Implementation of Three Text to Speech Systems for Kurdish Language

Anvar Bahrampour[1], Wafa Barkhoda[2], and Bahram Zahir Azami[2]

[1] Department of Information Technology, Islamic Azad University,
Sanandaj Branch Sanandaj, Iran
bahrampour58@gmail.com
[2] Department of Computer, University of Kurdistan
Sanandaj, Iran
{w.barkhoda,zahir}@ieee.org

Abstract. Nowadays, concatenative method is used in most modern TTS systems to produce artificial speech. The most important challenge in this method is choosing appropriate unit for creating database. This unit must warranty smoothness and high quality speech, and also, creating database for it must reasonable and inexpensive. For example, syllable, phoneme, allophone, and, diphone are appropriate units for all-purpose systems. In this paper, we implemented three synthesis systems for Kurdish language based on syllable, allophone, and diphone and compare their quality using subjective testing.

Keywords: Speech Synthesis; Concatenative Method; Kurdish TTS System; Allophone; Syllable, and Diphone.

1 Introduction

High quality speech synthesis from the electronic form of text has been a focus of research activities during the last two decades, and it has led to an increasing horizon of applications. To mention a few, commercial telephone response systems, natural language computer interface, reading machines for blinds and other aids for the handicapped, language learning systems, multimedia applications, talking books and toys are among the many examples [1].

Most of the existing commercial speech synthesis systems can be classified as either formant synthesizers [2,3] or concatenation synthesizers [4,5]. Formant synthesizers, which are usually controlled by rules, have the advantage of having small footprints at the expense of the quality and naturalness of the synthesized speech [6]. On the other hand, concatenative speech synthesis, using large speech databases, has become popular due to its ability to produce high quality natural speech output [7]. The large footprints of these systems do not present a practical problem for applications where the synthesis engine runs on a server with enough computational power and sufficient storage [7].

Concatenative speech synthesis systems have grown in popularity in recent years. As memory costs have dropped, it has become possible to increase the size of the

E. Bayro-Corrochano and J.-O. Eklundh (Eds.): CIARP 2009, LNCS 5856, pp. 321–328, 2009.

acoustic inventory that can be used in such a system. The first successful concatenative systems were diphone based [8], with only one diphone unit representing each combination of consecutive phones. An important issue for these systems was how to select, offline, the single best unit of each diphone for inclusion in the acoustic inventory [9,10]. More recently there has been interest in automation of the process of creating databases and in allowing multiple instances of particular phones or groups of phones in the database, with the selection decided at run time. A new, but related problem has emerged: that of dynamically choosing the most adequate unit for any particular synthesized utterance [11].

The development and application of text to speech synthesis technology for various languages are growing rapidly [12,13]. Designing a synthesizer for a language is largely dependent on the structure of that language. In addition, there can be variations (dialects) particular to geographic regions. Designing a synthesizer requires significant investigation into the language structure or linguistics of a given region.

In most languages, widespread researches are done on Text-to-Speech systems and also, in some of these languages commercial versions of system are offered. CHATR [14, 15] and AT&T NEXT GEN [16] are two examples offered in English language. Also, in other languages such as French [17,18], Arabic [1,4,7,19,20], Norwegian [21], Korean [22], Greek [23], Persian [24-27], etc, much effort has been done in this field.

The area of Kurdish Text-to-Speech (TTS) is still in its infancy, and compared to other languages, there has been little research carried on in this. To the best of our knowledge, nobody has performed any serious academic research on various branches of Kurdish language processing yet (recognition, synthesis, etc.) [28, 29].

Kurdish is one of the Iranian languages, which are a sub category of the Indian-European family [30]. Kurdish has 24 consonants, 4 semi vowels and 6 vowels. Also /ح/, /ع/, and /غ/ entered Kurdish from Arabic. Also, this language has two scripts: the first one is a modified Arabic alphabet and the second one is a modified Latin alphabet [31]. For example “trifa” which means “moon light” in Kurdish, is written as /تریفه/ in the Arabic script and as “tirîfe” in the Latin. Whereas both scripts are in use, both of them suffer some problems (e.g., in Arabic script the phoneme /i/ is not written; also both /w/ and /u/ are written with the same Arabic written sign /و/ [32,33], and Latin script does not have the Arabic phoneme /ڵ/, and it does not have any standard written sign for foreign phonemes [31]).

In concatenative systems, one of the most important challenges is to select an appropriate unit for concatenation. Each unit has its own advantages and disadvantages, and appropriate for a specific system. In this paper we develop three various concatenative TTS systems for Kurdish language based on Syllable, Allophone, and Diphones, and compare these systems in intelligibility, naturalness, and overall quality.

The rest of the paper is organized as follows: Section 2 introduces the allophone based TTS system. Section 3 and 4 presents syllable and diphone based systems respectively, and finally, comparison between these systems and quality test results are presented in Section 5.

2 Allophone Based TTS System

In this part, a Text-To-Speech system for Kurdish language, which is constructed based on concatenation method of speech synthesis and use allophones (several pronunciation of a phoneme [33]) as basic unit will be introduced[28,29]. According to the input text, proper allophones from database have been chosen and concatenated to obtain the primary output.

Differences between allophones in Kurdish language are normally very clear; therefore, we preferred to explicitly use allophone units for the concatenative method. Some of allophones obey obvious rules; for example if a word end with a voiced phoneme, the phoneme would lose the voicing feature and is called devoiced [34]. However, in most cases there is not a clear and constant rule for all of them. As a result, for extracting allophones we used a neural network. Because their learning power, neural networks can learn from a database and can recognize allophones properly [35].

Fig. 1 shows the architecture of the proposed system. It is composed of three major components: a pre-processing module, a neural network module and an allophone-to-sound module. After converting the input raw text to the standard text, a sliding window of width of four is used as the network input. The network detects second phoneme's allophone, and the allophone waveform is concatenated to the preceding waveform.

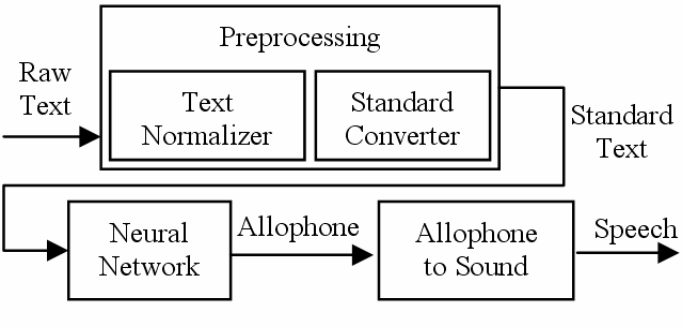

Fig. 1. Architecture of the proposed Kurdish TTS system

The pre-processing module includes a text normalizer and a standard converter. The text normalizer is an application that converts the input text (in Arabic or Latin script) to our standard script; in this conversion we encountered some problems [30,32,34,36].

Finally, in standard script, 41 standard symbols were spotted. Notice that this is more than the number of Kurdish phonemes, because we also include three standard symbols for space, comma and dot. Table 1 shows all the standard letters that are used by the proposed system. Table 2 shows the same sentence in various scripts. Also, the standard converter performs standard text normalization tasks such as converting digits into their word equivalents, spelling out some known abbreviations, etc.

In the next stage, allophones are extracted from the standard text. This task is done using a neural network. Kurdish phonemes have about approximately 200 allophones, but some of them are very similar, and non-expert people cannot detect the differences [34]. As a result, it is not necessary for our TTS system to include all of them (for simplicity, only 66 important and clear instances have been included; see Table 3). Also the allophones are not divided equally between all phonemes (e.g., /p/ is presented by five allophones but /r/ has only one allophone [34]). However, the neural network implementation is very flexible as it is very simple to change the number of allophones or phonemes.

Major Kurdish allophones (more than 80%) are dependent only on the following phonemes. Others (about 20%) are dependent on one preceding and two succeeding phonemes [34]. Hence, we employed four sets of neurons in the input layer, each having 41 neurons for detection of the 41 mentioned standard symbols. A sliding window of width four provides input phonemes for the network input layer. Each set of input layer is responsible for one of the phonemes in the window. The aim is to recognize the relevant allophone to the second phoneme of the window. The output layer has 66 neurons (corresponding to the 66 Kurdish allophones used here) for the recognition of the corresponding allophones and the middle layer is responsible for detecting language rules and it has 60 neurons (These values are obtained empirically); (See Fig. 2). The neural network accuracy rate is equal to 98%. In Table 4, neural network output and desired output are compared.

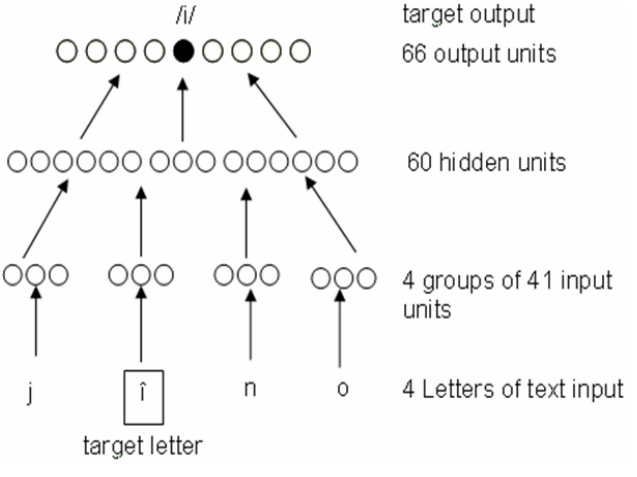

Fig. 2. The neural network structure

After allophone recognition, corresponding waveform of allophones should be concatenated. For each allophone we selected a suitable word and recorded it in a noiseless environment. Separation of allophones in waveforms was done manually by using of Wavesurfer software. The results of this system and comparison between it and other systems are presented in Section 5.

3 Syllable Based TTS System

Syllable is another unit which is used for developing a text-to-speech system. Various languages have different patterns for syllable. In most of these languages, there are many patterns for syllable and therefore, the number of syllables is large; so usually syllable is not used in all-purpose TTS systems. For example, there are more than 15000 syllables in English [6]. Creating a database for this number of units is a very difficult and time-consuming task.

In some languages, the number of syllable patterns is limited, so the number of syllables is small, and creating a database for them is reasonable; therefore this unit can be used in all purpose TTS systems. For example, Indian language has CV, CCV, VC, and CVC syllable patterns, and the total number of syllables in this language is 10000. In [37], some syllable-like units are used; the number of this unit is 1242.

Syllable is used in some Persian TTS systems, too [26]. This language has only CV, CVC, and CVCC patterns for its syllables and so, its syllable number is limited to 4000 [26].

Kurdish has three main groups of syllables that are Asayi, Lekdraw, and Natewaw [36]. Asayi is most the important group and it includes most of the Kurdish syllables. In Lekdraw group, two consonant phonemes occur at the onset of syllable. For example, in /pshu/ two phonemes /p/ and /sh/ make a cluster and the syllable pattern is CCV. Finally, Natewaw group occurs seldom, too. Each group is divided into three groups, Suk, Pir, and Giran [36]. Table 5 shows these groups with corresponding patterns and appropriate examples.

According to Table 5, Kurdish has 9 syllable patterns; but two groups Lekdraw and Natewaw are used seldom and in practice, three patterns, CV, CVC, and CVCC are the most used patterns in Kurdish language. According to this fact, we can consider only Asayi group in implementations, and so the number of database syllables are less than 4000. In our system, we consider these syllables and extend our TTS system using them.

4 Diphone Based TTS System

Nowadays diphone is the most popular unit in synthesis systems. Diphones include a transition part from the first unit to the next unit, and so, have a more desirable quality rather than other units. Also, in some modern systems, a combination of this unit and other methods such as unit selection are used.

Kurdish has 37 phonemes, so in worst case, it has 37×36=1332 diphones. However, all of these combinations are not valid. For example, in Kurdish two phonemes /ع/ and /غ/ or /خ/ and /ح/ do not succeed each other immediately. Also, vowels do not form a cluster. So, the number of serviceable diphones in Kurdish is less than 1300.

After choosing the appropriate unit, we should choose the suitable instance for each unit. For this reason, we chose a proper word for each diphone and then extract its corresponded signal using COOL EDIT. Quality testing results are discussed in Section 5.

5 Quality Testing Results

In this paper, we have developed three synthesis systems based on allophone, syllable, and diphone. In order to assess the quality of the implemented systems and to have a comparison of them, a subjective test was conducted. A set of seven sentences was used as the test material. The test sets were played to 17 volunteer native listeners (5 female, and 12 male). The listeners were asked to rate system's intelligibility, naturalness and overall voice quality on a scale of 1 (bad) to 5 (good). The obtained test results are shown in Table 6.

The allophone based system has the worst quality and in practice, we cannot use it in all-purpose system. In this system we use only 66 most important allophones and so, we can improve its quality using more units in the database.

The syllable based system has intermediate overall quality and high intelligibility. In fact, syllable is a large unit, therefore, its prosody is constant and naturalness is intermediate. On the other hand, because of large size of this unit, this system intelligibility is high.

The diphone based TTS system has best quality between these three systems. Intelligibility and naturalness is high and overall quality is acceptable. Diphones include transition part between a specific phoneme and its next phoneme, so using this unit, we have smooth and pleasant output signal. Hence, diphone is most appropriate unit for developing an all purpose TTS system and in most modern TTS systems use of it as main unit.

Table 1. List of the proposed system standard letters

Arabic	غ	ع	ش	س	ژ	ز	ږ	ر	د	خ	ح	چ	ج	ت	پ	ب	ا	ډ
Latin	-	-	Ş	s	j	z	rr	r	d	x	-	ç	c	t	p	b	a	-
Standard	X	G	S	s	j	z	R	R	d	x	H	C	c	t	p	b	a	A

Arabic	ی	ی	ه	-	ێ	ھ	وو	ۆ	و	و	ن	م	ڵ	ل	گ	ک	ق	ڤ	ف
Latin	y	î	e	i	ê	h	û	o	U	w	n	m	ll	l	g	k	q	v	f
Standard	y	I	e	i	Y	h	U	o	U	w	n	m	L	l	g	k	q	v	f

Table 2. The same sentence in various scripts

Arabic Format	دلوپ دلوپ باران گول ئه نووسیته وه و نمه نمه یش چاوانم تو
Latin Format	Dillop dillop baran gull enûsêtewe û nime nimeyş çawanim to
Standard Format	diLop diLop baran guL AenUsYtewe U nime nimeyS Cawanim to

Table 3. List of phonemes and their corresponding allophones as used in the proposed system

Phoneme	P	b	t	d	K	g	Q	F	s	S	z	J	G	A
Allophones	PpO*&	bEB	t@T	d!D	k?K	G%g	Qq	FVf	s	$S	zZ>	Jj	^	A

Phoneme	C	h	H	m	X	X	n	v	y	l	r	L	R	Y
Allophones	Cc	h	H	mWM	X	X	nN	v	y	l	r	L	R	Y

Phoneme	E	a	N	U	u	o	w	I	i	C	Ü	.	,
Allophones	E	a	#	U	u	o	w	I	i	~	_	.	,

Table 4. A comparison between neural network output and desired output

NN Output	DiLo&_DiLoP_baraN_GuL_AenUsY@_U,_nime_nimeyS_~awaniM_to
Desired Output	DiLo&_DiLo&_baraN_GuL_AenUsY@_U,_nime_nimeyS_~awaniM_to

Table 5. Kurdish syllable patterns

		Suk	Pir	Giran
Asayi	Syllable Pattern	CV	CVC	CVCC
	Example	De, To	Waz, Lix	Kurt, Berd
Lekdraw	Syllable Pattern	CCV	CCVC	CCVCC
	Example	Bro, Chya	Bjar, Bzut	Xuast, Bnesht
Natewaw	Syllable Pattern	V	VC	VCC
	Example	-i	-an	-and

Table 6. Subjective testing results for various systems

	Intelligibility	Naturalness	Overall Quality
Allophone Based TTS System	2.71	2.31	2.45
Syllable Based TTS System	3.35	2.85	3.02
Diphone Based TTS System	3.9	3.37	3.51

References

1. Al-Muhtaseb, H., Elshafei, M., Al-Ghamdi, M.: Techniques for High Quality Arabic Speech Synthesis. In: Information sciences. Elsevier Press, Amsterdam (2002)
2. Styger, T., Keller, E.: Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts. In: Keller, E. (ed.) State of the Art, and Future Challenges Formant synthesis, pp. 109–128. John Wiley, Chichester (1994)
3. Klatt, D.H.: Software for a Cascade/Parallel Formant Synthesizer. Journal of the Acoustical Society of America 67, 971–995 (1980)
4. Hamza, W.: Arabic Speech Synthesis Using Large Speech Database. PhD. thesis, Cairo University, Electronics and Communications Engineering Department (2000)
5. Donovan, R.E.: Trainable Speech Synthesis. PhD. thesis, Cambridge University, Engineering Department (1996)
6. Lemmetty, S.: Review of Speech Synthesis Technology. M.Sc Thesis, Helsinki University of Technology, Department of Electrical and Communications Engineering (1999)
7. Youssef, A., et al.: An Arabic TTS System Based on the IBM Trainable Speech Synthesizer. In: Le traitement automatique de l'arabe, JEP–TALN 2004, Fès (2004)
8. Olive, J.P.: Rule synthesis of speech from diadic units. In: ICASSP, pp. 568–570 (1977)
9. Syrdal, A.: Development of a female voice for a concatenative text-to-speech synthesis system. Current Topics in Acoust. Res. 1, 169–181 (1994)
10. Olive, J., van Santen, J., Moebius, B., Shih, C.: Multilingual Text-to-Speech Synthesis: The Bell Labs Approach, pp. 191–228. Kluwer Academic Publishers, Norwell (1998)
11. Beutnagel, M., Conkie, A., Syrdal, A.K.: Diphone Synthesis using Unit Selection. In: The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis, ISCA (1998)
12. Sproat, R., Hu, J., Chen, H.: Emu: An e-mail preprocessor for text-to-speech. In: Proc. IEEE Workshop on Multimedia Signal Proc., pp. 239–244 (1998)
13. Wu, C.H., Chen, J.H.: Speech Activated Telephony Email Reader (SATER) Based on Speaker Verification and Text-to- Speech Conversion. IEEE Trans. Consumer Electronics 43(3), 707–716 (1997)
14. Black, A.: CHATR, Version 0.8, a generic speech synthesis, System documentation. ATR-Interpreting Telecommunications Laboratories, Kyoto, Japan (1996)

15. Hunt, A., Black, A.: Unit selection in a concatenative speech synthesis system using a large speech database. In: ICASSP, vol. 1, pp. 373–376 (1996)
16. Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y., Syrdal, A.: The AT&T NEXT-GEN TTS System. In: Joint Meeting of ASA, EAA, and DAGA (1999)
17. Dutoit, T.: High Quality Text-To-Speech Synthesis of the French Language. Ph.D. dissertation, submitted at the Faculté Polytechnique de Mons (1993)
18. Dutoit, T., et al.: The MBROLA project: towards a set of high quality speech synthesizers free of use of non commercial purposes. In: ICSLP 1996, Proceedings, Fourth International Conference, IEEE (1996)
19. Chouireb, F., Guerti, M., Naïl, M., Dimeh, Y.: Development of a Prosodic Database for Standard Arabic. Arabian Journal for Science and Engineering (2007)
20. Ramsay, A., Mansour, H.: Towards including prosody in a text-to-speech system for modern standard Arabic. In: Computer Speech & Language. Elsevier, Amsterdam (2008)
21. Amdal, I., Svendsen, T.: A Speech Synthesis Corpus for Norwegian. In: lrec 2006 (2006)
22. Yoon, K.: A prosodic phrasing model for a Korean text-to-speech synthesis system. In: Computer Speech & Language, Elsevier, Amsterdam (2006)
23. Zervas, P., Potamitis, I., Fakotakis, N., Kokkinakis, G.: A Greek TTS based on Non uniform unit concatenation and the utilization of Festival architecture. In: First Balkan Conference on Informatics, Thessalonica, Greece, pp. 662–668 (2003)
24. Farrohki, A., Ghaemmaghami, S., Sheikhan, M.: Estimation of Prosodic Information for Persian Text-to-Speech System Using a Recurrent Neural Network. In: ISCA, Speech Prosody 2004, International Conference (2004)
25. Namnabat, M., Homayunpoor, M.M.: Letter-to-Sound in Persian Language Using Multy Layer Perceptron Neural Network. Iranian Electrical and Computer Engineering Journal (2006) (in persian)
26. Abutalebi, H.R., Bijankhan, M.: Implementation of a Text-toSpeech System for Farsi Language. In: Sixth International Conference on Spoken Language Processing (2000)
27. Hendessi, F., Ghayoori, A., Gulliver, T.A.: A Speech Synthesizer for Persian Text Using a Neural Network with a Smooth Ergodic HMM. ACM Transactions on Asian Language Information Processing, TALIP (2005)
28. Daneshfar, f., Barkhoda, W., Azami, B.Z.: Implementation of a Text-to-Speech System for Kurdish Language. In: ICDT 2009, Colmar, France (2009)
29. Barkhoda, W., Daneshfar, F., Azami, B.Z.: Design and Implementation of a Kurdish TTS System Based on Allophones Using Neural Network. In: ISCEE 2008, Zanjan, Iran (2008) (in persian)
30. Thackston, W.M.: Sorani Kurdish: A Reference Grammar with Selected Reading. Iranian Studies at Harvard University, Harvard (2006)
31. Sejnowski, J.T., Rosenberg, R.: Parallel Networks that Learn to Pronounce English Text, pp. 145–168. The Johns Hopkins University, Complex Systems Inc. (1987)
32. Rokhzadi, A.: Kurdish Phonetics and Grammar. Tarfarnd press, Tehran (2000)
33. Deller, R.J., et al.: Discrete time processing of speech signals. John Wiley and Sons, Chichester (2000)
34. Kaveh, M.: Kurdish Linguistic and Grammar (Saqizi accent), 1st edn. Ehsan Press, Tehran (2005) (In Persian)
35. Karaali, O., et al.: A High Quality Text-to-Speech System Composed of Multiple Neural Networks. In: Invited paper, IEEE International Conference on Acoustics, Speech and Signal Processing, Seattle (1998)
36. Baban, S.: Phonology and Syllabication in Kurdish Language, 1st edn. Kurdish Academy Press, Arbil (2005) (In Kurdish)
37. Rao, M.N., Thomas, S., Nagarajan, T., Murthy, H.A.: Text-to-Speech Synthesis using syllable-like units. In: National Conference on Communication, India (2005)