Text Independent Speaker and Emotion Independent Speech Recognition in Emotional Environment

A. Revathi and Y. Venkataramani

Abstract It is well known fact that the accuracy of the speaker identification or speech recognition using the speeches recorded in neutral environment is normally good. It has become a challenging work to improve the accuracy of the recognition system using the speeches recorded in emotional environment. This paper mainly discusses the effectiveness on the use of iterative clustering technique and Gaussian mixture modeling technique (GMM) for recognizing speech and speaker from the emotional speeches using Mel frequency perceptual linear predictive cepstral coefficients (MFPLPC) and MFPLPC concatenated with probability as a feature. For the emotion independent speech recognition, models are created for speeches of archetypal emotions such as boredom, disgust, fear, happy, neutral and sad and testing is done on the speeches of emotion anger. For the text independent speaker recognition, individual models are created for all speakers using speeches of nine utterances and testing is done using the speeches of a tenth utterance. 80 % of the data is used for training and 20 % of the data is used for testing. This system provides the average accuracy of 95 % for text independent speaker recognition and emotion independent speech recognition for the system tested on models developed using MFPLPC and MFPLPC concatenated with probability. Accuracy is increased by 1 %, if the group classification is done prior to speaker classification with reference to the set of male or female speakers forming a group. Text independent speaker recognition is also evaluated by doing group classification using clustering technique and speaker in a group is identified by applying the test vectors on the GMM models corresponding to the small set of speakers in a group and the accuracy obtained is 97 %.

Keywords Clustering · GMM · Speech recognition · Probability · MFPLPC · Emotions · Quantization

A. Revathi (✉) · Y. Venkataramani
Saranathan College of Engineering, Tiruchirappalli, India
e-mail: revathidhanabal@rediffmail.com

J.K. Mandal et al. (eds.), *Information Systems Design and Intelligent Applications*,
Advances in Intelligent Systems and Computing 339,
DOI 10.1007/978-81-322-2250-7_5

1 Introduction

In addition to the linguistic information, the speech signal contains the information regarding age, gender, social status, accent and emotional state of a speaker. It has become a challenging task to recognize a speaker, speech and emotion from emotional speeches. Each speaker expresses different emotions in different ways. Speech recognition on emotional speeches has found applications in call centers. People working in call centers may not behave in same manner when attending calls of the customers. When a customer experiences a negative emotion, the system has to adjust itself to the needs of the customer or pass the control to the human agents for giving alternate convenient reply to the customers. It also has found applications in controlling the hazardous processes where physical presence of humans is not possible. These systems can also be applied in health care systems for which treatments could be extended to the patients with depression and anxiety. Nwe et al. [1] have used short-time log frequency power coefficient as a feature and discrete HMM as a classifier in evaluating the performance of the emotion recognition system. Morrison et al. [2] have compared accuracy of emotion recognition system evaluated by using different classification techniques. Modulation spectral feature is used as a new feature by Wu et al. [3] for emotion recognition. Lee et al. [4] have used hierarchical binary classifier and acoustic & statistical feature for emotion recognition. Vogt and Andr [5] have used combination of pitch, energy and MFCC as feature for emotion recognition and they have done gender detection. Rao et al. [6] have used MFCC and GMM for recognizing emotions. Sapra et al. [7] has used modified MFCC feature and NN classifier for emotion recognition. Speaker identification in emotional environment has been done by Sahini [8] and he has used log frequency power coefficients as feature and evaluated the system using HMM, CHMM and SPHMM. Koolakudi et al. [9] have used MFCC and GMM for speaker recognition in emotional environment. This work is mainly focused on the use of features such as MFPLPC and MFPLPC concatenated with probability and iterative clustering modeling technique for recognizing speech and speaker by using the emotional speech database. Emotion independent speech recognition and text independent speaker recognition systems are evaluated by applying features of the test speeches on the clustering models developed using both features. Recognition rate is computed by considering the test speech being correctly identified for any one of the models corresponding to the feature for both text independent speaker recognition and emption independent speech recognition ant it is better as compared to the testing on individual models.

2 Feature Based on Cepstrum

The short-time speech spectrum for voiced speech sound has two components: (1) harmonic peaks due to the periodicity of voiced speech (2) glottal pulse shape. The excitation source decides the periodicity of voiced speech. It reflects the characteristics of speaker. The spectral envelope is shaped by formants which reflect the resonances of vocal tract. The variations among speakers are indicated by formant locations and bandwidth. MFCC is the feature widely used in speech recognition system. It represents the source characteristics of speech signal and based on the known variation of the human ear's critical bandwidth with frequencies, filters spaced linearly at low frequencies and logarithmically at high frequencies preferred to extract phonetically important characteristics of speech. perceptual linear predictive cepstrum (PLP) speech analysis method [10–12] is for modeling the speech auditory spectrum by the spectrum of low order all pole model. This perceptual feature mainly emphasizes the need for critical band analysis which integrates the energy spectral density in the frequency range (0–8) kHz to get the speech auditory spectrum. Loudness equalization is done to emphasize the spectrum in the upper and middle frequencies and cube root compression is performed to reduce the dynamics of the speech spectrum. Then inverse fast Fourier transform is done to get the signal in time domain. Autocorrelation method is used to find linear prediction coefficients. These prediction coefficients are converted into cepstral coefficients by using recursive procedure. Critical band analysis is done using 47 critical bands, when the frequencies are spaced in mel scale. The relationship between frequency in Mel and frequency in Hz is specified as in (1)

$$f(mel) = 2595 * \log(1 + f(Hz)/700) \tag{1}$$

3 Characteristics of Emotional Speech—Frequency Domain Analysis

The semantic part of the speech contains linguistic information which reveals the characteristics of the pronunciation of the utterances based on the rules of the language. Paralinguistic information refers to the implicit messages such as emotional state of the speaker. Speeches of the emotions such as anger, fear and happy are displaying the psychological behavior of the speaker such as high blood pressure and high heart rate. These speeches are loud, fast and enunciated with strong high frequency energy. On the other hand, speeches of the emotion sad reveal the characteristics of the speaker such as low blood pressure and low heart rate. These speeches are slow, low volume and possess little high frequency energy. This fact is demonstrated in Fig. 1.

Frequency analysis is done on the emotions of the speaker uttering the same sentence. Speech signals are converted into frames and frequency analysis is done on the frames. Frequencies are calculated on the basis of choosing the frequency bin which has high spectral energy. From the plot shown in Fig. 2, it is indicated that emotions such as anger, fear and happy have more number of frames with high frequency energy and the emotion sadness has very few frames with high frequency energy.

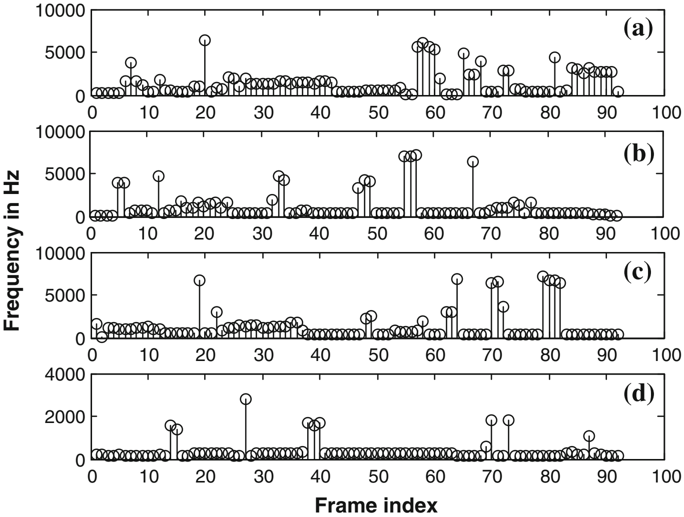
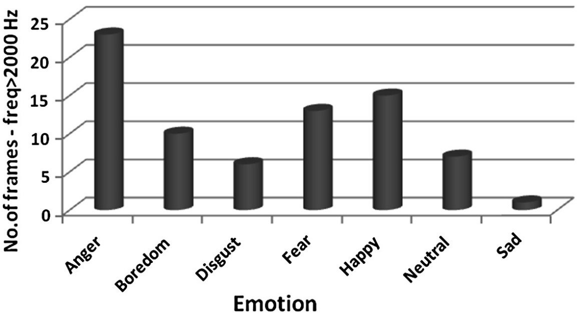


Fig. 1 Frequency distribution of speech in different emotions—**a** Anger. **b** Fear. **c** Happy. **d** sad

Fig. 2 Frequency analysis on emotions

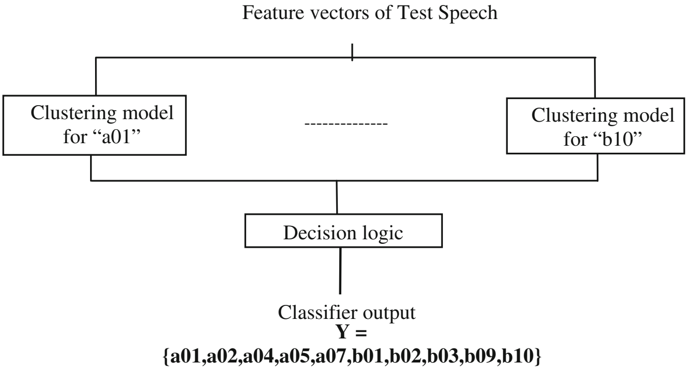


Similarly, frequency analysis is done on the speeches uttered by one speaker and the speech uttered by different speakers in the emotion anger.

4 Speech/Speaker Recognition Using Clustering Technique and GMM

Emotional speech database considered in this work is a Berlin database which contains about 500 utterances spoken by actors in happy, angry, anxious, fearful, bored and disgusted way as well as in a neutral version. Utterances are chosen from

10 different actors and ten different texts. Ten emotional utterances are collected from five male and female speakers respectively in the age ranging from 21 to 35 years. They are required to utter ten different utterances in Berlin in seven different emotions such as anger, boredom, disgust, fear, happy, neutral and sad. Speech recognition system generally involves the realization of speech signal as the message encoded as the sequence of one or more symbols. This is considered as recognizing the underlying sequence of symbols given a spoken utterance, the continuous speech signal is converted into the sequence of equally spaced discrete parameter vectors. For creating a training model, speech signal is first pre-emphasized using a difference operator. Hamming window is applied on differenced speech frames of 16 ms duration with overlapping of 8 ms. Then the MFPLPC features are extracted. In this work, probability as a feature is extracted by counting the number of samples whose spectral energy is greater than or equal to the average spectral energy of the frame and this feature is concatenated with MFPLPC. For each training model corresponding to continuous speeches, training set of K utterances are used, where each utterance constitutes an observation sequence of some appropriate spectral or temporal representation. The performance of speech or speaker recognition system based on perceptual features is evaluated by applying test speech vectors to the training models corresponding to the speakers or speeches. MFPLPC feature extraction is dealt in many literatures [10–12]. The usage of Clustering technique [9] for recognizing emotional speeches is depicted in Fig. 3. After the conventional preprocessing, features are extracted. Subsequently, training models are developed using clustering technique. During testing feature vectors are extracted and applied to the training models developed for the speeches. Average of minimum distances is computed for each model. Test speech is identified corresponding to the model which provides minimum of averages. Similarly, feature vectors of test speech are applied to 12 mixture GMM models, and log likelihood values are calculated for each model. Test speech is identified corresponding to the


Feature vectors of Test Speech
Clustering model for “a01”

Clustering model for “b10”
Decision logic
Classifier output
Y =
{a01,a02,a04,a05,a07,b01,b02,b03,b09,b10}

Fig. 3 Speech recognition phase using clustering technique

model which provides the largest log likelihood value. Accuracy can be improved for text independent speaker recognition by creating group models for Female and male speakers and testing is done after group is correctly identified.

5 Results and Discussion

The performance of emotion independent speech recognition is evaluated by considering the ten utterances spoken by ten actors. Training models are developed for speeches of emotions such as boredom, disgust, joy, fear, sad and neutral. Testing is done on the speeches of anger emotion. Text independent speaker recognition is done by developing training models for the speeches of nine utterances and testing is done on the speeches of tenth utterance with respect to all emotions. Figure 4 indicates the parallel group classifier for classifying speakers using emotional

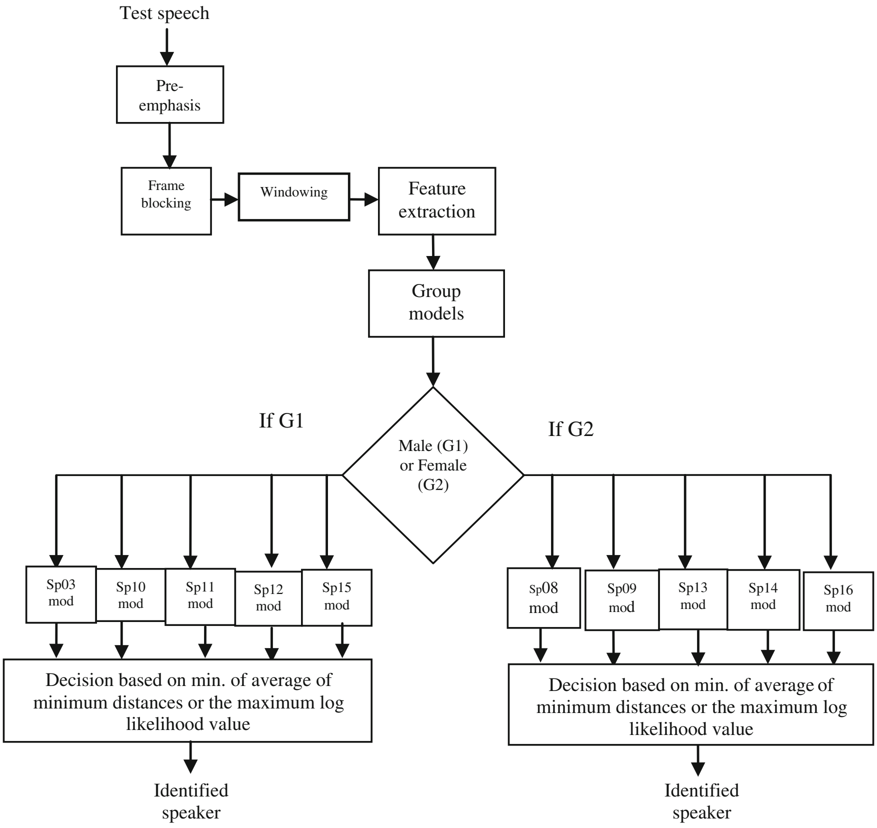


Fig. 4 Parallel group and specific emotional speaker classifier

Table 1 Accuracy—emotional speech and speaker recognition

Speaker	% RA			Speech	% RA		
	MFPLPC	MFPLPC + Prob	MFPLPC, MFPLPC + Prob		MFPLPC	MFPLPC + Prob	MFPLPC, MFPLPC + Prob
Sp03	80.9	82.59	86.52	a01	83.33	77.52	85.66
Sp08	91.63	94.42	94.89	a02	96.3	99.3	100
Sp09	100	87.88	100	a04	99.2	100	100
Sp10	98.72	85.9	100	a05	90.17	90.89	94.72
Sp11	82.38	82.86	92.86	a07	82.63	87.29	88.14
Sp12	100	100	100	b01	93.99	91.4	94.27
Sp13	91.86	82.56	95.35	b02	98.27	90	98.27
Sp14	77.02	79.5	83.85	b03	89.98	86.77	94.71
Sp15	99.26	94.03	99.26	b09	91.71	92.38	96.19
Sp16	75.69	91.44	94.18	b10	89.11	94.31	95.3
% Ave. RA	88.12	88.12	94.7	% Ave. RA	90.986	90.986	94.73

Table 2 Comparison chart—clustering and GMM—group models

Speaker	% RA					
	Clustering			GMM		
	MFPLPC	MFPLPC + Prob	MFPLPC, MFPLPC + Prob	MFPLPC	MFPLPC + Prob	MFPLPC, MFPLPC + Prob
Sp03	84.27	87.08	91.01	95.51	78.65	98.32
Sp08	91.63	94.42	94.89	86.51	90.7	93.94
Sp09	100	89.4	100	100	90.91	100
Sp10	98.72	85.9	100	100	58.98	100
Sp11	84.29	84.76	93.81	100	92.86	100
Sp12	100	100	100	94.05	95.24	98.81
Sp13	94.19	83.72	96.51	95.93	76.75	95.93
Sp14	77.64	82.61	84.47	78.26	78.89	87.58
Sp15	99.26	94.03	99.26	100	100	100
Sp16	82.2	92.81	95.81	86.65	78.43	88.02
% Ave. RA	91.22	89.473	95.576	93.691	84.141	96.56

database. First, clustering models for the training vectors corresponding to the female speakers and male speakers are developed. Group is correctly identified by first computing average of minimum distances for both the models and group classification is done based on minimum of averages using clustering technique. Subsequently, speaker is identified by applying the test vectors on the clustering models or GMM [10] models developed for the small set of female or male speakers in a group. Speaker is identified by computing the average of minimum distances for the clustering technique and classification is done based on the comparison of all values and identified speaker is the one with minimum of averages. Speaker is identified by first computing the log likelihood values for all the GMM models corresponding to the set of speakers in a group. Then, classification is done with reference to the model whose model log likelihood is the highest. Block diagram of a parallel group classifier and parallel specific emotional speaker classifier is shown in Fig. 4.

Emotion independent speech recognition is performed by creating training models corresponding to the speeches of the speaker of all emotions except anger. Testing is done on the speeches of a speaker concerned with anger emotion. Performance evaluation of emotion independent speech recognition and text independent speaker recognition is shown in Table 1. For some of the speeches or speakers, the accuracy is obtained as 100 %.

Text independent speaker recognition using clustering technique and GMM is evaluated by first doing group classification with reference to the set of female and male speakers. Accuracy is increased by 1 %, if group classification is done prior to respective speaker classification and it is indicated in Table 2.

6 Conclusions

We have proposed the use of clustering technique for emotion independent speech recognition and text independent speaker recognition using emotional database in Berlin language. MFPLPC features are extracted from the emotional speeches and clustering models are developed for speeches and speakers. Features of test speeches are applied to the models corresponding to the speech recognition and the accuracy is found to be good. This basic feature is concatenated with probability and models are developed. Overall accuracy is found to be same as that of basic feature. Overall accuracy is further improved by checking the correct identification of the test segment corresponding to any of the feature. Text independent speaker recognition is evaluated by considering the group models corresponding to the set of female and male speakers. Evaluation is done by developing clustering and GMM training models and the accuracy is found to be slightly better for GMM as compared to that of the system tested on clustering models. Accuracy can be further improved by using the database containing the large number of speech samples.

References

1. Nwe, T.L., Foo, S.W., De Silva, L.C.: Speech emotion recognition using hidden Markov models. Speech Commun. **41**, 603–623 (2003)
2. Morrison, D., Wang, R., De Silva, L.C.: Ensemble methods for spoken emotion recognition in call-centres. Speech Commun. **49**, 98–112 (2007)
3. Wu, S., Falk, T.H., Chan, W.-Y.: Automatic speech emotion recognition using modulation spectral features. Speech Commun. **53**, 768–785 (2011)
4. Lee, C.-C., Mower, E., Busso, C., Lee, S., Narayanan, S.: Emotion recognition using a hierarchical binary decision tree approach. Speech Commun. **53**, 1162–1171 (2011)
5. Vogt, T., Andr, E.: Improving automatic emotion recognition from speech via gender differentiation. In: Proceedings of Language Resources and Evaluation Conference (LREC 2006), Genoa (2006)
6. Rao, K.S., Kumar, T.P., Anusha, K., Leela, B., Bhavana, I., Gowtham, S.V.S.K.: Emotion recognition from speech. Int. J. Comput. Sci. Inf. Technol. **3**(2), 3603–3607 (2012)
7. Sapra, A., Panwar, N., Panwar, S.: Emotion recognition from speech. Int. J. Emerg. Technol. Adv. Eng. 3(2):341-345 (2013). ISSN 2250-2459, ISO 9001:2008 (Certified Journal)
8. Shahin, I.: Speaker identification in emotional environments. Iran. J. Electr. Comput. Eng. **8**(1), 41–46 (2009)
9. Koolagudi, S.G., Sharma, K., Rao, K.S.: Speaker recognition in emotional environment. Commun. Comput. Inf. Sci. **305**, 117–124 (2012)
10. Reynolds, D.A., Rose, R.C.: Text independent speaker identification using Gaussian mixture models. IEEE Trans. Speech Audio Process. **3**(1), 72–83 (1995)
11. Hermansky, H., Margon, N.: RASTA Processing of Speech. IEEE Trans. Speech Audio Process. **2**(4), 578–589 (1994)
12. Revathi, A., Venkataramani, Y.: Use of perceptual features in iterative clustering based Twins Identification System. Proceedings of International Conference on Computing, Communication and Networking, pp.1–6, (2008)