CREATION OF CONNECTED WORD SPEECH CORPUS FOR BANGLA SPEECH RECOGNITION SYSTEMS

ABSTRACT

A new speech corpus of connected Bangla words derived from newspapers text corpus BdNC01. This has been designed for various research activities related to speaker-independent Bangla speech recognition. The database consists of speech of 100 speakers, each of them speaking 52 sentences as connected words for training database. Another 50 new speakers were employed to speak all the list of speech to construct a test database. Every utterance was repeated 5 times in various days to avoid time variation of speaker property. A total of 62 hours of recording makes the corpus largest in its type, size and application area. This paper describes the motivation for the corpus and the processes undertaken in its construction. The paper concludes with the usability of the corpus.

Keywords: Bangla, Speech Corpus, Vocabulary, Connected Word, BdNC01.

1. INTRODUCTION:

The construction of standard speech database is an important prerequisite for contemporary research activities in speech recognition and understanding. A continuous growth of research efforts in corpus-based speech and natural language processing techniques among the researchers is seen in various institutes over the world. Since corpus-based methods are found as very efficient in most language and speech systems, the reliability of using these methods are also increasing with progressive development of language technology. In recent decades systems development using language technology has been uplifted by various laboratories, industries and also by governments in almost all influential languages. But the history of

corpus generation and corpus based Bangla speech recognition are not so far and limited within past few years. However Bangla is one of the influential languages spoken by about 260 million people around the world and is 8th most popular language. Till date among the few creations of Bangla speech corpora, probably the first step was taken by the Center for Development of Advanced Computing (CDAC) of India by creating Bangla Katha Bhandar released in 2005 [1]. It was a collection of Annotated Speech Corpus for Bangla. Another step of similar work was done by the Center for Research on Bangla Language Processing in BRAC University of Bangladesh in 2010 [2]. In between these two, a research project financed by the MOSICT of Bangladesh was completed in June, 2008. Under this project a large scale speech corpora were recorded in SIPL of Islamic University [3]. The distinction of The SIPL speech corpora from other two is that it was designed especially for Bangla speech recognition. As the continuation of the project results organizing, labeling and similar other processing are still ongoing. A couple of articles have published [4, 5] and several others are in processing on these ongoing works. This paper describes the design and development processes of the connected word speech corpus. After the basics of speech corpora, a brief description of BdNC01 text corpus has been discussed to understand the selection of words for speech database design. In the next subsections, speech recording, editing processes and final outcome are discussed. The paper concludes with the usability of the corpus.

2. SPEECH CORPUS FUNDAMENTALS

Corpus is a collection of written text or recorded speech of a language to discover the units and relation among the units of the language. Modern corpora are collected and stored in electronic form for efficient statistical analysis using software tools. Corpora are collected according to some external criteria to represent a language or language variety so that it can be used as a source of data for linguistic research [6]. Speech corpus is created by audio files and text transcriptions in a structured database that can be used to train automated system

which can then be used as a part of speech recognition engine [7]. Speech Corpora may be classified in two types as below:

1. Read Speech - This includes part of Books, Newspaper contents, Broadcast news, Lists of words and numbers etc.

2. Spontaneous Speech - This includes naturally occurred dialogs between two or more people, Narratives such as a person telling a story, Class lectures and discussions such as two people try to find a common meeting time based on individual schedules.

There are also some special kinds of speech corpora such as non-native speech databases that contain speech with foreign accent or dialect database.

The Speech corpus is frequently used as the basis for analyzing the characteristics of speech signal and the result of analysis then become useful for developing speech generation and recognition systems. The speech corpora are growing more complicated and larger in size day by day. This is because the computation power is increasing and various robust methods are developing in speech technology. One of the selection methods of speech content of a corpus is to use the analytical result from a text corpus. For example, a speech corpus of British English WSJCAM0 has been recorded at Cambridge University from the Wall Street Journal text corpus [8]. An important step before recording a speech corpus is to select popular words such that it becomes a representative vocabulary of the language in consideration. Since each unknown word causes an average recognition error usually between 1.5 and 2 [9]. Therefore the recognizer vocabulary is usually designed with the goal of maximizing lexical coverage for the expected input. A most popular approach is to choose the most frequent words from a text corpus which means that the reliable vocabulary is highly dependent upon the representativeness of the training data [10].

The Influential parameters to categorize a speech recognition system are speech types, speaker dependency, vocabulary size, etc. The importance of these parameters is context

sensitive. It depends on the design considerations of a recognition system to be used for a specific application or task [11]. There are three types of speech usually feed to the speech recognition systems. These are isolated, connected, or continuous speech. Isolated speech requires a significant pause between words, may be 250 milliseconds. In isolated speech system, one speech file may contain an utterance of a single word or a short string of several isolated words. In continuous speech recognition systems, continuous speech flows with a rhythm and the words are overlapped each other thus making recognition harder. In between these two, connected speech recognizers do not require the intermediate pause between inputs, but are able to detect word boundaries within a string of connected speech. However it requires careful utterance of each word like a dictation. Though many relevant literatures describe connected words and continuous words as alternative terms, but because of vast diversity of application it is required to define connected words separately. In fact the way to classify ``connected words'' and ``continuous speech'' is somewhat technical. A connected word recognizer uses words as recognition units, which can be trained in an isolated word mode. Specific and efficient applications of connected word recognizers are found in dictation and voice command recognition. Speech recognition systems can be classified further as either speaker-dependent or speaker-independent systems. In speaker-dependent systems, each speaker enters several samples of each word of expected vocabulary to form the reference templates [12]. Other important parameter to design a speech corpus is the vocabulary size. The words small, medium and large are usually applicable to vocabulary sizes of the order of 100, 1000 and (over) 5000 words, respectively. A typical small vocabulary recognizer can recognize only ten digits and a typical large vocabulary recognition system can recognize 20000 words [11]. A limited capability automatic dictation machine was proposed by Gould, Conti, and Hovanyecz [12] in 1983. They named the machine as listening typewriter. The machine was simulated by letter writing task with

isolated and connected speech databases using various vocabulary sizes. In their experiment the performance of the voice recognizer were estimated for a 1000 word vocabulary and various unlimited vocabulary. The 1000 word vocabulary was composed of the 1000 high frequent English words. The conclusion of the work indicated that roughly 75% of the words used in the letter writing task were available in the 1000 word vocabulary. Therefore in dictation and voice command recognition medium size vocabulary may be estimated enough for satisfactory performance.

3. BdNC01 CORPUS AND DATABASE DESIGN

BdNC01 corpus [13] is a text corpus collected from web editions of several influential Bangla newspapers during 2005-2011. BdNC01 contains a large amount of Bangla text including more than 11 million word tokens. As a requirement of this work, a program was developed using C Language to parse and sort the text in BdNC01 corpus. The resulting output of the program was a list of words with their frequency of occurrence in the text. The objective of this processing was to select a list of high frequent 1000 or more words so that it becomes a good representative of the language in consideration to construct a significant connected speech database. A part of the list is shown in Table-1 and top frequent 1000 words were selected to find some practical Bangla sentences. From three issues of daily newspapers picked randomly, 52 sentences were selected such that they include high frequent words as above. The list of sentences was accepted for a small-medium vocabulary speech database and includes 252 different words in 343 places. The special characteristic of this list is that some words are in multiple places with different context.

Table-1: Words are organized with Frequencies count

Words	Frequencies
ও	150919
করে	98271
এ	85107
থেকে	69838
করা	68858
না	67453
হয়	63512
এবং	59724
হয়েছে	52701
হবে	48345
জন্য	47501
এই	43988
বলেন	40123
করতে	35992
একটি	34197
করেন	34172
এক	32298
সঙ্গে	29507
হয়ে	29490
মধ্যে	29223

4. SPEECH ACQUISITION

The First step in this level is to select good speakers. A notice was published among the departments of Islamic University inviting speakers in this regard. A large amount of interested students both male and female applied to do the job. An audition was arranged to check their comparative efficiency of correct utterance/pronunciation. Depending on their performance, 75 male and 75 female speakers were selected to finalize the speaker list. The major content of the list was the students of language and computer related departments. The selected speakers were very young with age range of 18-25 years. Finally the selected speakers were attended a two day workshop. The objective of the workshop was to concern

the speakers about the theory, methods and work plan of the project. The speakers were given practical training of speech acquisition such as headset setting, loudness of utterance etc.

Speech data was recorded in Laboratory environment by a close-talking directly connected to the computer. The speakers were asked to read the text in standard pronunciation as well as possible. The whole recording was done in the presence of the researcher who controlled the recording. The control includes running the recording script, starting the recording session, breaking the recording, playing back the recorded speech, re-recording the sentence if required and moving to the next sentence. After the entire session for a reader was finished, all the utterances were listened to by the reader and the researcher for corrections. Furthermore, as only 8-10 speakers were scheduled to read per day, the researcher was able to listen again to the recorded speech at the end of each day to check its quality. There were some console operators to help the researcher. Each operator was on charge of a workstation and very careful to identify the hesitation or erroneous utterances. In such conditions, he was asked the speaker to repeat the utterance until the speaker made correct utterance. The recorded speech data were taken with 8 kHz sampling with 8 bit quantization. The recorded speech was stored as *wav* file format in various lengths depending on the speaker's ability to speak over a length of continuous time. Connected Words were uttered with a minimum pause between two consecutive words. The speakers were very much conscious about the context dependence of word utterances. The final result was that the speaker speaks normally but with a word-short pause-word style which ensure a minimum separation between two consecutive words. Depending on the speaker's ability, the total text was recorded and stored as *wav* file format in two or three primary unedited files. The speakers were trained and directed continuously to maintain the similar rhythm in text reading to ensure the same utterance of text.

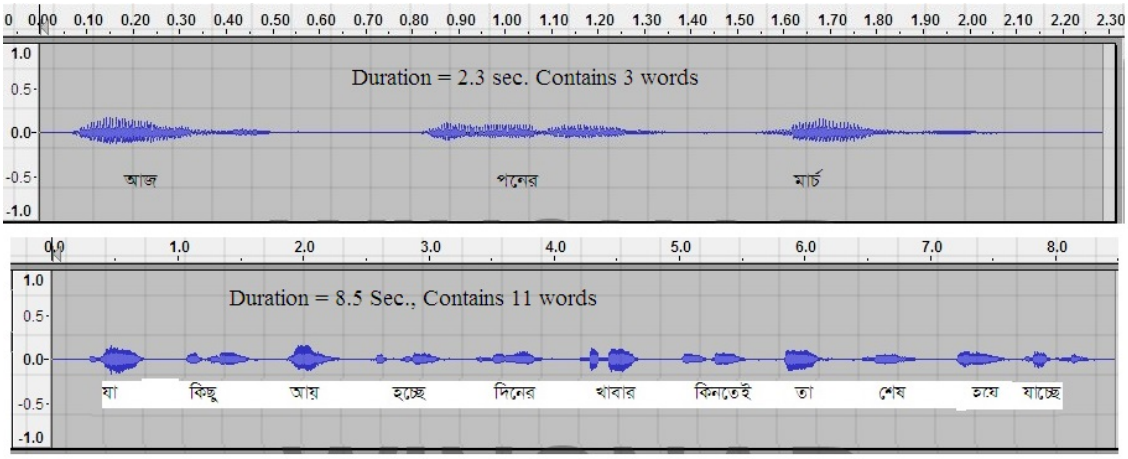

Figure-1: Two edited and finalized files with three (minimum) and eleven (maximum) words.

5. SPEECH EDITING AND LABELLING

It was necessary to check the recorded data from the following points: difference between the utterance and the utterance list, degree of dialectal accent, speech rate, clarity of pronunciation, recording level, noise, etc. There are some speakers with very low speech level and it was possible to magnify the amplitude of such speech data to a suitable level. However, the recording was carried out in an environment, which was not truly noiseless. The recoding instruments were also produced a little noise in some cases. All types of problems were identified and corrected during editing phase. Noiseless clean speech files were separated from the noisy speech files. Noisy files were cleaned using various filters and tagged with a comment. The original unedited recorded files were edited so that each file contains one complete sentence. As the HMM Toolkit developed by Cambridge University Engineering Department has already proved its efficiency by being used frequently by most of the research workers, the database was labeled by following the specified format of speech data, this made it ready for use in HMM Toolkit for evaluation [14]. Two edited and finalized files are shown in figure-1. As shown in the figure minimum length file contains 3 words with duration of 2.3 seconds and maximum length file contains 11 words with duration of 8.5

seconds. The estimated average time required for each file was 5.4 seconds. The total recording time for connected words database was about 62 hours after editing and labeling.

6. RESULTING CORPUS

Four types of speech corpus resulting from the text corpus were recorded and the summary of the developed speech corpus are given below table-2.

Table-2: Corpus Description

Contents	Vocabulary	No. of words in each file	No. of Training Files	No. of Test Files	Total no. of Files	Total recording time (Approx.)
Connected Words	252	3-11	26000	15600	41600	62 Hours

Test Speakers: 50 Male and 50 Female Students of 18-25 ages

Training Speakers: 25 Male and 25 Female Students of 18-25 ages

Recording environment: Recorded in a laboratory environment with a close talking microphone.

Recording media: Recorded on computer HDD (8 kHz sampling, 8 bit quantization), Copied later into CD's and DVD's for distribution.

7. DISCUSSION AND CONCLUSION

One of the advantages of newspaper corpus is that it reflects the current tradition of a language. Therefore the speech database with most frequent words from BdNC01 corpus is reasonably representative and covered the current tradition of Bangla language uses. Before recording a speech corpus, careful selection of vocabulary is important to maximize the lexical coverage from the expected input language. A straightforward approach is to choose the N most frequent words in the training data and from section-2, it may be concluded that 1000 most frequent words may be estimated enough for satisfactory performance in dictation and voice command recognition systems. From section-3, 52 sentences were selected such

that the corpus includes words only from 1000 high frequent words of BdNC01 text corpus. A standard ASR system is based on a set of acoustic models that link the observed features of the voice signal to the expected phonetics of the hypothesis sentence. The most typical implementation of this process is probabilistic, namely Hidden Markov Models (HMM) [15]. The evaluation of constructed speech database is imminent as it is formatted to use with HMM toolkit and the necessary next steps is ongoing. With the best of our knowledge these are the first speech corpus in Bangla language in its size, type and coverage. The achievement from this work will construct a fundamental base in speech recognition research in Bangla especially in dictation and command processing.

REFERENCES

1. Center for Development of Advanced Computing, India, April, 2005, available at: http://www.cdac.in/html/press/2q05/prs_rl165.aspx **.** retrieved on 24th January, 2011.
2. Firoj Alam, S. M. Murtoza Habib, Dil Afroza Sultana and Mumit Khan, "Development of Annotated Bangla Speech Corpora", Spoken Language Technologies for Under-resourced language (SLTU'10), Universiti Sains Malaysia, Penang, Malasia, May 3 - 5, 2010.
3. Md. Farukuzzaman Khan, Md. Babul Islam and Md. Mizanur Rahman, Construction and Analysis of Large-Scale Bangla Corpus for Bangla Speech Recognition, Project Report, Ministry of Science and Information and Communication, Bangladesh, 2008.
4. Md. Farukuzzaman Khan and M. Abdus Sobhan, "*Language Modeling for Error Identification in Bangla Speech Recognition*", Islamic University Studies: Journal of Applied Science and Technology, Vol. 8 No. 2, December 2011, ISSN 2218-841X.
5. Md. Farukuzzaman Khan, M. Abdus Sobhan, "Construction of Large Scale Isolated Word Speech Corpus in Bangla", Global Journal of Computer Science and Technology, Volume XVIII, Issue II, Version I, Page No: 21-26, ISSN: 0975-4172, 2018.
6. John Sinclair, Corpus and Text: Basic Principle, Tuscan Word Center, 2004, http:// www.ahds.ca.uk/litangling, retrieved on 18th Jan., 2011.
7. Wikipedia, Speech Corpus, available at: http://en.wikipedia.org/wiki/Speech_corpus, retrieved on 18th Jan., 2011.

8. Tony Robinson, Jeroen Fransen, DavidPye, JonathanFoote and Steve Renals, "WSJCAM0: A British English Speech Corpus For Large Vocabulary Continuous Speech Recognition", In Proc. of ICASSP 95, Detroit, Michigan, 1995.
9. Lean-Lac Gauvain and Lori Lamel, "Large Vocabulary Speech Recognition Based on Statistical Methods", Pattern recognition in speech and language processing, CRC press, New York, USA, 2003.
10. D.S. Pallett, J.G. Fiscus, W.M. Fisher, J.S. Garofolo, B.A. Lund, A.F. Martin and M.A. Przybocki, "1994 Benchmark Tests for the ARPA Spoken Language Program, Proc. ARPA Spoken Language System Technology Workshop, 5-36, Austin, TX, January 1995.
11. Dafydd Gibbon, Roger Moore, Richard Winski, "Handbook of Standards and Resources for Spoken Language Systems", Walter de Gruyter Publishers, Berlin & New York, 1997. Retrieved on 20th October, 2017. http://wwwhomes.uni-bielefeld.de/gibbon/Handbooks/gibbon_handbook_1997/node303.html
12. Sherry P. Casali , "The Effects of Recognition Accuracy and Vocabulary Size of A Speech Recognition System on Task Performance and User Acceptance" , M.Sc. Thesis, Virginia Polytechnic Institute and State University, pp. 5, May 1988. Retrieved on 20th October, 2017. https://vtechworks.lib.vt.edu/bitstream/handle/10919/43383/LD5655.V855_1988.C382.pdf
13. Md. Farukuzzaman Khan, Afraza Ferdousi , M. Abdus Sobhan ,"Creation and Analysis of a New Bangla Text Corpus BDNC01", International Journal for Research in Applied Science and Engineering Technology (IJRASET), Volume 5, Issue XI, Page No: 260-256 , ISSN: 2321-9653, November 2017.
14. HTK Speech Recognition Tool, website: http://htk.eng.cam.ac.uk/, retrieved on 9th May 2012.
15. Rabiner, L.R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, No. 2, Vol. 77, (March, 1989), pp: 257-289, ISSN: 0018-9219.