Phonetically balanced Bangla speech corpus

S.M. Murtoza Habib[1] **Firoj Alam**[1] **Rabia Sultana**[1] **Shammur Absar Chowdhur**[1,2] **Mumit Khan**[1,2]

{murtoza, firojalam, ummi.ulab.bu}@gmail.com,
{shammur, mumit}@bracu.ac.bd

[1] Center for Research on Bangla Language Processing, BRAC University
[2] Department of Computer Science and Engineering, BRAC University

Abstract

This paper describes the development of a phonetically balanced Bangla speech corpus. Construction of speech applications such as text to speech and speech recognition requires a phonetically balanced speech database in order to obtain a natural output. Here we elicited text collection procedure, text normalization, G2P[1] conversion and optimal text selection using a greedy selection method and hand pruning.

Index Terms — Phonetics, Balanced corpus, Speech Synthesis, Speech Recognition.

1 Introduction

The goal of this study is to develop a phonetically balanced Bangla speech corpus for Bangladeshi Bangla. With nearly 200 million native speakers, Bangla (exonym: Bengali) is one of the most widely spoken languages of the world (it is ranked between four[2] and seven[3] based on the number of speakers). However, this is one of the most under-resourced languages which lack speech applications. Some sparse work has been done on speech applications of this language. General speech applications such as speech synthesis and speech recognition system require a phonetically balanced speech corpus. One of the great motivations of this work is to develop such a corpus, which will in turn help the people to develop speech synthesis and speech recognition applications. However, corpus designing is a long and tedious task and therefore optimization is necessary, since recording every possible speech unit is not pragmatic. The "phonetic coverage" (Santen and Buchsbaum, 1997) is appropriate when we say phonetically balanced corpus. This coverage can be defined by the concept of phonetic unit. Sentence containing phonetic units according to their frequency of occurrence in a given language is called "phonetically balanced sentence" (Dafydd et al., 2000). Consequently, the corpus containing the phonetically balanced sentences is called "phonetically balanced corpus". The phonetic units can be phone, biphone, triphone, or syllable. Many studies (Kominek and Black, 2004), (S. Kasuriya et al., 2003) and (Patcharika et al., 2002) showed that, the use of biphone as the basic unit for speech corpus is feasible during text selection process. Therefore we wanted to develop a phonetically balanced speech corpus based on biphone coverage using the widely used greedy selection algorithm (Santen and Buchsbaum, 1997) (François and Boëffard, 2002) (François and Boëffard, 2001). The study of allophones and other spoken realizations such as phonotactic constraint are beyond the scope of this paper.

The work presented here is the design, construction, and characterization of text selection procedure optimized in terms of phonetic coverage. We hope that this work will be the preliminary step to develop a speech corpus which will provide necessary resources for speech synthesis and speech recognition of Bangla. Due to the lack of this resource, experts in this field have been unable to develop speech applications for this language.
A brief literature review is given in section 2, followed by a description of the development procedure in section 3. The analytical results are presented and discussed in section 4, where we have also presented a comparison between our technique and the technique used in arctic Komi-

[1] Grapheme to Phoneme

[2] http://www2.ignatius.edu/faculty/turner/languages.htm, Last accessed April, 2011.

[3] http://en.wikipedia.org/wiki/List_of_languages_by_total_speakers, Last accessed April, 2011.

nek and Black, (2004) database. A summary and conclusion of the study are given in section 5.

2 Literature review

It is already well established that the success of speech research mostly depends on speech corpora. Since speech corpora contain the variation of real phenomena of speech utterances, thus we are able to analyze the phenomena from speech corpus. Speech related research such as phonetic research, acoustic model and intonation model can be drawn from a speech corpus. Research on speech synthesis shows great improvement on the quality and naturalness of synthesized speech which adapted speech corpus (Kominek and Black, 2004) (Gibbon, 1997). Likewise, the development of a speech recognition system largely depends on speech corpus. The inspiration of this work came from (Kominek and Black, 2004) (Fisher et al., 1986) (Yoshida et al., 2002) (Patcharika et al., 2002) (Radová and Vopálka, 1999) where significant amount of work have done for different languages. There is a claim by LDC-IL[4] that, a phonetically balanced Bangla speech corpus is available. Besides that, CDAC[5] has also developed speech corpora which are publicly available through the web. There is a speech corpus publicly available for Bangladeshi Bangla - CRBLP[6] speech corpus (Firoj et al., 2010), which is a read speech corpus. The CRBLP speech corpus contains nine categories of speech but it is not phonetically balanced. Such categories are Magazines, Novels, Legal documents (Child), History (Dhaka, Bangladesh, Language movement, 7th March), Blogs (interview), Novel (Rupaly Dip), Editorials (prothom-alo) and Constitution of Bangladesh. However, to the best of our knowledge there is no published account. In addition, due to the differences in the writing style as well as the phonetic structure between Indian and Bangladeshi Bangla we have decided to compile a phonetically balanced Bangladeshi Bangla speech corpus based on phone and biphone coverage, which will be the first of its kind for Bangladeshi Bangla.

3 Development procedure

This section presents the methodology of text selection procedure for phonetically balanced speech corpus. We measured the phonetic coverage based on biphone in the phonetically transcribed database. In addition to phonetic coverage, we have also tried to maintain a prosodic and syntactic variation in the corpus for future works. Some sparse work has been done on phontactic constraint of Bangla. This is one of the obstacles to define an optimized biphone list. Therefore, we used the whole list of biphones in this study. Here, optimized means phonetically constrained biphones need to be omitted from the list. The system diagram of the whole development process is shown in figure 1.

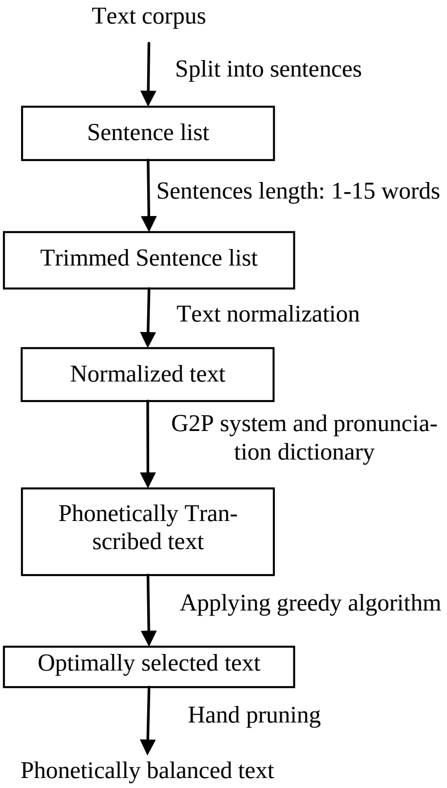
Text corpus
Split into sentences
Sentence list
Sentences length: 1-15 words
Trimmed Sentence list
Text normalization
Normalized text
G2P system and pronunciation dictionary
Phonetically Transcribed text
Applying greedy algorithm
Optimally selected text
Hand pruning
Phonetically balanced text

Figure 1: System diagram of phonetically balanced text selection

3.1 Text collection and normalization

Text selection from various domains is one of the most frequently used techniques for development of a speech corpus. However, it is one of the most time consuming phase, since a lot of manual work needs to be done, such as selecting different categories of text, proof reading and manual correction after text normalization. Therefore some constrains were considered dur-

[4] http://www.ldcil.org/

[5] http://www.kolkatacdac.in/html/txttospeeh/corpora/corpora_main/MainB.html

[6] CRBLP - Center for Research on Bangla Language Processing

ing text selection. The text was collected from two different sources such as prothom-alo news corpus (Yeasir et al., 2006) and CRBLP speech corpora (Firoj et al., 2010). The prothom-alo news corpus has 39 categories of text such as – general news, national news, international news, sports, special feature, editorial, sub-editorial and so on. Table 1 shows the frequency analysis of the initial text corpus.

Corpus	Sentences	Total tokens	Token type
Prothom-alo news corpus	2,514,085	31,158,189	529,620
CRBLP read speech corpus	10,896	1,06,303	17,797
Total:	**2,524,981**	**31,264,492**	**547,417**

Table 1: Frequency distribution of the corpus

Starting with our initial text corpus consisting ~31 millions words and ~2.5 millions sentences we used a python script to split the corpus into sentences based on punctuation marks such as ?, । and !. It is observed that the length of some sentences is too long i.e. more than 15 words, even more than 25 words. Study of Kominek and Black, (2004) explained and our recording experience claimed that, sentences longer than 15 words are difficult to read aloud without making a mistake. With respect to the length, we maintained the constraints and filtered out sentences that are not between 1-20 words. Table 2 shows the frequency analysis of the corpus after the filtering.

Corpus	Sentences	Total tokens	Token type
Prothom-alo news corpus	2,014,032	21,177,137	487,158
CRBLP read speech corpus	9,130	68,306	13,270
Total:	**2,023,162**	**21,245,443**	**500,428**

Table 2: Frequency distribution of the corpus after filtering

The text contains large number of non-standard words (NSW) (Sproat et al., 2008) such as number, date and phone number which need to be normalized to get full form of the NSW's. It is then normalized using a text normalization tool Firoj et al., (2009). There are some ambiguous NSW tokens such as year-number and time-floating point number. For example: (i). the token ১৯৯৯ (1999) could be considered as a year and at the same time it could be considered as number and (ii). the token ১২. ৮০ (12.80) could be considered as a floating point number or it could be considered as a time. In case of these ambiguous tokens the accuracy of the tool is 87% (Firoj et al., 2009). The 13% error was solved in the final text selection procedure. On the other hand, the accuracy rate of non-ambiguous token is 100% such as date (e.g: ০২-০৬- ২০০৬), range (e.g: ১০- ১২), ratio (e.g: ১/ ২), roman (e.g: I, II,), ordinal number (e.g: ১ম, ২য়, ৩য়) and so on.

Example of token	Normalized form
১২১	একশত একুশ
১ম	প্রথম
০২৯৫৬৭৪৪৭	শুন্য দুই নয় পাঁচ ছয় সাত চার চার সাত

3.2 Phoneme set

Defining the phoneme set is important for a phonetically balanced corpus. We considered the biphone as a unit for phonetically balanced corpus. The phoneme inventory used in this study is the one found in Firoj et al., (2008 a) and Firoj et al., (2008 b). The phoneme inventory consists of 30 consonants and 35 vowels including diphthongs. Since a biphone is the combination of two phones and Bangla phone inventory has 65 phones (Firoj et al., 2008 a) (Firoj et al., 2008 b), so the total number of biphones consist 65X65 = 4225. However, all biphones would not belong to the language in terms of phonotactic constraints.

Since no notable work has been done on phonetic constraint and as it is beyond our scope, we have not optimized the biphone list in this work.

3.3 Phonetic transcription

The phonetically transcribed text is needed to represent the phonetic coverage. Therefore, the text has to be phonetized so that the distribution of the phonetic unit can be analyzed. To perform phonetic transcription each sentence is tokenized based on 'white space' character. Then each word is passed through the CRBLP pronunciation lexicon (2009) and a G2P (Grapheme to Phoneme) converter (Ayesha et al., 2006). The system first checks the word in the lexicon, if the word is not available in the lexicon then it is passed through the G2P system for phonetic transcription. The CRBLP pronunciation lexicon contains 93K entries and the accuracy of the G2P system is 89%. So there could be errors in phonetic transcription due to the low accuracy rate of G2P system which is unavoidable. Manual correction of every word is not practical so a decision had been made that this problem would be solved in the "hand pruning" stage. In phonetic transcription we used IPA[7], since IPA has been standardized as an internationally accepted transcription system. A phonetic measurement has been conducted after the text has been phonetically transcribed. The phonetic measurement of phone, biphone and triphone is shown in table 3.

Pattern type	Unique	Total in the corpus
Phones	65	119,068,607 (~119 millions)
Biphones	3,277	47,360,819 (~47 millions)
Triphones	274,625	115,048,711 (~115 millions)

Table 3: Phonetic measurement of the corpus

Though the corpus contains all the phones, it does not cover all the biphones. There could be several reasons:

1. Trimming the main sentence list to 20 words per sentence.
2. The frequency of these missing biphones is too low in the spoken space of this language.

3.4 Balanced Corpus Design

The greedy selection algorithm (Santen and Buchsbaum, 1997) has been used in many studies of the corpus design. This is an optimization technique for constructing a subset of sentences from a large set of sentences to cover the largest unit space with the smallest number of sentences. Prior to the selection process, the target space i.e. biphone is defined by the unit definition, mainly the feature space of a unit. A detail of the algorithm is shown below:

Algorithm

```
Step 1: Generate a unique biphone list from the corpus.
Step 2: Calculate frequency of the biphone in the list form the corpus.
Step 3: Calculate weight of each biphone in the list where weight of a biphone is inverse of the frequency.
Step 4: Calculate a score for every sentence. The sentence score is defined by the equation (1).
Step 5: Select the highest scored sentence.
Step 6: Delete the selected sentence from the corpus.
Step 7: Delete all the biphones found in the selected sentence from the biphone list.
Step 8: Repeat from Step 2 to 6 until the biphone list is empty.
```

$$Score = \left(\sum_{i}^{Np} \left\{ \frac{1}{Pfi} \right\} \right)$$

Equation 1: Sentence score

SC - sentence score

N_p - the number of phonemes in each sentence

Pf_i - the i^th phoneme frequency of the sentence in the corpus.

This algorithm successively selects sentences. The first sentence is the sentence, which cover largest biphone count.

Our text corpus contains ~2 millions sentences with 47 millions biphones. Based on experiment, it took 26 hours 44 mins of CPU time in a Core i5 due 2.4 GHz PC equipped with 3 GB memory to run the greedy selection process.

It is observed that, the results of automatic selection are not ideal due to accuracy rate of text normalizer and G2P system. So a hand pruning (Kominek and Black, 2004) is required. A visual inspection was made by considering several criteria such as awkward grammar, confusable homographs and hard to pronounce words. Next, the phonetically transcribed text was visually inspected, as our text normalization and G2P system produced some errors. Finally, a phonetically balanced text corpus was developed.

4 Recording

The next issue is the recording, which relates to selecting speaker, recording environment and recording instrument. Since speaker choice is perhaps one of the most vital areas for recording so a careful measure was taken. A female speaker was chosen who is a professional speaker and aged 29.

As far as recording conditions are concerned, we tried to maintain as high quality as possible. The recording of the utterances was done using the Nundo speech processing software. A professional voice recording studio was chosen to record the utterances. The equipment consisted of an integrated Tascam TM-D4000 Digital-Mixer, a high fidelity noise free Audiotechnica microphone and two high quality multimedia speaker systems. The voice talent was asked to keep a distance of 10-12 inches from the microphone. Optionally a pop filter was used between

[7] IPA- International Phonetic alphabet

the speaker and the microphone to reduce the force of air puffs from bilabial plosive and other strongly released stops. The speech data was digitized at a sample rate 44.1 kHz, sample width 24-bit resolution and stored as wave format. After each recording, the moderator checked for any misleading pronunciation during the recording, and if so, the affected utterances were re-recorded.

There were a few challenges in the recording. First, speaker was asked to keep the speaking style consistent. Second, speaker was supervised to keep the same tone in the recording. Since speaking styles vary in different sessions, monitoring was required to maintain the consistency. To keep the consistency of the speaking style, in addition to Zhu et al. [29] recommendation the following specifications were maintained:

1. Recording were done in the same time slot in every session i.e 9.00 am to 1.00 pm.
2. A 5 minutes break was maintained after each 10 minutes recording.
3. Consistent volume of sound.
4. Normal intonation was maintained without any emotion.
5. Accurate pronunciation.

Pre-recorded voice with appropriate speaking style was used as a reference. In each session, speaker was asked to adjust his speaking style according to the reference voice.

5 Annotation

The un-cleaned recorded data was around 2 hours 5 minutes and it had a lot of repetition of the utterances. So in annotation, the recorded wave was cleaned manually using a proprietary software wavlab which tends to reduce the recorded data to 1 hours 11 minutes. Then, it was labeled (annotated) based on id using praat (Firoj et al., 2010). Praat provides a textgrid file which contains labels (in our case it is wave id) along with start and end time for each label. A separate praat script was written to split the whole wave into individual wave based on id with start and end time. We used id instead of text in labeling.

The structure of the corpus was constructed in a hierarchical organization using the XML standard. The file contains meta-data followed by data. The metadata contains recording protocol, speaker profile, text, annotation and spoken content. The data contains sentences with id, orthographic form, phonetic form and wave id.

6 Results

It was our desire to design a speech corpus which will exhibit good biphone coverage. The information of phonetic coverage of the corpus is shown in table 4.

Pattern type	**No of unique units**	**Total found in the corpus (unique)**	**Coverage**
Phones	65	65	100%
Biphones	4225	3,277	77.56%
Triphones	274,625	13911	5.06%

Table 4: Phonetic coverage information of the corpus

The percentages for biphone and triphone coverage are based on a simple combination. Thus the number of possible biphone is 4,225 and triphones is 274,625. This corpus covers nearly 100% phone and 77.56% biphone. A natural speech synthesizer achieved with 80% coverage of biphone as shown in Arctic of Kominek and Black, (2004). So it is hoped that better speech applications could be achievable by using this corpus.

We have also performed a frequency analysis of the phonemes on the whole phonetic corpus. And during the analysis, we observed an interesting phenomenon about the missing 22.44% biphone. That is, the phonemes whose frequency is less frequent (<0.11%) in the phonetic corpus, and surprisingly their combination came in the missing 22.44% biphone list. Observation also says that this combination basically came from diphthongs and a few nasals. So based on our empirical study we can claim that our findings of speech corpus are balanced and it would cover all of our everyday spoken space. This analysis leads to another assumption that this missing biphone list could be part of phonotactic constraint of Bangla language. This means that this combination possibly, may never occur in the spoken space of Bangla language. An effort needs to be done about the missing diphone using dictionary words and linguistic experts.

6.1 Comparison between festvox and our approach

We have made a comparison between the techniques that is used in festvox “nice utterance selection” (Kominek and Black, 2004)(Black and Lenzo, 2000) and the approach used in S. Kasuriya et al., (2003) and Patcharika et al., (2002) that we followed in this study. The difference between these two approaches is that festvox technique uses most frequent words to select the text in the first step. The rest of the steps are the same in both approaches. We experimented festvox nice utterance selection tool (*text2utts*) using our data and got the result that is shown in table 5. A comparison between two techniques is shown in table 5. According to Firoj et al., (2008 a) and Firoj et al., (2008 b) Bangla has 65 phonemes including vowels and consonants. This leads to 65X65 = 4225 biphones in Bangla. So in selecting corpus, a corpus would be best when it covers maximum number of biphones. So in festvox approach, the selected corpus covers 1,495 biphones which is 35.38% of the whole biphone set. On the other hand in our approach we got 3277 biphones of the selected text which results 77.56% coverage.

Here, in fextvox experiment we used most frequent 50,000 words in the first step to select the text. This approach limits the festvox *text2utts* tool to select the maximum number of utterances.

Pattern	**Festvox approach**	**Approach used in this study**
No. of sentences	677	977
No. of biphones	26,274	70,030
No. of phones	26,914	71,007
Biphone coverage	35.38% (1495 biphones)	77.56% (3276 biphones)
Phone coverage	84.61% (55 phones)	100% (65 phones)

Table 5: Comparison between festvox and our technique

7 Conclusion and future remarks

In this paper we presented the development of a phonetically balanced Bangla speech corpus. This speech corpus contains 977 sentences with 77.56% biphone coverage. It needs more sentences to cover all biphones. To do that, more text corpora may be required. However, finding out all biphones is pragmatically impossible due to the linguistic diversity and phonotactic constraint of a language. Besides, a significant amount of effort is needed to be able to use this resource in real speech applications. The efforts include recording voices by number of male and female voice talents for speech synthesis in a professional recording environment. Speech recognition application requires more recording data in different environments which includes recording the voice by huge number (>50) of male and female voice talents.

References

A. W. Black and K. Lenzo, 2000. *Building voices in the Festival speech synthesis system,* http://festvox.org/bsv.

Ayesha Binte Mosaddeque, Naushad UzZaman and Mumit Khan, 2006. *Rule based Automated Pronunciation Generator* , Proc. of 9th International Conference on Computer and Information Technology (ICCIT 2006), Dhaka, Bangladesh, December

C. W. Patcharika, C. Wutiwiwatchai, P. Cotsomrong, and S. Suebvisai, 2002. *Phonetically distributed continuous speech corpus for thai language,* Available online at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi =10.1.1.1.7778

CRBLP pronunciation lexicon, 2009. CRBLP, Available: http://crblp.bracu.ac.bd/demo/PL/

Dafydd Gibbon, Inge Mertins, Roger K. Moore, 2000. *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation* (The Springer International Series in Engineering and Computer Science), Springer, August 31,

Firoj Alam , S. M. Murtoza Habib and Professor Mumit Khan, 2008 a. *Acoustic Analysis of Bangla Consonants* , Proc. Spoken Language Technologies for Under-resourced language (SLTU'08), Vietnam, May 5-7, page 108-113.

Firoj Alam, S .M. Murtoza Habib, and Mumit Khan, 2008 b. *Research Report on Acoustic Analysis of Bangla Vowel Inventory* , Center for Research on Bangla Language Processing, BRAC University,

Firoj Alam, S. M. Murtoza Habib and Mumit Khan, 2009. *Text Normalization System for Bangla* , Conference on Language and Technology 2009 (CLT09), NUCES, Lahore, Pakistan, January 22-24,

Firoj Alam, S. M. Murtoza Habib, Dil Afroza Sultana and Mumit Khan, 2010. *Development of Annotated Bangla Speech Corpora* , Spoken Language Technologies for Under-resourced language (SLTU'10), Universiti Sains Malaysia, Penang, Malasia, May 3 - 5,

Fisher, William M.; Doddington, George R. and Goudie Marshall, Kathleen M. 1986. *The DARPA Speech Recognition Research Database: Specifications and Status* , Proceedings of DARPA Workshop on Speech Recognition. pp. 93–99.

François, H. and Boëffard, O., 2002. *The Greedy Algorithm and its Application to the Construction of* *a Continuous Speech Database* , Proc. of LREC, Las Palmas de Gran Canaria, Spain,

François, H. and Boëffard, O., 2001. *Design of an Optimal Continuous Speech Database for Text-To-Speech Synthesis Considered as a Set Covering Problem* , Proc. of Eurospeech, Aalborg, Denmark,

Gibbon, D., Moore, R., Winski, R. (eds.), 1997. *Handbook of Standards and Resources for Spoken Language Systems* , Mouton de Gruyter, Berlin New York

J. Kominek and A. Black, 2004. *The cmu arctic speech databases* , 5th ISCA Speech Synthesis Workshop, pp. 223-224,

S. Kasuriya, V. Sornlertlamvanich, P. Cotsomrong, S. Kanokphara, and N. Thatphithakkul, 2003. *Thai speech corpus for thai speech recognition* , The Oriental COCOSDA 2003, pp. 54-61. [Online]. Available online at: http://www.tcllab.org/virach/paper/virach/colips2004_final.rtf

Sproat R., Black A., Chen S., Kumar S., Ostendorf M., and Richards C, 2008. *Normalization of Non-Standard Words: WS'99 Final Report* , CLSP Summer Workshop, Johns Hopkins University, 1999, Retrieved (June, 1, 2008). Available: www.clsp.jhu.edu/ws99/projects/normal

V. Radová and P. Vopálka, 1999. *Methods of sentences selection for read-speech corpus design* , in TSD '99: Proceedings of the Second International Workshop on Text, Speech and Dialogue. London, UK: Springer-Verlag, pp. 165-170. [Online]. Available: http://portal.acm.org/citation.cfm?id=720594

Van Santen, J P. H. and Buchsbaum, A. L., 1997. *Methods for optimal text selection* , Proc. of Eurospeech, p. 553-556, Rhodes, Greece,

Yeasir Arafat, Md. Zahurul Islam and Mumit Khan, 2006. *Analysis and Observations From a Bangla news corpus* , Proc. of 9th International Conference on Computer and Information Technology (ICCIT 2006), Dhaka, Bangladesh, December

Yoshida Yoshio, Fukuroya Takeo, Takezawa Toshiyuki, 2002. *ATR Speech Database* , Proceedings of the Annual Conference of JSAI, VOL.16th, 124-125, Japan