A Flexible Architecture for Urdu Phonemes-Based Concatenative Speech Synthesis

MUHAMMAD RIZWAN AHMAD*, AND MUHAMMAD JUNAID ARSHAD*

ABSTRACT

TTS (Text-to-Speech) synthesis systems are extensively used across the world to intensify the accessibility of information and to make it possible for the handicapped to be involved directly with computers to get the benefits from this high technology revolution. Various TTS synthesis techniques have been used with their own advantages and limitations. There is not a concatenative synthesis strategy based architecture for Urdu TTS synthesis system for handling the homographs and to avoid the unnatural robot sounding speech produced due the use of di-phones. In this paper, we propose a flexible architecture for Urdu TTS synthesis system that uses concatenative synthesis strategy because this approach has the ability to join together the small corpus of speech to generate natural and intelligible sound. The main aspiration of this research is to disambiguate the homographs in the Urdu language and to avoid the unnatural robot sounding speech. Finally, the effectiveness of the system is tested in terms of intelligibility and acceptability on word and sentence level. The intelligibility rate is near to 80% and 65% while acceptability rate for the naturalness is 95% (75% natural, 20% acceptable).

Key Words: Articulatory, Text-to-Speech, Formant, Concatenative, Natural Language Processing, Waveforms, Speech Units, Phonemes, Speech Synthesis.

1.   INTRODUCTION

A system that converts text to voice is called speech synthesizer. The main objective of TTS synthesis system is to provide textual information through voice messages by helping machines to convert arbitrary texts to speech. In communications, key TTS applications include text-based messaging with voice rendering such as vocalizing fax, email and daily journals for handicapped, as well as text/visual information voice rendering for web pages. In some cases, these systems can also provide voice output for the information stored in the database. This information can include contact numbers, navigation information and addresses etc. The TTS system is aimed to convert all printed texts into speech. There are different types of TTS synthesis methods i.e. Formant Synthesis, Concatenative Synthesis and Articulatory Synthesis. Most powerful approach to synthesis is the Formant Synthesis. Jilka, et. al. [1] that described these synthesis systems are based on rules that control the source-filter model which is highly simplified. It assumes that source and filters are completely independent. Control parameters such as bandwidths and formant frequencies, determine

*Department of Computer Science, University of Engineering& Technology, Lahore.

Mehran University Research Journal of Engineering & Technology, Volume 35, No. 3, July, 2016 [p-ISSN: 0254-7821, e-ISSN: 2413-7219]

373

the filter. This approach is extremely flexible and these systems generate highly intelligible speech. Although the speech generated by formant synthesis systems is not natural sounding, but low memory footprint is the main advantage of such systems. Formant synthesizers are also called rule based synthesizers and are mostly used by phonologists and phoneticians because they constitute a cognitive phonation mechanism.

According to Mohan and Schroeter [2], articulatory synthesis systems are based on speech production biomechanical models. These include the models for moving the vocal tract and for generating the aspiration and periodic excitation. Ideally, such synthesizers would be controlled by the simulated muscular actions of articulators such as glottis, the tongue and the lips, by solving time dependent differential equations for computing the output of the synthetic speech [3-4]. Although such synthesis systems involve high computational requirements, but the speech generated by such systems is not natural-sounding and fluent.

Concatenative synthesis systems are based on recorded speech units [5]. These speech units include Urdu language phones, di-phones which contain phones with consonant-consonant, consonant-vowel and consonant-glide format. The stored speech snippets with different sizes can affect the quality, the intelligibility and the speed of the synthesized voice. These systems select speech units from voice database, concatenate these units and then after necessary decoding, outputs the resultant speech signal. The speech generated by these systems are natural sounding because of the recorded speech snippets. However, differences between the automatic segmentation of the waveforms and natural variations in the speech can cause audible glitches in output. Table 1 shows the summary of weaknesses of different TTS synthesis techniques as per [1, 2, 5]:

In this paper, we propose a flexible architecture for Urdu TTS synthesis system that uses concatenative synthesis strategy because this approach has the ability to join

together the small corpus of speech to generate natural and intelligible sound. The main aspiration of this research is to disambiguate the homographs in the Urdu language and to avoid the unnatural robot sounding speech produced due the use of di-phones. Homographs are the words that are spelled the same but pronounced differently and having different meanings.

The organization of this paper is as follows: Section 2 presents Urdu language orthography which includes alphabets, optional vocalic contents in the Urdu language, variations in Unicode, syllabification general principles and then finally the Urdu text tokenization; Section 3 presents the literature review; Section 4 presents the methodology; Section 5 describes the results and discussions of the methodology and section 6 concludes the paper.

2. URDU LANGUAGE ORTHOGRAPHY

2.1 Alphabets of Urdu Language

Urdu alphabets use vowels, numerals, punctuations, consonant letters, superscript signs and diacritic marks. Total alphabets are 39 [5]. There are more than one form for the graphical representation of each of the alphabet based on the position and its context in the word. Some consonants in the Urdu language have same phonetic sounds. These are called homonyms. Following are the examples of such letters that have similar phonetic sound: "SE" (ث, Arabic letter "THEH"), "SEEN" (س, Arabic letter "SEEN") and "SAD" (ص, Arabic letter "SAD"). Some characters do not join at both ends because they do not have middle shape. For example و,ا,د etc. The vowels in

TABLE 1. SUMMARY OF WEAKNESSES OF TTS SYNTHESIS TECHNIQUES

TTS Synthesis Techniques	Weaknesses
Formant Synthesis	Generated speech is not natural-sounding.
Articulatory Synthesis	Generated speech is not natural-sounding and fluent.
Concatenative Synthesis	Intelligibility can be disturbed due to the audible glitches in output.

Mehran University Research Journal of Engineering & Technology, Volume 35, No. 3, July, 2016 [p-ISSN: 0254-7821, e-ISSN: 2413-7219]

374

the Urdu language are "ALIF" (ا), "WAW" (و), "HAMZA" (ء) and "YEH" (ی). Ijazet. al. [8] described that Urdu language uses Arabic script that is in the Nastaleeq style, written from right to left. There are some distinct characteristics of this script. For example, Perso-Arabic script join letters with each other and therefore, there are different forms of the letters as per position in ligature [6,7]. Urdu language comprises the alphabets as shown in Fig. 1 [8]:

Diacritic marks are used to emphasize a particular sound or to specify a vowel. These diacritic marks are taken from Arabic script and appear above or below the character. Most commonly used diacritical marks are "PESH","TASHDEED","ZABAR" and "ZER"[8]. Fig. 2 shows the diacritics in Urdu language.

Digits, in Urdu language, are represented from 0 to 9 as ۰۱۲۳۴۵۶۷۸۹ [8]. Urdu language has its own numerals. These numerals are written from left to right. "ASHARYA" (.) is a decimal separator in Urdu numerals. Some punctuation marks in Urdu, borrows from English, have been modified to follow the right to left script behavior [8]. For example, question mark is written flipped horizontally like the English language. Exclamation, Division sign and Sentence Dash are also the Urdu punctuation marks. **Fig. 3** shows the special symbols that occur in Urdu text [8].
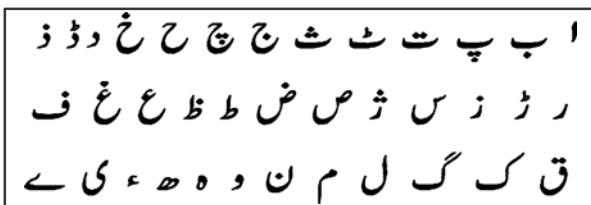

FIG. 1. URDU ALPHABETS

Diacritic	Symbol (Unicode)	Example
Zabar (Fatah)	(E064) ــَـ	اَب
Zair (Kasra)	(0650) ــِـ	شِعر
Paish (Zamma)	(F064) ــُـ	خُود
Tashdeed (Shad)	(0651) ــّـ	سَتّر
Tanween	(B064) ـًـ	یقیناً

FIG. 2. URDU DIACRITICS

2.2 Optional Vocalic Content

In Urdu language, diacritics are optional and letters are normally used to write Urdu. However, the consonantal contents of the string are represented by the letters. In some cases, vocalic contents are also represented by the letters. Diacritics with letters may be used to specify the vocalic content optionally or completely discussed in [9]. Every word consists of a diacritics set, however, word can also be written with or without diacritics. It is, therefore, permitted to omit the diacritics completely or partially. In some cases, if the diacritics are removed, two dissimilar words, having dissimilar pronunciations, may have identical form, but even in this case, it is permitted to write the words without diacritics. There are some Urdu language words that need minimum diacritics [10]. In the absence of these diacritics, these words are considered deficient and cannot be pronounced accurately. Fig. 4 shows some of these words.

2.3 Variations in Unicode

Unicode standard provides the complete support for Urdu language as per [11]. There are few discrepancies i.e. the character Hamza (ء) cannot be connected with the letter following it because it is a non-joiner character in Unicode. However words in Urdu language like قائل / Ka.il requires a Hamza to be joined with the following characters. In order to solve this issue, instead of ء (Hamza), Unicode provides a separate character ئ, joining Hamza, for such words. The character ے (Bari Yay) is also
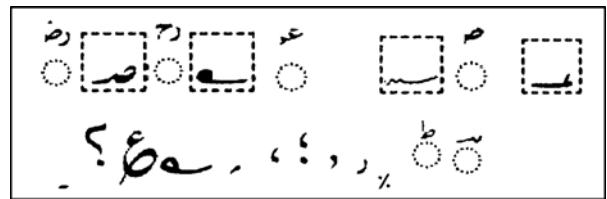

FIG. 3. SPECIAL SYMBOLS IN URDU

Actual word with diacritics	Actual word without diacritics	English translation	Urdu word with diacritics (correct)	Urdu word without diacritics (incorrect)
/a.la/	/a.li/	High quality	اعلیٰ	اعلی
/teq.ri.ban/	/teq.ri.ba/	almost	تقریباً	تقریبا

FIG. 4. URDU WORDS THAT REQUIRE DIACRITICS

a non-joiner character in Unicode, but the word بے کار /
be.kar/ (useless) is also written as بیکار/be.kar/ in Urdu
language. In order to write the latter, we have to put ی to
join Yay with Kaaf ک instead of ے. For complete Urdu
language support, Unicode standards need to resolve
these issues. Some characters have multiple Unicode
values in different keyboards like ی, ک etc. Depending
upon the character position within the word, one standard
character replaces such characters to normalize them
before any processing on them.

2.4   Syllabification General Principles

Syllable is a compact element of speech sound as per
[12]. There are three components of a syllable: the nucleus,
onset and the coda. A systematic process of splitting up
a syllable into its constituents is referred to as
syllabification. Syllabification principles are language
dependent, however, every language has some common
syllabification principles i.e.

- Uncovering the nucleus of the syllable.

- Finding the consonants and their syllabic
  affiliation.

- An abstract strategy for syllabification.

Every language forces its own parameters and constraints
on syllabification. Some of the languages take complicated
codas while other languages take complicated onsets.
Urdu language is very careful about the onset position
and it picks only a single consonant, but for the coda
position, it picks only two consonants.

2.5   Urdu Text Tokenization

According to Basit, et. al. [13], the tokenization process
isolates the words in the input Urdu text based on the
punctuation and space. This process also defines the
rules for some specific scenarios i.e. to isolate a text and
a number, connected together, into tokens. However, it is
not a good practice to use space for tokenization because
it can create errors in Urdu corpus.

3.   LITERATURE REVIEW

Shah, et. al. [5] developed a bi-lingual algorithm for TTS
synthesis of Sindhi and Urdu language text. This
algorithm uses a combination of hybrid rule based and
knowledge based approach along with concatenative
synthesis method for the conversion of Sindhi and Urdu
language text into speech. In this system, text analysis
module takes ASCII text as input and a series of phonetic
symbols with prosody targets is the output. This module
inspects the input text and expands the abbreviations
and non-alphabetic characters to their full representation.
Text is labeled by using the syntactic parser which
recognizes the part-of-speech for each word in sentence
by using the letter-to-sound rules. Word accents and
sentence phrasing is predicted by the prosody module.
Speech units are assembled together and then these
speech units are fed into the synthesizer which produces
speech waveform for the listener. Speech unit's inventory
for this bi-lingual TTS system consists of all the phones,
di-phones and some inter-word combinations. There are
352 speech units stored as wav files. This bi-lingual TTS
synthesis system, sometimes, produces unnatural robot
sounding speech because speech unit's inventory
contains recorded speech snippets up to di-phones. Also,
the homographs in the Urdu language are not handled by
this system. The proposed architecture also makes use
of concatenative synthesis strategy but it is distinguished
from the system in [5] as follows:

- It includes tri-phones in the speech unit's
  inventory to avoid the robot sounding speech
  produced due the use of di-phones.

- Text processing module removes the ambiguity
  of homographs in the Urdu text. The definite
  speech patterns for the homographs are stored
  in the speech unit's inventory.

4.   METHODOLOGY

A TTS synthesis system which is generally called as ASS
(Automated Speech System), involves conversion of Urdu
text into Urdu speech. Fig. 5 presents an architecture of

our methodology for Urdu TTS synthesis system using concatenative synthesis strategy. This architecture consists of a number of modules, with each module performs separate functionality. The next sections will describe the functions of each layer and their components in detail.

4.3   Text Analysis

This step involves the analysis of input text and internal representation of phonemes is generated. There are three modules to carry out this task that are used sequentially.

Text Processing: This module tokenizes Urdu sentences, handles acronym expansion and solves homograph disambiguation. Sentence tokenization module identifies various tokens in the input text and then separates the punctuation marks and whitespace characters from these identified tokens. It then map these tokens to words. Text normalization module performs some natural language processing tasks such as the number, date and time to text conversion. This module also expands abbreviations into full words before their pronunciation. Abbreviations are referred to as the non-standard words. Two steps are

FIG. 5. ARCHITECTURE OF PROPOSED SYSTEM

required to deal with these non-standard words: tokenization to identify potential abbreviations and then expand these abbreviations to map to full terms. Urdu language has some homographs. Text processing module removes the ambiguity of homographs in the Urdu text. The definite speech patterns for the homographs are stored in the speech unit's inventory. These speech patterns are marked by part of speech as this knowledge is required to remove the ambiguity of the homographs. A tagger is run to select a speech pattern for the specified homographs in the Urdu text.

Phonetic Analyzer: Normalized word strings are input to the phonetic analysis module. This module produces a speech pattern for all the words. This speech pattern consists not only of a list of phones, but also di-phones and tri-phones for some long words in Urdu language. The pronunciation also includes the lexical stress and a syllabic structure. Tri-phones in the pronunciation helps to avoid the unnatural robot sounding. This module uses letter to sound rules to find the pronunciation of a word. Urdu letter to sound rules convert the normalized text to a phonetic string.

Prosodic Analysis: This module estimates the sentence phrasing, word accents and based on those, it generates targets such as duration of phonemes and fundamental frequency. Syllabification performs the division of words regarded as a unit of pronunciation and containing one vowel sound [11]. This module marks the phonemic output with syllable boundaries which is required to make the Urdu sound change rules conditional.

4.4   Speech Production

A waveform is generated by this module using the phonetic descriptions and prosodic illustrations. These descriptions consist of a phones list with associated durations. The speech unit's analyzer will analyze the set of phonemes presented by the text analysis module and extracts the suitable speech units based on the phonemes Unicode from speech unit's inventory. The speech unit concatenation module concatenates the speech units. This module is responsible for more natural synthetic speech. Finally the concatenated speech units are sent to the speech synthesizer which generates the Urdu speech for listeners.
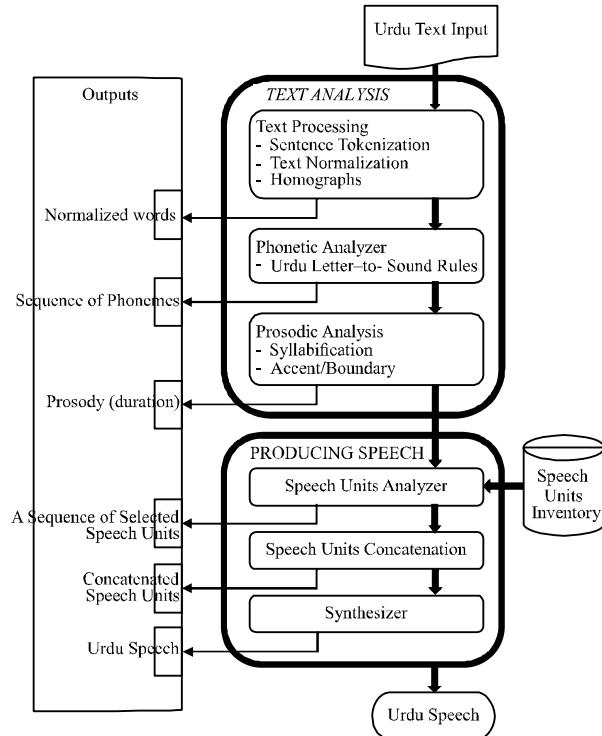
4.5 Speech Units Inventory

The speech unit's inventory is a set of recorded speech units for the Urdu language. This inventory contains speech units for all the possible combinations of Urdu language phones. We are using a total of 720 recorded speech units in our methodology. These speech units cover a combination of di-phones, tri-phones and some inter-word combinations. Tri-phones are used to avoid robot sounding speech produced due the use of di-phones. These speech units are stored as .wav files and are named according to the Unicode's of the Urdu phones. Recorded sound snippets are chosen to minimize the concatenation problems. The main challenge in the design of this inventory was to keep the number of recorded speech units small. A trained speaker has recorded these speech units in a sound proof environment to avoid the surrounding noise. These speech units are then edited to remove the extra gaps at the start and end of the speech units to minimize the audible glitches in the output to increase the intelligibility because in concatenative synthesis technique, intelligibility can be disturbed by these glitches in the output as mentioned in Table 1.

4.6 Dataset

The authors prepare a dataset for the evaluation of this methodology by randomly taking the Urdu words, sentences and paragraphs from Urdu newpapers, books and magazines. Fig. 6 shows an extract from the Urdu newspaper:

Fig. 7 shows an Urdu passage taken from an Urdu magazine:

Fig. 8 shows a list of Urdu words taken from the Urdu book and Urdu magazine. This list consists of Urdu words with and without diacritics to evaluate the intelligibility. Date format, abbreviation and number format are also a part of this word list.

There were 300 people, aged between 15 and 30 years, to carry out the performance evaluation by using the above data set. All of the participants were asked to judge 50-100 words and sentences.

5. RESULTS AND DISCUSSION

The results of the performance evaluation for TTS synthesis system are mentioned below. Effectiveness of the proposed system was evaluated with reference to intelligibility and acceptability.

FIG. 6. URDU NEWSPAPER PARAGRAPH

FIG. 7. URDU MAGAZINE PASSAGE

The first experiment was to test the intelligibility of the synthesized speech on sentence and word level. All the participants were instructed to write what they perceive. Fig. 9 shows the percentage of words and sentences that were correctly understood by the participants of the test. The intelligibility rate of the words and sentences was near to 80 and 65% respectively which shows the achieved milestone of the weakness of concatenative synthesis technique as discussed in Table 1.

Acceptability of the system was evaluated on sentence and word levels in our next experiment. The participants were asked some questions about the speed, naturalness and quality of the sound as per [5]. The speed factor of the TTS synthesis system is very important. If the speed of the synthesized speech is too slow or too fast, then the listeners may lose their concentration. About 75% of the listeners thought that the speed of the synthesized speech was normal. Fig. 10 shows the speed rate of the synthesized speech.

Fig. 11 shows the naturalness of synthesized speech and verifies the strength of concatenative synthesis technique as per [5].The query about the naturalness of the voice

was "Whether the synthesized voice is natural or not?". About 75% of the participants observed the synthesized voice was natural, 20% observed the level of naturalness was OK and 5% observed the synthesized voice was not natural.

The question about the quality of sound was "Do you think that the sound quality of the synthesized voice is good?" About 65% of the participants thought that quality of synthesized voice was high, 25% considered the quality was better and 5% each thought the quality was low and bad. Fig. 12 shows the sound quality of the voice.
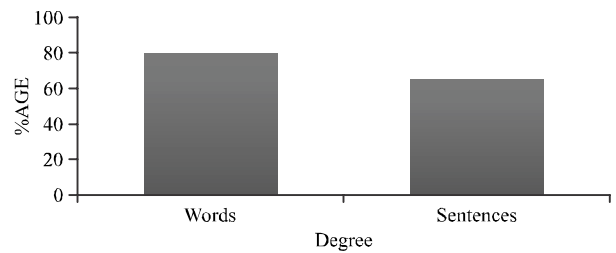
إستدراک	إستحصال	ارواح	ارفع
إيجاد	إطمينان	انبار	إستحجاب
جمال	ڈانٹ	تائب	بانگ درہ
اوجھل	بُھوت	ڈھونڈنا	جُھوٹ
گبری	صغری	اساعیل	إستحقام
وحیینا	خلوت	ثبوت	آخوت
جھانکنا	تشہیہ	پھیل	بُوڑھی
نگہبداشت	پھسلنا	پچھوندی	پُر نچھنا
رم جھم	زکوٰۃ	زراعت پیشہ	سنگسر
ذی شعور	فتین	ذی استعداد	حسب ذیل
ڈوڈھہ	ڈینگ مارنا	ذرہ	ذوالجلال
خلوت گاہ	عجیب الخلقت	دل بہلانا	ڈانٹ
حروف تہجی	حسب معمول	خلاف قاعدہ	خواب وخیال
۱۱ جون ۱۹۹۸	جھیل	چھیاسٹھ	حبُ الوطن
آپ	۱.۵۶	۲۰۰۲_۱۲_۰۱	۱۹۹۲
درخواست	صحرا	السلام علیکم	پاکستان

FIG. 8. URDU WORDS

FIG. 9. SYNTHESIZED SPEECH INTELLIGIBILITY
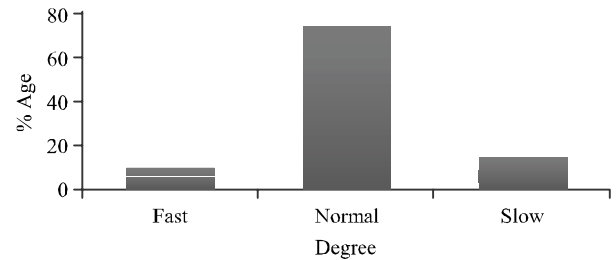

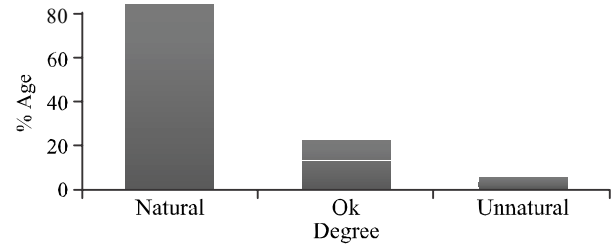*FIG. 10. SPEED OF SYNTHESIZED VOICE*
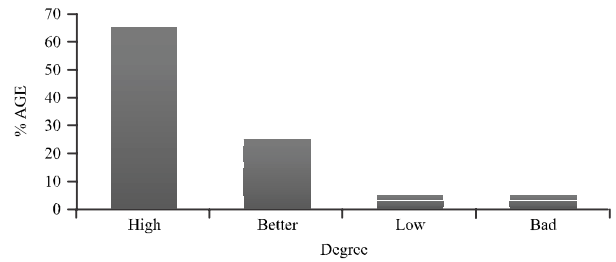

FIG. 11. NATURALNESS OF VOICE

FIG. 12. SOUND QUALITY

6.   CONCLUSION

Concatenative synthesis strategy has been used to implement the proposed TTS synthesis system for Urdu language. The proposed TTS system nicely disambiguates the homographs in the Urdu language and produces natural sound. The system is using the recorded speech units for Urdu phonemes. Speech unit's inventory contains a combination of Urdu phones, di-phones and tri-phones. Tri-phones are used to avoid robot sounding speech. Effectiveness of the proposed system was evaluated with reference to intelligibility and acceptability on sentence and word level. The intelligibility rate of the words and sentences was near to 80% and 65% respectively which shows the achieved milestone of the weakness of concatenative synthesis technique. The acceptability rate for the naturalness was 95% (75% natural, 20% acceptable). High quality Urdu speech is provided by this system. In future, Visual TTS methods can be incorporated in this system. Also the quality of the system can be improved by handling special characters such as (@, % etc…). This system can be further improved by improving the quality of the recorded speech units.

ACKNOWLEDGEMENT

REFERENCES

[1]   Jilka, M., Syrdal, A.K., Conkie, A.D., and Kapilow, D.A., "Effects on TTS Quality of Methods of Realizing Natural Prosodic Variations", Proceedings of International Congress of Phonetic Sciences, pp. 2549-2552, Barcelona, 2003.

[2]   Sondhi, M.M., and Schroeter, J., "Speech Production Models and Their Digital Implementations", The Digital Signal Processing Handbook, CRC Press LLC, pp. 44-1-44-21, Boca Raton, Florida, 1999.

[3]   Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y., and Syrdal, A., "The AT&T Next-Gen TTS System", Joint Meeting of ASA, EAA, and DAGA, pp. 18-24, Germany, 1999.

[4]   Black, A.W., and Kevin A.L., "Multilingual Text-to-Speech Synthesis", IEEE Proceedings of International Conference on Acoustics, Speech and Signal Processing, Volume 3, pp. 761-764, Canada, 2004.

[5]   Shah, A.A., Ansari, A.W., and Das, L., "Bi-Lingual Text-to-Speech Synthesis System for Urdu and Sindhi", National Conference on Emerging Technologies, pp. 20126-20130, Pakistan, 2004.

[6]   Soomro, H.K., Shah, A.A., and Shaikh, A.A.G., "Development of Computerized Sindhi to English and English to Sindh Dictionary", Mehran University Research Journal of Engineering & Technology, Volume 23, No. 4, pp. 289-296, Jamshoro, Pakistan, October, 2004.

[7]   Chowdhry, B.S., Memon, A.R., and Memon, N., "Voice Digitization for Sindhi Speech Synthesis", Mehran University Research Journal of Engineering & Technology, Volume 13, No. 3, pp. 1-8, Jamshoro, Pakistan, July, 1994.

[8]   Madiha, I., and Hussain, S., "Corpus Based Urdu Lexicon Development", Proceedings of Conference on Language Technology, University of Peshawar, Volume 73, Pakistan, 2007.

[9]   Hussain, S., "Letter-to-Sound Conversion for Urdu Text-to-Speech System", Proceedings of Workshop on Computational Approaches to Arabic Script-Based Languages, pp. 74-79, Switzerland, 2004.

[10]   Humayoun, M., Hammarström, H., and Ranta, A., "Urdu Morphology, Orthography and Lexicon Extraction", Master's Thesis, Chalmers University of Technology, Sweden, 2006.

[11]   Hussain, S., "Urdu Localization Project: Lexicon, MT and TTS (ULP)", Proceedings of Workshop on Computational Approaches to Arabic Script-Based Languages, pp. 80-81, Switzerland, 2004.

[12]   Ranjha, M.I., "Urdu Syllable: Templates and Constraints", Proceedings of Conference on Language & Technology, pp. 29-36, Pakistan, 2012.

[13]   Basit, R.H., and Hussain, S., "Text Processing for Urdu TTS System", Proceedings of Conference on Language & Technology, pp. 101-108, Pakistan, 2014.