The Development of Pashto Speech Synthesis System

Muhammad Akbar Ali Khan
Department of Computer Systems Engineering, University of Engineering & Technology, Peshawar, Pakistan.

Sahibzada Abdur Rehman Abid
Department of Computer Systems Engineering, University of Engineering & Technology, Peshawar, Pakistan.

Fatima Tuz Zuhra
Department of Computer Science, University of Peshawar, Pakistan.

Nasir Ahmad
Department of Computer Systems Engineering, University of Engineering & Technology, Peshawar, Pakistan

.

ABSTRACT
This paper presents a novel Pashto text-to-speech (TTS) synthesis system based on data driven techniques such as Classification and Regression Tree (CART), Bigrams, and Non Uniform Units (NUUs). A modular concatenative TTS system has been developed for the Pashto language. Speech synthesis is carried out through a series of steps with the intention to provide a gradually more absolute transcription of the text, from which the final speech signal is then generated. The steps can be divided into two modules; a Natural Language Processing (NLP) module and a Digital Signal Processing (DSP) module. These steps incrementally enhance the information derived from the input and put it on a generally accessible internal data structure. The goal is to obtain enough information on the internal data structure so as to be capable to obtain an intelligible and natural speech.

General Terms
Speech synthesis, Pashto speech synthesis, concatenative speech synthesis

Keywords
Pashto speech synthesis, Classification and Regression Tree, Non Uniform Units, Pashto TTS

1. INTRODUCTION
Speech synthesis is the process which takes a sequence of words as an input and converts them to an acoustic signal. It is the opposite process of speech recognition where speech is converted into corresponding text. The systems for automatically generating speech parameters from a linguistic representation (such as a phoneme string) were not available until the 1960s [1], and the systems for converting ordinary text into speech were first developed in the 1970s, with MITalk being the then most popular such system [2]. In the early days of synthesis, the research efforts were devoted mainly to simulating human speech production mechanisms, using basic articulatory models based on electro-acoustic theories. Though this modeling is still one of the ultimate goals of synthesis research, advances in computer science have widened the field of Text to Speech processing to include not only human speech production but also to model the text processing [3]. In [4], a TTS system for the Maltese language has been proposed, transforming arbitrary textual input into the spoken output.

In proposed Pashto speech synthesis the input is the Pashto text while the output is its corresponding speech signal. A number of applications can potentially take advantage of the Pashto speech synthesis system. Rest of the paper is organized as follows. Section 2 describes different methods of the speech synthesis. Section 3 explains the proposed Pashto speech synthesis system while section 4 concludes the findings of this work.

2. SPEECH SYNTHESIS METHODS
A number of methods for the speech synthesis have been proposed in literature. All of these method falls largely into one of the following three categories: articulatory synthesis, formant synthesis or concatenative synthesis. Each of these methods has their own advantages and disadvantages.

2.1 Articulatory Synthesis
Articulatory synthesizers are physical models based on the detailed description of the physiology of speech production and the physics of sound generation in the human vocal apparatus [5]. To make the computers to speak by articulatory synthesis, the human vocal apparatus is modeled by combining electrical, mechanical and electronic components and a robotic talking head is made that produces sound just similar to a person [6]. It is the most difficult approach as the physiology of human speech production is not yet fully explored. Recent progress in speech production imaging, articulatory control modeling, and tongue biomechanics modeling has led to significant improvements in the way articulatory synthesis is performed [7]. Articulatory synthesizers are computationally costly and difficult to debug. That is why they are far from practical applications.

2.2 Formant Synthesis
Formant synthesis is a descriptive acoustic-phonetic approach to the speech synthesis [3]. In formant synthesis parameters such as fundamental frequency and noise levels are varied over time to create a waveform of artificial speech. Formant synthesis is based on the source filter model of speech and is the most broadly used synthesis method and has two basic structures, cascaded and parallel. Synthesis of dissimilar voices and voice characteristics, and the modeling of emotive speech have kept research on formant synthesis active [8]. At least three formants are required to produce an intelligible speech; however up to five formants have been used for producing a higher quality speech. Each formant is usually modeled with a two pole resonator which enables both, the formant frequency and its bandwidth to be specified [9]. Rule-based formant synthesis is based on a set of rules determining the parameters necessary to synthesize a desired utterance [2].

Infinite number of sounds provided by the formant synthesis makes it more flexible than other synthesis methods.

2.3 Concatenative Synthesis

Concatenative synthesis is the generation of natural sounding synthesized speech waveforms by selecting and concatenating speech units from a large database [10]. It is the simplest way of producing natural and intelligible synthetic speech. Locating the correct unit is the most important factor in the concatenative synthesis. Shorter units need less memory, but the collection and labeling of the speech samples becomes complex and difficult. On the other hand longer units need more memory; however more naturalness, less concatenation points and a fine control of the co-articulation can be achieved. The units used can be words, syllables, demisyllables, phonemes, diphones, or triphones [11]. Word is perhaps the most natural unit for written text and a suitable unit for limited vocabulary synthesis system. Concatenation of words is relatively easy to perform and the co-articulation effects within a word are captured in the stored units. However, words uttered in isolation are greatly different from their utterance in continues sentences thus making the synthesized continuous speech sounding unnatural [2]. Phonemes are the most commonly used units in speech synthesis as they are the standard linguistic presentation of speech. Moreover, the inventory of fundamental units is usually between 40 and 50, which is clearly the minimum as compared with the other units [2].

3. PASHTO SPEECH SYNTHESIS SYSTEM

The transduction procedure for the Pashto speech synthesis is achieved through a sequence of steps, which gives a detailed transcription of the text, from which the corresponding speech is finally derived. These steps can be divided into two modules, the NLP module and the DSP module, as shown in Figure 1.

3.1 Natural Language Processing

The NLP module processes analyze the text to derive a more suitable phonetic transcription that can be finally used by the DSP module. The subtasks of the NLP module are discusses in more details in the following.

3.1.1 Pre-Processing

The preprocessor block transforms the text into processable input in the form of a words list. The function of the preprocessor is to divide the incoming sentences into tokens and determine punctuation ambiguity such as a full stop indicating the end of a sentence.

3.1.2 Morphological Analysis

Morphological analyzer uses lexical information to obtain a morphological parse for each word and thus recognizes its possible parts of speech category.

The parts of speech categories of the Pashto words can be expressed in the form of a morphological dictionary which gives a list of all words linked with their part of speech categories, as shown below in Table 1.

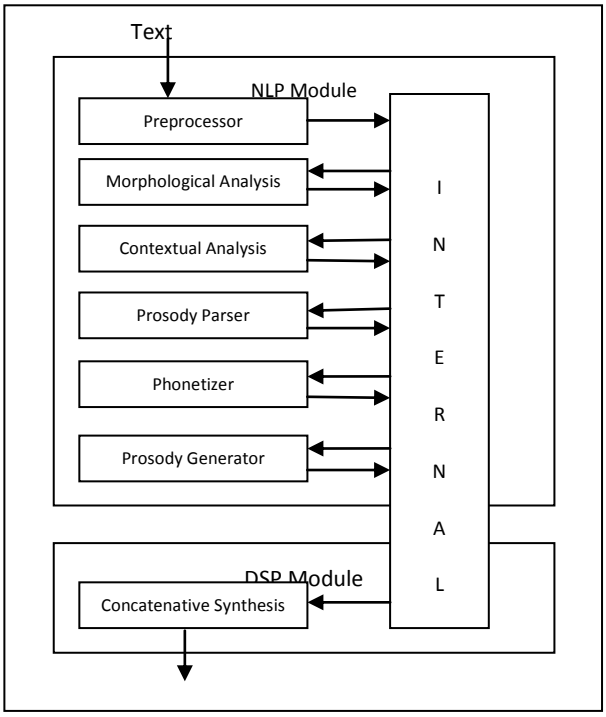

Fig 1: Pashto Text to Speech System

Table 1: Pashto words and corresponding parts of speech

Word	*POS*	*Word*	*POS*
tlwyzywn	*Noun*	*Iw*	*Noun*
bIa	*Adverb*	*zh*	*Pronoun*
hm	*Adverb*	*kE*	*Postposition*
slamwnh	*Noun*	*wRandE*	*Adverb*
kwm	*Verb*	*tasw*	*Pronoun*
ghg	*Intransitive verb*	*ahtram*	*Noun*
d	*Pronoun*	*aw*	*Conjunction*
,	*Punctuation*	*gwr@i*	*Verb*
twns	*Noun*	*ph*	*Preposition*
amnItI	*Adjective*	*wkR@i*	*Verb*
.	*Punctuation*	*hklh*	*noun*

3.1.3 Contextual Analysis

For contextual analysis of Pashto a bi-gram model has been used. In bi-gram model the probability of a tag depending on the pervious tag is considered. The bigram model is sketched by using a set of states that represent the part of speech categories based on the grammar. Every transition from state y to state x is associated with a transition probability P ($c_x|c_y$), which is the probability for a word of category c_y to be tracked by a word of category c_x. Transition probability is the probability of the part of speech word to follow the current part of speech word while emission probability is the probability of a word occurrence in the same category part of speech. A state dependent probability P ($w_x|c_y$) is calculated for each state and every word in the vocabulary, which shows the probability that category c_y appears as word w_x. The transition and emission probabilities are shown in table 2.

The emission and transition probabilities are determined by the words and tag combination appearances in a corpus. The emission probability $P(w_x|c_y)$ is estimated by the number of times w_x appears as c_x, divided by the total number of words with part of speech category c_x.

$$P(w_x|c_y) \approx \frac{\#(w_x, c_y)}{\#(c_y)} \qquad (1)$$

In the same way, the bigram transition probability between categories c_y and c_x is estimated by the number of times c_x appears after c_y, divided by the total number of words with part-of-speech category c_y,

$$P(c_x|c_y) \approx \frac{\#(c_x, c_y)}{\#(c_y)} \qquad (2)$$

Once emission and transition probabilities are estimated, getting the most excellent sequence of tags for a given sentence reduces to selecting the best sequence of part of speech tags for the sentence, i.e., the one with highest probability given the sequence of words and the bigram model.

Table 2: Emission and Transition Probabilities

Part of speech (POS)	*Words*	*Emission Probability*	*POS*	*Transition probability*
Adjective	*amnItI*	*0.2000*	*Noun*	*0.6000*
	chWr	*0.2000*	*Transitive*	*0.2000*
	mhm	*0.2000*	*Verb*	*0.2000*
	mtid	*0.2000*	*Verb*	
	pwrh	*0.2000*		
Adverb	*bIa*	*0.2500*	*Adverb*	*0.2500*
	hm	*0.2500*	*Pronoun*	*0.5000*
	nh	*0.5000*	*Verb*	*0.2500*
Conjunction	*aw*	*0.6667*	*Adjective*	*0.1667*
	chE	*0.3333*	*Postposition*	*0.1667*
			Pronoun	*0.5000*
			Verb	*0.1677*
Intansitive Verb	*ghg*	*0.5000*	*Noun*	*0.5000*
	Im	*0.5000*	*Punctuation*	*0.5000*
Noun	*afghanstan*	*0.0294*	*Adjective*	*0.0588*
	amrika	*0.0294*	*Adverb*	*0.0882*
	ghwrdzng	*0.0294*	*Conjunction*	*0.0294*
	hklh	*0.0294*	*Intransitive*	*0.0588*
	tlwyzywn	*0.0588*	*verb*	*0.4412*
	twns	*0.0588*	*Noun*	*0.0588*
	.	.	*Postposition*	*0.0294*
	.	.	*Preposition*	*0.0588*
	dzwakwnw	*0.0294*	*Pronoun*	*0.1765*
			Verb	
Postposition	*kE*	*0.3333*	*Adjective*	*0.3333*
	srh	*0.6667*	*Preposition*	*0.3333*
			Pronoun	*0.3333*
Preposition	*ph*	*0.6667*	*Conjunction*	*0.3333*
	th	*0.3333*	*Noun*	*0.3333*
			Pronoun	*0.3333*
Pronoun	*d*	*0.5333*	*Adjective*	*0.0667*
	dE	*0.0667*	*Noun*	*0.7333*
	xpl	*0.0667*	*Pronoun*	*0.1333*
	.	.	*Verb*	*0.0667*
	.	.		
	zmwng	*0.0667*		
Punctuation	,	*0.2000*	*Conjunction*	*0.2500*
	.	*0.8000*	*Noun*	*0.2500*
			Pronoun	*0.5000*
Transitive Verb	*chpawl*	*1.0000*	*Noun*	*1.000*
Verb	*awr@i*	*0.0833*	*Conjunction*	*0.2500*
	bh	*0.0833*	*Preposition*	*0.0833*
	.	.	*Pronoun*	*0.1667*
	.	.	*Punctuation*	*0.3333*
	wRandE	*0.1667*	*Verb*	*0.1667*

3.1.4 Prosodic Parser

In Pashto speech synthesis, prosodic phrases are identified with a rather trivial chinks 'n chunks algorithm [12]. In the proposed system it is considered that a prosodic phrase break is automatically set when a word belonging to the chunks group is followed by a word classified as a chink. Chinks are composed of conjunction preposition, pronoun, postposition; and chunks are composed of adjective, adverb, intransitive verb, noun, transitive verb, verb, and punctuation. The classes of chinks and chunks considered for the synthesis of Pashto speech are given in table 3.

Table 2: Pashto Chinks and Chunks

Pashto Chinks	*Pashto Chunks*
zh, xpl, aw, tasw, d, , srh, ph, kE, chE, dE, IE, lh, hghh, twlw, mwng, etc.	*Iw, dzl, ghg, bIa, Im sId, wRandE, sIlman, slamwnh, awnR@y, kwm, afghanstan, mhm, amrIka, ashna, twns, tlwyzywn, xbrwnh, gwr@i, awr@i, srprst, jmhwr, etc.*

3.1.5 Phonetizer

In the proposed synthesis system corpus based phonetizer has been developed and is implemented as a decision tree trained on the real data. In the construction of automatic phonetization, the characteristic utilized in the decision tree are only the letter being currently phonetized, the part of speech of the current word and the letters on the left and right of the current letter. In the Pashto training corpus phonetic transcription are given to each word and thus each letter of the word obtains its phonetic symbol. A phonetic character is given to each phoneme by choosing the phonetic symbol used in the corpus. The CART tree is implemented in MATLAB which repeats itself, accounting for the details so that building a tree from its top is the same as building a tree from any of its interior nodes. This phonetization was tested on the entire Pashto test corpus to get the part of speech details for each word from the corpus and no error was found.

3.1.6 Prosody Generator

Prosody is achieved as a result of unit selection from a large speech corpus. Phonetic features such as current and neighbouring phonemes, as well as linguistic features such as stress, position of the phoneme within its word, position of the word within its prosodic phrase, position of the prosodic phrase within the sentence and part-of-speech tag of the current word are used to find a sequence of speech segments or units taken from the speech corpus, whose features most closely match the features of the speech unit to be synthesized.

3.2 Digital Signal Processing

DSP module operates on the phonetic transcription obtained from the previous module and creates the speech waveform that can be reproduced audibly. In this work, the concatenative synthesis approach has been adopted. Twenty sentences of Pashto are stored in text corpus and the same are recorded and stored as .wav files. The HMM-based text-to-speech alignment system [13] is used to create the segmentation files. The content of the segmentation files is such that each line refers to a start point, an end point, and a phoneme name. Alignment, on the other hand, is trained by the degree of correspondence between the assumed phonemic transcription and the actual list of phonetic units produced. In some cases a difference between the assumed phonemic transcription and the actual list of phonetic units occurs due to the co-articulation which cannot be taken into account in the phonemic transcriptions. The segmentation files are check and corrected where needed using the Wavesurfer tool. A speech unit database is generated from the segmented speech, containing information about the current phoneme, previous phoneme, next phoneme, the index of the part of speech of the current word, the index of the current prosodic phrase within the current sentence, the number of prosodic phrases on the right until the end of the sentence, the index of the current word within the current prosodic phrase, the number of words on the right until the end of the current prosodic phrase, the index of the sentence containing the phoneme and the start and end point for the current phoneme in the related wav file. A few entries in the database are shown in Figure 2.

'_#In1113'	*[1]*	*[0]*	*[108]*
'I#Wn1113'	*[1]*	*[108]*	*[394]*
'WIDn1113'	*[1]*	*[394]*	*[699]*
'DWZn1122'	*[1]*	*[699]*	*[928]*
'ZDLn1122'	*[1]*	*[928]*	*[1324]*
'LZXn1122'	*[1]*	*[1324]*	*[1993]*

Fig 2: Database entries

NUU synthesis formats targets by appending them with the linguistic context features. It also checks for the accessible diphones in the speech unit database similar to the target diphones, and selects a maximum of 10 units per diphone to accelerate the search process. Viterbi algorithm finds the best order of units by minimizing the selection cost. At the end the selected diphones from the speech corpus are concatenated to produce the final synthetic speech.

4. CONCLUSION

Pashto speech synthesis system utilizing the Bigram model, CART and NUUs techniques has been presented. The emission and transition probability for each word in the Pashto words dictionary are calculated through the bigram model. CART is efficient due to its lower computational requirements and greater flexibility. NUUs checked for the available diphones in the speech unit database matching to the target diphones while viterbi algorithm finds the finest order of units. The Pashto speech synthesis system produces natural speech for the sentences in the speech corpus, while for other sentences it produced nearly natural audio results, with minor discontinuities. In the future work the problem of acronyms, abbreviations, and out of vocabulary words will be considered.

5. REFERENCES

[1] R. Sproat, and J. Olive, "Text-to-Speech Synthesis" in V. K. Madisetti and D. B. Williams (eds.), *Digital Signal Processing Handbook*, Ch. 46, CRC Press, 1998.

[2] J. Allen, M.S. Hunnicutt, and D. Klatt, *From Text to Speech*, Cambridge University Press, Cambridge, 1987.

[3] J. Allen, M.S. Hunnicutt, and D. Klatt, *From Text to Speech: the MITalk System*, Cambridge University Press, Cambridge, 1987.

[4] P. J. Farrugia "Text-To-Speech Technologies for Mobile Telephony Services", MSc thesis, Dept of Computer Science and AI, University of Malta 2005.

[5] S. Parthasarathy, and C. H. Coker, "Automatic estimation of articulatory parameters", *Computer Speech and Language*, vol. 6, no.1, pp. 37-75, 1992.

[6] B. Baxter, and W.J. Strong, "WINDBAG—a vocal-tract analog speech synthesizer", *Journal of the Acoustical Society of America*, vol. 45, no. 1, pp. 309, 1969.

[7] P. Birkholz, D. Jackel, and B.J. Kröger, "Construction and control of a three-dimensional vocal tract model", *ICASSP 2006*, Toulouse, France, pp. 873-876. 2006.

[8] R. Carlson, B. Granström, and I. Karlsson "Experiments with voice modelling in speech synthesis", *Speech communication*, vol. 10, pp. 481-490. 1991.

[9] R. Donovan, "Trainable Speech Synthesis", PhD. Thesis. Cambridge University Engineering Department, England, 1996.

[10] A. J. Hunt and A. W. Black, "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database", ATR Interpreting Telecommunications Research Labs. 2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan, 1996.

[11] S. Lemmetty, "Review of Speech Synthesis Technology", MSc Thesis, Helsinki University of Technology Department of Electrical and Communications Engineering, March 30, 1999.

[12] M.J. Liberman and K.W. Church, "Text Analysis and Word Pronunciation in Text-to-Speech Synthesis," in S. Furui and M.M. Sondhi, (eds.), *Advances in Speech Signal Processing*, pp. 791–831. Dekker, New York, 1992.

[13] F. Malfrere, O. Deroo, T. Dutoit, and C. Ris, "Phonetic Alignement: Speech-Synthesis-based versus Viterbi-based", *Speech Communication*, vol. 40, no. 4, pp. 503–517, 2003.