Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

Construction of Large Scale Isolated Word Speech Corpus in Bangla

Md. Farukuzzaman Khan[1]

[1] Islamic University

Received: 16 December 2017 Accepted: 5 January 2018 Published: 15 January 2018

Abstract

A new speech corpus of isolated words in Bangla language has been recorded including high frequent words from a text corpus BdNC01. It has been specifically designed for various research activities related to speaker-independent Bangla speech recognition. The database consists of speech of 100 speakers, each of them speaking 1081 words. 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 different days to avoid time variation of speaker property. The total 400 hours of recording makes the corpora largest in its type, size and language domain. This paper describes the motivation for the corpora and the processes undertaken in its construction. The paper concludes with the usability of the corpus.

Index terms— Bangla, speech corpora, BDNC01, vocabulary, isolated word, speech recognition.

1 Introduction

Abstract-A new speech corpus of isolated words in Bangla language has recorded including high frequent words from a text corpus BdNC01. It has designed specifically for various research activities related to speaker-independent Bangla speech recognition. The database consists of speech of 100 speakers, each of them speaking 1081 words. Another 50 new speakers were employed to speak all the list of words to construct a test database. Every utterance was repeated five times in different days to avoid time variation of speaker property. The total 375 hours of original recording makes the corpora largest in its type, size and language domain. This paper describes the motivation for the corpora and the processes undertaken in its construction. The paper concludes with the usability of the corpus.

With the continuous development of Speech corpora, it becomes one of the key issu es in speech technology development like text to speech and speech to text systems. Often read speech is used since it seems to be the easiest way to obtain a recorded speech corpus with highest control of the content [2]. From 1990 to 2000, several speech corpora were developed for various dimension of speech research. The OGI Multi-language Telephone Speech Corpus (1992) [3] and The Switchboard Telephone Speech **??**orpus (1990 **??**orpus (-2001)) [4] were two important development of speech corpora. In 1993, the most popularly used speech corpora TIDIGITS and TIMIT are published. TIMIT [5] was commissioned by DARPA and worked on by many sites, including Texas Instruments (TI) and Massachusetts Institute of Technology (MIT), hence the corpus name. The TIMIT corpus of read speech is designed to provide speech data for acousticphonetic studies and contains broadband recordings of 630 speakers of eight major dialects of American English. TIDIGITS [6] also designed and collected at Texas Instruments, Inc. (TI) for the purpose of designing and evaluating algorithms for speaker-independent recognition of connected digit sequences. The Buckeye Corpus including complete 2000 recording of conversational speech is created by a team of linguists and psychologists at Ohio State University, contains high-quality recordings from 40 speakers in Columbus, Ohio conversing freely with an interviewer [7]. A noisy speech corpus (NOIZEUS) was developed in 2007 to facilitate comparison of speech enhancement algorithms among research groups [8]. In the past several years, it is seen an explosion in corpus based methods in language technology development over the

horizon. Incremental computing power and invention of several efficient modeling techniques makes this research effort as popular field of natural language processing as well as artificial intelligence. Beyond of British and American English, corpora are creating frequently and using to design speech based technologies in all influential languages including Arabic, Chinese, Japanese and Indian languages.

Bangla is the seventh most spoken native language in the world by population [9]. In present days Bangla speakers are spread around the world, but majority of Bangla speakers are located in the countries of Bangladesh and India with a diversity of cultural and linguistic traditions. Though it is one of the influential world languages, the history of corpus generation and corpus based Bangla speech recognition are limited apid globalization of technology has made the world in society and cultural diversity of most populations is rapidly changing. This situation demands support to linguistic and cultural integration and hence needs an accurate, appropriate and robust language technology. It is also desirable for these technologies to be reproducible in new languages within limited time, expertise and monetary constraints. Fortunately, there are enormous and ever growing technologies to process digital text and speech freely available for many languages of the world [1]. But there are still needs for more tools and techniques to improve theory and practice of spoken language technologies in multiple language environments as the demand for opportunities to communicate with people speaking different languages are increasing. It is observed from contemporary texts and literature that the research interest in corpus-based techniques is growing steadily in the global environment during past decades. Corpusbased methods are found as seed point in almost all language and speech processing systems. The construction of standard speech database or corpus now becomes an obligatory necessity for the progressive development of speech processing, recognition and understanding systems.

2 Basics of Speech Corpus

Speech corpus is a large collection of audio recordings of spoken language. Most speech Corpora also have additional text files containing transcriptions of the words spoken and the time each word occurred in the recording. In a sense, Speech Corpora may be viewed in two types. Read Speech includes Book excerpts, Broadcast news, Lists of words, Sequences of numbers and Spontaneous Speech includes Dialogs, Narratives, Map-tasks, Appointment-tasks etc. Speech corpus is the basis for both analyzing the characteristics of speech signal and developing speech synthesis and recognition systems. The corpus content becomes more and more complicated and the size larger and larger with the development of computation power and the speech technology. To build a large scale speech corpus, the first task is to identify a large text corpus that has broadly representative distributions of words of the target language. Potential sources include online versions of news papers, web contents and books. For example, a speech corpus of British English WSJCAM0 has been recorded at Cambridge University from the Wall Street Journal text corpus [13]. Before recording a speech corpus, careful selection of vocabulary is important since on average each out-of-vocabulary word causes errors usually between 1.5 and 2 [14]. The recognizer vocabulary is usually designed with the goal of maximizing lexical coverage for the expected input. A straight forward approach is to choose the N most frequent words in the training data which means that the usefulness of the vocabulary is highly dependent upon the representativeness of the training data [15].

There are different features to characterize a speech corpus. Some of the influential features are speech types, speaker dependency, vocabulary size, etc. As the main application area of speech corpora is in the design process of speech recognition system, the importance of these parameters is based upon the typical design considerations of a recognition system, which may be closely related to a specific application. In terms of speech types, speech content of corpora may be isolated, connected or continuous speech. Isolated or discrete speech requires a significant pause between words, may be 250 milliseconds. A single utterance may consist of a single word or a short string of isolated words may contain ten words in best estimate. In continuous speech recognition systems, fluent or continuous speech flows with a normal rhythm and the words bump into each other thus making recognition harder. Speech recognition systems can also be classified further as either speaker-dependent or speaker-independent systems. The system may be designed to tolerate a large variety of speaker variability. In this case the system is speaker independent and has to deal with a large population of users, mainly from the general public. Other systems may be tuned to the voice of a particular speaker and thus the system is speakerdependent. We may also have a system that is adapted to the voice of a particular set of speakers called multispeaker system. Thus the speech corpora design varies with speaker condition for different context or application area. Another important consideration to design a speech corpus is its vocabulary size. The number of words that are recognized by a system may consist of a small set of words with small vocabulary of about ten words, a medium-size set with 10 to 100 words, a large set of words with 100 to 1000 words or very large set with over 1000 words [16]. The study of Gould, Conti, and Hovanyecz [17] to determine the feasibility of a limited capability automatic dictation machine which was simulated along with isolated and connected speech modes using various vocabulary sizes. In their experiment users composed and edited letters with the simulated voice recognizer which had either a 1000 word vocabulary or an unlimited vocabulary. The 1000 word vocabulary was composed of the 1000 most frequently used English words. An analysis afterward indicated that roughly 75% of the words used in the letter writing task were available in the 1000-word vocabulary. In another experiment with a 5000-word vocabulary, it was found that approximately 90% of words used by the letter writing tasks fall in the vocabulary. The result of the experiment indicates that large vocabulary increase the language coverage, but the most limiting problem of larger vocabulary sizes is the corresponding decrease in recognizer accuracy.

Thus as a compromise, an Isolated Word corpus of 1000 words may be satisfactorily applicable to evaluate a speech recognition system. within few years. Probably the first instance was Bangla Katha Bhandar, created and released by Center for Development of Advanced Computing (CDAC) of India in 2005 **??**10]. Another step of similar work was done by the Center for Research on Bangla Language Processing of Bangladesh in 2010 [11]. 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 was recorded in SIPL of Islamic University [12]. 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 is still ongoing. This paper describes the design and recording processes of Isolated word speech corpora. After the basics of speech corpora, a short 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. Possible benefits, limitation and extension possibilities of this work are discussed at the end of this article.

3 III. Word Selection and Database Design

BdNC01 corpus is a text corpus collected during 2005-2011. A large amount of Bangla text was compiled in BdNC01 from several influential Bangla newspapers including more than 11 million word tokens. To exploit the ease of collection, web editions of the dailies was used as the source of texts. For statistical processing of the corpus required software tools were developed using C Language. The output of the sorting function 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 large scale isolated speech database. A part of the list is shown in Table **??**1. Thus from this list high frequent 1081 words were separated and rearranged in the alphabetic order to finalized the database of isolated words.

4 Speaker Selection

The first step in this level is to select good speakers. Because the corpus was planned to use in speaker independent system, it was required to select speakers as required in number and quality. Therefore a notice was published among the departments of Islamic University inviting speakers in this regard. Huge amount of interested students both male and female were come with interest to do the job. So an audition was arranged to check their comparative efficiency of correct utterance or pronunciation of Bangla. Depending on their performance, 75 male and 75 female speakers were selected to finalize the speaker list. The list was included speakers from almost all dialect regions of Bangladesh but with an influence of majority of local students from Kushtia and Jhinaidah as shown in table-2. The selected speakers were very young in the age range of 18-25 years. Finally 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 and accuracy of utterance etc.

5 Rank

Table **??**: Distribution of speakers according to dialect regions V.

6 Speech Acquisition

Speech data were recorded in Laboratory environment by close-talking microphone directly connected to the computer. The speakers were asked to read the text in standard pronunciation as well as possible and the whole recording process was done under the supervision of theses authors. Every speaker was explained the purpose of the project and instructed to start the recording when she/he was ready to read. 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. Isolated Words were uttered one after another with a minimum pause between two consecutive words to avoid overlapping. The time to speak 1081 isolated words once was more than 30 minutes for each speaker and the total recording time for all 150 speakers to repeat all words in five different sessions was about (150x30x5) minutes or 375 hours.

7 VI.

8 Speech Editing and Labelling

It is 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. Especially, we found some speakers speak 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. As the HMM

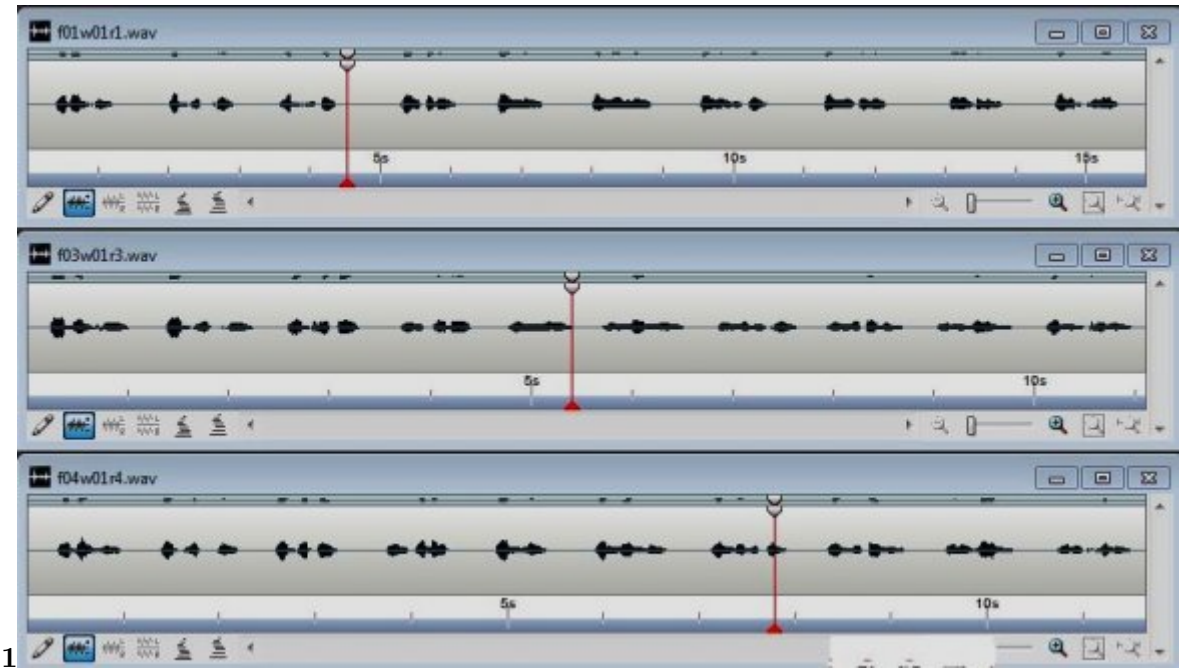
Toolkit developed by Cambridge University Engineering Department was already proved its efficiency and using frequently by most of the research workers, the database was labeled by following the specified format of speech data to make it ready for use in HMM Toolkit [18] for evaluation. For isolated word list, recorded files are divided to make files of equal number of words. In our work, ten words were decided to store as a wav file in isolated word database. A 300 ms pause on start and end positions of each file and 250 ms pause between two consecutive words were added. The time length of each file was 10-16 seconds or in average 13 seconds. The total time length of all the files in isolated words database was about (150x108x5x13) seconds or 292 hours after editing and labeling. Three edited and finalized files are shown in figure-1.

9 Resulting Corpora

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

10 Discussion and Conclusion

A standard ASR system is based on a set of socalled 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). Our database was not evaluated but ready to do the work because it has been formatted as required for using with HMM toolkit. 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. With the best of our knowledge these are the first speech corpora in Bangla language in its size, type and coverage. The evaluation of these corpora using HMM toolkit is left for future scope. We hope that the achievement from this work will construct a fundamental base in speech recognition research in Bangla especially in dictation and command processing. [1]



1

Figure 1: Figure 1 :

1

Figure 2: Table 1 :

[1] © 2018 Global Journals 1

Districts	No. of Speakers	Districts	No. of Speakers	Districts	No. of Speakers
Kushtia	30	Dhaka	4	Dinazpur	7
Jhenaidah	20	Manikganj	2	Nilphamari	3
Chuadanga	11	Mymensingh	2	Gaibandha	2
Meherpur	6	Jamalpur	3	Rangpur	3
Jessore	6	Faridpur	3	Kurigram	2
Magura	3	Madaripur	3	Rajshahi	4
Khulna	4	Gopalganj	2	Natore	3
Bagerhat	3	Razbari	4	Bogra	4
Satkhira	2	Jhalokati	3	Pabna	3
Cox's Bazar	2	Bhola	2	Sirazganj	

Figure 3:

3

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	1081	10	54000	27000	81000	292 hours

Total Recording

VIII.

Figure 4: Table 3 :

[Wikipedia (2017)] , Wikipedia . https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers 25 th October, 2017.

[Pallett et al. (1994)] 'Benchmark Tests for the ARPA Spoken Language Program'. D S Pallett , J G Fiscus , W M Fisher , J S Garofolo , B A Lund , A F Martin , M A Przybocki . *Proc. ARPA Spoken Language System Technology Workshop*, (ARPA Spoken Language System Technology WorkshopAustin, TX) 1994. January 1995. p. .

[Buckeye Corpus Collection and Recording, retrieved on 6 th (2011)] *Buckeye Corpus Collection and Recording, retrieved on 6 th*, http://buckeyecorpus.osu.edu/php/corpus.php Jan., 2011.

[Md et al. ()] 'Construction and Analysis of Large-Scale Bangla Corpus for Bangla Speech Recognition'. Md , Md Khan , Md. Mizanur Islam , Rahman . *Ministry of Science and Information and Communication*, (Bangladesh) 2008. (Project Report)

[Bozkurt et al.] *Corpus Building Using A Modified Greedy Selection*, Baris Bozkurt , Ozlem Ozturk , Thierry Dutoit , Asbl Multitel . (EUROSPEECH 2003 -GENEVA)

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

[Gopala Krishna Anumanchipalli et al. (2011)] 'Festvox: Tools for Creation and Analyses of Large Speech Corpora'. Kishore Gopala Krishna Anumanchipalli , Alan W Prahallad , Black . https://www.cs.cmu.edu/~awb/papers/vlsp2011_festvox.pdf *Workshop on very large scale phonetic research*, (UPenn) 2011. 25 th October, 2017.

[Lean-Lac Gauvain and Lamel ()] 'Large Vocabulary Speech Recognition Based on Statistical Methods'. Lori Lean-Lac Gauvain , Lamel . *Pattern recognition in speech and language processing*, (New York, USA) 2003. CRC press.

[Leonard and Doddington (1993)] 'Linguistic Data Consortium'. R , Gary Leonard , George Doddington . http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1 *TIDIGITS* 1993. Jan. 2011.

[Yeshwant et al. ()] 'Oshika The Ogi Multi-Language Telephone Speech Corpus'. K Yeshwant , Ronald A Muthusamy , Cole , T Beatrice . http://citeseerx.ist.psu.edu/viewdoc/summary *CiteSeerX* 1992. 2010. (retrieved on 28 th Dec.)

[Gibbon et al. ()] *Spoken Language System and Corpus Design*, Dafydd Gibbon , Roger Moore , Richard Winski . 1998. Walter de Gruyter. 38.

[Hu and Loizou ()] 'Subjective evaluation and comparison of speech enhancement algorithms'. Y Hu , P Loizou . *Speech Communication* 2007. 49 p. .

[Godfrey and Holliman (2010)] *Switchboard-1 Release 2, Linguistic Data Consortium, Philadelphia, 1997, retrieved on 28 th*, John J Godfrey , Edward Holliman . http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC97S62 Dec. 2010.

[Sherry and Casali (1988)] 'The Effects of Recognition Accuracy and Vocabulary Size of A Speech Recognition System on Task Performance and User Acceptance'. P Sherry , Casali . http://htk.eng.cam.ac.uk/,retrievedon9th *HTK Speech Recognition Tool* May 1988. 20 th October, 2017. March 2018. (5) . Virginia Polytechnic Institute and State University (M.Sc. Thesis)

[Garofolo (1993)] *TIMIT Acoustic-Phonetic Continuous Speech Corpus, Linguistic Data Consortium*, John S Garofolo . http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1 1993. Jan. 2011. Philadelphia.

[Robinson et al. ()] 'WSJCAM0: A British English Speech Corpus For Large Vocabulary Continuous Speech Recognition'. Tony Robinson , Jeroen Fransen , Jonathan Davidpye , Steve Foote , Renals . *Proc. of ICASSP 95*, (of ICASSP 95Detroit, Michigan) 1995.