Bengali Spoken Numerals Recognition by MFCC and GMM Technique

Bachchu Paul, Somnath Bera, Rakesh Paul, and Santanu Phadikar

Abstract Speech is the standard vocalized communication media. Speech is one of the comfortable way for humans to communicate with each other. Similarly, speech recognition system is eagerly necessary to communicate with computer through voice. Speech recognition in English language already helps us to operate English voice command-based applications. But in rural and semi-urban areas, due to lack of knowledge in English in India, it is necessary to implement automatic speech recognition in regional languages. Here, we have built a Gaussian Mixture Model (GMM)-based Bengali (also called Bangla) isolated spoken numerals recognition system where mel frequency cepstral coefficients denoted as MFCC is taken for feature extraction. The proposed system achieved 91.7% correct prediction for the Bangla numeral data set of 1000 audio samples for 10 classes which is satisfactory for previous Bangla spoken digit recognition.

Keywords ASR · Zero crossing · FFT · MFCC · HMM · DTW · GMM

B. Paul (✉) · S. Bera · R. Paul
Department of Computer Science, Vidyasagar University, Midnapore, West Bengal 721102, India
e-mail: ableb.paul@gmail.com

S. Bera
e-mail: somnathh.beraa@gmail.com

R. Paul
e-mail: rakeshpaul470@gmail.com

S. Phadikar
Department of Computer Science and Engineering, Maulana Abul Kalam Azad University of Technology, Kolkata 700064, India
e-mail: sphadikar@yahoo.com

P. K. Mallick et al. (eds.), *Advances in Electronics, Communication and Computing*,
Lecture Notes in Electrical Engineering 709,
https://doi.org/10.1007/978-981-15-8752-8_9

1 Introduction

Speech is the primary communication media in our every moment of life to communicate with each other. Anyone who is not computer professionals can communicate with computer through speech because speech is normally easy and more comfortable for communication with human. As all over the India, people are not totally literate and the major percentage of people is illiterate and semi-illiterate, so speech recognizable application will be more benefitable and suitable for them [1]. Communicating with computer through speech can be possible through speech recognition. For this application of speech, it is seen that already speech recognition becomes a demanding and interesting subject for research purpose. Generally, speech recognition system recognizes the speech and converts it into text format and finally made it into a format that a machine can read it easily [2]. Each country has multiple regional languages. Bangla is a regional language which has been considered in this paper. More than 215 million people all over the world speak in Bangla as their native language [3]. But very few research works have been done on regional language. So, there is a good opportunity to us to do more research work in ASR of Bangla language to improve more. The proposed work based on spoken Bangla numeral recognition. This research work can help to those people who are interested to do their research in Bangla language.

There are many applications of spoken numerals recognition. It is used in ATM machines, biometric system, cellular phone, computer, smart wheel chair, etc. In railway system, announcement of train number of arriving, or departure trains, this system can be used. In our paper, we try to build a Bengali spoken numerals recognition system using GMM where MFCC can be used as feature extraction technique.

2 Literature Review

Karpagavalli and Chandra [1] developed phoneme which is speaker independent and also developed word-based speech recognition system in Tamil language using hidden markov tool kit. They took MFCC to extract features and also used HMM for developing the acoustic model. For estimating the state emission probabilities, they have used multi-variant Gaussian Mixture Model to build acoustic model. They choose 10 speaker who used Tamil language and made 50 words vocabulary for building and testing the model. After analyzing, they discussed about the accuracy of identification and word error rate (WRR) of this model. Taking the small data set, it is seen that the accuracy of recognition is high where the word error rate is too minimum which will finally treated as negligible.

Gamit and Dhameliya [4] carried out their research on isolated word recognition using artificial neural network. In this paper, combination of MFCC and LPC both are used for feature extraction. They used a classifier, i.e., back propagation

neural network to separate unvoiced speech samples from the voiced speech samples. Speech database contains the speech uttered by 28 speakers in which 14 speakers are males, and 14 speakers are females. After evaluation, they got 51.25% accuracy by using only MFCC whereas by using both MFCC and LPC, they got 85% accuracy.

Patil et al. [5] proposed an isolated word recognition system in Hindi language. They took MFCC for feature extraction and used vector quantization with GMM for isolated Hindi word recognition. The Hindi words were taken from some male and female speaker and used KNN for matching the pattern. They used KNN classifier for classification of sample feature of training and testing. Finally, in result, they shown some performance parameter and presented graphical representation of classification. Their implementation will be helpful to disabled, illiterate people in communication, education sector, etc.

Hammami et al. [6] proposed automatic spoken Arabic digit recognition based on GMM. They used $\Delta\Delta$MFCC for feature extraction. It has seen that accuracy level of GMM is average 99.31%, whereas accuracy level of CHMM is 98.41%. This paper shows the result and says GMM is most appropriate and attractive for this system. From the recognition result, it is seen that comparable rate of automatic speech recognition system is too high and also it is too much better than other reported results.

Chauhan et al. [7] carried out their research on speech-to-text conversion using GMM. They used MFCC for extracting the feature of speech signals and also used GMM to train the audio files for speech recognition. They experimented on multiple isolated words and got near about 71% accuracy to recognize those words. The only drawback of this system is that it is not suitable for high ambient noisy environment.

Ali et al. [8] proposed a technique to recognize Bengali words. They proposed four different models for words recognition system. In model 1, to extract features, they have used MFCC as a feature extraction technique and they used dynamic time warping for the purpose of matching. In model 2, they used LPC. Linear predictive coefficients are also calculated to extract the features and dynamic time warping for matching. In model 3, as previous MFCC was used to extract features and GMM was used to get the probability function for the purpose of matching. In model 4, LPC compressed MFCC for extracting the features and dynamic time warping for matching purpose. Finally, in this paper, they got 84% accuracy to recognize the speech whereas they took 100 Bangla words and they took general room environment to complete the Bengali word recognition purpose.

After careful studying, some of the existing system we focused a simple Bengali spoken digit recognition system by GMM. In Fig. 1, the proposed method is given properly.

This paper is expressed as: Sect. 3 discusses the dataset that are used and the pre-processing phase, Sect. 4 discusses feature extraction phase, Sect. 5 describes how GMM classify a speech and also shows the outcome of the proposed method and at last in Sect. 6 conclusion of the above work is discussed here.

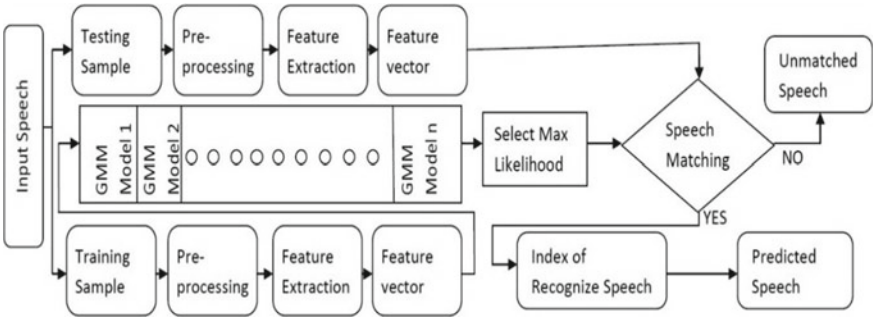

Fig. 1 Block diagram of the proposed model

3 Dataset and Preprocessing

3.1 Dataset Used

For the proposed work of isolated Bangla spoken digit recognition, we have taken a small data set of 10 Bangla digits zero to nine (pronounced as 'sunno' to 'noi'), uttered by 10 speakers, among them five male and five female with the age group from twenty to forty. Each word is uttered by ten times for each speaker with normal room environment. We used the audacity software with sampling frequency of 16 kHz and 32 bit mono channel. The data set contains 1000 audio samples of 10 classes. The whole data set has been used as training data set of GMM. Then each audio sample has been tested for most accurate match.

3.2 Preprocessing

In this stage, the voiced activity zone is detected from each of the uttered word. This is done by framing the signal of 25 ms with 50% overlapping. Then for each of the frame, the average energy and average zero crossing have been computed by the formula given in Eqs. 1 and 2, respectively. The energy of a frame calculates how much information it holds and zero crossing takes decision for a noise or noiseless frame with some threshold [9, 10].

$$E_n = \sum_{m=-\infty}^{\infty} [X(m) - W(n-m)]^2 \tag{1}$$

where $X(.)$ is the frame and $W(.)$ is the windowing function.

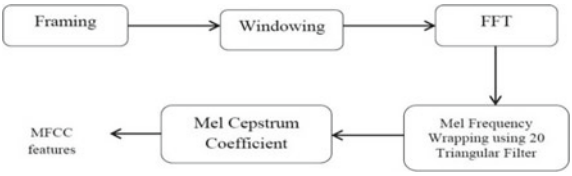

Fig. 2 Steps taken to calculate MFCC

$$\text{ZCR} = \frac{1}{2N} \sum_{j=i-N+1}^{i} |\text{sgn}(x(j)) - \text{sgn}(x(j-1))| w(i-j) \tag{2}$$

where,

$$\text{sgn}(x(j)) = \begin{cases} 1, \text{ if } x(j) \geq 0. \\ 0, \text{ if } x(j) < 0. \end{cases}$$

4 Feature Extraction

For each of the voiced frame, we have computed the first 13 coefficients which are taken as the MFCC coefficients and taken as our feature vector. Here, the feature extraction technique MFCC is computed in the following steps given in Fig. 2.

In Fig. 2, we discussed the steps of MFCC which we used for finding MFCC from speech signal.

4.1 Framing

The voiced section for each audio sample detected in Sect. 3.2 is segmented into 25 ms frame with 50% overlap. A single frame contains 400 samples, i.e., 80 frames per second.

4.2 Windowing

Since speech is an aperiodic signal, so the [6, 8] same size hamming window multiplied with signal because of maintaining the continuity at two extreme ends of a frame, Here, the hamming window equation is expressed by Eq. 3.

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \tag{3}$$

4.3 Fast Fourier Transform (FFT)

Here time domain is converted into frequency domain by using FFT [4]. It is generally used to measure the energy distribution over frequencies. The FFT is calculated using the Discrete Fourier Transform (DFT) formula given in Eq. 4.

$$S_i(k) = \sum_{n=1}^{N} s_i(n) e^{\frac{-j2\pi kn}{N}} \quad 1 \leq k \leq K \tag{4}$$

K is the DFT length.

4.4 Mel-Frequency Wrapping

Here, power spectrum is mapped onto mel-scale using 20 triangular band pass filter. There exist a relationship between frequency (f) and mel (m) is given in Eq. 5.

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{5}$$

4.5 Mel Cepstrum Coefficient

The frequency domain into time domain of the signal is converted by discrete cosine transform (DCT) using Eq. 6.

$$C_m = \sum_{k=1}^{M} \cos\left[m\left(k - \frac{1}{2}\right)\frac{\pi}{M}\right] E_k \tag{6}$$

Here, M is the length of filter bank which is 20 in our case; $1 \leq m \leq L$ is the number of MFCC coefficients.

Thus, for a single frame, the 13 numbers of features as our feature vector.

5 Construction of GMM

An acoustic model for each utterance of individual word can identify the word. Since, we know the sounds are produced by different shape of vocal track and different frequency. But it encounters a problem if we want to match the same word uttered

by another person even the same person in later time. If we see, the power spectral density (PSD) shape of the same word spoken by different speakers, it changes, since the human vocal track change from person to person. This can be solved by the GMM, where one spectral feature commonly very robust is MFCC calculated from each utterance of the same class. Combining all such features, we developed a multidimensional probability density function (PDF) for the particular class of Bangla numeral. For ten Bangla numeral (zero to nine), ten such model is developed.

GMM is a probabilistic model expressed as a weighted sum of Gaussian component densities. It is a probability density function that can be used as a parametric model to measure the features in biometric system [6]. GMM evaluates mean and variance using iterative expectation maximization (EM) algorithm. Mean calculates frequency of power spectrum and variance measure how distance number spread out. These features are here extracted through MFCC [8, 9]. Multiple Gaussian distributions are mixed up and finally create the Gaussian Mixture Model. There are two types of Gaussian distribution. First one is uni-variant Gaussian distribution given as,

$$G(X|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\Pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{7}$$

Here, μ is denoted as mean, and σ is the standard deviation. σ^2 is the variance of distribution. Second one is multi-variant Gaussian distribution given as,

$$G(X|\mu, \Sigma) = \frac{1}{\sqrt{2\Pi|\Sigma|}} \exp^{\left(-\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu)\right)} \tag{8}$$

Where, Σ is the covariance matrix. So, GMM can calculate the mean and variance using EM algorithm [5]. If x is a d-dimensional feature vector, then for a K-cluster problem, the probability distribution of the MFCC and obtained from cluster i, $i = 1, 2, \ldots, K$ is modeled as a mixture of N component probability densities as follows:

$$p(x|\lambda_i) = \sum_{j=1}^{N} p_{ij} f_i\left(x|\theta_{ij}\right), \sum_{j=1}^{N} p_{ij} = 1 \tag{9}$$

where for the ith speaker, P_{ij} is the prior probability for the jth component of the mixture. $\lambda_i = \{P_{ij}, \theta_{ij}, j = 1, 2, \ldots, N\}$ is the collection of unknown parameters and $f(x|\theta_{ij})$ is the probability density of x

$$p(x|\lambda_i) = p_{ij} \frac{1}{(2\Pi)^{\frac{d}{2}}|\Sigma_{ij}} e^{-\frac{1}{2}(x-\mu_{ij})^T \sum_{ij}^{-1}(x-\mu_{ij})} \tag{10}$$

where

$$\left\{\theta_{ij} = \mu_{ij,} \sum_{ij}\right\} \quad i = 1, 2, \ldots, K, j = 1, 2, \ldots, N$$

During testing phase, the MFCC feature is calculated for the test audio sample. Then, the maximum likelihood is calculated with the posterior probability of all GMM. The index of the maximum log-likelihood value is the recognized digit.

5.1 Result and Analysis

The PSD of the Bangla numeral one and four (Bengali pronunciations 'ek' and 'char,' respectively) for three different speakers is shown in Figs. 3 and 4, respectively, using the all-pole filter of Yule-Walker parametric spectral estimation technique.

It is clear that, the number of peaks for each of the digit is same, but different from others. The voiced portion boundary of the utterance of numeral 'ek' (English equivalent one) is given in Fig. 5. The accuracy on different class is given in the confusion matrix of Table 1. To justify the performance of the given proposed technique depends on the True Positive says as TP, True Negative says as TN, False Positive says as FP, and False Negative denoted as FN. So,

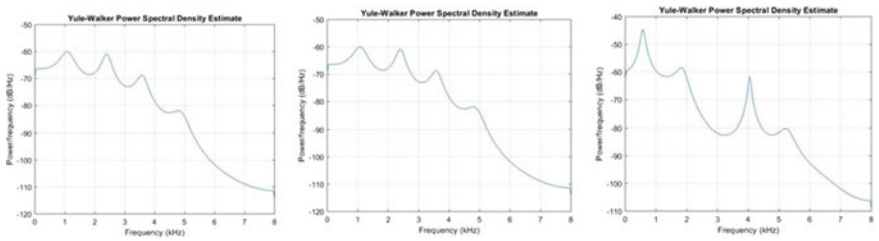

Fig. 3 Three different PSD of the numeral 'Ek'

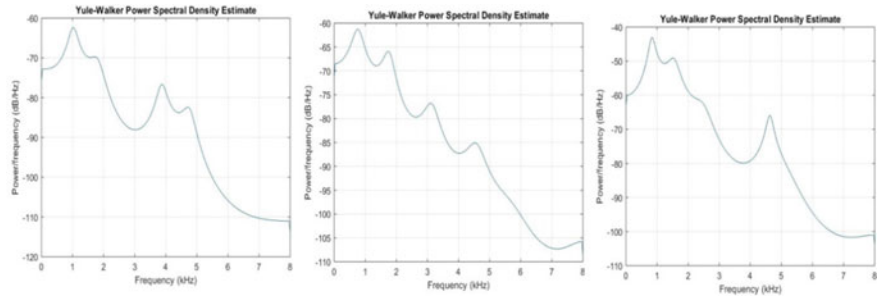

Fig. 4 Three different PSD of the numeral 'Char'

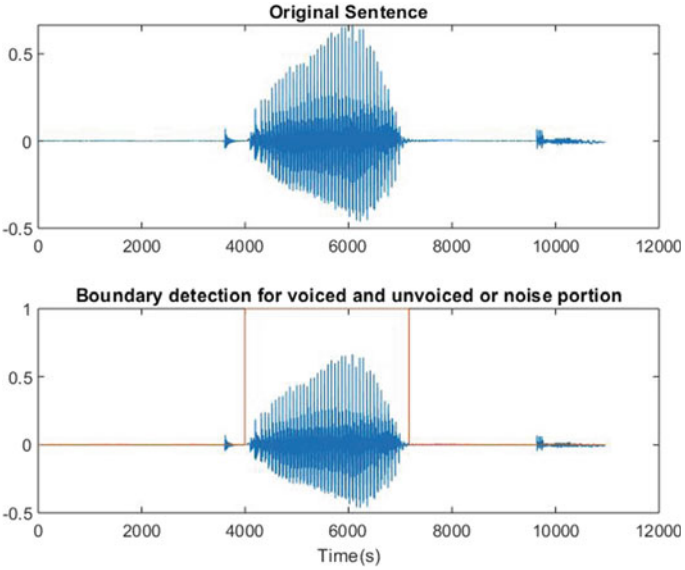

Fig. 5 Boundary detection in voice section

Table 1 Confusion matrix

	Predicted class										
True Class		0	1	2	3	4	5	6	7	8	9
	0	**99**	0	0	0	0	0	0	0	0	1
	1	4	**88**	0	0	0	0	0	1	6	1
	2	0	0	**98**	1	0	0	0	0	1	0
	3	6	0	3	**90**	0	0	0	0	0	1
	4	3	0	0	0	**84**	0	0	7	6	0
	5	0	0	0	2	0	**91**	0	0	5	2
	6	2	0	3	0	0	0	**88**	0	0	7
	7	2	2	0	0	0	0	0	**84**	8	4
	8	0	0	0	0	0	0	0	2	**98**	0
	9	0	0	1	0	0	0	2	0	0	**97**

i. Recall expressed as RE = TP/(TP + FN)
ii. Precision expressed as PR = TP/(TP + FP)
iii. Specificity expressed as SP = TN/(TN + FP)
iv. False Positive rate expressed as FPR = FP/(FP + TN)
v. False Negative rate expressed as FNR = FN/(TP + FN)
vi. Percentage of wrong classifications says as PWC = 100 * (FN + FP)/(TP + FN + FP + TN)

Table 2 Evaluation metrics

Evaluation metrics	Class										Mean
	0	1	2	3	4	5	6	7	8	9	
RE	0.99	0.88	0.98	0.9	0.84	0.91	0.88	0.84	0.98	0.97	0.917
PR	0.8534	0.97778	0.9333	0.96774	1	1	0.978	0.894	0.79	0.858	0.92524
SP	0.9811	0.99778	0.9922	0.99667	1	1	0.998	0.989	0.971	0.982	0.99078
FPR	0.0189	0.00222	0.0078	0.00333	0	0	0.002	0.011	0.029	0.018	0.00922
FNR	0.01	0.12	0.02	0.1	0.16	0.09	0.12	0.16	0.02	0.03	0.083
PWC	1.8	1.4	0.9	1.3	1.6	0.9	1.4	2.6	2.8	1.9	1.66
F-score	0.9167	0.92632	0.9561	0.93264	0.913	0.953	0.926	0.866	0.875	0.911	0.91757

vii. F-Score: (2 * Precision * Recall)/(Precision + Recall)

where, True Positive is highly predictable and False Negative is regrettable. The values of the all these traditional efficiency parameters compared with the ground truth and the outcome is given in Table 2.

6 Conclusion

Speech recognition is basically the highly preferable research work to researchers. A very progressive result on English, French, and Chinese like languages, but not satisfactory result in local or regional language. Our proposed work of isolated word recognition focused on Bangla language using GMM and can recognize a spoken Bangla numeral satisfactorily. Here, it has been observed from the confusion matrix that a misclassification between 'choy' and 'noy' similarly between 'sat' and 'aat' have occurred because their PSD mostly matches. The proposed works fine for a small data set but the performance degrades with large number of class and data set. It is also highly biased for a speaker dependent system. In our future work of isolated word recognition for Bangla language, a hybrid model of both feature extraction process and multiple classifiers such as DTW, SVM together with GMM to improve the accuracy and evaluation metrics.

References

1. Karpagavalli, S., & Chandra, E. (2015). Phoneme and word based model for Tamil speech recognition using GMM-HMM. In *2015 International Conference on Advanced Computing and Communication Systems* (pp. 1–5). IEEE.
2. Gupta, A., & Sarkar, K. (2018). Recognition of spoken bengali numerals using MLP, SVM, RF based models with PCA based feature summarization. *International Arab Journal of Information Technology, 15*(2), 263–269.
3. Muhammad, G., Alotaibi, Y. A., & Huda, M. N. (2009). Automatic speech recognition for Bangla digits. In *2009 12th International Conference on Computers and Information Technology* (pp. 379–383). IEEE.
4. Gamit, M. R., & Dhameliya, K. (2015). Isolated words recognition using MFCC, LPC and neural network. *International Journal of Research in Engineering and Technology, 4*(6), 146–149.
5. Patil, U. G., Shirbahadurkar, S. D., & Paithane, A. N. (2016). Automatic speech recognition of isolated words in Hindi language using MFCC. In *2016 International Conference on Computing, Analytics and Security Trends (CAST)* (pp. 433–438). IEEE.
6. Hammami, N., Bedda, M., & Farah, N. (2012). Spoken Arabic digits recognition using MFCC based on GMM. In *2012 IEEE Conference on Sustainable Utilization and Development in Engineering and Technology (STUDENT)* (pp. 160–163). IEEE.
7. Chauhan, V., Dwivedi, S., Karale, P., & Potdar, S. M. (2016). Speech to text converter using Gaussian Mixture Model (GMM). *International Research Journal of Engineering and Technology (IRJET), 3*(5), 160–164.

8. Ali, M. A., Hossain, M., & Bhuiyan, M. N. (2013). Automatic speech recognition technique for Bangla words. *International Journal of Advanced Science and Technology, 50*.
9. Padmanabhan, J., & Johnson Premkumar, M. J. (2015). Machine learning in automatic speech recognition: A survey. *IETE Technical Review, 32*(4), 240–251.
10. Permanasari, Y., Harahap, E. H., & Ali, E. P. (2019). Speech recognition using dynamic time warping (DTW). In *Journal of Physics: Conference Series* (Vol. 1366, No. 1, p. 012091). UK: IOP Publishing.