2016 Conference of The Oriental Chapter of International Committee
for Coordination and Standardization of Speech Databases and Assessment Technique (O-COCOSDA)
26-28 October 2016, Bali, Indonesia

Comparison of Urdu Text to Speech Synthesis using Unit Selection and HMM based Techniques

Farah Adeeba ,Tania Habib
Department of Computer Science & Engineering
University of Engineering & Technology
Lahore, Pakistan
fadeeba@gmail.com

Sarmad Hussain, Ehsan-ul-haq, Kh. Shahzada Shahid
Centre for Language Engineering
KICS, UET
Lahore, Pakistan
firstname.lastname@kics.edu.pk

***Abstract*—This paper presents the development of Urdu text to speech system using Festival framework. Unit selection based, and HMM-based speech synthesizer are developed using 10 hours manually annotated speech data for Urdu text to speech synthesis. The architecture of the Urdu text to speech system is also discussed. The objective assessment of the synthesized speech is evaluated by using automatic speech recognition system. The results show that the speech produced by HMM-based synthesizer has the highest recognition accuracy.**

Keywords—Urdu speech synthesis; speech corpus; Unit-selection based speech synthesis; HMM based speech synthesis;speech recognition;Festival

I. INTRODUCTION

Urdu belongs to Indo-Aryan language family and written in Persio-Arabic script using Nastaliq style. Urdu is spoken by more than 104 million people[1] and national language of Pakistan. With the tremendous increase in digital data, access to information is becoming important for today's age. However, due to low literacy rate of this population, access to modern information is a significant challenge. According to the Pakistan Bureau of Statistics, the literacy rate for Pakistan is 58% [2], resulting into barrier of information access for about half of the population. There is need to develop a mechanism, to enable information access orally i.e. Urdu Text-to-Speech system. In addition, Urdu TTS integrated in Urdu screen reader will resolve a multitude of access problems, for the visually impaired community of Pakistan.

A Text to Speech (TTS) system transforms given text into speech. Two of the well-known and widely used speech synthesis techniques are unit selection and Hidden Markov Model (HMM) based speech synthesis. In unit selection speech synthesis, suitable pre-recorded units are concatenated to obtain the speech corresponding to the given text. The units (word or subword) with optimal concatenation and joining costs are selected for concatenation. The naturalness of the synthesized speech depends upon the size , context of the speech unit, and number of concatenation points i.e. naturalness is preserved with the selection of longer units and less number of concatenation points. Ideally for more natural speech, each speech unit should be present multiple times in all possible context in the speech database.

Development of an appropriate speech corpus for unit selection based speech synthesis is laborious and time consuming. Hidden Markov Model(HMM) based speech synthesis, can be used to minimize the barrier of such speech corpus. HMM-based synthesis is a statistical parametric based speech synthesis technique. Spectral and excitation features from speech corpus are extracted to form a parametric representation of speech [4] Given text is transformed into speech by using this parametric representation. The main advantage of this parametric representation is that only statistics are stored rather than original speech waveforms, resulting into small footprint. Previously, Nawaz and Habib [5] developed an HMM-based speech synthesizer for the Urdu language, using 30 minutes of speech data. During the subjective testing of the system, 92.5% words were correctly recognized. This HMM-based synthesizer was trained using only 30 minutes speech data and it was not integrated in text to speech system. In the current work, 10 hours of manually annotated speech data is used for the development of HMM-based Urdu speech synthesizer. Furthermore, a unit selection-based speech synthesizer is also developed using the same data, and quality of the synthesized speech is also evaluated. Automatic speech recognition system is utilized for the objective intelligibility assessment of the generated speech. While, Urdu speech recognizer is also developed for the evaluation.

This paper is structured as follows: Section II discuss the speech corpus used for TTS development. Section III contains the discussion of the development of Urdu TTS system, modification of Festival for Urdu, and speech synthesis engines for speech generation. Section IV is dedicated to system evaluation by employing the objective assessment of the synthesized speech. The current study with a discussion and its proposed future applications are concluded in Section VI.

978-1-5090-3516-8/16/$31.00 ©2016 IEEE

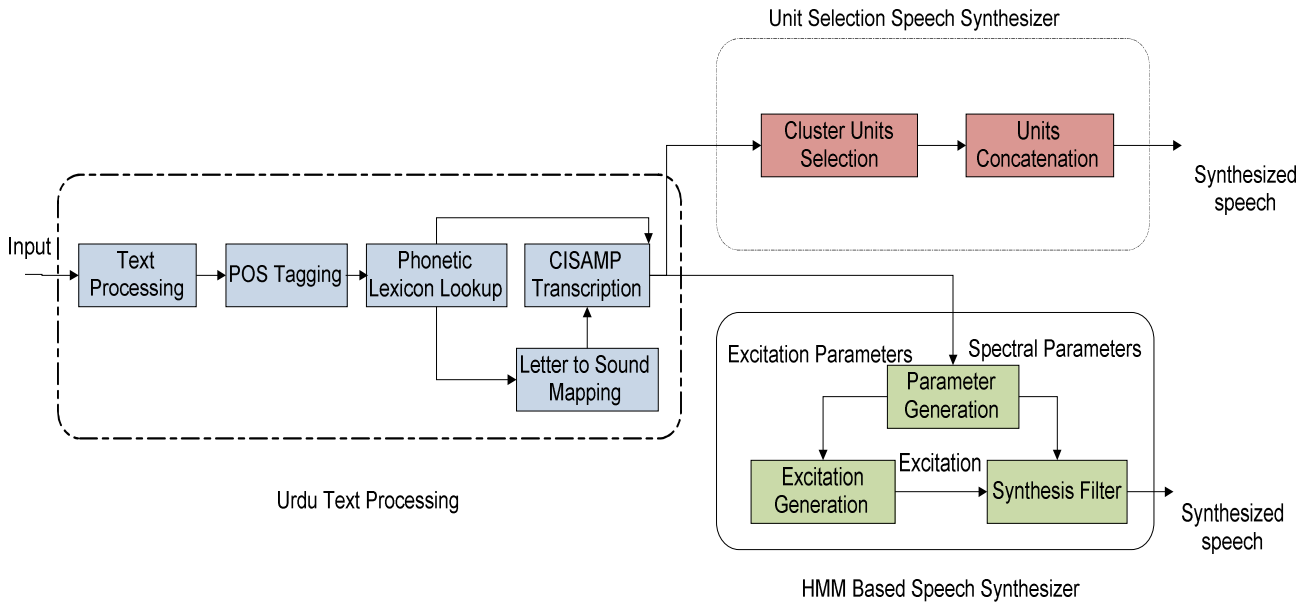

Figure 1. Urdu TTS architecture

II. URDU SPEECH CORPUS

Designing a speech corpus is one of the core issues in the development of high quality text to speech synthesis systems[6]. For the development of Urdu speech synthesizers, a text comprising of 8,081 sentences and 130,163 syllables is optimally extracted from three different Urdu text corpora [7] using greedy algorithm [8] to include the maximum phonetic coverage of Urdu bi-phones and tri-phones. Then, considering various factors such as dialect, voice quality, natural speaking rate, natural pitch of the speaker and intonational range of speaker, a professional female speaker was hired to record these sentences. She carried out the recordings for this corpus in multiple sessions of 25 to 30 minutes in an anechoic chamber at the sampling rate of 48kHz.

The speech corpus is being annotated manually at segment and word levels [9] and automatically at syllable, stress and phrase levels using Urdu syllabification algorithm, acoustic stress marking algorithm [10] and Urdu phrase identification algorithm, respectively. Furthermore, the quality of this corpus has been assessed using multiple quality assessment utilities such as phoneme label comparison, phoneme boundary comparison, word pronunciation comparison and syllable template matching analysis [9]. Based on testing results, only error free files are selected.

A total of 10 hours speech data is used to build open-domain Urdu TTS, and comprised of 8081 sentences, 82,049 words, and 130,163 syllables.

III. URDU TTS DEVELOPMENT

The TTS for Urdu is developed by using an open source software Festival [11]. Given Urdu text is transformed into speech using by executing several modules as shown in Fig. 1. In first step, text analysis module process input text and change all non-standard words into standard words, and grapheme-to-phoneme conversion is performed on the basis of information in the written text. Text analysis method is discussed in detail in subsequent section. In prosody modeling, phase intonation and duration models are applied. Phonemes can be synthesized using any of two available synthesis engines i.e. Unit selection based speech synthesizer or HMM-based speech synthesizer. Different models of these speech synthesizers are built using a phonetically balanced Urdu speech corpus.

A. Urdu Text Processing

The text is first transformed into its corresponding pronunciation to generate speech from Urdu text. Text analysis module of festival is modified to accomplish this task. Updated text analysis module takes Urdu text as input and transform into corresponding CISAMPA [12] transcription. Fig. 2 demonstrates a pictorial view of the components involved in the text analysis.

The first component i.e. Text Processing which pre-process and clean the input text and output normalized text. This normalized text is forwarded to the Part of Speech(POS) tagger, and POS tags are assigned. These tags help to resolve the ambiguity of multiple transcription of word during lexicon lookup. For example, Urdu word چور(ʧu:r/ʧo:r) should be spoken as T_SU_UR(exhaust) if

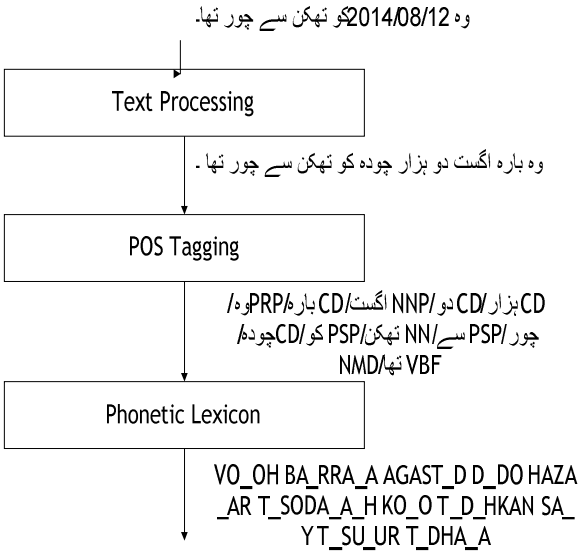

Figure 2. Text analysis for Urdu

used as noun modifier rather than T_SO_OR(thief). This phonological processing of string is done using pronunciation lexicon lookup. If a word is not found in lexicon, Urdu letter to sound mapping rules are used. Details of aforementioned modules are present in subsequent sections.

1) Text Processing: Text processing modules takes text as input and change it into normalized text. Text processing module performs following steps (1) sentence segmentation, (2) word boundary marking, (3) classification of digits, dates, and symbols on the basis of context , and (4) text generation by using the earlier study[13] . Sentences are segmented on the basis of punctuation mark (full stop, question mark),and line break. Each sentence is further tokenized at word level by using the space and punctuation marks as word boundary. Firstly, semantic tagger analyze each token and converts each token into text. Because, in the written form of a language, numbers, dates, symbols, and abbreviations can exist, that are pronounced differently when used in different context. Hence, a semantic tagger is responsible for identification of the token type in the context.

2) Part Of Speech(POS) Tagging: This modules takes normalized text as input and assign POS tags to the words by using the trigram language model and POS lexicon. A trigram language model is generated by part of speech tags and the probabilities of words given their part of speech tag and POS lexicon contains these probabilities. To enable the festival built in POS tagger for Urdu text, Urdu trigram language model and POS lexicon has been used. The data for the generation of trigram model and POS lexicon has been extracted from a manual POS tagged corpus consisted of one hundred thousand Urdu words [14]. The trigram language model was generated using 'Good-Turing' method as smoothening technique. Words that are not present in the POS lexicon are marked as out-of vocabulary. Four basic POS tags are assigned to the out-of vocabulary words: NN (Noun), NMD(Nominal Modifier), PR(Pronoun), and VB(Verb). Trigram language model is used to select among the above mentioned tags for the out-of vocabulary words.

3) Phonetic Lexicon: Pronunciation lexicon is an important part of natural language processing as it is used in assigning phonemic transcription to the words. The developed pronunciation lexicon for Urdu consist of three parts; one is the Urdu word, second part is its POS tag and the third field is its phonemic transcription in CISAMPA. Words have been extracted from different sources and the IPA symbols are used for transcription of these words. A utility has been developed to convert IPA transcription into CISAMPA. The lexicon is passed through the utility to verify that all the symbols are CISAMPA compatible. This manually generated pronunciation lexicon is then automatically transformed to a specific lexicon format defined by the festival system. Currently, the pronunciation lexicon consists of 70,597 Urdu words only.

*4) Letter to Sound Mapping:*In TTS system, a lexicon is required to give pronunciations of words, though a lexicon will never be sufficient enough to capture all words of a language. So, there is requirement for a mechanism for giving pronunciation of words not found in the lexicon. To handle these words rule based letter to sound conversion is implemented [15].The system assigns CISAMPA based transcriptions to the unknown words for synthesizing them. Generated transcription is automatically divided into syllables using the algorithm [16]. After generating the syllables, the algorithm for assigning lexical stress to the syllables has also been added [16].

Generated trasncription of given input text is routed to the prosody modeling module that is later on forwarded to the speech synthesis engine.

B. Synthesis Engines

In Urdu text to speech system two different synthesis engines have been taken into consideration for speech generation: Unit selection based speech synthesizer, and HMM based speech synthesizer.

1) Unit Selection based Speech Synthesizer: Clunits method [17] is employed for unit selection based speech synthesizer. During the training phase, units (phonemes) from the speech database are automatically clustered on the basis of acoustic, prosodic and phonetic context. During synthesis, appropriate units with minimum joining cost are selected using Viterbi search.

The training of synthesizer is carried out by using ten hours annotated speech data. During the training phase,

contextual and acoustic features are extract from the database. Speech units in database are clustered on the basis of aforementioned features, by using the speech tools library[1] . To classify the units into clusters, Urdu phonemes specific questions are defined. The default cluster size of 20 and prune limit of 40 is used .

During synthesis, the synthesizers utilize the information related to units, and pick the most appropriate unit based on the target cost and the concatenation cost. On the basis of target cost, best match units in the database are identified, whereas the joining cost choose the units that can be concatenated smoothly. The best optimal selected units are concatenated and speech is synthesized.

*2) HMM-based Speech Synthesizer:*In addition to the unit selection speech synthesis, HMM based speech synthesis is also enabled in Urdu TTS system. HMM-based speech synthesis is statistical parametric based synthesis technique, which uses the recorded speech for training the synthesis parameters. Its footprint size is less than unit selection because once the synthesis parameters are trained HMM,make use of these parameters for synthesis.

During the training phase excitation and spectral parameters are extracted as feature vectors. These feature vectors are modeled by the HMM framework. In the synthesis part, it takes context dependent sequence of phones as input and generates excitation and spectral parameters using parameter generation algorithm. These generated parameters are fed to the synthesis filter to reconstruct the output waveform.

HTS [18] toolkit is used for the development of statistical models for speech synthesis. Ten hours speech data is used for models training. In the training phase, wave files along the utterance structure files (derived through festival) are used for parameter extraction and training is carried out by following the process previously discussed [19]. At the synthesis stage, waveforms are generated by selecting the appropriate trained models suggested by the utterance structure generated using festival. These models along with excitation parameters are subjected to the synthesis filter and final waveform is generated.

IV. SYSTEM EVALUATION

A text-to speech system is evaluated on the intelligibility and naturalness of synthesized speech. To ensure good quality of synthesized speech, the system is evaluated during the development process. There are different ways for the assessment of Text-to-Speech system. TTS system can be evaluated by using human perception. In subjective testing human subjects are involved for the assessment of speech. The intelligibility of the synthesized speech is evaluated by different tests and naturalness is evaluated. Native language speakers or phoneticians are required for this assessment. Evaluation of a TTS system can be done automatically without human involvement, this process of evaluation is known as objective testing.

Current Urdu TTS is also subjectively evaluated by humans and Mean Opinion Score(MOS) is computed by 23 subjects, subjective tests and their results are discussed in [20]. During the subjective evaluation of the system it is observed that speech generated through HMM-based synthesize is more intelligible as compared to the speech generated through unit selection based synthesizer. Whereas unit selection based synthesis is more natural as compared to the HMM-based synthesis.

In this paper, incremental evaluation of the synthesized speech is carried out, to judge the affect of training dataset on speech quality. The synthesized speech is objectively assessed with the help of Urdu Automatic Speech Recognition(ASR) system. The ASR system is trained using the 10 hours of original human speech consisting of 8081 utterances and vocabulary size of 9267 words. Urdu speech recognizer is developed by using Sphinx[21], an open source software. Three state HMM with the mixture of continuous density output was used for training of the ASR system, and described by diagonal covariance matrices. Moreover, ten passes of the Baum-Welch algorithm are used for the training of models. MFCC features are used for training and testing of the ASR system. MFCC vectors are calculated at every 10ms using window size 25ms. ASR is decoded by back-off trigram language model with language weight equal to 17. Trained ASR system is tested before using it for objective evaluation , accuracy of the ASR on the training data is measured i.e. around 97.76%.

For objective evaluation of the TTS systems, 607 phonetically balances sentences are extracted from the 1M corpus [7], 37M corpus, and online news. In addition to phonetic coverage, much attention has been paid to cover three different sentence types: declarative, interrogative and exclamatory. Handcrafted semantically unpredictable sentences are also included for the intelligibility test. The evaluation has been performed on speech generated through unit selection and HMM based synthesizers. Speech recognition accuracy for Unit Selection and HMM based generated speech is shown in Fig. 3. It shows gradual improvements in ASR accuracy from 1-10 hours for HMM based synthesis (HTS). ASR accuracy is also improving for Unit Selection(US) based synthesis system.

However, there is a drop in accuracy for US system corresponding to the point where it was trained on 7-hours speech data. ASR error analysis shows that this drop in accuracy is mainly due to the misrecognition of silence/pause in the synthesized speech generated using 7-hour data.

This objective evaluation of TTS system shows that intelligibility of synthesized speech generated through HMM based synthesis system is better than generated through unit selection based system.

[1]http://www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0/

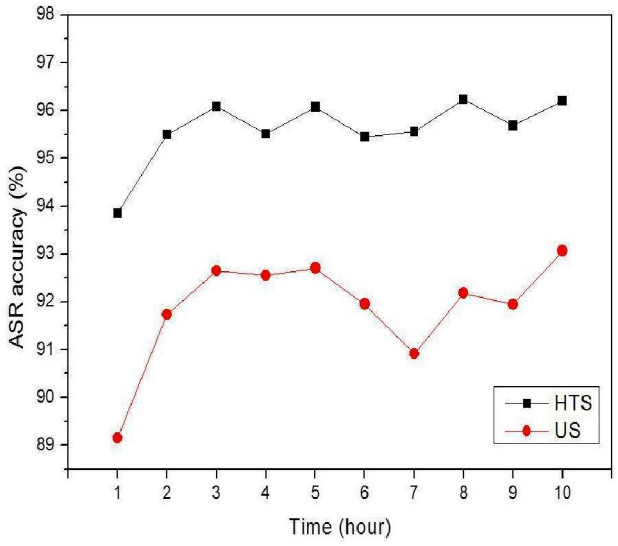

Figure 3. ASR Accuracy of synthesized speech

V. CONCLUSION

In this study, we discussed the development of open domain Urdu text to speech system using Festival. For Urdu text processing, modules of Festival are updated. In addition, Urdu phonetic lexicon consists of 70,597 words is also developed. A total of 10 hours speech data is used for the development of Unit selection based and HMM based speech synthesizers. The manual annotation of the speech signal at phoneme, word and syllable level is carried out.

Urdu synthesized speech generated using the HMM based and unit selection based synthesis engine is evaluated objectively by automatic speech recognition. The objective assessment results exhibited improvement of accuracy in speech recognition of HMM over unit selection speech of 4%. This shows that HMM generated speech is more intelligible than the speech generated through unit selection based synthesizer.

ACKNOWLEDGMENT

This work has been conducted under the project titled "Enabling Information Access through Mobile Based Dialog Systems and Screen Readers for Urdu". The authors acknowledge the financial support provided by National ICTR&D Fund, Ministry of Information Technology, Islamabad, Pakistan.

The presented Urdu text to speech system is available at: http://182.180.102.251:8080/UrduTTS/

REFERENCES

[1] (2016). *Urdu*. Available: http://www.omniglot.com/writing/urdu.htm

[2] "Pakistan Social And Living Standards Measurement Survey (PSLM) 2013-14 National / Provincial," 2014.

[3] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," presented at the Proceedings of the Acoustics, Speech, and Signal Processing, 1996. on Conference Proceedings., 1996 IEEE International Conference - Volume 01, 1996.

[4] A. W. Black, H. Zen, and K. Tokuda, "Statistical Parametric Speech Synthesis," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, 2007, pp. IV-1229-IV-1232.

[5] O. Nawaz and T. Habib, "Hidden Markov Model (HMM) based speech synthesis for Urdu language," presented at the Conference on Language and Technology, 2014.

[6] B. Bozkurt, O. Ozturk, and T. Dutoit, "Text Design for TTS Speech Corpus Building Using a Modified Greedy Selection," in *INTERSPEECH*, Geneva, 2003.

[7] S. Urooj, S. Hussain, F. Adeeba, F. Jabeen, and R. Parveen, "CLE Urdu digest corpus," in *Conference on Language and Technology(CLT)*, Lahore, Pakistan, 2012, pp. 47-53.

[8] W. Habib, R. Hijab, S. Hussain, and F. Adeeba, "Design of speech corpus for open domain Urdu text to speech system using greedy algorithm," in *Conference on Language and Technology(CLT)*, Karachi, Pakistan, 2014.

[9] B. Mumtaz, A. Hussain, S. Hussain, A. Mahmood, R. Bhatti, M. Farooq*, et al.*, "Multitier annotation of Urdu speech corpus," in *Conference on Language and Technology (CLE)*, Karachi, Pakistan, 2014.

[10] B. Mumtaz, S. Urooj, S. Hussain, and W. Habib, "Stress annotated Urdu speech corpus to build female voice for TTS," in *Oriental COCOSDA/CASLRE Conference*, Shanghai, 2015.

[11] (2015, December). *Festival*. Available: http://www.festvox.org/packed/festival/2.1/

[12] (2015, December). *Center for Language Engineering Web site*. Available: http://www.cle.org.pk/resources/CISAMPA.pdf

[13] R. H. Basit and S. Hussain, "Text processing for Urdu text," in *Conference on Language and Technology (CLT)*, Karachi, Pakistan, 2014.

[14] T. Ahmed, S. Urooj, S. Hussain, A. Mustafa, R. Parveen, F. Adeeba*, et al.*, "The CLE Urdu POS tagset," in *Language Resources and Evaluation Conference(LREC)*, Reykjavik, Iceland, 2014.

[15] H. Sarmad, "Letter-to-sound conversion for Urdu text-to-speech system," in *Workshop on Computational Approaches to Arabic Script*, Geneva, Switzerland, 2004.

[16] H. Sarmad, "Phonological processing for Urdu text to speech system," ISBN 99946-57-69-0, 2005.

[17] A. W. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *Eurospeech*, 1997, pp. 601-607.

[18] T. N. Heiga Zen, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan W Black, Keiichi Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Sixth ISCA Workshop on Speech Synthesis(SSW6)*, Bonn, Germany, 2007, pp. 294-299.

[19] O. Nawaz and T. Habib, "Hidden Markov Model (HMM) based speech synthesis for Urdu language," in *Conference on Language & Technology(CLT)*, Karachi, Pakistan, 2014.

[20] T. H. Kh.Shahzada Shahid, Benazir Mumtaz, Farah Adeeba and Ehsan Ul Haq, "Subjective Testing of Urdu Text-to-Speech (TTS) System," in *Conference on Language & Technology*, Lahore, 2016.

[21] K.-F. Lee, "Automatic speech recognition: The development of the Sphinx recognition system," *Springer Science & Business Media,* vol. 62, 1989.