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

Authors: Kh. Shahzada Shahid, Tania Habib, Benazir Mumtaz, Farah Adeeba and Ehsan Ul Haq

Presenter: Ehsan ul Haq

Outline

- Motivation
- Background
- Urdu TTS Architecture
- Design of Subjective Test
- Experiment
- Results
- Summary

Motivation

- Text-to-Speech (TTS) system plays important role in various fields.
- Research on the development of (Text-to-Speech)TTS system for the Urdu Language, which is a national language of Pakistan and is spoken by more than 162 million people worldwide [1,] is still in its earlier stages [2.]
- To Assess the speech quality of recently developed Urdu TTS system [3.]

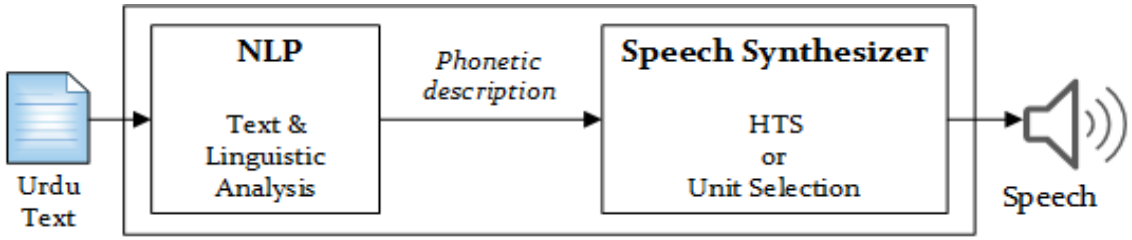
[1] G. F. S. Lewis M. Paul and C. D. F. (eds.), Eds., Ethnologue: Languages of the World, 19th ed. Dallas, Texas: SIL International,
[2] S. Hussain, "Phonological Processing for Urdu Text to Speech System," in Contemporary Issues in Nepalese Linguistics (eds. Yadava, Bhattarai, Lohani, Prasain and Parajuli), 2005, vol. ISBN 99946.
[3] "Online Urdu TTS." 2016.

Background

- Text-to-Speech(TTS) System is used for converting given input text to speech.
- **Speech Quality**
  - **Naturalness:** It means how close the synthesized speech is to the human voice.
  - **Intelligibility:** It means how clearly the synthesized speech is being understood.
- **Evaluation Methods**
  - **Subjective Evaluation:** Human users are involved.
  - **Objective Evaluation:** Different algorithms are used.
  - For measuring the naturalness and intelligibility of voice subjective methods are most commonly used.

Urdu TTS Architecture

- TTS system generally consists of two main modules, Natural Language Processor (NLP) and Speech Synthesizer.
- **NLP Module**
  - 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**
  - Speech Synthesizer then generates corresponding speech signal using the description provided by NLP.
  - Two different types of voices are used for speech synthesis
    - Hidden Markov Models Based voice(HTS)
    - Unit Selection based voice(US)



Design of Subjective Test

- The theme of this subjective test revolves around four questions:
  - Is the underlying message understandable?
    - This question addresses intelligibility.
  - Is Urdu TTS' voice closer to that of humans?
    - This question addresses naturalness.
  - Is it suitable for both the blind and non-blinds?
    - This question addresses usability.
  - Which one of the two speech synthesis approaches (HTS or US) is a better choice for Urdu TTS?
    - This question addresses quality comparison of HTS and US.

Design of Subjective Test

- **Intelligibility Tests**
  - Segmental Test
  - Sentence level Test
  - Comprehension Test
- **Naturalness Test**
  - Mean Opinion Score (MOS) Test

Design of Subjective Test

- **Segmental Test**
  - Smallest speech units, like phonemes
  - Consonants, being difficult to be recognized
  - **Diagnostic Rhyme Test (DRT)**
    - Word pairs which differ by a single acoustic feature in the initial consonant
  - **Modified Rhyme Test (MRT)**
    - Word pairs which differ by a single acoustic feature in the final consonant
  - **Segmental Test Design**
  - 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).

Design of Subjective Test

Phonemic features	Description	Pairs with different initial consonants	Pairs with different final consonants
Voicing	voiced - unvoiced	پات/pa:ʈ/ ، بات/ba:ʈ	باپ/ba:p/ ، باب/ba:b
Nasality	nasal - oral	مول/mol/ ، بول/bol	تام /ʈa:m/ ، تاب/ʈa:b
Aspiration	Aspirated – Non-Aspirated	بال/ba:l/ ، بھال/bʰa:l	باپ /ba:p/، باپھ /ba:pʰ
Sibilation	sibilated - unsibilated	چھال /ʧʰa:l/ ، کال/ka:l	ساز/sa:z/ ، ساتھ /sa: ʈʰ/

Design of Subjective Test

- These rhyme words are tested through following carrier sentence:
  - (1) ---- کیا آپ اردو لغت سے لفظ - - - - کا مطلب بتا سکتے ہیں؟
  - kæɑ: ɑ:p ʊrd̪u lʊɣət̪ se ləfz ----- kɑ: mət̪ləb bət̪ɑ: səkt̪e hæ:
  - What- kæɑ: you- ɑ:p Urdu- ʊrd̪u dictionary- lʊɣət̪ case marker-se word- ləfz ---- case marker- kɑ: meaning- mət̪ləb tell- bət̪ɑ: can- səkt̪e tense aux- hæ:
  - “Can you inform me the meaning of --- word from the dictionary?”

Design of Subjective Test

- **Sentence Level Test**
  - Intelligibility at sentence level is usually evaluated through transcription task of Semantically Unpredictable Sentences (SUS).
  - SUS sentences have grammatically correct syntax, however, they are unpredictable semantically.
    - Example of SUS
      - میز تیز رفتاری سے بیٹھ گیا
      - mez tezrɑːftɑːriː se bæt gəɑː
      - Table- mez speedily- tezrɑːftɑːri case marker-se sat: bæt tense-gəɑ
      - “Table sat down speedily”

Design of Subjective Test

- **Comprehension Test**
  - Correct reception of the underlying message rather than accuracy of individual sounds
  - Paragraph is presented, followed by a questionnaire about the content
  - Hundred percent segmental intelligibility is not needed to answer
  - Less familiar topics are selected

Design of Subjective Test

- **Naturalness Test**
  - **MOS Test**
    - Rating scale from 1 (bad) to 5 (excellent)
    - Questions:

1. How do you rate the quality of the sound?
2. What was the average speed of delivery?
3. Did you notice any anomalies in pronunciation?

Design of Subjective Test

- **Naturalness Test**
  - **MOS Test**
    - Meaningful sentences
    - Wide variety of sentence structures, e.g., sentences with definitions, date, time, contact numbers, and facts & figures are selected

اس دوران ترکی اور ایران کے مابین مجموعی تجارت کا حجم ۲۱۔۸ بلین ڈالر رہا

Is ɖo:ra:n ʈʊrki: ɔr æra:n ke ma:bæn mədʒmu:i: ʈədʒarəʈ ka hʊdʒəm a:Th se Ikki:s bIljən Dɔlər rəha:

Experiment

- **Setup**
  - 23 naïve subjects (3 female, 20 male)
  - Aged between 18 and 22
  - Out of 23 subjects 5 were blind males
  - All of them were native Urdu speakers
  - Experiments were conducted under control environment

Experiment

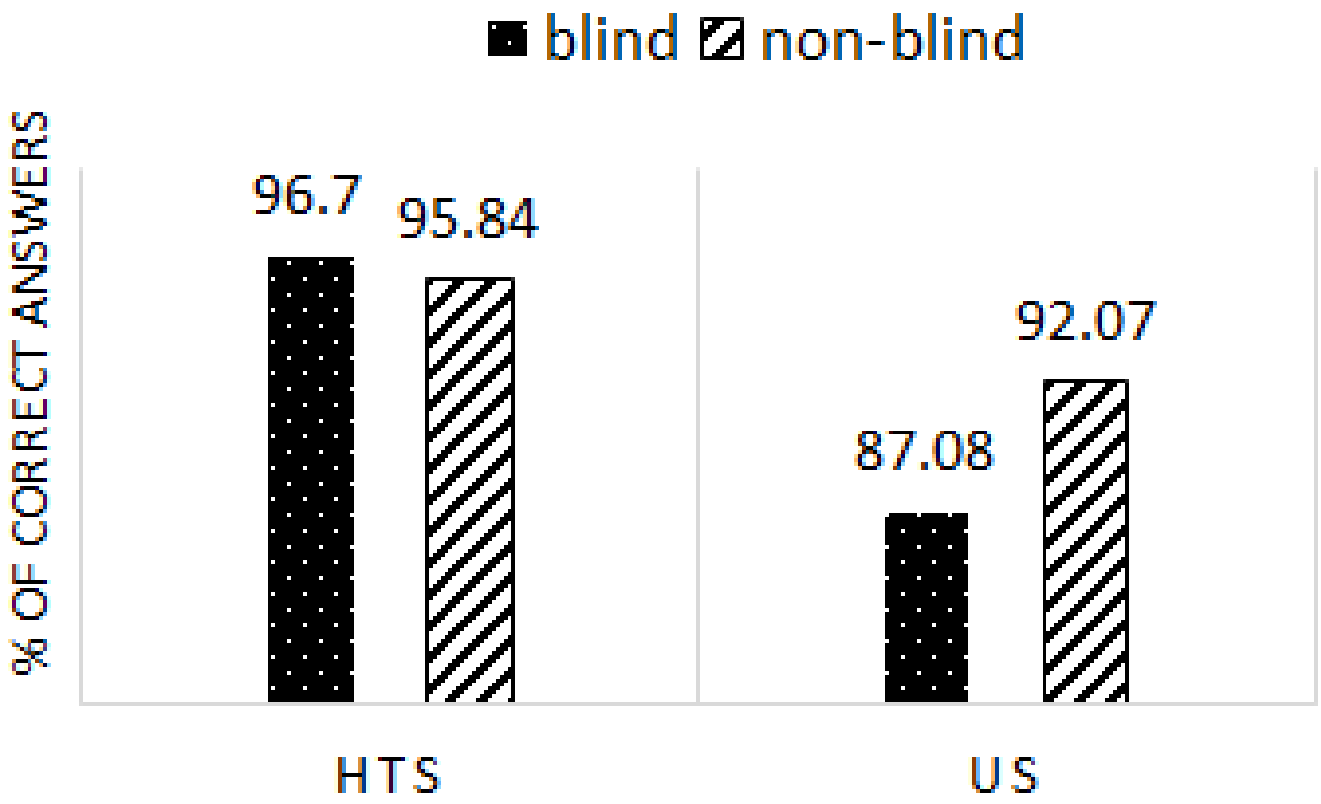
- **Procedure**
  - Test divided in four sections
  - Each sentenced is played in both voices (HTS & US)
  - Voices' identity was kept hidden
  - Subjects have to rate voices according to naturalness, speaking rate, and pronunciation
  - **Comprehension Test:** Subjects were allowed to listen twice if needed.
  - **Sentence level Test:** Subjects have to transcribe SUS sentences
  - **Segmental Test:** Subjects have to pick one of the two possible rhyme words against the played voice.

Results and Discussion

	Non-Blind				Blind			
	HTS		US		HTS		US	
	Word Initial	Word Final	Word Initial	Word Final	Word Initial	Word Final	Word Initial	Word Final
Voicing	89.6	65.3	73.5	64.6	72.5	67.5	75	63.75
Nasality	97.2	95.1	97.9	95.1	90	97.5	95	95
Aspiration	95.8	51.4	84.5	62.5	77.5	42.5	82.5	52.5
Sibilation	97.9	97.9	100	99.3	100	85	97.5	95

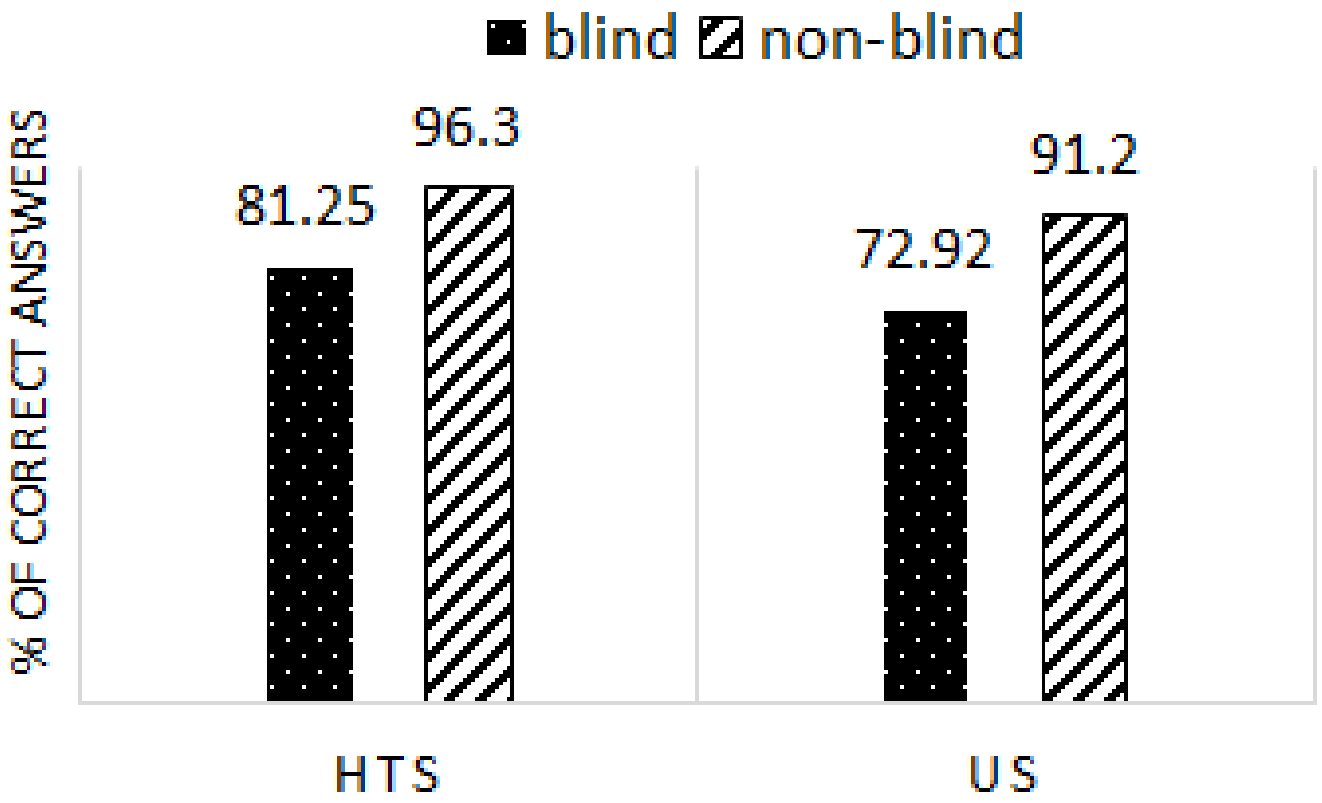
Segmental Test Results

Results



SUS Test Results

Results



Comprehension Test Results

Results

	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

MOS Test Results

Conclusion

- Both synthesized voices (HTS and US) are reasonably intelligible
- Comparatively HTS voice is better understood than US voice
- Naturalness point of view, US is preferable among both types of subjects (blind and non-blind)
- It pinpoints the shortcomings of Urdu TTS, e.g., weak aspiration model

Summary

- From the naturalness point of view, however, US is preferable among both types of subjects (blind and non-blind).
- 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.

Questions?