Analysis of a Small Vocabulary Bangla Speech Database for Recognition

Sumana Huque
Dept. of Applied Physics and Electronic Engineering,
University of Rajshahi, Rajshahi 6205, Bangladesh

Ahsan Habib Rasel
Dept. of Applied Physics and Electronic Engineering,
University of Rajshahi, Rajshahi 6205, Bangladesh

M. Babul Islam
Dept. of Applied Physics and Electronic Engineering,
University of Rajshahi, Rajshahi 6205, Bangladesh

ABSTRACT

To carry out any kind of research in the field of speech signal processing, a standard database is essential. There are many databases in different languages but not in Bangla language. Therefore, in this article, it has been attempted to develop and analysis a small vocabulary Bangla database for recognition. In this database 11 Bangla digits (/ak/, /dui/, /tin/, /chaar/, /panch/, /chhoy/, /shaat/, /aat/, /noy/, /zero/, /shunno/) have been used. The developed database consisted of two sets of data such as training and testing datasets. The training dataset contains 3824 utterances of 50 speakers, and testing dataset is subdivided into four groups (clean1, clean2, clean3 and clean4) and contains 1985 utterances of 52 speakers. All recordings have been done in a quiet room but not sound proof 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 wavepad. Finally, an HMM based recognizer is developed to evaluate the database. The word accuracy for test sets is found to be 98.05% on the average. In this recognition process Mel-LPC based front-end and as a reference recognizer HTK (Hidden Markov Model Toolkit) have been used.

General Terms

Bangla Speech Database Processing and Recognition.

Keywords

Bangla Speech Database, Bangla Speech Recognition, HMM, Mel-LPC

1. INTRODUCTION

At present, speech recognition is widely used research topic around the world. Though Bangla is one of the largely spoken languages in the world, only a few works on Bangla speech can be found in the literature, especially on Bangladeshi accented Bangla [1]. There are many databases in different languages but in Bangla language it is not enriched yet. The evaluation process of the constructed database for Bangla connected digit corpus involves in algorithm design to train an ASR system using the training set, and performance evaluation of the recognition system using the testing set. The evaluation scenarios are designed in the same way as AURORA-2 [2][3][4][5].

Therefore, a database as well as a recognition experiment is presented in this paper to obtain comparable recognition results for the speaker-independent recognition of connected sequences of Bangla digits. The database together with the definition of training and test sets can be taken to determine the performance of a complete recognition system. The HTK (Hidden Markov Model Toolkit) [6][7] is used to train the develop system and to evaluate the performance of the database.

In this paper, the Mel-LP based analysis technique is used as front-end since it incorporates auditory-like frequency resolution.

2. STATISTICAL APPROACH TO SPEECH RECOGNITION

Research in speech recognition has produced numerous algorithms and commercially available speech recognizers that all work to some extent. Among these, statistical approach, in particular, the Hidden Markov Model, is the most prevailing approach that has been proved its practical and theoretical soundness [7]. First the acoustic waveform is recorded by a microphone and sampled typically at 8 or 16 kHz to allow processing by a digital device. The acoustic front-end processor converts the sampled waveform into a sequence of observation vectors (frames), $O = \{o_1, o_2, \ldots, o_T\}$ by removing unimportant information such as pitch and noise. There is a considerable amount of variability in the observation sequences even if the same words were uttered by the same speaker. Hence a statistical approach is adopted to map the observation sequence into the most likely sequence of words. The speech recognizer has to choose the word sequence, $W = \{w_1, w_2, \ldots, w_n\}$ with the maximum posterior probability given the observation sequence as follows:

$$W = \arg\max_{W} P(W|O) = \arg\max_{W} \frac{P(W)P(O|W)}{P(O)} \qquad (1)$$

where $P(W)$ is the a-priori probability of observing the word sequence W independent of acoustic evidence. This probability is determined by a language model. The term $P(O|W)$ is the probability of observing *O* given word string *W*, and is determined by an acoustic model.

Bayes' formula [8] has been applied to obtain the final form. It should be noted that the likelihood of the observation sequence $P(O)$ may be omitted in the maximization process since it is independent of the word sequence. However, direct modeling of the probabilities, $P(W|O)$ is not feasible due to the observation sequence variability and the vast number of possible word sequences. The role of the recognizer is to affect a mapping between sequences of speech vectors and the wanted underlying symbol sequences as shown in Figure 1.

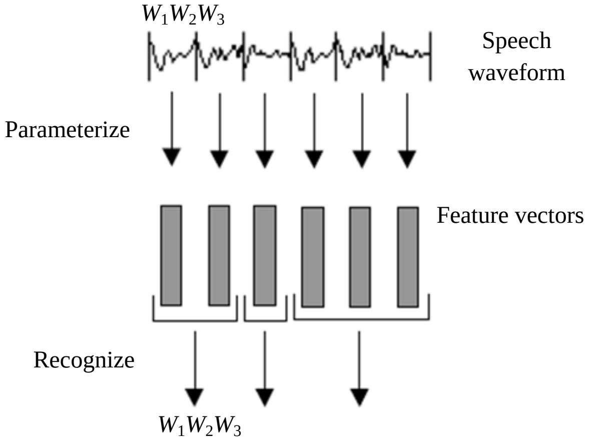


Fig. 1: Message encoding/decoding in an ASR.

2.1 Acoustic Model by HMM

Any acoustic unit, such as word, syllable, diphone or phone can be modeled by Hidden Markov Models (HMMs). An HMM is a finite state machine which can be viewed as a generator of random observation sequences according to probability density functions [8][9]. The model changes state once at each time step and at time *t* a state *j* is entered a speech vector o_t is generated from the probability density a_{ij}. The values of a_{ij} should satisfy

$$\sum_{j=1}^{N} a_{ij} = 1 \qquad (2)$$

where N is the number of states. Figure 2 shows an example of this process where five state model moves through the state sequence $X = 1, 2, 3, 4, 5$ in order to generate the sequence o_1 to o_5.

The joint probability that O is generated by the model M moving through the state sequence X is calculated as

$$P(O, X \mid M) = a_{12}b_2(o_1)a_{22}b_2(o_2)a_{23}b_3(o_3)a_{34}b_4(o_4)a_{44}b_4(o_5)a_{45} \qquad (3)$$

The likelihood of the model generating the observation is computed by summing over all possible state sequences $X = x(1), x(2), x(3), \ldots, x(T)$. That is,

$$P(O \mid M) = \sum_X a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)} \qquad (4)$$

where $x(0)$ is constrained to be the model entry state and $x(T+1)$ is constrained to be the model exit state. Alternatively, the likelihood can be approximated by only considering the most likely state sequence as follows:

$$\hat{P}(O|M) = \max_x \left\{ a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)} \right\} \qquad (5)$$

Therefore, for a given set of models M_i corresponding to word sequence W_i, Eq. 1 is solved by assuming that

$$P(O|W_i) = P(O|M_i) \qquad (6)$$

In the above discussion, it is assumed that the parameters $\{a_{ij}\}$ and $\{b_j(O_t)\}$ are known for each model M_i. For a given set of training examples corresponding to a particular model, the parameter of the model can be determined automatically by a robust and efficient re-estimation procedure.

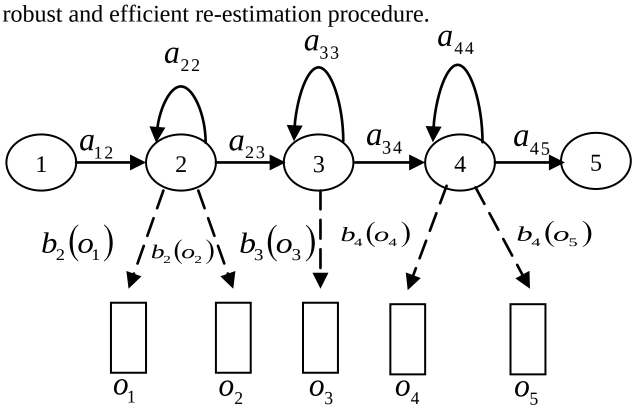


Fig.2: Markov generation model.

2.2 Language Model

The task of a language model (LM) is to estimate the probability of a word sequence $P(w_1 w_2 \ldots w_n)$. More specifically, a language model estimates a-priori probability $P(W)$ for a given word sequence $W = \{w_1, w_2, \ldots, w_n\}$ as defined in Eq. 1, which is given by

$$P(w_1 w_2 .... w_n) = \prod_{k=1}^{n} P(w_K \mid w_1 ...... w_{k-1}) \qquad (7)$$

The conditional probability defined by Eq. (7) is an exact computation. In practice, the LM provides an approximation to the true probability by assuming that the conditional probability of observing a word w_n at a position *n* is restricted to its immediate $N - 1$ predecessor words $w_{n-N+1} \ldots w_{n-1}$, that is

$$P(w_1 w_2 \ldots w_n) \approx \prod_{k=1}^{n} P(w_k \mid w_{k-N+1} \cdots w_{k-1}) \qquad (8)$$

The resulting model is that of Markov chain and is referred to as *N*-gram model [10]. The widely used models are bigram and trigram models. The language models are estimated from the training text in training session.

3. RECOGNIZER

A recognizer, also called decoder is a searching algorithm which finds the corresponding word sequence *W* for which the maximum probability estimate $P(W \mid O)$ has been found as a consequence of a spoken utterance *O* and its corresponding word string. The following diagram (Figure 3) shows the used network levels for recognitions [11].

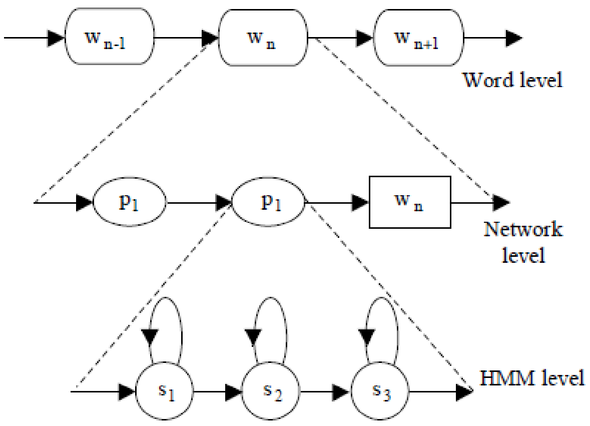


Fig. 3: Recognition network levels.

3.1 Back-End Design

The reference recognizer was based on HTK (Hidden Markov Model Toolkit). The evaluation framework is designed as follows:

The HMM was trained on clean condition using the training dataset. The digits are modeled as whole word HMMs. In the recognition, a standard pronunciation dictionary and recognition grammar based on EBNF syntax notation are defined as shown in Figure 4.

$ digit = ak| dui | tin | chaar |panch|
chhoy | shaat | aat |
noy | zero |shunno |;
([sil] < $digit [sp] > [sil])

Fig. 4: Grammar written in EBNF.

Each digit HMM had 18 states with 16 states output distributions since according to HTK notations, there are two dummy states at the beginning and end. In this evaluation process a simple left-to-right model without skips over states has been used. A mixture of 3 Gaussians per state is used. Only the variances of all acoustic coefficients (no full covariance matrix) are used.

Two pause models ‘sil’ and ‘sp’ are defined. The ‘sil’ model consists of 3 states which illustrates in Figure 5. This HMM shall model the pauses before and after the utterance. A mixture of 6 Gaussians models each state. The second pause model ‘sp’ is used to model pauses between words. It consists of a single state, which is tied with the middle state of the ‘sil’ model.

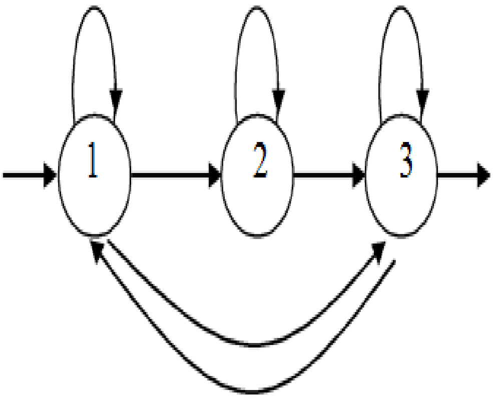


Fig. 5: Possible transition in the 3-state pause model ‘sil’.

The training is done in several steps by applying the embedded Baum-Welch estimation scheme using the HTK tool HERest – Initialize all word models and the 3-state pause model with the global means and variances (determined by HcompV). Word and pause models contain only 1 Gaussian per state in this initialization stage. Three iterations of Baum-Welch re-estimation with the pruning option –t of HERest set to 250.0 150.0 1000.0. Introduce the inter-word pause models, increase the number of Gaussians to 2 for the 3-state pause model and apply three further iterations of Baum-Welch re-estimation. Increase the number of Gaussians to 2 for all states of the word models, increase the number of Gaussians to 3 for all states of the pause model and apply three further iterations of Baum-Welch re-estimation. Increase the number of Gaussians to 3 for all states of the word models, increase the number of Gaussians to 6 for all states of the pause models and apply seven further iterations of Baum-Welch re-estimation. During recognition an utterance can be modeled by any sequence of digits with the possibility of a ‘sil’ model at the beginning and at the end and a ‘sp’ model between two digits.

3.2 Front-End Design

3.2.1 Mel-LPC based Speech Analysis

The frequency-warped signal $\tilde{x}[n]$ $(n = 0,\ldots,\infty)$ obtained by the bilinear transformation of a finite length windowed signal $x[n]\,(n = 0, 1,\ldots, N-1)$ [12][13][14] is defined by

$$\tilde{X}(\tilde{z}) = \sum_{n=0}^{\infty} \tilde{x}[n]\tilde{z}^{-n} = X(z) = \sum_{n=0}^{N-1} x[n]z^{-n} \qquad (9)$$

where $\tilde{z}^{-1}$ is the first-order all-pass filter,

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha . z^{-1}} \qquad (10)$$

where $0 < \alpha < 1$ is treated as frequency warping factor.

The phase response of $\tilde{z}^{-1}$ is given by

$$\tilde{\lambda} = \lambda + 2 \cdot \tan^{-1}\left\{\frac{\alpha \sin\lambda}{1 - \alpha \cos\lambda}\right\} \qquad (11)$$

This phase function determines a frequency mapping. As shown in Fig. 1, $\alpha = 0.35$ and $\alpha = 0.40$ can approximate the mel-scale and bark-scale at the sampling frequency of 8 kHz respectively.

Now, the all-pole model on the warped frequency scale is defined as

$$\tilde{H}(\tilde{z}) = \frac{\tilde{\sigma}_e}{1 + \sum_{k=1}^{p} \tilde{a}_k \tilde{z}^{-k}} \qquad (12)$$

where $\tilde{a}_k$ is the *k*-th mel-prediction coefficient and $\tilde{\sigma}_e^2$ is the residual energy.

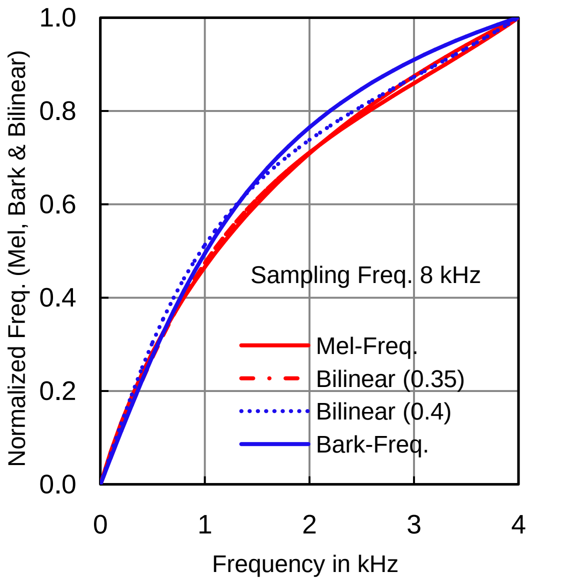


Fig. 6: The frequency mapping function by bilinear transformation.

On the basis of minimum prediction error energy for $\tilde{x}[n]$ over the infinite time span, $\tilde{a}_k$ and $\tilde{\sigma}_e$ are obtained by Durbin's algorithm from the autocorrelation coefficients $\tilde{r}[m]$ of $\tilde{x}[n]$ defined by

$$\tilde{r}[m]=\sum_{n=0}^{\infty}\tilde{x}[n]\tilde{x}[n-m] \qquad (13)$$

which is referred to as mel-autocorrelation function.

The mel-autocorrelation coefficients can easily be calculated from the input speech signal $x[n]$ via the following two steps. First, the generalized autocorrelation coefficients are calculated as

$$\tilde{r}_\alpha[m]=\sum_{n=0}^{N-1}x[n]x_m[n] \qquad (14)$$

where $x_m[n]$ is the output signal of an m-th order all pass filter $\tilde{z}^{-m}$ excited by $x_0[n]=x[n]$. That is, $\tilde{r}_\alpha[m]$ is defined by replacing the unit delay z^{-1} with the first order all-pass filter $\tilde{z}(z)^{-1}$ in the definition of conventional autocorrelation function as shown in Figure 7.

:

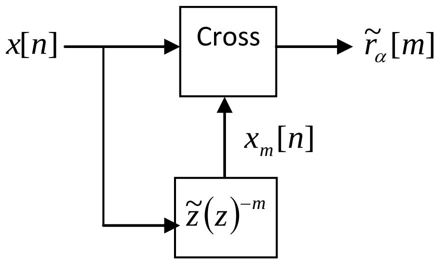


Fig.7: Generalized autocorrelation function.

Due to the frequency warping, $\tilde{r}_\alpha[m]$ includes the frequency weighting $\tilde{W}(e^{j\tilde{\lambda}})$ defined by

$$\tilde{W}(\tilde{z})=\frac{\sqrt{1-\alpha^2}}{1+\alpha\tilde{z}^{-1}} \qquad (15)$$

which is derived from

$$\frac{d\lambda}{d\tilde{\lambda}}=\left|\tilde{W}(e^{j\tilde{\lambda}})\right|^2 \qquad (16)$$

Thus, in the second step, the weighting is removed by inverse filtering in the autocorrelation domain using $\left\{\tilde{W}(\tilde{z})\tilde{W}(\tilde{z}^{-1})\right\}^{-1}$

As feature parameters for recognition, the Mel-LP cepstral coefficients can be expressed as:

$$\log\tilde{H}(\tilde{z})=\sum_{n=0}^{\infty}c_k\tilde{z}^{-n} \qquad (17)$$

where $\{c_k\}$ are the mel-cepstral coefficients.

The mel-cepstral coefficients can also be calculated directly from mel-prediction coefficients $\{\tilde{a}_k\}$ [15] using the following recursion:

$$c_k=-\tilde{a}_k-\frac{1}{k}\sum_{j=1}^{k-1}(k-j)\tilde{a}_k c_{k-j} \qquad (18)$$

It should be noted that the number of cepstral coefficients need not be the same as the number of prediction coefficients.

4. FUNCTIONAL BLOCK DIAGRAM

The functional block diagram (Figure 8) followed by the research experiment is given below

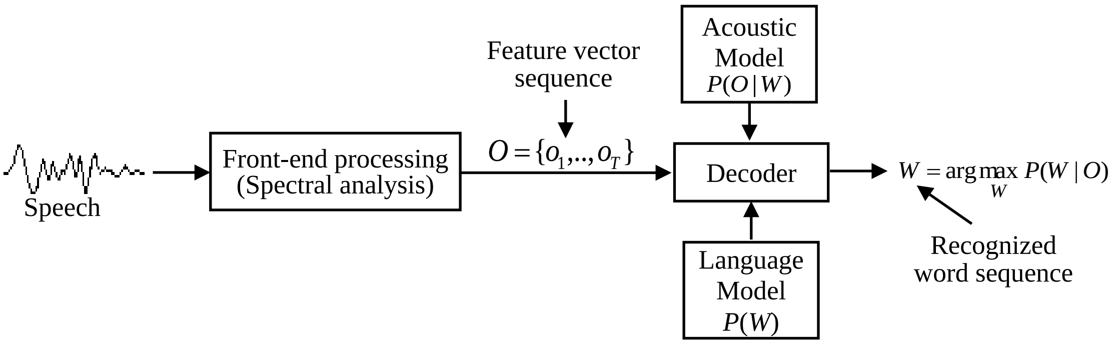


Fig. 8: The functional block diagram of a HMM (Hidden Markov Model) based ASR.

5. EXPERIMENTAL SETUP

All the recognition experiments were conducted with a 12th order Mel-LPC analysis. The speech signal was windowed using Hamming window of length 20 ms with 10 ms frame period. The frequency warping factor was set to 0.30. As front-end, 14 cepstral coefficients and their delta coefficients including 0th terms were used. Thus, each feature vector size is 28. The recognition accuracy (*Acc*) is evaluated as follows:

$$Acc = \frac{N - D - S - I}{N} \times 100\% \qquad (19)$$

where, *N* is the total number of words. *D*, *S* and *I* are deletion, substitution and insertion errors respectively.

5.1 Recognition Performance

5.1.1 *Word Accuracy*

As shown in Table 1, the word accuracy for the four datasets clean1, clean2, clean3 and clean4 are found to be 98.11%, 98.15%, 98.08% and 97.84%, respectively. The highest word accuracy is obtained for set clean2; on the contrary, the minimum word accuracy is obtained for clean4.

5.1.2 *Different Errors*

There are three types of errors such as deletion, substitution and insertion are considered to evaluate the performance of recognition process and database. A comparison between deletion and substitution errors for each test set is presented in Figure 9. An overall comparison among three errors obtained from HTK reference recognizer for whole test sets is shown in Figure 10.

Table 1. 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

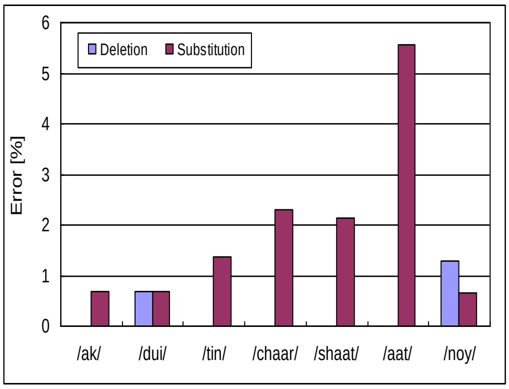
5.2 HTK Result Analysis

The detail results obtained from HTK reference recognizer are presented in this subsection. The HTK software produces analysis of a recognition result using a confusion matrix. The confusion matrix for test set clean1 is given in Table 2.

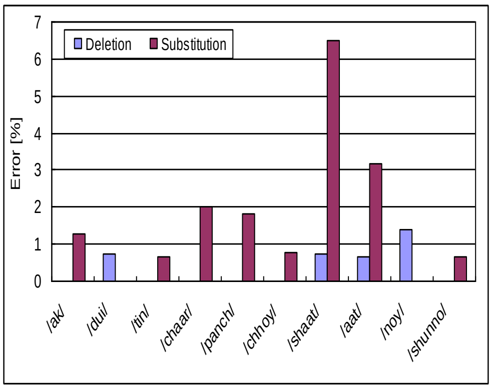
6. CONCLUSION

In this research work, a small vocabulary Bangla speech database of connected digit sequences is prepared for the recognition of speaker-independent connected digit sequences. The developed database consists of two sets – one is training and other is testing dataset which is dialectically balanced since the speakers are selected from various regions of Bangladesh.

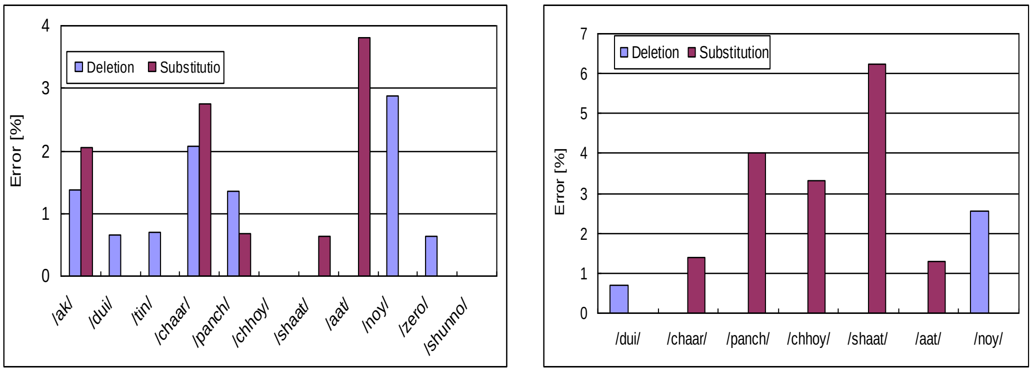
In evaluation process of the database a recognition experiment has been conducted using an HMM based back-end with a front-end based on Mel-LPC. The recognition result shows the effectiveness of the database for speaker-independent recognition of connected digit sequences which has been found to be 98.05% word accuracy on the average.



Test set clean1



Test set clean2



Test set clean3 **Test set clean4**

Fig. 9: Graphical representation of deletion and substitution errors for all test sets.

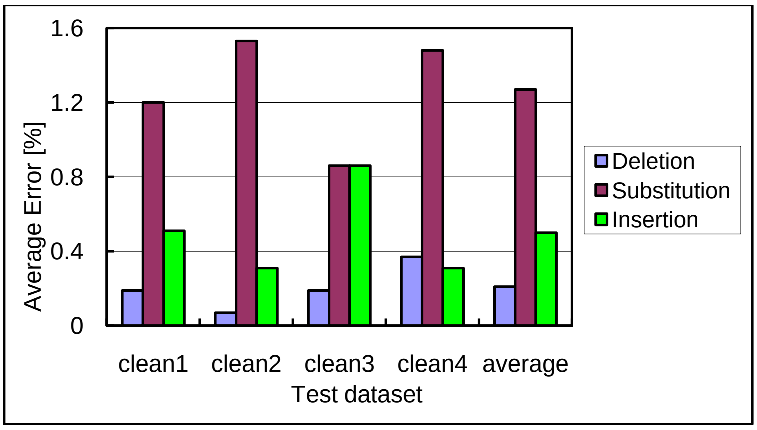


Fig. 10: Comparison among Deletion, Substitution and Insertion errors for entire test sets.

Table 2. Details results for test set clean1 obtained from HTK reference recognizer

```
condition: clean
====================== HTK Results Analysis ====================
SENT: %Correct=95.60 [H=478, S=22, N=500]
WORD: %Corr=98.62, Acc=98.11 [H=1567, D=3, S=19, I=8, N=1589]
------------------------ Confusion Matrix ----------------------
---
       a   d   t   c   p   c   s   a   n   z   s
       k   u   i   h   a   h   h   a   o   e   h
           i   n   a   n   h   a   t   y   r   u
                   a   c   o   a           o   n
                   r   h   y   t               n Del [ %c / %e]
  ak 145   0   0   1   0   0   0   0   0   0   0   0 [99.3/0.1]
 dui   0 146   0   1   0   0   0   0   0   0   0   1 [99.3/0.1]
 tin   0   0 144   0   0   0   0   0   0   2   0   0 [98.6/0.1]
chaa   1   0   0 128   1   0   0   1   0   0   0   0 [97.7/0.2]
panc   0   0   0   0 149   0   0   0   0   0   0   0
chho   0   0   0   0   0 157   0   0   0   0   0   0
shaa   0   0   0   2   0   1 137   0   0   0   0   0 [97.9/0.2]
 aat   0   0   0   0   8   0   0 136   0   0   0   0 [94.4/0.5]
 noy   0   0   0   0   0   1   0   0 151   0   0   2 [99.3/0.1]
zero   0   0   0   0   0   0   0   0   0 125   0   0
shun   0   0   0   0   0   0   0   0   0   0 149   0
Ins    0   0   0   0   0   1   0   4   1   2   0
    ========================================================
```

6.1 Limitation of Research

The research has addressed some limitations due to many reasons. To carry out the research no digital studio was used to record the database rather a quiet but not sound proof laboratory was used which might decrease the quality of recorded data. In addition, the development of Bangla speech database was based on limited speaker's participations

6.2 Recommendation for further Research

The same research can be carried out in future from various aspects to develop Bangla speech database for recognition.

For instance, this research has been done in a quiet laboratory environment, so an extended research may be carried for noisy environment. Moreover, in this research, the database has been developed using Bangla digits only. So, further research can be carried out to develop Bangla speech database by considering 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.

7. REFERENCES

[1] Muhammad, G. et al. 2009. Automatic speech recognition for Bangla Digits. IEEE, 12th International Conference on Computers and Information Technology (ICCIT '09), Dhaka.

[2] Hirsch, H. G. and D. Pearce, 2000. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. Proc. ISCA ITRW ASR 2000: 181:188.

[3] E. T. S. Institute. 2000. Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms. ETSI Standard, vol. 1, 12, 2000-2004.

[4] Pearce, D. et al. 2000. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. Motorola Labs, UK.

[5] Nakamura, S. 2005. AURORA-2J: An Evaluation Framework for Japanese Noisy Speech Recognition. IEICE Transactions on Information and Systems. E88-D, 3, 535-544.

[6] Moreno, A. et al. 1998. SPEECH DAT CAR. A Large Speech Database For Automotive Environments. Universidad Politécnica de Cataluña, Barcelona, Spain.

[7] Young, S. et al. 1999. The HTK Book, USA: Microsoft Corporation.

[8] Weisstein, A. E. 2013. Hidden Markov Model Manual v1.0. Washington University and Truman State University.

[9] Weisstein's,E. W. E. 2010. Wolfram math world. MathWorld Book.

[10] Mooney, R. J. 1997. Natural Language Processing: N-Gram Language Models. University of Texas at Austin, Texas, USA.

[11] Entropic, 2011. General Principles of Recognition. [Online].

[12] Islam, M. B. 2007. Mel-Wiener Filter for Mel-LPC Based Speech Recognition. IEICE Transactions on Information and System. 90, 6, 30-35.

[13] Rahman, M. and Islam, M. B. 2010. Performance evaluation of MLPC and MFCC for HMM based noisy speech recognition. International Conference on Computer and Information Technology (ICCIT), Dhaka.

[14] Matsumoto, H. et al. 1998. An efficient Mel-LPC analysis method for speech recognition. Proc. ICSLP, 98, 1051-1054.

[15] Furui, S. 1981. Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust., Speech and Signal Processing, ASSP-29, 254-272.