Tamil Speech Emotion Recognition Using Deep Belief Network(DBN)

M. Srikanth(✉), D. Pravena, and D. Govind

Centre for Computational Engineering and Networking (CEN),
Amrita School of Engineering, Amrita Vishwa Vidyapeetham,
Coimbatore 641112, Tamilnadu, India
srikanthmurali3@gmail.com, d.pravena@gmail.com, d_govind@cb.amrita.edu
http://www.amrita.edu/campus/coimbatore

Abstract. The proposed system shows the effectiveness of Deep Belief Network(DBN) over Gaussian Mixture model(GMM). The development of the proposed GMM-DBN system is by modeling GMM for each emotion independently using the extracted Mel frequency Cepstral Coefficient(MFCC) features from speech. The minimum distance between the distribution of features for each utterance with respect to each emotion model is derived as Bag of acoustic features(BoF) and plotted as histogram. In histogram, the count represents the number of feature distributions that are close to each emotion model. The BoF is passed in to DBN for developing train models. The effectiveness of the emotion recognition using DBN is empirically observed by increasing the Restricted Boltzmann machine(RBM) layers and further by tuning available parameters. The motivation is by testing the Classical German Speech emotion database(EmodB) with the proposed GMM-DBN system which gives the performance rate increase by 5% than the conventional MFCC-GMM system by empirical observation. Further testing of the proposed system over the recently developed simulated speech emotion database for Tamil language gives a comparable result for the emotion recognition. The effectiveness of the proposed model is empirically observed in EmodB.

1 Introduction

The Speech Emotion recognition (SER) can bridge gap between the Human-Machine interaction. The significant empirical observations are made decades ago [1] provides information about the importance of vocal cues for the expression of speech. Information about the emotional status of the speaker can improve the communication between the listeners and derive more understanding, especially the exact meaning in between words. The identification of emotion state in speech is termed as the SER. The SER is practically applied in many growing areas like automobile industry, robotics, e-learning centers, etc. The lack of understanding of emotion intelligence due to the unavailability of actual emotion data and procedures to collect data are existing due to previous survey works [2].

S.M. Thampi et al. (eds.), *Advances in Signal Processing and Intelligent Recognition Systems*, Advances in Intelligent Systems and Computing 678,
DOI 10.1007/978-3-319-67934-1_29

The difficulty is present in understanding emotion at the levels of physiology and psychology is present even before going to the actual analysis of speech data. The understanding about the difficulties in analyzing emotion made us focus on stages like feature extraction and modeling. During the initial stage, the previous research works were motivated towards emotion dependent parameter analysis by keeping intact the pattern classifiers for emotion recognition [3]. The later works were focused towards development of classifiers and the combination of classifiers Gaussian Mixture Model (GMM) based, Support Vector Machine (SVM) based, Deep Belief Nets (DBN) based [4], combined classifiers [5] by keeping same emotion dependent features [6]. The machine learning and data mining is giving better performance for SER system, but using such methods needs improvement [2].

DBN is one of the deep learning tool used for pattern recognition, voice and speech analysis. The experimental study leads to deeper models and architectures with many visible and hidden layers are disconnected for learning [7]. Many layers and parameters are learnt using deep models [4,8]. The deep learning tools are less used when huge number of parameters are needed for complicated learning process. Training is trapped at local minima and it is time consuming when layers are increased [9]. Achieving acceptable results is difficult. The tool to deal with such problem is DBN by creating deeper networks using many hidden layers [4]. DBN can be used in learning feature and classification. Representation of data is significant in machine learning. So, work done for processing features, extracting features and learning features must be taken more into concern [10]. The feature learning and machine learning can be done using available emotion databases.

Speech emotion database consist of two types namely: simulated emotion speech database and spontaneous emotion speech database. In terms of simulated emotion speech database, the expression conveyed by a subject is generation of various singular feelings of the individuals. However, the speaker has prior knowledge of emotion on which speech elicitation is recorded. Spontaneous emotion contains multiple emotion recorded from a real situation or a genuine conversation. The simulated emotion database is considered due to limited spontaneous emotion database and also not cost effective to develop. The available simulated emotion databases are German emotional speech database(EmodB) and speech under simulated and actual stress (SUSAS) database are the databases available popularly in simulated emotion database. SUSAS database contains the sample emotions of the excluded words elicited by different speakers. SUSAS database has lack of speech recording of continuous sentences. The database utilized here is Classical German emotion database(EmodB) for continuous sentences. The SER system described by Zhou et al., utilizes German Berlin Speech Emotion database by automated feature extraction. The performance acquired using DBN was 65%. The German(EmoDb) is a popular and publicly available emotion database. The motivation to use this database is due to previous experimental observation made in terms of classifiers and accuracy [11]. Further more, the empirical observation described by Pravena et al. [12] on emotionally

biased utterance shows higher performance compared to the emotionally neutral utterance for recently developed simulated emotion for Tamil language. The German(EmodB) database and Tamil database are applied in GMM-DBN system, and empirically studied. Speaker emotion recognition is classified into: (1) Speaker Independent and (2) Speaker Dependent. Speaker Independent system does not have a pre-training system for SER. In Speaker dependent, the SER is developed with a training of speakers voice beforehand. The SER used here is for Speaker dependent systems.

The study of the performance for SER using the proposed model are comparable to state of the art classifiers.

This paper discusses on the Modeling of SER system using the available speech emotion database in Sect. 2. The experimental results and its performances are discussed in Sect. 3. The conclusion and future work are discussed in Sect. 4.

2 Emotion Database

2.1 German (EmodB) Speech Emotion Database

EmodB is one of the publicly available popular classical emotion database. The data type available are Speech wave and Electroglottograph(EGG). The speech waves are considered and the total emotion utterances are 535 recorded from five female and five male speakers. Each subject speaks about ten different sentences in seven different emotions anger, happy, sad, neutral, boredom, disgust and fear, out of which only the first 5 emotions are considered for developing the model, sampling frequency is 16 Khz.

2.2 Tamil Speech Emotion Database

The established database consist of three language (Tamil, English and Malayalam) and contains emotions namely anger, happy, sad and neutral utterances. The Tamil database consist of 11 speakers out of which six were female and five were male. The database contains 220 utterances, present for each emotion (anger, happy, sad, neutral) of 9680 sentences(4 emotions * 11 speakers * 220 sentences). The signal recording is on dual channel with a sampling frequency of 48 Khz down sampled at 16 KhZ. The speech and EGG was recorded simultaneously. The present work deals with the speech wave.

3 Modeling of Speech Emotion Recognition System

3.1 MFCC-GMM System for Emotion Recognition

Weighted sum of Gaussian component densities are the representation of parametric probability density function. GMMs are commonly used as a features in a bio-metric system, such as vocal-tract related spectral features in a SER system [13]. Initially the work is to develop a MFCC-GMM system for the existing

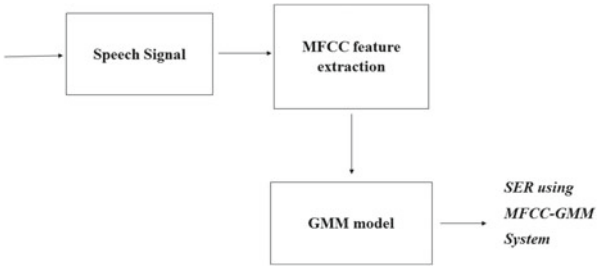

Fig. 1. MFCC-GMM system

German (EmodB) and Tamil speech emotion database. The Gaussian mixture model is a weighted sum of N component Gaussian densities as given by the equation,

$$p(y/\lambda) = \sum_{i=1}^{N} w_i \; g(y/\mu_i, \; \sum\nolimits_i) \tag{1}$$

where y is the features(D-dimensional continuous valued data vector), the mixture weights is $w_i, i = 1, \ldots, N$ and the component Gaussian densities

$$g\left(y/\mu_i, \Sigma_i\right), i = 1, ..., M \tag{2}$$

D-variate Gaussian function for each component density is,

$$g\left(y/\mu_i, \Sigma_i\right) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(y-\mu_i)'\Sigma_i^{-1}\left(y-\mu_i\right)\right\} \tag{3}$$

The proposed model here consist of state of the art Mel frequency cepstral coefficients(MFCCs). The MFCC 39 dynamic coefficients are extracted frame by frame from each speech signal. The frame is divided in terms of 20 ms frame with an overlap of 10 ms from each speech utterance, using Hamming window. Spectral leakage is avoided using Hamming window [12]. The GMM model is developed using 256 GMM for German(EmodB) database and 1024 GMM for Tamil database using Expectation maximization algorithm for all the emotions by taking these 39 dynamic MFCC coefficients consist of 13 MFCC features, 13 velocity(Δ) and 13 acceleration($\Delta - \Delta$) coefficients (Fig. 1).

3.2 GMM-DBN System for Emotion Recognition

The minimum distance between the distribution of features for each utterance with respect to each emotion model is averaged with over all emotion is represented as BoF using histogram plot. The BoF is passed in to Deep Belief Network (DBN) as a training vector for emotion recognition (Fig. 2).

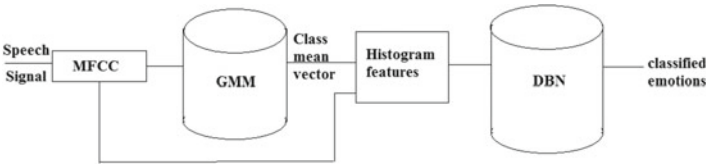

Fig. 2. GMM-DBN system

Experimental Analysis for GMM-DBN System

Let us consider the speech signals (S1, S2, ... Sn), n represents the total number of speech signals. The matrix contains the 39 dynamic MFCC coefficient vector frames F_i (f_1^i.....f_{39}^i), i represents the total number of frames for each speech emotion file.

$$S_1 = \begin{bmatrix} f_1^1 \ldots\ldots\ldots\ldots f_{39}^1 \\ f_1^2 \ldots\ldots\ldots\ldots f_{39}^2 \\ . \qquad\qquad . \\ . \qquad\qquad . \\ f_1^n \ldots\ldots\ldots\ldots f_{39}^n \end{bmatrix} \quad \ldots\ldots \quad \ldots\ldots \quad \ldots\ldots \quad S_n = \begin{bmatrix} f_1^1 \ldots\ldots\ldots\ldots f_{39}^1 \\ f_1^2 \ldots\ldots\ldots\ldots f_{39}^2 \\ . \qquad\qquad . \\ . \qquad\qquad . \\ f_1^n \ldots\ldots\ldots\ldots f_{39}^n \end{bmatrix}$$

The minimum distance is calculated for each frame of the speech emotion signal with respect to N-Gaussian mixture emotion (c1, c2, c3, c4) and averaged with over all N-Gaussian mixture emotion is represented using histogram (Fig. 3).

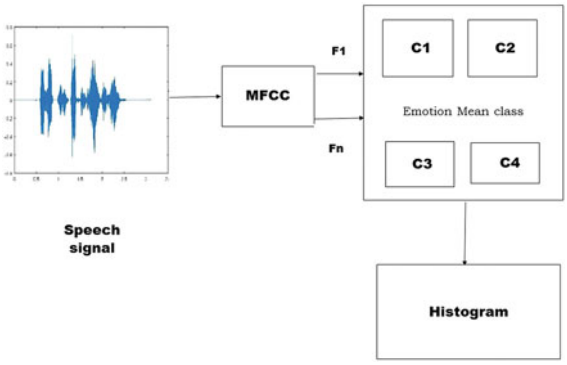

Fig. 3. Histogram

Deep Belief Network

DBN is a deep neural network consist of many hidden layers. DBN allows each RBM model in the sequence to receive different representation of the data. The proposed method is used to develop training in a supervised way [4]. The Restricted Boltzman machine network consist of set of visible units $v \in \{0,1\}^{n_v}$

and a set of hidden units $h \in \{0,1\}^{n_h}$ where n_v and n_h are the number of visible units and hidden units respectively. The connection between visible and hidden layers are disconnected as shown in Fig. 6.

The different types of RBMs can be modeled. The types defined are generative(data without labels) and discriminative(data with class lables). The sampling methods available are Gibbs, PCD(Persistent Contrastive Divergence), CD(Contrastive Divergence) and FEPCD(Free Energy in Persistent Contrastive Divergence) [14]. The hidden units are independent units due to the disconnection between them, the computation is by giving the training data v, the binary state h_i is set to 1 for each unit i and its probability is given by

$$P\left(h_i = 1/v\right) = g\left(b_i + \sum_j v_j w_{ji}\right) \tag{4}$$

where g(x) is log sigmoid function expressed as $g\left(x\right) = 1/\left(1 + \exp\left(-x\right)\right)$. The connection is not present in between hidden and visible units also, the computation is unbiased. The visible units sample state given the hidden unit vector is computed by,

$$P\left(v_j = 1/h\right) = g\left(a_j + \sum_i h_i w_{ij}\right) \tag{5}$$

Computation is difficult with large running time. So, the CD method is used [15]. In this method, the initialization of visible unit is made with respect to training data. The binary hidden units computation is according to the Eq. (4), the determination of binary hidden unit states leads to computation of v_j values according to the Eq. (5). There are some disadvantages noted that it is not exact due to imperfect gradient computation. This problem is overcome by using PCD method [4]. In this method, it uses the last updated step from the last chain state whereas in the CD method it uses training data as initial value for visible units. The imperfect gradient computation is reduced by using PCD. The FEPCD needs to run many times to obtain appropriate samples from the model and it is impossible. The PCD is being applied here in construction of the RBM layers.

4 Results and Discussion

The German speech emotion recognition for the proposed GMM-DBN system gives the characteristic of performance as shown in Fig. 4. The default 3 RBM based DBN model for which the performance is about 77.27% for PCD binary. The RBM layers are increased and through empirical observation the DBN architecture is constructed as shown in Fig. 5, and performance is about 78.45% as shown in Table 2 The proposed GMM-DBN system performance rate is increased by 5% with comparison to the conventional GMM model is tabulated in Table 1.

Similarly the recent developed simulated emotion for Tamil language is applied in the proposed GMM-DBN system gives the characteristics of performance as shown in Fig. 6. The empirical observation shows that the sampling

method Persistent Contrastive Divergence for the unit binary gives maximum performance with comparison to other parameters(CD, FEPCD), for both the databases by applying in the proposed GMM-DBN system. The default 3 RBM based DBN model performance is about 80.74%, comparable to state of the art classifiers.

Table 1. Comparison of MFCC-GMM system and GMM-DBN system performance for EmodB speech emotion database

Type	Performance
MFCC-GMM	73.28%
GMM-DBN	78.45%

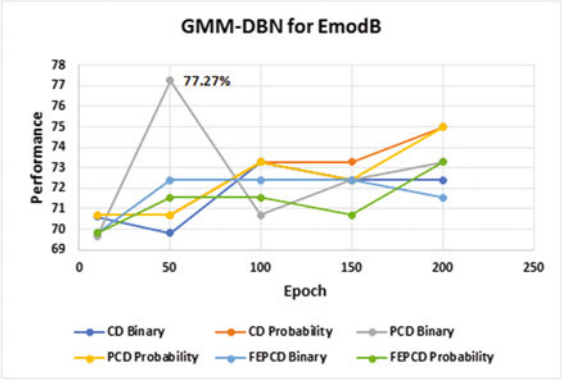

Fig. 4. Performance graph for GMM-DBN system

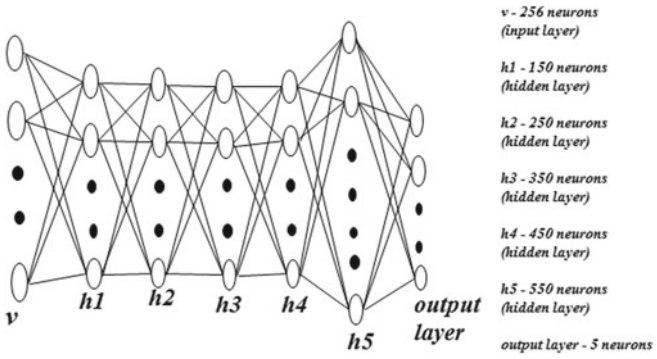

Fig. 5. DBN architecture for German(EmodB) database

Table 2. GMM-DBN system performance for EmodB database using different RBM layers

Layers	Performance
3	77.27%
5	78.45%

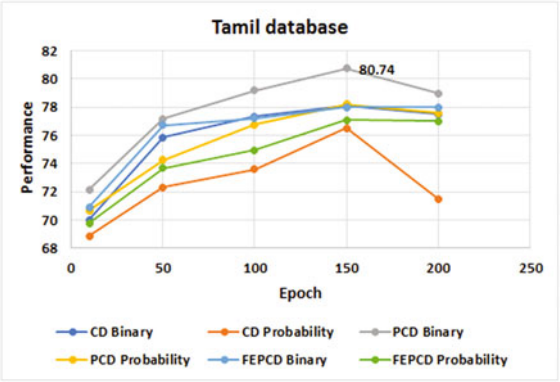

Fig. 6. Performance graph using Tamil speech emotion database for GMM-DBN system

5 Conclusion and Future Work

The proposed work here studies about the effectiveness of GMM-DBN system by testing on the Classical German database(EmodB). The empirical observation shows an emotion recognition performance rate of 5% increase than the conventional GMM model. The GMM-DBN system is also tested with the recently developed simulated emotion database for Tamil language and shows a comparable result for emotion recognition. The effectiveness of the proposed model is clearly observed in EmodB. The characteristics of the GMM-DBN system is empirically studied by various parameters and increase in RBM layers. The present study is used on Speaker dependent. However, the future work needs to concentrate on Speaker independent. Additionally, the present work needs to be improved for other Indian languages.

References

1. Schuller, B., Reiter, S., Muller, R., Al-Hames, M., Lang, M., Rigoll, G.: Speaker independent speech emotion recognition by ensemble classification pp. 864–867 (2005)
2. Scherer, K.R.: Vocal affect expression: a review and a model for future research. Psychol. Bull. **99**(2), 143 (1986)

3. Govind, D., Joy, T.T.: Improving the flexibility of dynamic prosody modification using instants of significant excitation. Circ. Syst. Sig. Process. **35**(7), 2518–2543 (2016)
4. Keyvanrad, M.A., Homayounpour, M.M.: A brief survey on deep belief networks and introducing a new object oriented toolbox (deebnet). arXiv preprint arXiv:1408.3264 (2014)
5. Ververidis, D., Kotropoulos, C.: A state of the art review on emotional speech databases. In: Proceedings of 1st Richmedia Conference, pp. 109–119. Citeseer (2003)
6. Williams, C.E., Stevens, K.N.: Emotions and speech: some acoustical correlates, vol. 52, pp. 1238–1250. ASA (1972)
7. Erickson, D.: Expressive speech: production, perception and application to speech synthesis. Acoust. Sci. Technol. **26**(4), 317–325 (2005)
8. Salakhutdinov, R., Hinton, G.: Deep Boltzmann machines. In: Artificial Intelligence and Statistics, pp. 448–455 (2009)
9. Liu, Y., Zhou, S., Chen, Q.: Discriminative deep belief networks for visual data classification. Pattern Recogn. **44**(10), 2287–2296 (2011)
10. Rong, J., Li, G., Chen, Y.-P.P.: Acoustic feature selection for automatic emotion recognition from speech. Inf. Process. Manage. **45**(3), 315–328 (2009)
11. Altun, H., Polat, G.: On the comparison of classifiers performance in emotion classification: critiques and suggestions. In: 2008 IEEE 16th Signal Processing, Communication and Applications Conference, SIU 2008, pp. 1–4. IEEE (2008)
12. Pravena, D., Govind, D.: Development of simulated emotion speech database for excitation source analysis. Int. J. Speech Technol. **20**, 327–338 (2017)
13. Reynolds, D.: Gaussian mixture models. In: Encyclopedia of Biometrics, pp. 827–832 (2015)
14. Tieleman, T.: Training restricted Boltzmann machines using approximations to the likelihood gradient. In: Proceedings of the 25th International Conference on Machine Learning, pp. 1064–1071. ACM (2008)
15. Carreira-Perpinan, M.A., Hinton, G.E.: On contrastive divergence learning. In: AISTATS, vol. 10, pp. 33–40. Citeseer (2005)