Pioneering Dzongkha Text-to-Speech Synthesis

Uden Sherpa, Dawa Pemo, Dechen Chhoeden
Department of Information Technology, Bhutan
{usherpa, dpemo, dchhoeden}@ dit.gov.bt

Anocha Rugchatjaroen,
Ausdang Thangthai, Chai Wutiwiwatchai
HLT Laboratory, National Electronics and Computer Technology Center, Thailand
{anocha.rug, ausdang.tha, chai.wut}
@nectec.or.th

Abstract

This paper describes a construction of Bhutanese or Dzongkha text-to-speech synthesis (TTS) using an HMM-based method. The procedure of creating the text-to-speech system consists of designing a phoneme set, building a text processing module, designing and collecting a speech database, training the HMM under the HMM-based Speech Synthesis System (HTS) toolkit framework, and integrating all components in a command prompt application. A simple text processor converts an input text to its corresponding phoneme sequence using a syllable pronunciation dictionary. A linguistics-driven decision tree is designed specifically for the Dzongkha phoneme set and is used in the HMM training step. A subjective test shows a 3.19 average mean opinion score compared to 3.93 given to natural speech.

1. Introduction

The national language of Bhutan is called Bhutanese or Dzongkha. Dzongkha and its dialects are native tongue of eight western districts of Bhutan which belong to Sino-Tibetan family of languages. The writing system of Dzongkha is very similar to Tibetan and so does the script. Moreover, the spoken form is different from Tibetan just only in the use of different set of vocabulary (George van Driem. 1992). In recent years in Bhutan, much effort has been put in to develop Dzongkha computing technology. A Dzongkha Linux operating system has already been launched under the first phase of the PAN Localization project[1]. Many sub-projects such as optical character recognition (OCR), Dzongkha lexicon, as well as text-to-speech synthesis (TTS), are on developing. TTS could be applied into many applications such as reading machines, talking web browser or talking dictionary. Figure 1 shows two sub-processes which are separated by "word". The left hand side says a path of changing a written signal to words, and the right encodes the words through phonemes to a speech signal (Paul Taylor. 2008).

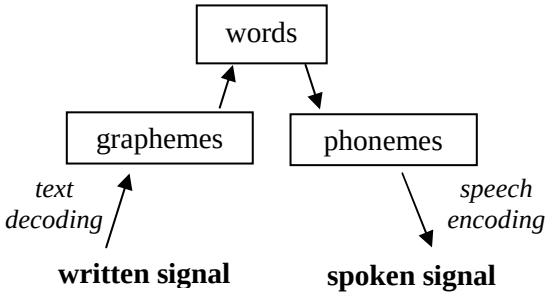


Figure 1: The common-form model of TTS (P. Taylor. 2008).

Totally, key features of this model are text analysis and speech synthesis. The task of analysis is to find the intermediate form (words or syllables) and the task of synthesis is to generate the speech signal from the form by any appropriate speech synthesizer. At the moment, two well-known approaches of synthesizer are presented in many publications, unit-selection based and Hidden Markov Model (HMM) based speech synthesis. In the way of HMM-based, a spoken signal is generated from acoustic parameters that are synthesized from context-dependent HMM models. Therefore, it gives a small footprint, stability, signal smoothness and speaker adaptability. This paper exploited the HMM-based Text-to-Speech toolkit (HTS) version 2.0[2] as a synthesizer which works correspondingly with MCEP (Mel-Cepstral Coefficients), log F0 and duration parameters.

The rest of this paper is organized as follows. Section 2 describes writing and sound systems of Dzongkha. Section 3 summarizes a phoneme design for Dzongkha grouped by consonants and vowels and also describes the tone in Dzongkha. Section 4 shows the Dzongkha TTS design and development. An evaluation and discussion is presented in Section 5, and a conclusion and future plan are given in the last section.

[1] PAN Localization Project, http://www.panl10n.net

[2] HMM-based Speech Synthesis System (HTS), *http://hts.sp.nitech.ac.jp/*, Nagoya, Japan.

2. Bhutanese Language

As the same family to Tibetan, Dzongkha writing is based on syllable units. Syllables are normally delimited by a delimiter called ‘tsheg’, while there is no inter-word space in Dzongkha. Each syllable contains a root letter (ming-zhi) and may additionally have any/or all of the following parts in the given order: a prefix, a head letter, a subjoined letter, a vowel, a suffix and a post-suffix, as shown in Figure 2.

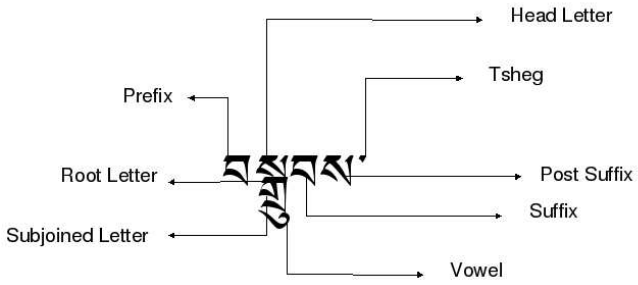


Figure 2: The writing system of a Dzongkha syllable.

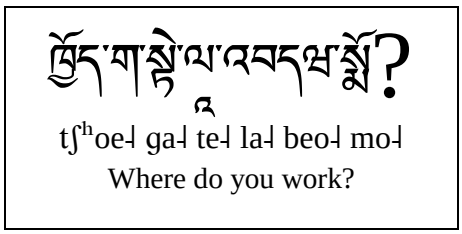


Figure 3: An example of Dzongkha script showing syllable delimiters ‘tsheg’ and a question marker at the end.

The **root letter** is the most important letter in a syllable which partly defines the initial consonant of the syllable.

The **head-letter** is stacked over root letter and works corporately to identify the initial consonant. Three letters, sounding 'r', 'l', and 's', can be the head-letter.

The **subjoined letter** is each of four letters, sounding 'y', 'r', 'l', and 'b' (written as 'w' but pronounced as 'b'). They can join a few root letters.

The **prefix** is unpronounced and located in the beginning of a syllable before the root letter. Though they are never pronounced, some of them modify the pronunciation of some root letters. There are five prefixes; sounding 'g', 'd', 'b', 'm' and 'h'.

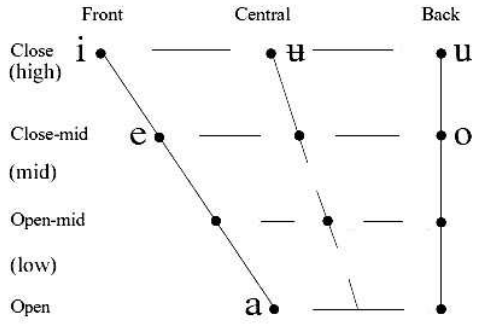
The **suffix** is a letter written after the root letter. Suffixes sounding 'hh' and 's' are never pronounced. The pronunciation of the rest of the suffixes depends on the combination with root letters. There are ten suffixes sounding 'g', 'ng', 'd', 'n', 'b', 'm', 'h', 'r', 'l' and 's'.

The **secondary-suffix** is a letter that comes after the suffix letter. It can also be called a post-suffix. The letters sounding 's' and 'd' are two possible post-suffixes. They are not pronounced at all in a syllable.

For ending a sentence or phrase, a ‘shed’ marker (།) or a question mark is placed. Figure 3 shows an example of Dzongkha script with syllable and sentence markers.

	Bilabial	Labio-velar	Alveolar	Palatal	Velar	Glottal
Stop	p pʰ b	t tʰ d			k kʰ g	ʔ
Nasal	m		n	ɲ	ŋ	
Fricative			s z	ʃ ʒ		
Approx.		w	ɹ	j		h
Affricative			ts tsʰ dzʰ	tʃ tʃʰ dʒ		

(a)



(b)

Figure 4: IPA tables for Dzongkha; (a) consonants and (b) vowels.

In the sound system, IPA tables of Dzongkha sounds are shown in Figure 4. Spoken Dzongkha is represented by initial consonants as shown in Figure 4 (a) and consonant clusters combining some single consonants, vowels as shown in Figure 4(b) and diphthongs. While pronouncing a single consonant, an inherent vowel sound 'a' is always attached. For example, ཀ is pronounced as 'ka' instead of only 'k'. Some vowels are modified when the root letter is combined with certain suffixes. While prefixes before the root letters are not pronounced, certain suffixes are pronounced along with root letters. There is a variety of a vowel combination or diphthongs that glide with a smooth movement of the tongue from one articulation to another. In this paper, ten diphthongs were defined for spoken Dzongkha. Moreover, spoken Dzongkha has four consonant clusters. They are written with

special letter conjuncts in which a letter is stacked over another (e.g, ག or 'r' is stacked under a root letter sounded 'g').

For the tonal system in Dzongkha, two tones were defined like Tibetan (Hu Tan. 1982). The low tone is common while the high tone is the modification of the low tone. Modification usually depends on following conditions:

- a combination of certain prefixes with a root letter,
- a combination of head letter with the root letter (superscripted),
- a subjoined letter stacked under a root letter (subscripted).

A few number of head letters and a few number of subjoined letters combined with particular root letters, will modify the tone of a syllable. Figure 5 demonstrates normalized F0 contours of the word 'lam' using the low tone (means “way” or “road”) and the high tone (means “lama” or “monk”).

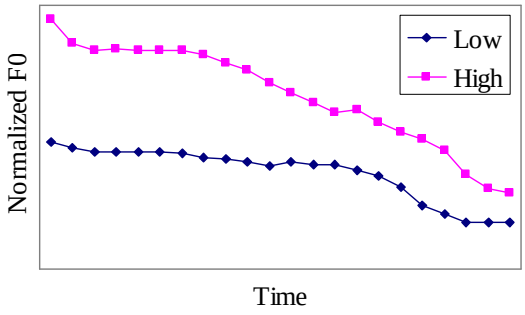


Figure 5: Normalized F0 contour of the syllable 'lam' in low and high tones.

3. Phoneme Design for TTS

In this section a phoneme set for Dzongkha is defined and grouped into initial consonants, initial consonant clusters, vowels, diphthongs, and final consonants. By observing the Dzongkha IPA table and trying to transcribe a sample set of Dzongkha scripts, 30 initial consonants, 5 initial consonant clusters, 10 vowels, 10 diphthongs, plus 2 tones were defined. Table 1 presents all phonemes defined for our TTS task. All single phonemes described in the Figure 4 including 30 consonants and 7 vowels were employed. 8 of 30 consonants can be placed as final consonants of syllables. 4 more vowels ('aa', 'ii', 'uu', and 'oo') were observed quantitatively different from the existing vowels and were separately defined. 5 consonant clusters, mostly combined with the 'r' sound, and 10 diphthongs were observed.

Specially, some vowel sounds are modified when combined with some certain suffixes. As shown in Table 2, suffixes 'n', 'r', 'd', and 's', connected to vowels 'a', 'u', and 'o', will change the vowel sounds to 'e', 'ue', and 'e', respectively. Suffixes 'd', 's', and 'hh' are not pronounced in syllables while the rest are all pronounced. These special characteristics are very crucial for making an accurate automatic G2P module, which is expected to be developed in the future.

Two tones presented in the Section 2 are given digit symbols '0' and '1' for low and high tones.

Table 1: Dzongkha phoneme inventory for TTS.

Type		**Symbol (IPA/Computerized)**
Initial consonant (*Ci*)	Single	k, k^h/kh, g/g, ŋ/ng, tʃ/c, $tʃ^h$/ch, dʒ/j, ɲ/ny, t, t^h/th, d, n, p, p^h/ph, b, m, ts, ts^h/tsh, dz, w, ʒ/zh, z, h^h/hh, j/y, ɹ/r, l, ʃ/sh, s, h, ʔ/@
	Cluster	dɹ/dr, tɹ, t^hɹ/thr, l^h/lhh, hɹ/hr
Vowel(*V*)	Single	a, i, u, e, o, ue, aː/aa, iː/ii, uː/uu, oː/oo
	Diphthong	ai, au, ae, ui, oi, ou, eu, ei, eo, iu
Final consonant (*Cf*)		g/g, ŋ/ng, n, b, m, ɹ/r, l, p
Tone (*T*)		˩/0, ˥/1

Table 2: Modification of vowels from suffix letters.

Vowel	**Suffix**										
	g	**ng**	**n**	**b**	**m**	**r**	**l**	**p**	**d**	**s**	**hh**
a			e n				e l		e	e	
i											
u			ue n				ue l		ue	ue	
e											
o			e n				e l		e	e	

4. TTS Design and Development

The proposed system contains 2 main modules, text analysis and speech synthesis. In this first phase, simply text analysis is proposed using dictionary based grapheme to phoneme (G2P) converter. The G2P module produces a phoneme string with tonal levels given an input Dzongkha text. Then, speech is synthesized correspondingly to the given phoneme sequence by using the HTS toolkit. Figure 6 shows an overall structure of the proposed system.

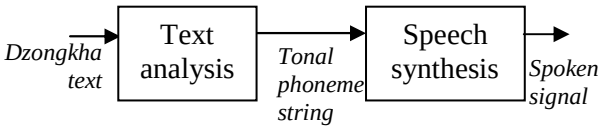


Figure 6: The proposed system structure.

Normally, G2P in the first part can be created using either rule-based or statistical approaches.

However, the presence of syllable-boundary marker in written scripts makes this process easier by using only a lookup dictionary of written syllables and their associated pronunciations. The dictionary was created by simply collecting distinct syllables from a Dzongkha text corpus. A text corpus of forty thousand sentences was collected from sample sentences in a Dzongkha dictionary and a large number of sentences in Dzongkha websites. The top 4,000 most frequently used syllables, appeared more than once in the text corpus, were included in the dictionary and transcribed into phoneme strings with tones.

For speech synthesis, a speech corpus was created for HMM training. Prompted sentences have to cover all 53 phonemes with 2 tones presented in the Table 1. This paper specifies those sentences by selecting from the 40,000-sentence text corpus using the criterion proposed in (C. Wutiwiwatchai. 2007). The idea is to gather iteratively a sentence having the most distinctive tonal-diphones (two connected tonal-phonemes) not presented in the selected set. The selection process is terminated when all tonal-diphones in the text corpus are all presented in the selected set. Finally, 509 sentences were selected and used for a speaker to speech recording. Recording was done in a controlled chamber with following conditions.

Table 3: Speech recording conditions.

Hardware	
Microphone	Plantronics monaural H-41
PC and sound card	Toshiba M600 (PM600 sound on board)
Software	
Recording software	Adobe Audition 1.5
Wave file format	16 kHz, 16 bits, Mono, PCM Wav
Speaking style	Fluent speaking

Table 4: shows Dzongkha speech corpus statistic.

No. of sentences	509
No. of syllables	5,404
No. of tonal diphones	6,048
No. of distinct tonal diphones	539

In the process of building the synthesizer, we extracted Mel-Cepstrum (MCEP), duration and Log Fundamental frequency (Log F0) parameters from each utterance in the speech corpus. Figure 7 shows an overall diagram of using the HTS toolkit for speech synthesis.

The HTS toolkit utilizes the Hidden Markov Model Toolkit (HTK) and the Speech Signal Processing Toolkit (SPTK). Its data format is mainly corresponding to the HTK format. Similar to those performed in the speech recognition area, HMMs could be trained in a flat-start fashion which requires no exact phoneme boundary tag, but only a phoneme transcription of each speech utterance. A clustering tree designed specifically for Dzongkha phonemes was used in HMM state tying. Tree questions are simply derived from the IPA table with additional questions regarding consonant clusters and diphthongs. Figure 8 shows a part of the clustering tree.

The HTS demonstration project provided in the toolkit website could be employed as a training script of the synthesizer. Also the "hts-engine" command provided in the toolkit could be simply used to synthesize speech given trained HMMs.

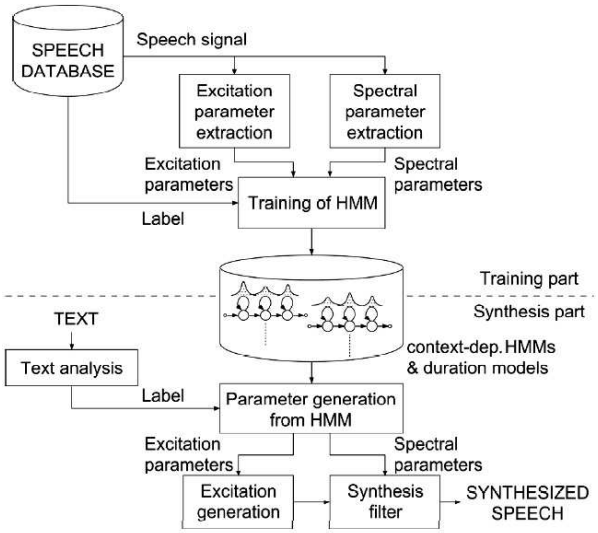


Figure 7: A diagram of HTS toolkit usage.

```
QS Left-InitialConsonants { "k_*","kh_*","ng_*", …
QS Left-FinalConsonants { "p^_*","t^_*","k^_*", …
QS Left-Voiced { "b_*","d_*","ng_*" }
QS Left-StopConsonants { "p_*","t_*","c_*", …
QS Left-Nasal { "m_*","n_*","h_*" }
QS Left-Fricative { "f_*","s_*" }
QS Left-Vowels { "a_*","aa_*","i_*","ii_*", …
QS Left-CloseVowels { "i_*","ii_*","v_*","vv_*" …
…
QS Right-InitialConsonants { "k_*","kh_*","ng_*", …
QS Right-FinalConsonants { "p^_*","t^_*","k^_*", …
…
```

Figure 8: A part of clustering tree used for HMM state tying.

5. Evaluation and Discussion

The evaluation was based on Mean Opinion Scoring (M. Viswanathan. 2005). Each speech utterance has was given a score 1 to 5, 1 for the

worst to 5 for the best. Fifteen Bhutanese evaluated 15 sentences which were shuffled between human and synthesized speeches. As a result, the human speech was rated 3.93 and synthesized speech was rated 3.19 in average. Incuriously, the synthesized speech is somewhat robotic and more work needs to be done to make it sound more natural.

In the future work, more speech utterances will be recorded. Enlarging the speech corpus makes several advantages including:

- larger coverage of diphone units, and so does the variety of phoneme context,
- a larger number of distinct syllables required by the G2P module,
- and broader prosodic phenomena required for making important prosody generation modules such as pausing between words and phrases, duration and F0 modeling.

6. Conclusion

This paper presented the overall procedure for pioneering a Dzongkha text-to-speech system, which consisted of designing a phoneme inventory set, building a text processing module, designing and creating a speech database, training HMM units under the HMM-based Speech Synthesis System (HTS) framework, and integrating all components in a command-prompt application. The experiment set up to subjectively compare synthesized speech and natural speech revealed the effectiveness of the overall designed system. Yet several modules are needed to be further developed in order to make synthesized speech more natural. Some of them include word and phrase boundary detection and better duration modeling.

7. Acknowledgements

This work has been partially supported by the PAN localization project. We would like to thank all DIT's volunteers who spent their valuable time on evaluating our synthesizer.

References

G. van Driem. 1992. The Grammar of Dzongkha. DDC.

H. Tan, A. Qu and L. Lin. 1982. Zangyu Lasahua Shengdiao Shiyan (Experimental studies on Lhasa Tibetan tone). Yuyan Yanjiu, 2: 18-38.

P. Taylor. 2008. Text-to-Speech Synthesis. Online: *http://svr-www.eng.cam.ac.uk/~pat40/book.html*. Cambridge.

M. Viswanathan. 2005. Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale. *Computer, Speech and Language19*. 55-83.

C. Wutiwiwatchai, A. Rugchatjaroen and S. Saychum. 2007. An Intensive Design of a Thai Speech Synthesis Corpus. *The Seventh International Symposium on Natural Language Processing*. Thailand.