Chapter 13
Assamese Vowel Speech Recognition Using GMM and ANN Approaches

Debashis Dev Misra, Krishna Dutta, Utpal Bhattacharjee, Kandarpa Kumar Sarma and Pradyut Kumar Goswami

Abstract This work focuses on the classification of Assamese vowel speech and recognition using Gaussian mixture model (GMM). The results are compared to the results obtained using artificial neural network (ANN). The training data is composed of a database of eight different vowels of Assamese language with 10 different recorded speech samples of each vowel as a set in noise-free and noisy environments. The testing data similarly is composed of the same number of vowels with each vowel containing 23 different recorded samples. Cepstral mean normalization (CMN) and maximum likelihood linear regression (MLLR) are used for speech enhancement of the data which is degraded due to noise. Feature extraction is done using mel frequency cepstral coefficients (MFCC). GMM and ANN approaches are used as classifiers for an automatic speech recognition (ASR) system. We found the success rate of the GMM to be around 81 % and that of the ANN to be above 85 %.

Keywords Cepstral mean normalization (CMN) · Maximum likelihood linear regression (MLLR) · Vowel · Gaussian mixture model (GMM) · Artificial neural network (ANN)

D.D. Misra (✉)
Royal School of Engineering and Technology, Guwahati, Assam, India
e-mail: debashish.dm@gmail.com

K. Dutta
NIT Nagaland, Dimapur, Nagaland, India
e-mail: krishnadutta54@gmail.com

U. Bhattacharjee
Department of Computer Science and Engineering, Rajiv Gandhi University, Doimukh, Arunachal Pradesh, India
e-mail: utpal.bhattacharjee@rgu.ac.in

K.K. Sarma
Department of Electronics and Communication Technology, Gauhati University, Guwahati 781014, Assam, India
e-mail: kandarpaks@gmail.com

P.K. Goswami
Assam Science and Technology University, Guwahati 781014, Assam, India
e-mail: pradyutgoswami@yahoo.com

K.K. Sarma et al. (eds.), *Recent Trends in Intelligent and Emerging Systems*, Signals and Communication Technology, DOI 10.1007/978-81-322-2407-5_13

13.1 Introduction

Vowels are voiced sound during the production of which sound obstruction occurs in the oral or nasal cavities. Voiced speech is a sound produced with the vibration of vocal cords. In the speech, vowels are produced by exciting an essentially fixed vocal tract, shaped with quasiperiodic pulses of air caused by the vibration of the vocal cords [1].

Assamese is an eastern Indo-Aryan language spoken by about 20 million people in the Indian states of Assam, Meghalaya, and Arunachal Pradesh, and also spoken in Bangladesh and Bhutan [2]. There are 11 vowels in Assamese language and are distinguished by the place of articulation (front, central or back) and the position of the tongue (high, mid or low). The way in which the cross-sectional area varies along the vocal tract determines the resonance frequencies of the tract (the formants) and thereby the sound that is produced. The vowel sound produced is determined primarily by the position of the tongue, but the position of the jaw, lips, and to a small extent, the velum also influence the resulting sound [1].

This work focuses on the classification of Assamese vowel speech and recognition using gaussian mixture model (GMM). The results are compared to the results obtained using artificial neural network (ANN). The training data is composed of a database of 8 different vowels of Assamese language with 10 different recorded speech samples of each vowel as a set in noise-free and noisy environments. The testing data similarly is composed of the same number of vowels with each vowel containing 23 different recorded samples. Cepstral mean normalization (CMN) and maximum likelihood linear regression (MLLR) are used for speech enhancement of the data which is degraded due to noise. Feature extraction is done using mel frequency cepstral coefficients (MFCC). GMM and ANN approaches are used as classifiers for an automatic speech recognition (ASR) system. We found the success rate of the GMM to be around 81 % and that of the ANN to be above 85 %. Some of the related literature are [1, 3–8].

The rest of the paper is organized as follows: in Sect. 13.2, we briefly discuss about the basic notions related to the work. The system model is described in Sect. 13.3. The experimental details and results are discussed in Sect. 13.4. The work is concluded in Sect. 13.5.

13.2 Theoretical Considerations

Here, a brief discussion about ANN and GMM is given.

13.2.1 ANN

ANN: ANNs are bio-inspired computational tools that provide human-like performance in the field of ASR. These models are composed of many nonlinear computational elements called perceptrons operating parallel in patterns similar to the

biological neural networks [9]. ANN has been used extensively in ASR field during the past two decades. The most beneficial characteristics of ANNs for solving ASR problem are the fault tolerance and nonlinear property. The earliest attempts involved highly simplified tasks, e.g., classifying speech segments as voiced/unvoiced or nasal/fricative/plosive. Success in these experiments encouraged researchers to move on to phoneme classification. The basic approaches to speech classification using ANN are static and dynamic.

In static classification, the ANN accepts the input speech and makes a single decision. By contrast, in dynamic classification, the ANN considers only a small window of the speech. This window slides over the input speech while the ANN generates decisions. Static classification works well for phoneme recognition, but it scales poorly to the level of words or sentences. In contrast, dynamic classification scales better. Either approach may make use of recurrent connections, although recurrence is more often found in the dynamic approach [4, 5].

13.2.2 GMM

The GMM is a density estimator and is one of the most commonly used types of classifier. In this method, the distribution of the feature vector x is modeled clearly using a mixture of M Gaussians. A GMM is modeled by many different Gaussian distributions. Each of the Gaussian distribution has its mean, variance, and weights in the GMM. A Gaussian mixture density is a weighted sum of M component densities (Gaussians) as depicted in following figure and given by equation.

$$p(\overrightarrow{x}|\lambda) = \sum_{i=1}^{M} p_i x_i(\overrightarrow{x}) \tag{13.1}$$

where x is a L dimensional vector, p_i are mixture weights, and $b_i(x)$ are component densities with $i = 1M$. Each component density is a L variate Gaussian function of the form,

$$b_i(x) = \frac{1}{(2\pi)^{L/2}|\sum i|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)' \sum_{i=1}^{-1}(x-\mu_i)\right) \tag{13.2}$$

where μ_i is the mean and $\sum i$ is covariance matrix. The mixture weights satisfy the constraint that $\sum_{i=1}^{M} p_i$. T is the total number of feature vectors or total number of frames. T is the total number of feature vectors or total number of frames. The mean vectors, covariance matrices, and mixture weights of all Gaussians together represent a speaker model and parameterize the complete Gaussian mixture density. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system, such as vocal tract-related

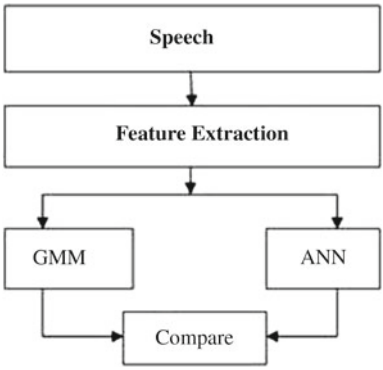

Fig. 13.1 System model

spectral features in a speaker recognition system. GMM parameters are estimated from training data using the iterative expectation-maximization (EM) algorithm or maximum a posteriori (MAP) estimation from a well-trained prior model [6, 7]. GMMs are often used in biometric systems, most notably in speaker recognition systems, due to their capability of representing a large class of sample distributions. One of the powerful attributes of the GMM is its ability to form smooth approximations to arbitrarily shaped densities.

13.3 System Model

In this study, we concentrate on Assamese vowel speech recognition using GMM and compare the results obtained using ANN. CMN and MLLR are used for speech enhancement of the data degraded due to noise. Feature extraction is done using MFCC. GMM and ANN approaches are used as classifiers for an ASR system for Assamese speech.

The system model is shown in Fig. 13.1. Feature extraction is the estimation of variables (feature vector) from the observation of a speech signal which contains different information such as dialect, context, speaking style, and speaker emotion. It estimates a set of features from the speech signal that represents some speaker-specific information. The aim is to transform the speech signal into a collection of variables that can preserve the signal information and that can be used to make comparisons [8].

13.4 Experimental Details and Results

Initially, we record certain number of vowel speech samples of Assamese language out of which some are retained in a clean form and a few are corrupted. In the proposed system, data is collected in 16 kHz sampling rate at 16 b mono format. Speech data collected is grouped into frame of 30 ms with one-third overlapping. It gives a frame

rate of 10 ms. After pre-emphasis, each frame is multiplied by Mel-filter bank with 20 filters and the MFCC coefficients are calculated. Here, 19 MFCC coefficients are considered along with their first-order derivatives as feature vector for each frame. Thus, we extract features using MFCC. The feature set contains samples which are clean and noise corrupted.

These are next modeled using GMM and applied to ANN for training. There is a training phase during which the GMM and ANN learns. The samples sets have the clean and noise-corrupted sets. Next, test and validation processes are performed during which the GMM and ANN demonstrate the decision-making role as part of the ASR. The speech samples derived from the inputs before feeding to GMM and ANN are enhanced by the CMN and MLLR approaches which contribute to the performance of the system. CMN has been done for cepstral coefficients extracted from the speech signal. After CMN, the model-based algorithm MLLR has been used for further noise elimination. MLLR is a model-based compensation method. It uses a mathematical model of the environment and attempts to use samples of the degraded speech to estimate the parameters of the model. In order to evaluate the clean speech in a real environment condition, the clean speech is deteriorated by adding the white Gaussian noise. The assumptions made are that the noise is additive and not correlated with the speech signal.

The noisy signal due to the addition of the white Gaussian noise is shown in the second plot of the below Fig. 13.4. The signal-to-noise ratio (SNR) is set at 5 dB. After applying CMN and MLLR to the noisy signal, the enhanced speech signal is shown below in Fig. 13.2. The original speech signal is plotted in blue, the signal checked for enframing and deframing is shown in black and the average noised removed signal using CMN and MLLR is shown in red.

A similar set of results are generated using a combination of CMN and MLLR shown in Fig. 13.3.

The training data is composed of a database of eight different Assamese vowels with 10 different recoded speech samples of each vowel as a set in noisy environments. The testing data similarly was composed of the same number of vowels with each vowel containing 23 different recorded samples (Fig. 13.4).

We have here used a recurrent neural network (RNN) of two hidden layers trained with error backpropagation through time (BPTT) algorithm [9]. The RNN is a special form of ANN with the ability to track time variations in input signals. The

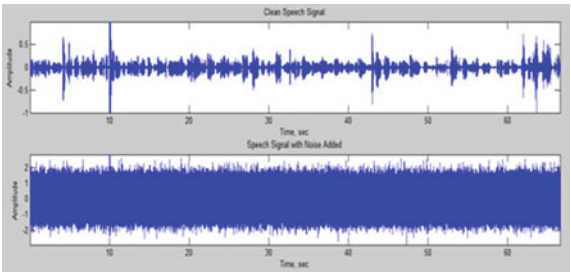

Fig. 13.2 Clean speech signal and speech signal with white Gaussian noise added

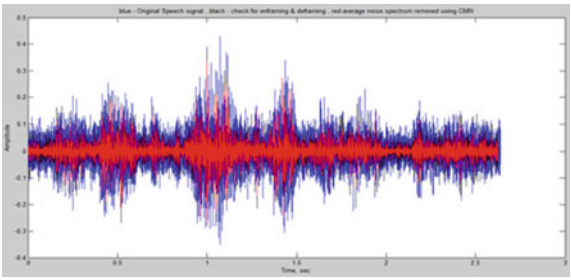

Fig. 13.3 Clean signal in *blue*, check signal in *black*, and enhanced signal in *red*

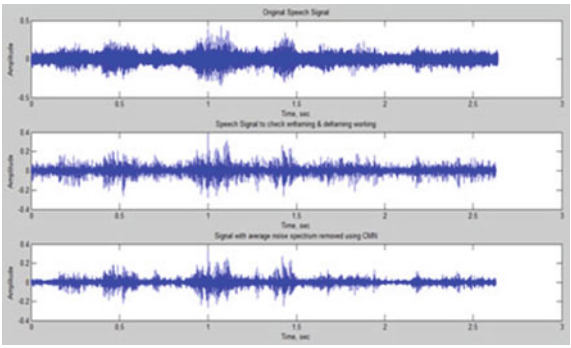

Fig. 13.4 Clean, check, and enhanced signal derived using CMN and MLLR

Table 13.1 Results derived using ANN

Sl. no.	Input class	Recognition rate (%)	False rejection rate (%)	False acceptance rate (%)
1	Class 1	87.60	10.04	2.35
2	Class 2	83.26	10.04	6.69
3	Class 3	91.95	5.69	2.34
4	Class 4	78.91	10.04	11.04
5	Class 5	91.95	1.34	4.69
6	Class 6	83.26	5.69	11.04
7	Class 7	91.95	5.69	2.34
8	Class 8	71.21	18.73	11.04
Overall		85.01	9.83	5.44

experimental results are shown in Tables 13.1 and 13.2. The GMM approach shows a success rate between 65.86 and 87.6 %, a rejection rate of 5.69–23.08 %, and a false acceptance rate between 2.35 and 15.39 %. The ANN, on the other hand, shows a success rate of 71.2–91.95 %, rejection performance between 1.34 and 18.73 %, and a false acceptance rate of 2.34–11.04 %. This improved performance of the ANN is due to its robustness, adaptive learning, and ability to retain the learning.

Table 13.2 Results derived using GMM

Sl. no.	Input class	Recognition rate (%)	False rejection rate (%)	False acceptance rate (%)
1	Class 1	83.26	14.39	2.35
2	Class 2	78.91	15.39	5.69
3	Class 3	87.60	10.04	2.35
4	Class 4	74.56	10.04	15.39
5	Class 5	87.60	5.69	5.69
6	Class 6	78.91	10.04	11.04
7	Class 7	91.95	5.69	2.35
8	Class 8	65.86	23.08	11.04
Overall		81.08	11.80	6.98

Table 13.3 Results showing computational complexity in GMM and ANN

Algorithm	GMM	ANN
Time (s)	64.29	50.45

The computational requirements of the two approaches recorded during training is shown in Table 13.3. It shows that the ANN takes lesser time to complete the processing. Thus, in terms of higher recognition accuracy, lower rejection, and false acceptance rates and reduced computational requirement, the ANN-based approach is superior compared to the GMM approach.

13.5 Conclusion

This work focuses on the classification of Assamese vowel speech and recognition using GMM and ANN. CMN and MLLR are used for speech enhancement of the data which is degraded due to noise. Feature extraction is done using MFCC. GMM and ANN approaches are used as classifiers for an ASR system. We found success rate of the GMM to be around 81 % and that of ANN to be above 85 %. The GMM approach shows a success rate between 65.86 to 87.6 %, a rejection rate of 5.69 to 23.08 % and a false acceptance rate between 2.35 to 15.39 %. The ANN, on the other hand, shows a success rate of 71.2 to 91.95 %, rejection performance between 1.34 to 18.73 % and a false acceptance rate of 2.34 to 11.04 %. The ANN further takes atleast 21 % lower computational time compared to the GMM approach. This improved performance of the ANN is due to its robustness, adaptive learning and ability to retain the learning.

References

1. Sarma M, Sarma KK (2013) Vowel phoneme segmentation for speaker identification using an ANN-based framework. J Intell Syst 22(2):111–130
2. Gait EA (1905) A history of assam (1926), Rev. edn. Thacker Spink and Co, Calcutta
3. Zhao B, Schultz T (2002) Towards robust parametric trajectory segmental model for vowel recognition. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing
4. Lippmann RP (1989) Review of neural networks for speech recognition. Neural Comput Spring 1(1):1–38
5. Huang WY, Lippmann RP (1987) Neural net and traditional classifiers. In: Proceedings of IEEE conference on neural information processing systems—natural and synthetic, New York
6. Reynolds DA, Rose RC (1995) Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans, Speech Audio Process, p 3
7. Hasan R, Jamil M, Rabbani G, Rahman S (2004) Speaker identification using mel frequency cepstral coefficients. In: Proceedings of 3rd international conference on electrical and computer engineering (ICECE 2004), Dhaka, Bangladesh
8. Sarma M, Sarma KK (2013) An ANN based approach to recognize initial phonemes of spoken words of assamese language. Appl Soft Comput 13(5):2281–2291
9. Haykin S (2009) Neural network and learning machine, 3rd edn. PHI Learning Private Limited, New Delhi