Text to Speech synthesizer for Dzongkha Language

Achyut Nepal[1], Cheni Zangmo[2], Nidup Wangmo[3], Sangay Choden[4], Yeshi Wangchuk[5], Kamal Kr. Chapagai[6]

Department of Information and Technology [1,2,3,4,5] Department of Electronics and Communication Engineering[6], College of Science and Technology, Royal University of Bhutan.
* E-mail: [1]eit2011001@cst.edu.bt, [2]eit2011005@cst.edu.bt, [3]eit2011018@cst.edu.bt, [4]eit2011024@cst.edu.bt, [5]yeshi@cst.edu.bt, [6]kamal@cst.edu.bt

Abstract

A high quality speech synthesizer should be intelligent and produce natural speech. The quality of speech generated by Text to speech synthesizer also dependent on the amount of data used for training. This paper presents the development of Dzongkha TTS system using open source toolkit Hidden Marcov Model toolkit (HTK) and propose a method to increase the speech database for quality output. Every word in a language can be broken down into several phonemes which also means combination of phonemes would generate words. Therefore we suggest developing a corpus through phoneme concatenation which can increase the database for training TTS systems.

Key Words : NLP, TTS,Synthesizer, HMM, phoneme

1. INTRODUCTION

The automatic conversion of written to spoken language is commonly called Text-to-speech or simply TTS. If the input to a speech synthesizer is given with text, the system is called a text-to-speech (TTS) synthesizer. A text to speech synthesizer is a computer based system that can read text aloud automatically, regardless of whether the text is introduced by a computer input stream or a scanned input submitted to an Optical character recognition (OCR) engine. (Sasirekha and Chandra,2012). A text-to-speech synthesis system for a particular natural language is the technology for translating or converting a given typed or stored text input into its equivalent spoken waveform format. The concept of speech synthesis has been around for centuries, but only in recent decades has the process become available to the general public. It is the digitized audio rendering of computer text into speech. TTS software can read text from a document, Web page or e-Book, generating synthesized speech through a computer's speakers (Chauhan et al, 2011).

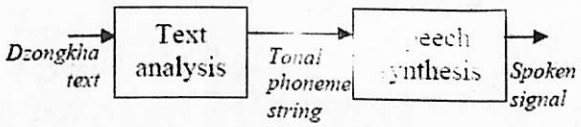


Fig 1:Generic TTS systems model (Chhoeden et al, 2011)

Dzongkha, the national language of Bhutan is widely spoken, however for citizens who are illiterate and cannot read and write in Dzongkha, it is necessary to have a system which will read out the written text. Various research has been carried out on Dzongkha TTS development. Dzongkha TTS prototype has been developed which is based on HMM (Sherpa et al). This paper presents TTS development for Dzongkha language using HTK toolkit version 2.0 and suggest a method to increase the speech database for better speech quality through concatenation.

The rest of this paper is organized as follows. Section 2 describes the need of TTS system. Section 3 focuses on literature review and section 4 describes the method we have followed and finally Section 5 presents conclusion and future scope.

1. The need for TTS System

Our language lacks behind in computerization as compared with other foreign languages due to late introduction of information technology & limited research carried out in the field of Natural Language Processing (NLP).

TTS system is primarily beneficial to illiterate people who are left inaccessible to information and resources available in the written form. It is also useful for people who are physically handicap and visually impaired. Using TTS system we can develop application like farm advisory, health advisory, and SMS readers. Moreover it can be used to promote safety driving, e-learning and education toys for kids.

2. The need for TTS System

Our language lacks behind in computerization as compared with other foreign languages due to late introduction of information technology & limited research carried out in the field of Natural Language Processing (NLP).

TTS system is primarily beneficial to illiterate people who are left inaccessible to information and resources available in the written form. It is also useful for people who are physically handicap and visually impaired. Using TTS system we can develop application like farm advisory, health advisory, and SMS readers. Moreover it can be used to promote safety driving, e-learning and education toys for kids.

3. Background and previous works

Different papers focusing on various complexity levels are reviewed to study the current state of art in TTS. We have found the following that is suitable for our Language. Chhoeden et al have described the development of advanced Dzongkha text-to-speech (TTS) system which is a marked improvement over the first Dzongkha TTS prototype using the Hidden Markov Model. For improving the quality of the synthesized Speech Advanced Natural Language Processing techniques like word segmentation and phrase boundary prediction were integrated with the earlier prototype (Chhoeden et al). Black A.W.(2006) said that the introduction of statistical parametric speech synthesis techniques has made it easier for building a voice in a language with fewer sentences and a smaller speech corpus According to Kishore et al and Rama et al(2006), developing a TTS system in a new language needs inputs for resolving language specific issues requiring close collaboration between linguists and technologists. Anumanchipalli et al said that corpus-based speech synthesis are very popular for its high quality and natural speech output. Alam et al built TTS using Festival by creating the voice data for festival and additionally extends festival using its embedded scheme scripting interface to incorporate Bangla language support. Their implementation uses two different kind of concatenative method supported in festival: unit selection and multisyn unit selection. S.P. Kishore et al have build a synthesis method using syllables as basic unit of concatenation and uses a database containing syllables along with other information such as Pitch, Duration and energy. The text to be synthesized is analyzed and broken into sequence of syllables to be selected from database and concatenated. Sproat R et al (2001) developed a taxonomy of NSWs on the basis of four rather distinct text types (news text, a recipes newsgroup, a hardware-product-specific newsgroup, and real-estate classified ads) and then investigated the application of several general techniques including n-gram language models, decision trees and weighted finite-state transducers to the range of NSW types. They demonstrated that a systematic treatment can lead to better results than have been obtained by the ad hoc treatments that have typically been used in the past.

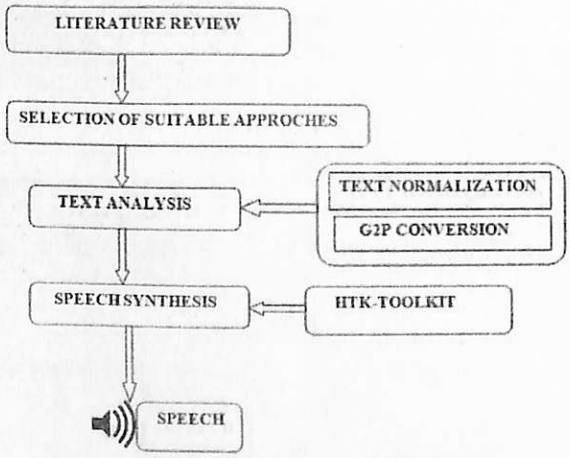
4. Methodology

This section describes the methodology adopted for our project. First we have done literature review on the existing TTS method and based on

their pros and cons we have selected Hidden Marcov model as our synthesizer. TTS consist of two phases, in the text analysis we have done the text normalization and grapheme to phoneme conversion. Text normalization basically consists of normalization of dates and numbers. G2P is created using dictionary based method where it contains the mapping of syllables and its corresponding pronunciation. The synthesis part is done using HTK toolkit version 2.0 as synthesizer.

The Hidden Markov Model Toolkit (HTK) is a portable toolkit for building and manipulating hidden Markov models. HTK is primarily used for speech recognition research, although it has been used for numerous other applications. HTK consists of a set of library modules and tools available in C source form. The tools provide sophisticated facilities for speech analysis, HMM training, testing, and results analysis. HTK was originally developed at the machine intelligence laboratory of Cambridge university engineering department (CUED) where it has been used to build CUED's large vocabulary speech recognition systems.

HMM-Based Synthesis is one of the widely applied methods in speech synthesis. HMM is a statistical model, which can be used for modeling the speech parameters extracted from a speech database, and then generating the parameters according to text input for creating the speech waveform. HMM-based speech synthesis systems



are able to produce speech in different speaking styles with different speaker characteristics and even emotions. They also benefit from better adaptability and clearly smaller memory requirement. However, the HMM-based TTS systems often suffer from degraded naturalness in quality compared to concatenative based speech synthesizers.

5. Phoneme Concatenation

The need for increasing the speech database is to improve the accuracy of speech. If we train the HTS toolkit with more data then the accuracy of speech will be better. So for that we have tried to concatenate the recorded phonemes which were provided by Dzongkha Development Commission (DDC).

We have 52 phonemes consisting of consonant vowels and diphthongs derived from the IPA table.

We have use matlab toolkit for concatenating the phonemes. Our main aim is to concatenate all of these phonemes and then generate word from it.

	Bilabial	Labio-velar	Dental	Postalveolar	Retroflex	Palatal	Velar	Glottal
Voiceless stop	p		t				k	ʔ
Aspirated stop	pʰ		tʰ				kʰ	
Voiced stop	b		d				g	
Voiced nasal	m		n			ɲ	ŋ	
Voiceless fricative			s	ʃ				h
Voiced fricative			z	ʒ				ɦ
Voiceless affricate			ʦ		ʧ	ʨ		
Aspirated affricate			ʦʰ		ʧʰ	ʨʰ		
Voiced affricate			ʣ		ʤ	ʥ		
Voiced approximant		w	ɹ			j		
Voiced lateral			l					
Voiceless lateral			l̥					

Table 1: Dzongkha consonant inventory

From the word we are going to generate the sentence which will be used for training HTS toolkit. For that first we should know the occurrences of each of these phonemes in order to form a word. Since we do not know the occurrences of each of these phonemes what we have done is like we tried to concatenate all 52 together and stored each of them in folder. In total we have 52x52 combinations of phonemes.

	Front		Central		Back	
	Unrounded	Rounded	Unrounded	Rounded	Unrounded	Rounded
Close	i	y				u
Close-mid	e	ø				o
Open-mid	ɛ					
Open			a			

Table 2: Dzongkha vowel inventory

6. Conclusion

The Development of Dzongkha TTS system work towards the computerization of Dzongkha which is syntactically a complex language. The project has been done in collaboration with DDC. This project has been accomplished within the specified scopes. The main area to be worked on is to increase the number of data for training the system.

For the future work we can increase the size of speech database in order to improve the accuracy of the speech. We can also develop web based interface so that Dzongkha TTS system can be available globally and also deve') an algorithm for Dzongkha pronunciation rule.

REFERENCES

Blake, A. et al (2007). Statistical parametric speech synthesis. proc. ICASSP, Honolulu, HI, vol IV, PP 1229-1232.

Blake, A et al (2001). The festival speech synthesis system: system documentation. University of Edinburgh.

Blake, A. & Taylor P., (1997).The festival speech synthesis system. Technical report HCRC/TR-83. University of Edinburgh. Scotland.

Chauhan, A. et.al. (2011). A text to speech system for Hindi using English language. IJCST vol. 2. Issue 3.

Chhoeden et.al. (2011). Pioneering Dzongkha text to speech synthesisChunku, C., & Rabgay, J. (2010), Building NLP resource for Dzongkha. A Tagset and A Tagged Corpus. Asian federation for Natural Language processing, 103-110.

Evermann, Young S. et.al (2009). The HTK Book. Microsoft Corpotation and Cambridge University Engineering Department.

Gopalakrishna, A. et.al. (n.d). Department of Indian Language Speech Database for large vocabulary speech recognition System.

Guo, Qing et. al (2010). High quality prosody generation in mandarin Text to Speech system. FujiTSu Sci.Tech Vol 46. No.1 pp. 40-46.

Hunt, A.J., & Blake, A.W., (1996). Unit selection in a concatenative speech synthesis system for a large database. Proceeding of IEEE int. conf. on Acoustic speech and signal processing, pp 373-376.

Kishore, S.P., et.al., (2002). A data driven synthesis approach for for Indian language using syllable as basic unit. International conference on National language pro sing (ICON). pp 311-316.

Liberman, M.Y., & Church, K.W., (n.d). Text analysis and word pronunciation in Text-to-speech synthesis. AT&T Bell Laboratory.

Rama, J et.al., (2002). A complete TTS system in Tamil. IEEE workshop on speech synthesis Sarkar, T., et. al. (2005). Building Bengali Voice using festvox. ICLSI Sasirekha, D., & Chandra, E., (2012). Text to Speech: A simple Tutorial. International journal of soft computing and engineering (IJSCE). Volume 2. Issue 1. ISSN 2231-2307.

Sproat, R., et.al (2001). Normalization of non-standard words. Computer speech and language. pp. 287-333.

Thinley, N., (n.d) Dzzongkha Segments and Tone. Dzongkha Development commission.