International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.9, No.8 (2016), pp.121-128
http://dx.doi.org/10.14257/ijsip.2016.9.8.10

Development of A Small Vocabulary Database for Bengali Speech Recognition

Sumana Huque[1], Md. Abdullah Al Asad[2*] and Md Rokonuzzaman[3]

[1]*Junior Engineer, Roots Communication Limited, Bangladesh*
[2]*Assistant Professor, Dept. of Applied Physics, Electronics & Communication Engineering* [2]*Bangabandhu Sheikh Mujibur Rahman Science & Technology University, Gopalgonj, Bangladesh*
[3]*Lecturer, Dept. of Electrical, Electronic and Communication Engineering,* [3]*Military Institute of Science and Technology (MIST), Mirpur Cant., Dhaka, Bangladesh*
[1]*sumanashimu@gmail.com,* [2]*asad_135uits@yahoo.com,* [3]*ragib_isme@yahoo.com*

Abstract

This paper describes a small vocabulary Bengali database development to evaluate the performance of speech recognition algorithms in clean conditions. The database is constructed by Bangla digit sequences (/ak/, /dui/, /tin/, /chaar/, /panch/, /chhoy/, /shaat/, /aat/, /noy/, /zero/, /shunno/) are used. The developed database is consisted of two sets of data which are training and testing datasets. The training dataset contains 3824 utterances of 50 speakers; on the other hand, the testing dataset is subdivided into four groups (clean1, clean2, clean3 and clean4) and contains 1985 utterances of 52 speakers. In both sets of data the speaker's age ranges from 19 to 25 years. All the recordings have been done in a quiet room but not soundproof with the A4Tech HS-60 headset microphone interfaced to an Intel Dual Core 2.0 GHz CPU. The software used to record and edit the speech file is wave-pad; ver. 3.05.The recognition experiment is presented in this thesis to obtain comparable recognition results for the speaker-independent recognition of connected sequences of Bangla digit. As the research results has proved that the words accuracy is average 98%.

Keywords*: Speech recognition, Speech, Bengali Database*

1. Introduction

The speech database is a primary element for any kind of research in the field of speech signal processing. There are many such databases in various foreign languages, for instance, TIMIT (English) [1], TI Digits (English) [2], AURORA (English, Japanese, Spanish, *etc.*,) [3-7], CENSREC (Japanese) [8] *etc.,* Such database is not available in Bangla language yet, thus the development of a Bangla speech database is significant. However, Bangla speech database is a primary component to perform the speech-related research. In this paper, a small vocabulary Bangla speech database is constructed in quiet laboratory environment for speaker-independent recognition of connected digit sequences. Since acoustic patterns depend on speaker's gender, age and mood, thus 102 adult speakers were selected to develop this database with age ranging from 19~25 years. After selecting the speakers, a training program is conducted on speech recording. The data were recorded with a microphone placed 1~1.5 inch in front of speakers mouth and digitized at 22.05 kHz. The recording software is wave pad, ver. 3.05 was used.

*Corresponding Author

ISSN: 2005-4254 IJSIP

Finally, a recognition experiment is performed using this database to evaluate the performance of the database. The comparison of developed Bangla database and others database are given below table.

Table 1. Comparison between Different Database Systems

Dataset		AURORA-2	TIMIT Corpus	Bangla Database
Training	No. Of speakers	Male: 55 Female: 55	Male: 176 Female: 150	Male: 25 Female:25
	SNR	Clean multi-condition (clean & noisy)	Clean	Clean
	Filter	G.712	---	--
Test	No. of speakers	Male: 52 Female: 52	Male: 170 Female: 145	Male: 26 Female: 26
	SNR	Clean, 20, 15, 10, 5, 0 & -5 dB	Clean digits sequences	Clean
	Filter	G.712 & MIRS	--	--

2. Research Design

The following steps are followed to carry out the above mentioned tasks to develop the research.

Figure 1. The Step by Step Process of Research

3. Research Methodology

To construct this speech database a text corpus has been developed which includes the list of recorded words for a pre-prepared small vocabulary. Then the speech recording is performed using at least 50 adult speakers. The recorded database is edited using speech editing software to standardize and make useable for the researcher who would like to use it for automatic recognition, synthesis or any kind of further processing. After the successful completion of database construction, an experiment is performed for automatic recognition of Bangla speech using computer to evaluate the performance of the database.

3.1. Description of the Speakers

The number of speakers for constructing this database is 102. The speaker's descriptions for both training and testing datasets are given below table.

Table 2. Number and Age Ranges of Speakers

Dataset	No. of Speakers		Age Range (Years)
	Male	Female	
Training	25	25	19-25
Test	26	26	19-25

Moreover, to make the datasets dialectically balanced, the speakers were selected from different region of Bangladesh presented below table.

Table 3. Distribution of Participated Speakers from Several Regions for Training Dataset

Name of region (Division)	Number of speakers
Dhaka	5
Rajshahi	27
Khulna	9
Barisal	1
Rangpur	8
Total	50

3.2. Digits Pronunciations

The database is constructed as the format of AURORA-2 with some differences. The digit strings for each speaker are identical. The table below shows the pronunciations of eleven digits in AURORA-2 and Bangla database.

Table 4. Pronunciations of Bangla Digits

Digit	AURORA-2	Bangla Pronunciation
1	One	/ak/
2	Two	/dui/
3	Three	/tin/
4	Four	/chaar/
5	Five	/ panch/
6	Six	/chhoy/
7	Seven	/shaat/
8	Eight	/aat/
9	Nine	/noy/
0 (Z)	Zero	/zero/
0 (O)	Oh	/shunno/

Speakers were requested to pronounce digits as specified in this table. These pronunciations are assigned considering the occurrence frequency in uttering digits file.

3.3. Data Collection

This section describes about the utterance patterns, text corpus construction, the speech data collection, different errors finding and editing processes.

3.3.1. Text Corpus: Speech corpus is an important requirement for developing any ASR system, thus seventy-seven sequences of these digits were collected from each speaker for training set as shown in Table 5.

Table 5. Digit Sequences for Each Speaker of Training Set

No. of digits in a sequence	No. of sequences for each speaker
Isolated digit	22 (two tokens of each of the eleven digits)
Two-digit	11
Three-digit	11
Four-digit	11
Five-digit	11
Seven-digit	11
Total sequences	77

Hence, each speaker provided 253 digits for training dataset. Therefore, the text corpus for training dataset contains 3850 lines and each line contains the AURORA-2 digit sequence and corresponding Bangla digit sequence. An example list of text corpus for a speaker is shown in Table 6.

Table 6. Example of Text List

Line No.	AURORA-2 Digit Sequence	Bangla Digit Sequence
1	9 4 OH	Noy chaar shunno
2	3 2 1	tin dui ak
3	3 5 8 9	tin panch aat noy
4	2 6 3 4 ZERO	Dui chhoy tin chaar zero
5	7 4 OH	Shaat chaar shunno
…	…	…

3.3.2. Data Recording: During recording sessions, speech data were recorded consecutively 10 utterances with sufficient gap among the utterances in a file through a microphone at a distance of 1~1.5 inch in front of speaker's mouth. A waveform of such a file containing 10 utterances is shown in Figure 2. During the recording session, speakers were seated on a chair with a headset microphone (Connected to desktop) and were requested to speak the utterances from the supplied text corpus. A well configured desktop PC with Intel Dual Core 2.0 GHz CPU is used to record speech through an A4Tec HS-60 Headset microphone.

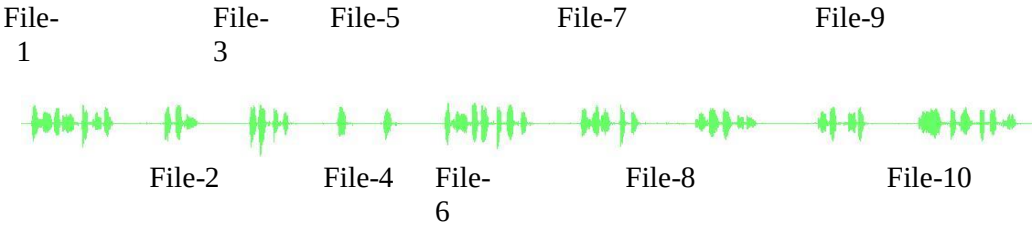

Figure 2. Waveform of 10 Utterances of Digit Sequences

During recording session the following parameters (Table 7.) of the wave file has been maintained throughout the recording process:

Table 7. Recording Parameters

Sampling rate of the audio	22.05 kHz
Bit rate (bits per sample)	16
Channel	Mono (Single)

The speech corpus is recorded in a quiet room but not sound-proof and took nearly three months to complete starting from October 2011 until January 2012. This speech corpus is enriched with varieties of speakers taking for recording and recoding environment and technologies used for recording. Moreover, all the recordings are performed by using the software wave pad, ver. 3.05, because the software is useful for professional recordings and edit audio easily [9].

For this work 22.05 kHz sample rate has been chosen because it provides more accurate high frequency information and 16 bit per sample divides the element position in to 65536 possible values. Each of the selected speakers spent an amount of time about 15~30 minutes during recording session. The session conductor and speaker decided on hand signals to communicate during the recording. After recording, the splitting of the audio files per sentence has been done manually using the same software and saved in a .wav format. It should be noted that each splitting file is down sampled to 8 kHz.

3.4. Preparation of the Database for the Recorded Pronunciations

To prepare the database for use, some meaningful filenames are assigned to the data files, the data are divided into training and testing subsets, and the data were copied to a CD. Before preparing the database all of the 43 utterances which contained speaker errors were corrected. However, corrections for some errors were not possible and were deleted from the database. Therefore, total number of utterances is 5809 including 3824 utterances for training and 1985 utterances testing.

The database is divided into two subsets, one is training dataset consists of 3824 utterances of 50 speakers to be used to train an ASR system, and the other one is testing dataset consists of 1985 utterances of 52 to be used for evaluating the performance of ASR system. The testing dataset is further divided into four groups, such as clean1, clean2, clean3 and clean4 containing 500, 500, 490 and 495 utterances, respectively.

Speech data are recorded consecutively 10 utterances in a file for a speaker, each recorded file is split up into 10 individual files by allowing sufficient silence (300 ~ 500 ms) at the beginning and end of each file. During the recording, the sampling frequency was kept at 22.05 kHz but in the splitting time speech signal was down sampled to 8 kHz. Sampling frequency is same for training and test datasets. The file naming is done considering three characteristics – first one represents the speaker category (M for Male and F for Female) and last two characters represent the speaker identity and digit sequences with. For examples: FHP_ 5Z28Z56A.08.wav, where first character represents speaker category, *i.e.* FHP for female speaker, then digit sequences, .08 represents the sampling frequency and finally, extension of the file. Cutting process is done with the software wave pad, version 3.05. The same steps are followed to prepare the test files.

3.5. Raw Training and the Preparation of the Testing Data

The design of the training dataset is same as AURORA-2. But in this database, only clean-training dataset is prepared which can be used in research purpose. The advantage of training on clean data only is the modelling of speech without distortion by any type of noise. Such models should be suited best to represent all available speech information.

To prepare the raw training data the followings steps were followed:

- Skip the header of the wave files.

- Read the data of a wave file as short and write the data to another file with same name leaving the extension .wav. As for example, if a wave file's name is FAH_23A.08.wav, the raw data file name will be FAH_23A.08.

The preparation process of raw test data is same as that of raw training data. Therefore, we have four groups of raw test dataset such as clean1, clean2, clean3 and clean4 containing 500, 500, 490 and 495 raw data files, respectively.

4. Finding and Results

A database as well as a recognition experiment is presented in this thesis to obtain comparable recognition results for the speaker-independent recognition of connected sequences of Bangla digit. The database together with the definition of training and test sets can be taken to determine the performance of a complete recognition system.

As shown in Table 8. The word accuracy for the four datasets clean1, clean2, clean3 and clean4 are found to be 98.11%, 98.15%, 98.08%, 97.84%; respectively. The highest word accuracy is obtained for set clean2; on the contrary, the minimum word accuracy is obtained for clean4.

Table 8. Word Accuracy for Mel-LPC Based Speech Recognition

Group	Word accuracy (%)
Clean 1	98.11
Clean 2	98.15
Clean 3	98.08
Clean 4	97.84
Average	98.05

5. Conclusion

In this research, a small vocabulary Bangla database of connected digit sequences is prepared. The developed database consists of two sets – one is training and testing dataset which is dialectically balanced. The recording has been done in quiet laboratory environment with necessary technologies. The preparation of recorded pronunciations of speaker made error free as sufficient gap among the utterances given during saving the recorded files. And, all the errors of recorded data were removed and each recorded data file was saved into sub files thus the database contains the raw data without header. Finally, the constructed database was successfully developed and used in an experiment of speech recognition where the word accuracy found around 98%.

5.1. Recommendation for Further Research

The same research can be carried out in future from various aspects to develop Bangla database for speech recognition. For instance, in this research, Bangla database has been developed in small contents of Bangla digits only so further research can carry out to develop Bangla database to consider the large number of vocabulary. In addition, the research has conducted in the basis of limited number of speakers where their age ranging from 19-25 years old. So, the research could be applied for different age groups to enhance viability of new database system. Also, the research can be conducted to large geographical area to balance more of dialects of different region of Bangladesh.

Acknowledgments

The authors would like to thank the concerned authority of the University of Rajshahi, Rajshahi, Bangladesh for providing us laboratory facilities for the purpose of completing this research work. The financial support of the ministry of information and communication technology, Bangladesh (Fellowship 2010-2011) is gratefully acknowledged.

References

[1] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett and N. L. Dahlgren, "Darpa Timit Acoustic-Phonetic Continuous Speech Corpus CD-ROM," The National Institute of Standards and Technology (NIST) is an agency of the U.S. Department of Commerce, **(1993)** Febuary 1.
[2] R. Leonard, "A database for speaker-independent digit recognition", IEEE International Conference on ICASSP '84., vol. 9, no. 1, **(1984)**, pp. 328 - 331.
[3] H. G. H. and D. Pearce, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions", Ericsson Eurolab Deutschland Gmbh, Paris, **(2000)**.
[4] E. T. S. Institute, "Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms", ETSI Standard, vol. 1, no. 12, **(2000)**, pp. 2000-2004.
[5] H. G. H. and D. Pearce, "The Aurora Experimental Framework For The Performance Evaluation Of Speech Recognition Systems Under Noisy Conditions", Motorola Labs, UK, Beijing, China, **(2000)**.
[6] S. Nakamura, "AURORA-2J: An Evaluation Framework for Japanese Noisy Speech Recognition", IEICE Transactions on Information and Systems , vols. E88-D , no. 3, **(2005)**, pp. 535-544.
[7] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukry and S. Euler, "SPEECH DAT CAR. A Large Speech Database For Automotive Environments", Universidad Politécnica de Cataluña, Barcelona, Spain, **(1998)**.
[8] C.-2. Database, "IPSJ SIG-SLP Noisy Speech Recognition", Google Documents , **(2005)** July5. [Online]. Available: http://www.slp.cs.tut.ac.jp/CENSREC/data/CENSREC-2-data-E.pdf. [Accessed 4 May 2010].
[9] R. Iniguez, "WavePad Sound Editor," Informer Technologies, Inc., 10 January 2011. [Online]. Available: http://wavepad-sound-editor.software.informer.com/. [Accessed 7 January 2011].

Authors

Sumana Huque, She is currently serving as Jr. Engineer, Roots Communication Limited an IOS and IGW Company, Bangladesh. Former Engineer at Solution Art limited a reputed Software development and Solution Provider Company, Mirpur-12, Dhaka about six months. She has achieved B.Sc. in Applied Physics & Electronics Engineering Department, from Rajshahi University in 2009 and M.Sc. (Thesis) in Applied Physics & Electronics Engineering Department, from Rajshahi University in 2010. Her research area is in Digital Signal Processing and Pattern recognition, Image Processing, Speech Signal Processing and Performance Recognition.

Md. Abdullah Al Asad, He is currently serving as Assistant Professor, Dept. Applied Physics, Electronics & Communication Engineering, Bangabandhu Sheikh Mujibur Rahman Science & Technology University, Gopalgonj, Bangladesh. Former lecturer in University of Information Technology and Sciences (UITS), Baridhara, Dhaka. He has achieved B.Sc. in Applied Physics & Electronics Engineering Department, from Rajshahi University in 2009 and M.Sc. (Thesis) in Applied Physics & Electronics Engineering Department, from Rajshahi University in 2010. His

research area is in Magnetic materials, Spintronic device, Solar cell, Nanotechnology and Solid state device and already has published more than two of research papers in different national and international Journal.

Md Rokonuzzaman, He received the B.Sc. in Electrical and Electronic Engineering (EEE) degree from University of Information Technology and Sciences (UITS); Currently a M.Sc. student and Faculty in Dept. of Electrical, Electronic and Communication Engineering at Military Institute of Science and Technology (MIST), Dhaka. Former adjunct lecturer in Bangladesh University of Professionals, lecturer in Bangladesh Institute of Science and Technology (BIST), Lab Demonstrator/Instructor in UITS. His research area is in Renewable Energy, Smart Grid, Power system and Power Electronic Devices. By profession he is member of IEEE, PES, BES and American Center, Bangladesh.