoneme, word, syllable, stress and break index levels only and the prosodic information, which is essential for naturalness effect in synthetic speech, still has not been incorporated. In future, the prosodic structure of Urdu language for various types of sentences and role of grammatical and prosodic information in the high-quality speech synthesis should also be investigated.

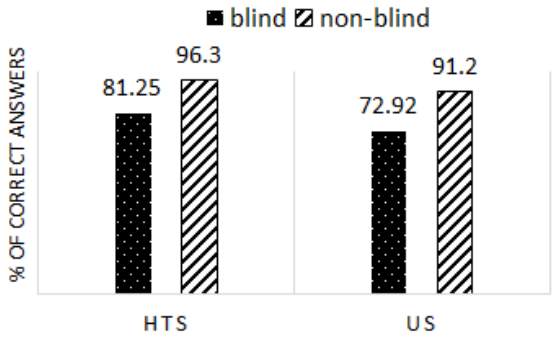

Figure 3 Comprehension test results.

7. Acknowledgment

This work has been conducted as part of the project, Enabling Information Access for Mobile based Urdu Dialogue Systems and Screen Readers, supported through a research grant from ICT RnD Fund, Pakistan.

8. References

[1] S. Arndt, J.-N. Antons, R. Gupta, R. Schleicher, S. Moller, T. H. Falk, and others, "Subjective quality ratings and physiological correlates of synthesized speech," in *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on*, 2013, pp. 152–157.

[2] T. Dutoit, *An introduction to text-to-speech synthesis*, vol. 3. Springer Science & Business Media, 1997.

[3] M. W. Macon, A. Kain, A. Cronk, H. Meyer, K. Mueller, B. Saeuberlich, and A. W. Black, "Rapid prototyping of a german tts system," 1998.

[4] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, 1996, vol. 1, pp. 373–376.

[5] A. N. S and S. T, "Article: Text to Speech Synthesis of Hindi Language using Polysyllable Units," *IJCA Proc. Natl. Conf. Power Syst. Ind. Autom.*, vol. NCPSIA 201, no. 3, pp. 15–20, Dec. 2015.

[6] G. F. S. Lewis M. Paul and C. D. F. (eds.), Eds., *Ethnologue: Languages of the World*, 19th ed. Dallas, Texas: SIL International, 2016.

[7] S. Hussain, "Phonological Processing for Urdu Text to Speech System," in *Contemporary Issues in Nepalese Linguistics (eds. Yadava, Bhattarai, Lohani, Prasain and Parajuli)*, 2005, vol. ISBN 99946.

[8] "Online Urdu TTS." 2016.

[9] U. Jekosch, "Speech quality assessment and evaluation," in *Third European Conference on Speech Communication and Technology*, 1993.

[10] A. Mariniak, "A global framework for the assessment of synthetic speech without subjects," in *Third European Conference on Speech Communication and Technology*, 1993.

[11] S. Suryawanshi, R. Itkarkar, and D. Mane, "High quality text to speech synthesizer using phonetic integration," *Int. J. Adv. Res. Electron. Commun. Eng.*, vol. 3, no. 2, pp. 77–82, 2014.

[12] J. Allen, M. S. Hunnicutt, D. H. Klatt, R. C. Armstrong, and D. B. Pisoni, *From text to speech: The MITalk system*. Cambridge University Press, 1987.

[13] S. Lemmetty, "Review of Speech Synthesis Technology." 1999.

[14] I. Rec, "P. 85. A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices," *Int. Telecommun. Union, Geneva*, 1994.

[15] I. McLoughlin, "Subjective intelligibility testing of Chinese speech," *Audio, Speech, Lang. Process. IEEE Trans.*, vol. 16, no. 1, pp. 23–33, 2008.

[16] R. A. J. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Commun.*, vol. 49, no. 4, pp. 317–330, 2007.

[17] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0.," in *SSW*, 2007, pp. 294–299.

[18] Nawaz Omer; Habib Tania, "Hidden Markov Model (HMM) based Speech Synthesis for Urdu Language," in *the Proceedings of Conference on Language and Technology 2014 (CLT14), Karachi, Pakistan*, 2014.

[19] F. Adeeba, S. Hussain, T. Habib, Ehsan-Ul-Haq, and K. S. Shahid, "Comparison of Urdu Text to Speech Synthesis using Unit Selection and HMM based Techniques," in *the Proceedings of Oriental COCOSDA Conference 2016, Bali, Indonesia (accepted)*.

[20] V. J. van Heuven, R. van Bezooijen, and others, "Quality evaluation of synthesized speech," *Speech coding Synth.*, vol. 21, pp. 707–738, 1995.

[21] T. Ojala, "Auditory quality evaluation of present Finnish text-to-speech systems," Ph.D. thesis, HELSINKI UNIVERSITY OF TECHNOLOGY, 2006.

[22] M. A. Redford and R. L. Diehl, "The relative perceptual distinctiveness of initial and final consonants in CVC syllables," *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1555–1565, 1999.

[23] M. Goldstein, "Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener," *Speech Commun.*, vol. 16, no. 3, pp. 225–244, 1995.

[24] L. C. W. Pols and others, "Multi-lingual synthesis evaluation methods," 1992.

[25] Y.-Y. Chang, "Evaluation of TTS systems in intelligibility and comprehension tasks," in *Proceedings of the 23rd Conference on Computational Linguistics and Speech Processing*, 2011, pp. 64–78.

Appendix A

Table A - 1 Phonetic characteristics at word initial and final

Phonemic features	Description	consonant pairs to be tested	Pairs with different initial consonants	Pairs with different final consonants
Voicing	voiced - unvoiced	/p/-/b/	پات/pɑ:ʈ/ ، بات/bɑ:ʈ/	باپ/bɑ:p/ ، باب/bɑ:b/
			پول/po:l/ ،بول/bo:l/	آپ/ɑ:p/ ، آب/ɑ:b/
		/ʈ/-/ɖ/	تار/ʈɑ:r/ - دار/ɖɑ:r/	رت/roʈ/ ، رد/roɖ/
			تاب/ʈɑ:b/ ، داب/ɖɑ:b/	دات/ɖɑ:ʈ/ ، داد/ɖɑ:ɖ/
		/t/-/d/	ٹاٹ/tɑ:t/ ، ڈاڈ/dɑ:t/	چاٹ/ʧɑ:t/ ، چاڈ/ʧɑ:d/
			ٹال/tɑ:l/ - ڈال/dɑ:l/	روٹ/rot/ ، روڈ/rod/
		/ʧ/-/ʤ/	چار/ʧɑ:r/ ، جار/ʤɑ:r/	سیڈج/səʤ/ ، سیٹچ/səʧ/
			چاٹ/ʧɑ: t/ ، جاٹ/ʤɑ:t/	بیڈج/bəʤ/ ، بیٹچ/bəʧ/
Nasality	nasal - oral	/m/-/b/	مل/mɪl/ ،بل/bɪl/	تام/ʈɑ:m/ ، تاب/ʈɑ:b/
			مول/mol/ ،بول/bol/	آم/ɑ:m/ - آب/ɑ:b/
		/m/-/p/	مان/mɑ:n/ ، پان/pɑ:n/	کم/kɪm/ ،کپ/kɪp/
			پول/pol/ ،مول/mol/	آم/ɑ:m/ - آپ/ɑ:p/
		/n/-/ʈ/	نان/nɑ:n/ ،تان/ʈɑ:n/	ران/rɑ:n/ ،رات/rɑ: ʈ/
			نال/nɑl/ ،تال/ʈal/	پان/pɑn/ ،پات/pɑʈ/
		/n/-/ɖ/	نام/nɑ:m/ ،دام/ɖɑ:m/	بان/bɑ:n/ ،باڈ/bɑ:ɖ/
			نال/nɑ:l/ ،دال/ɖɑ:l/	بن/bɪn/ ،بڈ/bɪɖ/

Table A - 2 Phonetic characteristics at word initial and final

Phonemic features	Description	consonant pairs to be tested	Pairs with different initial consonants	Pairs with different final consonants
Aspiration	Aspirated - Non-Aspirated	/p/-/ pʰ/	پٹ/pət/ ، پھٹ/pʰət/	باپ/bɑ:p/ ،باپھ/bɑ:pʰ/
			پل/pəl/، پھل/pʰəl/	-
		/b/-/ bʰ/	بال/bɑ:l/، بھال/bʰɑ:l/	لوب/lob/ ،لوبھ/lo bʰ/
			بات/bɑ:t/، بھات/bʰ ɑ:t/	گاب/gɑ:b/، گابھ/gɑ:bʰ/
		/ ʧ/-/ ʧʰ /	چاپ/ʧɑ:p/ ، چھاپ/ʧʰɑ:p/	مچ/moʧ/ ،مچھ/moʧʰ/
			چیڑ/ʧiɽ/ ، چھیڑ/ʧʰiɽ/	پاچ/pɑ:ʧ/،پاچھ/pɑ:ʧʰ/
		/ ɖ/-/ ɖʰ /	دانی/ɖɑ:ni:/، دھانی/ɖʰɑ:ni:/	سگنڈ/sɪŋɖ/ ، سگنڈھ/sɪŋɖʰ/
			دات/ɖɑ:ʈ/ ، دھات/ɖʰɑ:ʈ/	گد/gɪɖ/،گدھ/gɪɖʰ/
Sibilation	sibilated - unsibilated	/ ʧʰ/-/K/	چال/ʧʰɑ:l/ ، کال/kɑ:l/	بچ/bɪk/ ،بک/bɪʧʰ/
			چان/ʧʰɑ:n/ ، کان/kɑ:n/	باچ/bɑ:ʧʰ/ ،باک/bɑ:k/
		/Z/ - / ʈʰ /	زال/zɑ:l/، تھال/ʈʰ ɑ:l/	ساز/sɑ: z/ ، ساتھ/sɑ: ʈʰ/
			زم/zəm/، تھم/ʈʰəm/	رزز/rəzz/، رتھ/rəʈʰ/
		/S/-/ ʈ/	ساس/sɑ:s/، طاس/ʈɑ:s/	ساس/sɑ:s/، سات/sɑ: ʈʰ/
			سان/sɑ:n/، تان/ʈɑ:n/	راس/rɑ:s/، رات/rɑ: ʈ/
		/ʃ/-/k/	شام/ʃɑ:m/ ، کام/kɑ:m/	نوش/noʃ/ ،نوک/nok/
			شان/ʃɑ:n/، کان/kɑ:n/	شاک/ʃɑ:k/ ، شاش/ʃɑ:ʃ/