# Feature Extraction from Speech Data for Emotion Recognition

S. Demircan and H. Kahramanlı

***Abstract*—In recent years the workings which requires human-machine interaction such as speech recognition, emotion recognition from speech recognition is increasing. Not only the speech recognition also the features during the conversation is studied like melody, emotion, pitch, emphasis. It has been proven with the research that it can be reached meaningful results using prosodic features of speech. In this paper we performed pre-processing necessary for emotion recognition from speech data. We extract features from speech signal. To recognize emotion it has been extracted Mel Frequency Cepstral Coefficients (MFCC) from the signals. And we classified with k-NN algorithm.**

***Index Terms*—Speech processing, speech recognition, emotion recognition, MFCC.**

## I. Introduction

It is well known that emotional conditions such as anger, sadness and delight can have effect on speech sound. This effect can be observed mainly in the suprasegmental features, such as F0, intensity and temporal characteristics of speech. Since muscle tension may be raised in some emotional conditions, there is a possibility that some segmental features are also influenced by the speaker's emotional conditions [1].

Studies of signal processing of emotional in speech recently have been investigated:

In reference [2], X. M. Cheng and his friends analysis the feather of the time, amplitude, pitch and formant construction involved such four emotions as happiness, anger, surprise and sorrow in their paper. Through comparison with non-emotional quiet speech signal, they sum up the distribution law of emotional feather including different emotional speech. Nine emotional features were extracted form emotional speech for recognizing emotion. They introduce two emotional recognition methods based on principal component analysis and the results show that the method can provide an effective solution to emotional recognition.

In Reference [3], D. Ververidis and friend work they introduce a more fine-grained yet robust set of spectral features: statistics of Mel-Frequency Cepstral Coefficients computed over three phoneme type classes of interest – stressed vowels, unstressed vowels and consonants in the utterance. They investigate performance of their features in the task of speaker-independent emotion recognition using two publicly available datasets. Their experimental results clearly indicate that indeed both the richer set of spectral features and the differentiation between phoneme type classes are beneficial for the task. Classification accuracies are consistently higher for their features compared to prosodic or utterance-level spectral features. They show that, while there is no significant dependence for utterance-level prosodic features, accuracy of emotion recognition using class- level spectral features increases with the utterance length.

Nwe and friends [4], a text independent method of emotion classification of speech is proposed in their paper. The proposed method makes use of short time log frequency power coefficients (LFPC) to represent the speech signals and a discrete hidden Markov model (HMM) as the classifier. The emotions are classified into six categories. Performance of the LFPC feature parameters is compared with that of the linear prediction Cepstral coefficients (LPCC) and mel-frequency Cepstral coefficients (MFCC) feature parameters commonly used in speech recognition systems. Results show that the proposed system yields an average accuracy of 78%and the best accuracy of 96% in the classification of six emotions. Results also reveal that LFPC is a better choice as feature parameters for emotion classification than the traditional feature parameters.

This paper is organized as follows. In Section II, we give an overview of the emotion recognition system. MFCC is described in Section III. In Section IV, extracted features and experimental results are presented. Finally, discussions and conclusions are given in Section V.

## II. Speech Recognition

In this section we first briefly review how the speech signal recognition is becoming. It is known that the speech signal is one of the most complex signals to recognize. First of all the signal get through some pre-processing for analyzing.

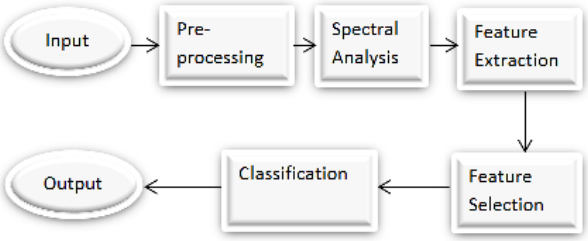


Fig. 1. Speech Recognition.

A block diagram of the speech recognition is shown as Fig. 1 [5].

In theory it should be possible to recognize speech directly from the signal. However, because of the large variability of the speech signal, it is a good idea to perform

Manuscript received July 9, 2013; revised October 1, 2013.

S. Demircan is with the Department of Computer Engineering in the University of Selcuk, Konya, Türkiye (e-mail: semiye@selcuk.edu.tr).

DOI: 10.7763/JACN.2014.V2.76

some form of feature extraction to reduce the variability [6].

Feature extraction is the most important stage of the recognition. There are many kinds of feature extraction methods. Some of the parametric representations are The mel-frequency cepstrum coefficients (MFCC), the linear-frequency cepstrum coefficients (LFCC), the linear prediction coefficients (LPC), the reflection coefficients (RC), and the cepstrum coefficients derived from the linear prediction coefficients (LPCC) [7].

After the feature extraction if necessary (generally according to classifying method) feature selection is performed.

The last stage of the recognition is classifying.

In our paper we extracted the features using mel-frequency cepstrum coefficients (MFCC) method (see Fig. 2).

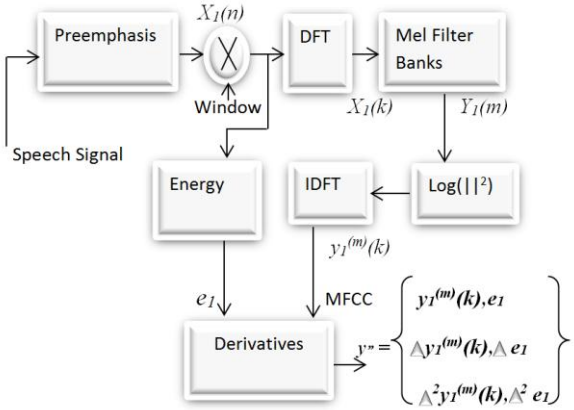


Fig. 2. MFCC.

## III. MFCC

The major stages of MFCC can be summarized as follows [8]:

*Preemphasis:* A preemphasis of high frequencies is therefore required to obtain similar amplitude for all formants. Such processing is usually obtained by filtering the speech signal with a first order FIR filter whose transfer function in the *z*-domain is:

$$H(z) = 1-\alpha.z^{-1} \tag{1}$$

$\alpha$ being the preemphasis parameter. In essence, in the time domain, the preemphasized signal is related to the input signal by the relation:

$$x'(n)=x(n)-\alpha\, x(n-1) \tag{2}$$

*Windowing:* Traditional methods for spectral evaluation are reliable in the case of a stationary signal. For voice, this holds only within the short time analysis can be performed by "windowing" a signal $x'(n)$ into a succession of windowed sequences $x_t(n)$ $t$=1, 2,…,$T$, called frames, which are then individually processed:

$$x'(n) \equiv x'(n-t.Q), \quad 0\le n<N \quad 1\le t\le T \tag{3}$$
$$x_t(n) \equiv w(n).\, x'(n)$$

where $w(n)$ is the impulse response of the window. Each frame is shifted by a temporal length $Q$. If $Q$=$N$, frames do not temporally overlap while if $Q$<$N$, $N$–$Q$ samples at the end of a frame $x'(n)$ duplicated at the end of the following frame $x'_{t+1}(n)$.

*Spectral analysis:* The standart methods for spectral analysis rely on the Fourier transform of $x_t(n)$:$X_t(e^{j\omega})$. Computational complexity is greatly reduced if $X_t(e^{j\omega})$ is evaluated only for a discrete number of $\omega$ values. If such values are equally spaced, for instance considering $\omega=2\pi k/N$, then the discreate Fourier Transform (DFT) of all frames of the signal is obtained:

$$X_1(k) = X_t(e^{j2\pi k/N}),\ k=0,\ldots,N\text{-}1 \tag{4}$$

*Filter bank processing:* Spectral analysis reveals those speech signal futures which are mainly due $t$ the shape of the vocal tract. Spectral futures of speech are generally obtained as the exit of filter banks, which properly integrate a spectrum at defined frequency ranges. A set of 24 band-pass filters is generally used since it simulates human ear processing.

There are many methods to implement such filters. A computationally inexpensive method consists of performing filtering directly in the DFT domain. DTF responces of of the filters are simply shifted and frequency warped versions of a triangular window $U_{\Delta m}(k)$:

$$U_{\Delta m}(k)=\begin{cases} |k| < \Delta_m \longrightarrow 1-|k|/\Delta_m \\ |k| \ge \Delta_m \longrightarrow 0 \end{cases} \tag{5}$$

where $k$ is the DFT domain index, and $2\Delta_m$ is the size of the $m$-th filter bank triangular window. The $m$-th filter bank output is given by:

$$Y_1(m) = \sum_{k=b_m-\Delta_m}^{b_m+\Delta_m} X_1(k)U_{\Delta m}(k+b_m) \tag{6}$$

The central frequency may be computing according to $b_m = b_{m-1} + \Delta_m$ , and , for $\frac{\omega}{2\pi Tc} = f$>1 kHz , $\Delta_m$ is chosen so that 10 uniformly spaced filters are obtained. For $f$>1 kHz, the following approximation can be used: $\Delta_m = 1.2 \times \Delta_{m-1}$.

*log energy computation:* the previous procedure has the role of smooting the spectrum, performaing a processing that the similar to that is similar to that executed by human ear. The next step consisit of computing the algorithm os square magnitude of the coefficients $Y_1(m)$ obtained with (Eq. 6). This reduces the simply computing the logarithm of magnitude of the coefficients, because of logarithm algebraic property which brings back the logarithm of a power to a multiplication by a scaling factor.

*Mel frequency cepstrum computation:* the final procedure for the Mel frequency cepstrum computation (MFCC) consist of performing the inverse DFT on the logarithm of the magnitude of the filter bank output:

$$y_t^{(m)}(k)= \sum_{m-1}^{M} log\{|Yt(m)|\}.cos\left(k\left(m-\tfrac{1}{2}\right)\tfrac{\pi}{M}\right),\ k=0,\ldots,L \tag{7}$$

In our application, VOICEBOX: Speech Processing Toolbox for MATLAB package is used for coding MFCC.

## VI. KNN

The most basic instance-based method is the *k*-NEAREST NEIGHBOR algorithm. This algorithm assumes all instances correspond to points in the *n*-dimensional space $\Re^n$. The nearest neighbors of an instance are defined in terms of the standard Euclidean distance. More precisely, let an arbitrary instance *x* be described by the feature vector

$$\langle a_1(x), a_2(x), \ldots a_n(x), \rangle \quad (8)$$

where $a_r(x)$ denotes the value of the $r_{\text{th}}$ attribute of instance *x*. Then the distance between two instances $x_i$ and $x_j$ is defined to be d ($x_i$, $x_j$), where

$$d(x_i,x_j) \equiv \sqrt{\sum_{r=1}^{n}\left(a_r(x_i) - a_r(x_j)\right)^2} \quad (9)$$

In nearest-neighbor learning the target function may be either discrete-valued or real-valued [10].

## V. FEATURE EXTRACTION AND CLASSIFICATION

In this paper we used Berlin Database [9]. In Berlin Database there are 7 emotional conditions. This emotions are Ager, Boredom, disgust, Anxiety(Fear), Happiness, Sadness, Normal. Ten different texts vocalized from 10 different actors. The speech data is divide two part (train %80 and test % 20).


Fig. 3. The changing of the data size

Firstly 16 mel cepstrum coefficients extracted from data. Each data size is reduced by a certain amount (window size=128). Although a certain degree of decreasing the data, data is still a great size for us to classifying. Each property of each Speech data is obtained seven statistical values. Statistical values are minimum, maximum, mean, standard deviation, median, skewness, and kurtasis.

We can show that the changing of the data size by using one data in Fig. 3.

AS shown in Fig. 3 at the beginning of the process one data size was 24933×1. After the MFCC the size was decreased 388×16. After the statistical process the size is become 7×16.

We classified our data with *k*-Nearest Neighbor Algorithm. Classification success is found %50.

## REFERENCES

[1] S. Katari, *Handbook of Neural Network for Speech Processing*, Artecch House,2000.
[2] X. M. Cheng, P. Y. Cheng, and L. Zhao, "A study on emotional feature analysis and recognition in speech signal," in *Proc. International Conference on Measuring Technology and Mechatronics Automation*, 2009, IEEE, pp. 418-420.
[3] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, Sep. 2006.
[4] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, 2003.
[5] M. W. Bhatti, Y. W. Y. Wang, and L. G. L. Guan, "A neural network approach for human emotion recognition in speech," in *Proc. 2004 IEEE International Symposium on Circuits and Systems IEEE Cat No04CH37512*, vol. 2, pp. 0–3, 2004.
[6] D. B. Roe and J. G. Wilpon, *Voice Communication Between Humans and Machines,* National Academy Press, 1994, pp.177.
[7] S. B. Davis, P. Mermelstein, D. F. Cooper, and P. Nye, "Comparison of Parametric Representations for Monosyllabic Word Recognation" vol. 61, 1980.
[8] C. Becchetti and L. P. Ricotti, *Speech Recognition;theory an C++ Implementation,*3rd ed., John Wiley &Sons, 2004, pp.125– 135
[9] Berlin database of emotional speech. [Online]. Available: http://pascal.kgw.tu-berlin.de/emodb/index-1280.html
[10] T. M. Mitchell, *Machine Learning*, McGra-Hill Companies, 1997, pp.231–232.

**S. Demircan** was born on July 26, 1980 in Konya, Türkiye. She received the B. engineering degree and the M. Sc. Degree in the department of the Computer Engineering from Selcuk University, Konya, Türkiye in 2002 and 2009, respectively. Currently, she is a PhD student in the same department. She is a research assistant in the Department of computer engineering in the University of Selcuk, Konya, Türkiye. Her research interests are multi-agent systems, intelligent agents, optimization, artificial immune system and speech recognition.