Developing a Text to Speech System for Dzongkha

Yeshi Wangchuk[1*] Kamal K Chapagai[2] Pema Galey[1] Yeshi Jamtsho[1]
1.Information Technology Department, College of Science and Technology, Royal University of Bhutan, Rinchending, Phuentsholing, 21002, Bhutan
2.Electronics and Communication Engineering Department, College of Science and Technology, Royal University of Bhutan, Rinchending, Phuentsholing, 21002, Bhutan
* E-mail of the corresponding author: yeshiwangchuk.cst@rub.edu.bt

Abstract
Text to Speech plays a vital role in imparting information to the general population who have difficulty reading text but can understand spoken language. In Bhutan, many people fall in this category in adopting the national language 'Dzongkha' and system of such kind will have advantages in the community. In addition, the language will heighten its digital evolution in narrowing the digital gap. The same is more important in helping people with visual impairment. Text to speech systems are widely used in talking BOTs to news readers and announcement systems. This paper presents an attempt towards developing a working model of Text to Speech system for Dzongkha language. It also presents the development of a transcription or grapheme table for phonetic transcription from Dzongkha text to its equivalent phone set. The transcription tables for both consonants and vowels have been prepared in such a way that it facilitates better compatibility in computing. A total of 3000 sentences have been manually transcribed and recorded with a single male voice. The speech synthesis is based on a statistical method with concatenative speech generation on FESTIVAL platform. The model is generated using the two variants CLUSTERGEN and CLUNITS of the FESTIVAL speech tools FESTVOX. The development of system prototype is of the first kind for the Dzongkha language.

Keywords: Natural Language processing (NLP), Dzongkha, Text to speech (TTS) system, Statistical speech synthesis, phoneme, corpus, transcription
DOI: 10.7176/CEIS/12-1-04
Publication date: January 31st 2021

1. Introduction

Natural language processing is key to understanding the human language in depth as it facilitates better analysis and is also effective in performing analysis on a large amount of data in very little time. People can have access to a wide range of information through computerization and automation of a language (Isewon *et al.* 2014; Vyas & Virparia 2020). Text to speech system is a part of NLP where a string of text is processed by a system to generate speech signal. TTS is broadly divided into two parts, text processing and speech generation. While the two can be studied differently, the end result is combined.

Text processing and speech generation can be performed in two methods, with large database (Black & Hunt 1996) or statistical modelling method (Black 2007; Zen 2007). Database method requires more computation with more data set but gives better results even though it breaks when new words or sentences are provided (Jamtsho & Muneesawang 2020). Statistical method requires lesser data to create a model of the language and is dynamic in adapting to newer words and sentences. A good TTS system employs a combination of the two methods. In this paper, a method using a large corpus of Dzongkha text and speech, and using statistical method based on speech tools in FESTIVAL platform is used to analyse and extract features and synthesize speech. With the parameter generated from the text and speech, a phoneme concatenative method is used to get to the desired string of spoken words and sentences.

With TTS system, the application like farm advisory, health advisory, and SMS readers can be developed. In addition, the system can be used to promote safety driving, e-learning and education toys for kids.

2. Dzongkha: The National Language of Bhutan

Dzongkha is Bhutan's national language which is widely spoken by all Bhutanese across the country. It is also the official language of Bhutan. The language is a syllable-based language and derives its form and phonetic based on the 'Uchen' script of classical Tibetan through many centuries of independent linguistic evolution on Bhutanese soil (Sherpa *et al.* 2008). It also shares its phonetic base with many other languages like Tibetan, Hindi and Nepali. It has 30 consonants and five vowels. It also has three head letters (r, l, s) and four tail letters (r, l, s, w). The vowels (e, a, o) goes on top of the root letter while the vowel (u) goes at the tail of the root consonant. A syllable of Dzongkha can have consonants ranging from one to six.

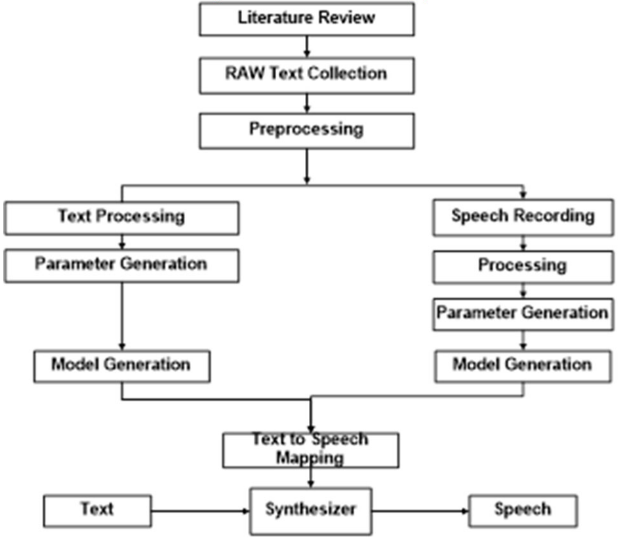
The consonants would be a pre-letter (Nyen-ju), a root letter (Min-Zhi), a head letter (go), a tail letter (tak), a post letter (Jen-Ju) and/or a post-post letter (Yan-Ju). All consonants forming the syllable do not play a role in the pronunciation. While the root letter in combination with the vowel is pronounced, the post letter determines if a vowel is long, short or glottal. Each syllable will have at least one and at most three phonemes. Each syllable is separated by a marker called "Tseg" with one or more syllables constituting a word. The sentence ends with a "shed". One difficulty in Dzongkha segmentation is that it does not have word limiter but a space in between phrases implies a pause while speaking or a statement marker. Thus Jamtsho & Muneesawang. (2020) have proposed syllable tagging method using deep learning for the efficient segmentation task. Further a complete reference of the language can be found in (Driem 1992; Dorji 1990).

Not much work has been done other than Sherpa *et al.* (2008); Chhoeden *et al.* (2011); Chungku & Rabgay (2010); Jamtsho & Muneesawang (2020) toward Computerization of the language. With Dzongkha key-in for different operating platforms, mails and browsers to Dzongkha dictionary by DDC, there is also work performed on HMM based TTS system (Sherpa *et al.* 2008) and analysis of phonetic set of Dzongkha (Chhoeden *et al.* 2011). POS tagged corpus has also been developed under the PAN project by Chungku & Rabgay (2010). Efforts are continuing in the improvement of computerized Dzongkha and this work is a step towards that and begins with development of basic dictionary of corpus.

3. Methodology

Text to Speech system is broadly divided into two parts, a text processing unit and a Speech generation unit. Figure 2 shows the methodology followed in the current work to produce TTS system for Dzongkha language.

Figure 2: Methodology



The raw texts were collected from different sources such as news, stories, regulation and guidelines, political and current affairs. A manual pre-processing of the collected Dzongkha texts was performed; breaking down the raw text into sentences, mapping the graphemes to phonemes by referring to the transcription table in the Table 4 and creating and indexing individual utterance to form a dataset for training and testing. The text normalization is also taken into consideration such as conversion of numbers, date and currency into its equivalent word form.

The transcribed and pre-processed text utterances along with the speech signal is fed to the speech tool of FESTIVAL system as there are no automatic text to phonemes (T2P) conversion system is available for the language. The FESTIVAL and FESTVOX tools are used to build the speech model and to generate the speech signal for the given input text.

4. Text Corpus and Analysis

A wide range of text documents are available online as well as in hard copies. But all are not usable as processing of these data requires the computer to understand the text and be able to analyse them. Therefore, raw texts were manually converted into unique sentences (utterances). A total of 3000 utterances were collected and pre-processed to give approximately 30,000 words with 2000 unique words. These utterances were manually transcribed to give meaning to the UTF-8 or UTF-16 characters.

4.1 Transcription table for Phonetic transcription

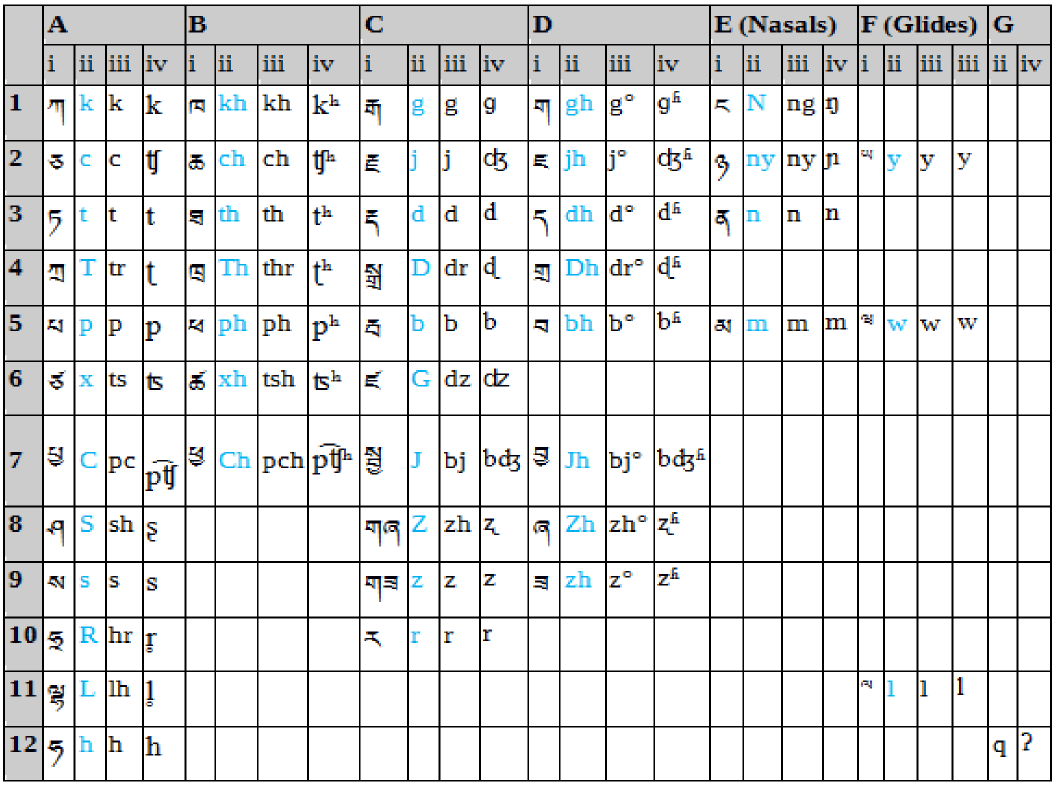
There are two methods of encoding or Romanising Dzongkha; transcription and transliteration as shown in Table 1. A transcription is a process of encoding Dzongkha text into Romanised form by way of how the word is pronounced, whereas transliteration is translation of character into its written form.

Table 1: Transcription and Transliteration of Dzongkha

Dzongkha	Transcription	Transliteration
རྒྱལ་པོ	gAp	rgyel po
སྦུལ	bU	Sbul
གཡུས	`U	gyus
ཞབས	Zap	zhabs

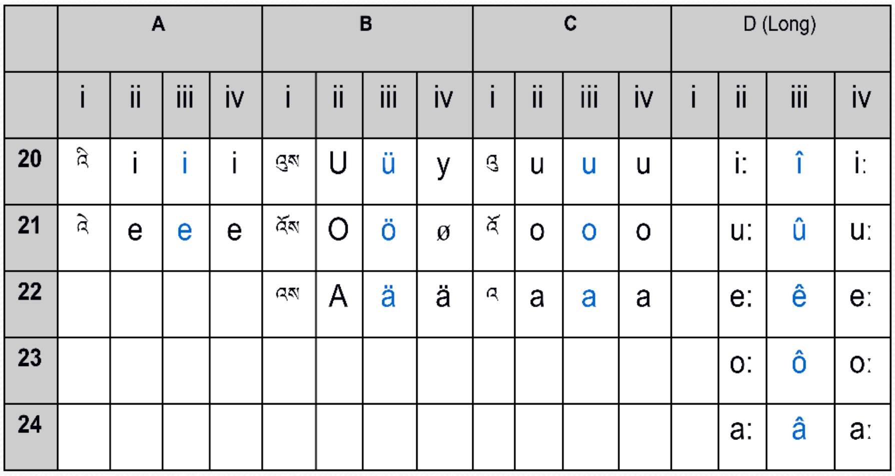
The transcription table has been modified from the Van Driem table by removing special characters to minimize conflict with programming languages; the number of characters to represent each phone is also decreased to optimize the computation. A transcription table for consonants and vowels are presented in Table 2 and 3 respectively. In the transcription table, there are four columns categorizing A, B, C, D, E, F, and G. The first column contains Script of Dzongkha, second column represents phone set that was used in transcribing the Dzongkha text, third column is a Van Driem grapheme, and forth column is the International Phonetic Alphabets (IPA) representation.

Table 2: Transcription table for consonants



	A				B				C				D				E (Nasals)				F (Glides)				G	
	i	ii	iii	iv	i	ii	iii	iv	i	ii	iii	iv	i	ii	iii	iv	i	ii	iii	iv	i	ii	iii	iii	ii	iv
1	ཀ	k	k	k	ཁ	kh	kh	kʰ	ག	g	g	ɡ	ག	gh	g°	ɡʱ	ང	N	ng	ŋ						
2	ཅ	c	c	ʧ	ཆ	ch	ch	ʧʰ	ཇ	j	j	ʤ	ཇ	jh	j°	ʤʱ	ཉ	ny	ny	ɲ	ཡ	y	y	y		
3	ཏ	t	t	t	ཐ	th	th	tʰ	ད	d	d	d	ད	dh	d°	dʱ	ན	n	n	n						
4	ཀྲ	T	tr	ʈ	ཁྲ	Th	thr	ʈʰ	སྒྲ	D	dr	ɖ	གྲ	Dh	dr°	ɖʱ										
5	པ	p	p	p	ཕ	ph	ph	pʰ	བ	b	b	b	བ	bh	b°	bʱ	མ	m	m	m	ཝ	w	w	w		
6	ཙ	x	ts	ʦ	ཚ	xh	tsh	ʦʰ	ཛ	G	dz	ʣ														
7	པྱ	C	pc	p͡ʧ	ཕྱ	Ch	pch	p͡ʧʰ	སྦྱ	J	bj	bʤ	བྱ	Jh	bj°	bʤʱ										
8	ཤ	S	sh	ʂ					གཞ	Z	zh	ʐ	ཞ	Zh	zh°	ʐʱ										
9	ས	s	s	s					གཟ	z	z	z	ཟ	zh	z°	zʱ										
10	ཧྲ	R	hr	r̥					ར	r	r	r														
11	ལྷ	L	lh	l̥																	ལ	l	l	l		
12	ཧ	h	h	h																					q	ʔ

Table 3: Transcription table for vowels



	A				B				C				D (Long)			
	i	ii	iii	iv	i	ii	iii	iv	i	ii	iii	iv	i	ii	iii	iv
20	ཨི	i	i	i	ཨུས	U	ü	y	ཨུ	u	u	u		i:	î	i:
21	ཨེ	e	e	e	ཨོས	O	ö	ø	ཨོ	o	o	o		u:	û	u:
22					འས	A	ä	ä	འ	a	a	a		e:	ê	e:
23														o:	ô	o:
24														a:	â	a:

The Dzongkha text is not explicitly used as the festival tools cannot understand Dzongkha. With reference to the transcription table, manual transcription of the 3000 sentences was performed and used as training and testing data of FESTIVAL speech tools. An example of transcription of Dzongkha text to its equivalent phone sets is shown in Table 4.

Table 4 Transcription of Dzongkha text into Romanized Phone sets

Dzongkha Text	རྒྱལ་ཁབ་ནང་གི་ སྲིད་དོན་ཚོགས་པ་ གསརཔ་ཚུ་གི་དོན་ལས་ ཤེས་ཚད་ཅན་དང་ ཉམས་མྱོང་ཅན་གྱི་ འདེམས་ངོ་འཚོལ་ཐོབ་ནི་འདི་ གདོང་ལེན་ཅན་ཅིག་སྦེ་ ཡོདཔ་ཨིན་མས།
Transcribed Text	gA khap naN ghi `si dhOn xhok pa sa:p xhu ghi dOn lA Se: xhe cen dhaN nyam yoN cen ghi dhem No xhOl thop ni dhi dhoN len cen ci be yOp im mA

4.2 Parser development

Keying-in of Dzongkha has not been implemented yet, and the phone set developed here is different than the phone set in English. Hence, a Perl script is written for mapping Dzongkha phone sets and generates all the unique set of phones and their combinations from the 3000 utterances. An error check is performed to compare the list of phone sets present in the text file and the phone set in the transcription table. Errors are manually rectified to provide correct transcribed text data for training of Festival/Festvox system.

A total of 57 unique phone sets including combinations of consonants and vowels were found. These phones sets match with the number of phone set in the transcription table.

5. Data Processing

Two flavours of FESTVOX tool were used to develop TTS system: one based on unit selection (CLUNITS) by Black *et al.* (2014) and the other based on statistical parametric approach (CLUSTERGEN) by Black *et al.* (2014). An automatic matching of text and speech was performed using EHMM and the resulting alignments (phone labels) were used to build catalogue (in case of CLUNITS) and model (in case of CLUSTERGEN).

5.1 Text Processing

Of the 3527 utterances, 1177 utterances did align to the audio signal making them usable data set. These utterances are formatted in a text file with the file name txt.done.data. The file containing utterances are shown below:

(dzonkha_0001 " gA khap naN ghi si dhOn xhok pa sa:p xhu ghi dOn lA Se: xhe cen dhaN nyam nyoN cen ghi dhem No xhOl thop ni dhi dhoN len cen ci be yOp im mA ")

(dzonkha_0002 " Duk gA khap naN lu lo Na dhe ci ghi naN khO lu ra U phoN ghi ka NAl dhi Che: dhe ci ghi ma phap be dhe yOp iN ")

(dzonkha_0003 " na: bha mi waN Du gel sum pa cho ghi ThuN kar dhU chen im lA gA khap naN ghi ZuN ghi yi xha dhaN lobh JoN pe khaN dhe lA lop Da xhu NA so bhe Za na:m iN ")

…

(dzonkha_1177 " te nyi:m dhi lop Da xhu naN lu lO bO ghi nyi:m bhe yaN xi: suN bhe dho yOp dha nyi:m dhi kha lA rim dhi xhu gha ra lop Thu xhu ghi bhe gho Dhen thap te lO bO xhu lu ka Din sam dho yOp im mA ")

The actual transcribed texts are embedded within the pair of double quotes where there is a space after opening quote and before closing quote. There is also a space after opening parenthesis '(' and before closing parenthesis ')'. For these utterances, annotation corrections are performed manually to eliminate diphthong, wrong usage of phone(s) and errors during transcription. Finally, these utterances were fed into the system for building the voice.

5.2 Speech Processing

The speech of a single speaker in a monotone voice was recorded in a quiet room with a portable device namely Zoom H4nSP with 16bit, PCM. The speech of each recorded utterances is processed with FESTVOX tool to create word level and phoneme level datasets, and then their parameters were extracted.

6. Building a Unit Selection and Statistical Voice

The steps to develop cluster units and statistical parametric synthesizer are adapted from (Black *et al.* 2014). Prior to building the voice, the prompt file is created. The file named "txt.done.data" contains prompts formatted in a specific format as can be read by the FESTVOX tool. The same format is applied to rename the recorded utterances. The parametric analysis extracts both F0 parameters and MCEP features. The F0 parameters and MCEP parameters

are combined into a single parameter file for each utterance in the database.
The training and test set are generated next. For every 10th prompt in the test set, the other 9 are in the training set.

7. Result Discussion

The TTS system built for Dzongkha language can accept the input in transcribe text and generate the output in voice. The input data are fed manually and the generated output voices have been tested initially by the core working team. The voice tested with the unit selection (CLUNITS) is more natural and intelligent than the statistical generated one (CLUSTERGEN) when the data set is less. However, the large data set will determine the actual naturalness and intelligence of the synthesized voice by using these approaches.

Another approach of evaluation of the intelligent and naturalness of the output is tested and verified by a team of researchers from DDC (Dzongkha Development Commission), Bhutan. The DDC is a government organization which looks after the development of national language (i.e. Dzongkha) in the country. The team were given to evaluate and identify the person's voice after listening to the system generated speech. All members have been able to identify the person's voice with the good rating of naturalness. Hence, the System called Text to Speech version 1 (TTv1) has been released and the system with the input data are handed over to the DDC.

8. Future Work

With the development of TTS system using both variants; CLUNITS and CLUSTERGEN of Festival tool, the first TTS synthesizer of Dzongkha has been achieved. Since the pre-processing of text is tedious and time consuming, the Text to Phonemes (T2P) conversion module is of priority in order to minimize the effort and errors, and to accelerate the accuracy of transcription. Other fields of recommendation to improve and enhance the system are to develop the user interface integrating T2P module, Word segmentation, Normalization, and automatic annotation correction.

Another scope of the future work is to; increase the dataset in order to produce better result, enhance the phone set by incorporating diphthongs and pitch accents, aligning the text and the audio signal to improve acoustic modelling and suitably post processing of the speech signals by noting glitches. The approach of the Mean Opinions Score Test (MOST) can be conducted by a group of native Dzongkha speaker to evaluate the naturalness and intelligent of the system.

References

Black, A. & Hunt, A. (1996), "Unit selection in a concatenative speech synthesis system using a large speech database", *Proceedings of ICASSP'96,* 373–376.

Black, A. & Lenzo, K. (2014), "Building Synthetic Voices" available at: http://festvox.org/festvox/c3174.html.

Black, A. & Lenzo, K. (2014), "Building Synthetic Voices", available at: http://festvox.org/festvox/c2645.html#AEN2665.

Black, A., Zen, H. & Tokuda, K. (2007), "Statistical parametric speech synthesis", *Proceedings of ICASSP'07,* 1229–1232.

Chhoeden, D., Sherpa, U., Pemo, D. & Choejey, P. (2011), "Dzongkha Phonetic Set Description and Pronunciation Rules," Conference on Human Language Technology for Development, Alexandria, Egypt, 2-5 May 2011.

Chungku, C. & Rabgay, J. (2010), "Building NLP resources for Dzongkha:A Tagset and A Tagged Corpus", *Proceedings of the 8th Workshop on Asian Language Resources'10*, 103–110, Beijing, China, 21-22 August 2010.

Dorji, S. (1990), "Dzongkha Rabsel Lamzang (Dzongkha Phrasebook)", Dzongkha Development Commission, Royal Government of Bhutan, 1st Edition, ASIN B004S65BJC.

Driem, V. (1992), "A Grammar of Dzongkha", Dzongkha Development Commission, Royal Government of Bhutan.

Isewon, I., Oyelade, J. & Oladipupo, O. (2014), "Design and Implementation of Text to Speech Conversion for Visually Impaired People", International Journal of Applied Information Systems (IJAIS), Foundation of Computer Science, New York, USA, Volume 7-No. 2, April 2014.

Jamtsho, Y., & Muneesawang, P. Dzongkha Word Segmentation using Deep Learning. In 2020 12th International Conference on Knowledge and Smart Technology (KST) (pp. 1-5). IEEE.

Sherpa, U., Pemo, D. & Chhoeden, D. (2008), "Pioneering Dzongkha Text-to-Speech Synthesis," available at: https://www.dit.gov.bt/sites/default/files/OCOCOSDA2008_final.pdf

Vyas, H. A., & Virparia, P. V. (2020). Template-Based Transliteration of Braille Character to Gujarati Text—The Application. In Rising Threats in Expert Applications and Solutions (pp. 437-446). Springer, Singapore.

Zen, H., Nose, T., Yamagishi, J, Sako, S., Masuko, T., Black, A.W. & Tokuda, K. (2007), "The HMM-based Speech Synthesis System (HTS) Version 2.0", 6th ISCA Workshop on Speech Synthesis, Bonn, Germany, August 22-24, 2007.