Emotion Recognition from Speech Signal Using Mel-Frequency Cepstral Coefficients

Onur Erdem Korkmaz[1,2], Ayten Atasoy[2]

[1]Ataturk University, Department of İspir Hamza Polat Vocational College, Erzurum, Turkey
onurerdem.korkmaz@atauni.edu.tr
[2]Karadeniz Technical University, Department of Electrical and Electronics Engineering, Trabzon, Turkey
ayten@ktu.edu.tr

Abstract

In this paper, mel-frequency cepstral coefficients are investigated for emotional content of speech signal. The features are extracted from spoken utterance. When these features are extracted, speech signal is divided small frames and each frame overlap a part of previous frame. The purpose of this overlap operation is to provide a smooth transition from one frame to the other and, to prevent information loss in the end of the frame. The length of frame and scroll time is important for emotion recognition applications. Also, we investigated the effects of different length frames and scroll times on the classification success of four emotions which are defined as happy, angry, neutral and sad. Those emotions were classified by using Support Vector Machine and k-Nearest Neighbors algorithms. In this study to determine the classification success, 10-Fold Cross Validation method was used and the maximum success rate was obtained as 98.7 %.

1. Introduction

Speech signal is one of the most natural communication methods among humanity. For this reason researchers start to work about speech signals to make productive and speed human computer interaction. Thus, the studies led the excellent computers that can recognize voice. Recently, in spite of significant developments about speech recognition, still we are considerably far natural interaction between human and computer due to unrecognizing emotion of human of computers. Therefore, studies which are taking out features of emotion mood of speaker and recognizing sense from speech become more important. It is believed that studies which are recognizing sense from speech are increased to performance of speaker recognition systems because of acquiring sense information from speech signals [1].

Especially, the emotion recognition process from speech is used on web films which required natural interaction in human-computer and instructional computer applications that are depend on perceived feeling of users [2]. Also, it is used automobile applications which adjust security systems depended on mental mood, aircraft cockpits and automatic translation systems when sense mood of speaker is important. Moreover, sense recognition process from speech is used in call center systems and communication applications [3].

Emotions don't have an scientific definition which is widely accepted [4]. However, people know how they feel emotions. Therefore researchers define emotions in different aspects. One of the most important problems in sense recognition systems is that there are a lot of the sense classes. In the results of many researches, it is clarified that linguists defined feelings with too many classes. But, it is hard to make classification between feeling classes [5]. For example, besides the primary colors, there are a lot of their shadows too. Nevertheless the colors are defined with primary colors. Like this, feeling classes have six different groups such as angry, hate, fear, happy, sad and confused.

Studies which are related sense recognition from voice have three steps in basically [6]. In the first step, data base which are very important in these studies are determined. It is important to use the available data in order to save time and thus, it is easy to compare the results of the studies with each other. In the second step, it actualizes feature extraction from speech. The last step is classification process of the emotions with different techniques. In this study, Support Vector Machine (SVM) and k-Nearest Neighbors (k-NN) classifiers algorithms are used for recognizing the four different emotions defined as happy, angry, neutral and sad.

2. Database

In this paper we used new database which name is EmoSTAR. This database contains four different emotional classes and each class has English and Turkish sample speech. These are Angry, Happy, Neutral and Sad. EmoSTAR has prepared by using TV-internet and it has total 393 samples [7]. Neutral samples were taken newsreader. Happy samples were taken award ceremony as Oscar and Golden Globe. These utterances both Happy and Neutral are neutral samples. Angry samples were taken movie and sad samples were taken video on internet. Also these utterance both neutral and sad are unnatural [7]. This database relating to information can be seen Table 1.

Table 1. Emotion classes and samples on EmoSTAR (E: English, T: Turkish)

	Angry	Neutral	Happy	Sad
Men	33 E – 30 T	35 E – 34 T	45 E	12 E
Women	40 E	37 E – 20 T	37 E	51 E – 19 T
Total	103	126	82	82

The other databases in literature are Berlin emotional speech database [8], Danish emotional speech database [9], Sensitive Artificial Listener (SAL) [10], Airplane Behaviour Corpus (ABC) [11], Speech Under Simulated and Actual Stress (SUSAS) [12], Audiovisual Interest Corpus (AVIC) [13], SmartKom [14] and FAU AIBO [15].

3. Feature Extraction

In this paper, the features used are mel-frequency cepstral coefficients (Mfcc). Mel frequency cepstral coefficient is one of the best distinctive features of emotion recognition problems [16]. Mfcc is often used in the area of speech signal processing because this feature imitate the hearing of the human ear. Mfcc extraction steps are framing, windowing, Fast Fourier transform, mel-frequency conversion and cepstral.

Firstly the voice is obtained and divided into frame because signals stable small time intervals. Also, in this study, for the different frame lengths, the effects on the success of emotion recognition process are investigated. As seen in Figure 1, voice signal is divided into frames and each frame overlap a portion of a previous frame. The purpose of these overlap operations ensure a smooth transition from frame to other frame and prevent loss of information the end of the frame.

Second step of obtaining mel-frequency cepstral coefficient is the windowing. Voice signal is windowed because of resolving at the beginning and end of the frame. Another object reduces spectral effects [17], so the central region of the voice signal is strengthened, the edge regions are attenuated. Widely used window functions are Hamming, Hanning, Blackman, Gauss, Rectangle and Triangle.

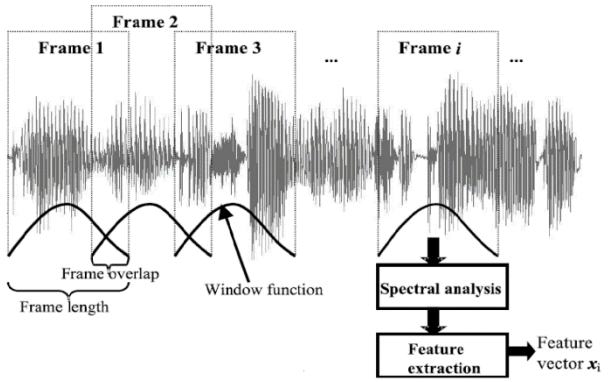

Fig. 1. Framing and windowing operations.

The most commonly used among them is Hamming window function. Hamming and other window function equations are expressed as follows.

Hamming:

$$w\,[k+1] = 0.54 - 0.46\cos\left(2\pi\frac{k}{N-1}\right) \quad (1)$$

Hanning:

$$w[k+1] = 0.5\left(1-\cos\left(2\pi\frac{k}{N-1}\right)\right) \quad (2)$$

Gauss:

$$w[k+1] = e^{\frac{1}{2}(a\frac{k-\frac{N}{2}}{\frac{N}{2}})^2} \quad (3)$$

Rectangle:

$$w[k+1] = 1 \quad (4)$$

Third step is fast Fourier transform (FFT). In this stage, N samples of each frame are applied FFT and translated frequency space. Discrete Fourier transform for each frame with N samples, are given as in equation 5;

$$X_n = \sum_{k=0}^{N-1}\left(x_k\, e^{-2\pi jkn/N}\right) \quad (5)$$

Fourth step is the mel-frequency conversion stage. Research in reference [18] has shown that 1 kHz to minor signal is linear but greater than 1 kHz is logarithmic. Conversion between the Mel scale and frequency scale is applied according to equation 6;

$$mel\,(f) = 2595 * \log\left(1+\frac{f}{700}\right) \quad (6)$$

The final step is to obtain cepstral coefficients. Cepstral definition is expressed as follows;

$$c(m) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\log\left|\mathrm{S(w)}e^{jwm}\mathrm{d(w)}\right| \quad (7)$$

4. Classifier

In this paper, extracted features from voice signals were analyzed by using SVM and k-NN classification algorithms.

4.1. Support Vector Machine

A support vector machine classifier separates the different class with a maximal margin. SVM during learning optimizes decision boundary between classes for distance to maximum. SVM creates a model using the training set. A new sample classifier decides by looking at the models. SVM algorithm is classified the data by using linear or non-linear function. This method based on prediction.

4.2. k Nearest Neighbors

k-NN method that is one of the commonly used classification algorithms in wide application fields. This method determines a class to a new observation. This new observation joins one of known classes in the sample set [19].

K nearest neighbors classifier makes classification looking at each other at close to the training examples and it is one of the most basic pattern recognition methods. This classifier makes classification looking at the value given k (e.g. 3 or 5) neighbors. In this study value of k was selected 3. The classification is performed using the known class of a feature vector so it is a sample-based method [19]. For measuring the distance between neighbors, the Euclidean distance is used.

5. Measurement

The success results of this study were evaluated by using 10-fold cross validation method and this process was repeated 25 times. Then success rates of SVM and k-NN classifiers are given in tables as the best success result and average of the repeated 25 times results.

The feature vector was occurred in different sizes because of mel frequency cepstral coefficient feature. Its feature for using frame and every voice sample since different size eventually line number 12 or different number (e.g. 13, 14) and column number will be different depending on the number of frames, which appeared vector. Because of the using different statistical

calculations, voice signals are represented by one-dimensional line matrix. Statistical calculations used are mean, median, standard deviation, skewness, kurtosis, maximum, minimum and range. Each statistical calculation result is line with 12, 13 or 14 dimensional column matrix. Such as 1x12 or 1x13, each feature vector represents the voice.

Also energy (+1) and zero cepstral coefficient (+1) are added to the feature vectors and eventually 1x14-dimensional feature vector was obtained. Moreover the first (1x14) and second (1x14) derivative of the 14-dimensional vector was calculated and 42-dimensional vector was obtained.

Both mel frequency cepstral coefficient feature vectors of different sizes and different frame size-shifting time are investigated using support vector machine and k nearest neighbors classifier methods.

From Table 2 and Table 3, statistical values of the Mfcc and energy (E), zero cepstral coefficient (0), first derivative (d), second derivative (D) of the Mfcc can be seen and extracted features were analyzed by using SVM and k-NN classifiers respectively. Looking at each table, it is seen that the success of SVM classifier is better than the success of k-NN classifier.

In the last part of the study different frame length and shift duration was investigated for two classifiers. The success rate of emotion recognition process with SVM and k-NN classifiers are given in Table 4 and Table 5 respectively.

As seen from Table 4, the best frame duration for support vector machine is 30 ms, best shift duration is 10 ms and optimal number of cepstral coefficient is 12. Also as given in Table 5, the best frame duration for k nearest neighbors is 30 ms and best shift duration is 10 ms also optimal number of cepstral coefficient is 14.

Consequently, the success results of SVM and k-NN methods for emotion recognition process were compared in terms of success and the experimental results showed that SVM is better than k-NN for EmoSTAR database.

Table 2. The results obtained using Support Vector Machine – Mel Frequency Cepstral Coefficient

Features	Number of Feature	Best Results (%)	Average Results (%)
Mfcc_mean (12) Mfcc_median (12) Mfcc_skewness (12) Mfcc_kurtosis (12) Mfcc_max (12) Mfcc_min (12) Mfcc_range (12)	96	97.5	96.9
MfccE_mean (13) MfccE_median (13) MfccE_skewness (13) MfccE_kurtosis (13) MfccE_max (13) MfccE_min (13) MfccE_range (13)	104	97	96.6
Mfcc_E0dD_mean (42)	42	94	92.6

Table 3. The results obtained using k Nearest Neighbors Classifier – Mel Frequency Cepstral Coefficient

Features	Number of Feature	Best Results (%)	Average Results (%)
Mfcc_mean (12) Mfcc_median (12) Mfcc_skewness (12) Mfcc_kurtosis (12) Mfcc_max (12) Mfcc_min (12) Mfcc_range (12)	96	93.3	92.3
MfccE_mean (13) MfccE_median (13) MfccE_skewness (13) MfccE_kurtosis (13) MfccE_max (13) MfccE_min (13) MfccE_range (13)	104	89.3	88.4
Mfcc_E0dD_mean (42)	42	81.9	81

Table 4. The results of different frame length (ms) and shift duration (ms) on Mfcc for Support Vector Machine

Features	Number of Feature (%)	Best Result (%)	Average Results (%)
Mfcc_30_10	12	99.5	98.7
Mfcc_20_10	12	98.5	98.3
Mfcc_30_10	14	98.5	98.2
Mfcc_30_20	12	99	97.8
Mfcc_50_30	12	99	97.7
Mfcc_40_10	14	98.5	96.9

Table 5. The results of different frame length (ms) and shift duration (ms) on Mfcc for k Nearest Neighbors

Features	Number of Feature	Best Result (%)	Average Results (%)
Mfcc_30_10	14	93.3	92.3
Mfcc_40_10	14	93.3	92.1
Mfcc_30_10	12	91.6	90.6
Mfcc_50_30	12	91.3	90.4
Mfcc_20_10	12	91.1	90
Mfcc_30_20	12	91.1	90

5. Conclusion

In this study, mel-frequency cepstral coefficients are investigated for emotional content of speech signals. Statistical values of the Mfcc and energy, zero cepstral coefficient, first derivative, second derivative of the Mfcc are investigated and extracted features are analyzed by using SVM and k-NN classifiers respectively. Then, the results of different frame length and shift duration on Mfcc are investigated for SVM and k-NN algorithms. The best results obtained for 12 dimension feature vector of Mfcc. Consequently, the best score is 97,5 % for EmoSTAR database. Experimental results showed that SVM method is better than k-NN method on the emotion classification.

6. References

[1] J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion recognition in speech using neural networks," *Neural Computing & Applications,* vol. 290-296, 2000.

[2] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine - Belief Network architecture," *2004 Ieee International Conference on Acoustics, Speech, and Signal Processing, Vol I, Proceedings,* pp. 577-580, 2004.

[3] D. Q. Sun, J. Pan, Q. Y. Cao, T. Li, and F. Yang, "Ubiquitous computing service model based on SPKI/SDSI," *Dynamics of Continuous Discrete and Impulsive Systems-Series B-Applications & Algorithms,* vol. 13e, pp. 2218-2223, Dec 2006.

[4] P. R. Kleinginna Jr and A. M. Kleinginna, "A categorized list of emotional definitions with suggestions for a consensual definition, Motivation Emotion," *5,* vol. 4, pp. 345-379, 1981.

[5] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz*, et al.*, "Emotion recognition in human-computer interaction," *Ieee Signal Processing Magazine,* vol. 18, pp. 32-80, Jan 2001.

[6] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition,* vol. 44, pp. 572-587, Mar 2011.

[7] C. Parlak, B. Diri, and F. Gürgen, "A Cross-Corpus Experiment in Speech Emotion Recognition " *International Workshop on Speech, Language and Audio in Multimedia (SLAM 2014),* pp. 58-61, 2014.

[8] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German Emotional Speech," *Interspeech* pp. 1517-1520, 2005.

[9] I. Engbert, S. and A. Hansen, V., "Documantetion of the danish emotional speech database (DES)." *Center for Person Kommunikation, Tech. Rep., Denmark,* 2007.

[10] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie*, et al.*, "The HUMAINE database: Addresing the Collection and Annotation of Naturalistic and Induced Emotional Data. ," *Affective Computing and Intelligent Interaction, Proceedings,* p. 12, 2007.

[11] B. Schuller, M. Wimmer, D. Arsic, G. Rigol, and B. Radig, "Audiovisual behavior modeling by combined feature spaces " *ICASSP 2007,* pp. 733-736, 2007.

[12] J. Hansen and B.-G. S., "Getting started with susas: A speech under simulated and actual stress database," *EUROSPEECH-97,* vol. 4, pp. 1743-1746, 1997.

[13] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer*, et al.*, "Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application " *Image and vision Computing Jornal (IMAVIS),* 2009b.

[14] S. Steininger, F. Schiel, O. Dioubina, and S. Raubold, "Development of user state conventions for the multimodal corpus in smartkom.," *Workshop on Multimodal Resources and Multimodal System Evaluation,* pp. 33-37, 2002.

[15] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt*, et al.*, "Combining Efforts for Improving Automatic Classification of Emotional User States.," *IS-LTC 2006,* pp. 240-245, 2006.

[16] N. Sato and Y. Obuchi, "Emotion Recognition using Mel-Frequency Cepstral Coefficients," *Journal of Natural Language Processing,* pp. 83-96, 01/2007.

[17] L. R. Rabiner and B. H. Juang, "Fundamentals of speech recognition " *Prentice Hall, Englewood Cliffs,* 1993.

[18] S. Karasartova, "Metinden bağımsız konuşmacı tanıma sistemlerinin incelenmesi ve gerçekleştirilmesi Yüksek Lisans Tezi," Ankara Üniversitesi Elektronik Mühendisliği Anabilim Dalı, 2011.

[19] H. Küçük, C. Tepe, and İ. Eminoğlu, "K-En Yakın Komşu Algoritması ve Destek Vektör Makinesi Yöntemleri İle EMG İşaretlerinin Sınıflandırılması," *IEEE,* 2013.