IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

Wavelet and Fourier Features Based Emotion Recognition of Speech Signals

Nisha Beegum S

M.Tech Scholar, Department of ECE, MZCE, Kadammanitta, Pathanamthitta, Kerala

Abstract: The studies have been performed on harmony features for speech emotion recognition. The first- and second-order differences of harmony features play an important role in speech emotionrecognition. Propose a new Fourier parameter model using theperceptual content of voice quality and the first- and second-order differences forspeaker-independent speech emotion recognition. Experimental results show thatthe proposed Fourier parameter (FP) features are effective in identifying variousemotional states in speech signals. They improve the recognition rates over themethods using Mel frequency cepstral coefficient (MFCC) features by 16.2, 6.8and 16.6 points on the German database (EMODB), Chinese language database(CASIA) and Chinese elderly emotion database (EESDB). In particular, whencombining FP with MFCC, the recognition rates can be further improved on theaforementioned databases by 17.5, 10 and 10.5 points respectively. Neural network classifier can be used to improve the classification of different emotions.

Keywords: MFCC, CASIA, EMODB, FP model, Speech emotion recognition.

I. INTRODUCTION

The Speech Signal is created at the vocal cords travels through the vocal tract and produced at speaker's mouth. It gets to the listeners ear as a pressure wave. Speech Signal is non-stationary, but can be divided to sound segments which have some common acoustic properties for a short time interval. The vocal tract is the cavity between the vocal cords and the lips, and acts as a resonator that spectrally shapes the periodic input, much like the cavity of a musical wind instrument.

Speech emotion recognition is defined as extracting the emotional states of a speaker from his or her speech. It is believed that speech emotion recognition can improve the performance of speech recognition systems and is thus very helpful for criminal investigation, intelligent assistance, surveillance and detection of potentially hazardous events, and health care systems. Speech emotion recognition is particularly useful in man-machine interaction. To effectively recognize emotions from speech signals, the intrinsic features must be extracted from raw speech data and transformed into appropriate formats that are suitable for further processing. It is a longstanding challenge in speech emotion recognition to extract efficient speech features. Researchers have performed many studies. First, it is found that continuous features including pitch related features, formants features, energy-related features, and timing features deliver important emotional cues.

In addition to time-dependent acoustic features, various spectral features such as linear predictor coefficients (LPC), linear predictor cepstral coefficients (LPCC) and mel-frequency cepstral coefficients (MFCC) play a significant role in speech emotion recognition. Bou-Ghazale and Hansen found that the features based on cepstral analysis, such as LPCC and MFCC, outperform the linear features of LPC in detecting speech emotions. Next, the Teager energy operator (TEO), introduced by Teager and Kaiser can be used to detect stress in speech. There are also other TEO based features proposed for detecting neutral versus stressed speech. Although the abovementioned features are useful for recognizing specific emotions, there is no sufficiently effective feature to describe complicated emotional states. Voice quality features are related to speech emotions.

According to an extensive study by Cowie, the acoustic correlations with voice quality can be grouped into voice level, pitch, phrase and feature boundaries and temporal structures. There are two popular approaches for determining voice quality terms. The first approach depends on the fact that speech signals can be modeled as the out-put of a vocal tract filter excited by a glottal source signal; hence, voice quality can be measured by removing the filtering effect of the vocal tract and by measuring the parameters of the glottal signal. However, the glottal signal must be estimated by exploiting the characteristics of the source signal and the vocal tract filter because neither of them is known. In the second approach, voice quality is represented by the parameters estimated from speech signals. Voice quality was represented by jitter and shimmer. The system for speaker-independent speech emotion recognition used the continuous hidden Markov model (HMM) as a classifier to detect some selected speaking styles: angry, fast, question, slow and soft. The baseline accuracy was 65.5percentage when using MFCC features only. The classification accuracy was improved to 68.1percentage when MFCC was combined with jitter, 68.5percentage when MFCC was combined with shimmer and 69.1percentage when MFCC was combined with both of them.

The voice quality parameters were estimated by spectral gradients of the vocal tract compensated speech signal for classifying utterances from the Berlin emotional data base to improve speaker-independent emotion classification. To

DOI 10.17148/IJARCCE.2016.51110

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

the best of our knowledge, Yang and Lugger first proposed a set of harmony features, which came from the well-known psychoacoustic harmony perception in music theory, for automatic emotion recognition. The following emotions were selected for classification: anger, happiness, sadness, boredom, anxiety, and neutral. The accuracy was 70.9percent when using voice quality features and standard features. Despite these contributions, further study regarding voice quality in delivering emotions is needed. Acoustic interpretation explains that the unique quality (tone) of each instrument is due to the unique content and structure of a harmonic sequence. According to music theory, the harmony structure of an interval or chord is mainly responsible for producing a positive or negative impression on listeners. In this paper, we propose a set of harmonic sequences, named Fourier parameter (FP) features, to detect the perceptual content of voice quality features rather than the conventional ones.
The new FP features will be evaluated on different speech databases. It is one of the first attempts to apply a new set of FP features, in particular, with the first- and second-order differences for speaker-independent speech emotion recognition. Both Bayesian classification and support vector machines (SVM) are evaluated.

A wavelet is a wave-like oscillation with an amplitude that begins at zero, increases, and then decreases back to zero. It can typically be visualized as a brief oscillation like one might see recorded by a seismograph or heart monitor.In mathematics, a waveletseries is a representation of a square-integrable (real- or complex-valued) function by a certain orthonormal series generated by a wavelet**.** Nowadays, wavelettransformation is one of the most popular of the time-frequency-transformations. A wavelet is a mathematical function useful in digital signal processingand image compression. The use of wavelets for these purposes is a recent development. The principles are similar to those of Fourier analysis, which was first developed in the early part of the 19th century. In signal processing, wavelets make it possible to recover weak signals from noise . This has proven useful especially in the processing of X-ray and magnetic-resonance images in medical applications. Images processed in this way can be cleaned up without blurring or muddling the details. Wavelet compression works by analyzing an image and converting it into a set of mathematical expressions that can then be decoded by the receiver.

II. EXISTING SYSTEM

A. FOURIER PARAMETER MODEL OF SPEECH

Fourier series is one of the most principal analytical methods for mathematical physics and engineering. Fourier analysis has been extensively applied for signal processing, including filtering, correlation, coding, synthesis and feature extraction for pattern identification. In Fourier analysis, a signal is decomposed into its constituent sinusoidal vibrations. A periodic signal can be described in terms of a series of harmonically related (i.e., integer multiples of a fundamental frequency) sine and cosine waves. In other words, a speech signal can be represented as the result of passing a glottal excitation waveform through a time-varying linear filter, which models the resonant characteristics of the vocal tract. A speech signal x(m) that is divided into l frames can be represented by a combination of an FP model as

$$x(m) = \sum_{k=1}^{M} Hk\left(\cos\left(2\Pi \frac{fk}{Fs} m\right) + \Phi k\right) \quad (1)$$

where F_s is the sampling frequency of x(m), H_k and Φk are the amplitude and phase of the k^{th} harmonic's sine component, l is the index of the frame, and M is the number of speech harmonic components.
The harmonic part of the model is a Fourier serial representation of a speech signal's periodic components. When a non-periodic component is sampled, its Fourier transform becomes a periodic and continuous function of frequency. The discrete Fourier transform (DFT) is derived from sampling the Fourier transform of a discrete-time signal at N discrete frequencies, which correspond to the integer multiples of the fundamental sampling interval $2\Pi/N$. For a finite duration discrete-time signal x(m) of length N samples, DFT is defined as

$$H(k) = \sum_{m=0}^{N-1} x(m) e^{-j(2\Pi/N)mk} \quad k=0,1,2\ldots.N-1 \quad (2)$$

B. FOURIER PARAMETER FEATURES FOR SPEECH EMOTION ANALYSIS

Here a new model is proposed, with special attention on three speech emotion databases in two different languages, to extract FP features.

1) Emotion Databases:
Three databases are considered: a German emotional corpus (EMODB), a Chinese emotional database (CASIA) and a Chinese elderly emotional speech database (EESDB), which is summarized as follows. EMODB was collected by the Institute of Communication Science at the Technical University of Berlin. It has been used by many researchers as a standard data set for studying speech emotion recognition. EMODB comprises 10 sentences that cover seven classes of emotion from everyday communication, namely, anger, fear, happiness, sadness, disgust, boredom and neutral. They could be interpreted in all emotional contexts without semantic inconsistency. EMODB is well annotated and publicly available.

CASIA was released by the Institute of Automation, Chinese Academy of Sciences. It is composed of 9,600 wave files that represent different emotional states: happiness, sadness, anger, surprise, fear, and neutrality. Four actors (two females and two males) simulated this set of emotions and produced 400 utterances in six classes of different emotions.

The EESDB database includes seven classes of emotions (angry, disgust, fear, happy, neutral, sadness and surprise). The sources of this database came from a part of Chinese TV statements presented by 11 elderly people over 60 years old (five females and six males). In the first step, two speech emotion databases, EMODB and CASIA, are employed to validate the method for extracting FP features.

DOI 10.17148/IJARCCE.2016.51110

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

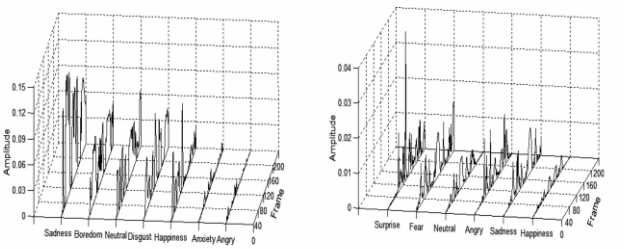
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

2) Emotion recognition on EMODB:
There are a number of recent contributions that implement emotion recognition on EMODB. In Yang and Lugger, the authors propose a set of harmony features. They are based on the psychoacoustic perception of pitch intervals and apply the theory of chords from music. Harmony features are derived from the pitch contour to characterize the relationship between different pitches, such as two-pitch intervals and chords involving more than two pitches. Harmony features are used in conjunction with energy, pitch, duration, formants, ZCR, and voice quality features. 306 statistical values of the fore mentioned features are computed. Sequential floating forward selection identifies the 50 most informative features, which are fed as input to a Bayesian classifier that exploits GMMs. Speaker-independent experiments are carried out. 6 emotional classes are considered, namely: happiness, boredom, neutral, sadness, anger, and anxiety.

In this approach a more exhaustive feature computation is available although in both works statistical values of features are computed and feature selection is applied in order to retain a small number of features. Nonetheless, in Yang and Lugger disgust is dismissed. The approach proposed in Ruvolo et al. combines selection and hierarchical aggregation of features aiming to combine short, medium, and long time scale features. Considering short time scale features, MFCCs, sones, and linear predictive cepstral coefficients are used. Medium time scale features are computed by spectro-temporal box-filters, while long time scale features include phase, sampling interval, moment, energy, and summary statistics like mean value and quantiles. Next, Gentle Boost is used to simultaneously select the best performing features and build the classifier. Speaker-independent experiments are performed. In specific, 63 binary classifiers are applied, each of which consists of 15 spectro-temporal box-filters selected by the Gentle Boost. Finally, multinomial ridge logistic regression is applied to the continuous outputs of the 63 binary classifiers.

The idea of calculating various features along with their corresponding statistics is also applied by the authors, although the categorization of features is not the same one. Moreover, the authors of this paper exploit feature selection and classification separately. Hierarchy is also applied in both approaches. However, in Ruvolo et al. hierarchical aggregation of features is tested whereas in this approach the psychologically-inspired binary cascade classification schema employs a hierarchy on emotional descriptors. Class-level spectral features for emotion recognition are proposed in Bitouk. The authors define 3 phoneme type classes: stressed vowels, unstressed vowels, and consonants in the utterance. MFCC class-conditional means and standard deviations for each class are aggregated into one feature vector, by using the phoneme-level segmentation of the utterance. The average duration of the phoneme classes is appended to the feature vector. Moreover, 24 utterance level prosodic features are computed. The aforementioned features are related to statistics of fundamental frequency, first formant, voice intensity, jitter, shimmer, and relative duration of voiced segments. This results to a total of 261 features which are fed as input to a linear SVM classifier. A speaker-independent scenario is applied, whereas 6 emotional classes are taken into account, namely: anger, anxiety, disgust, happiness, sadness, and neutral. At a second set of experiments, feature selection is applied. Inspired by this approach, we performed sets of experiments with and without feature selection, as well. In the approach proposed by the authors of this paper, statistics of fundamental frequency, first formant, MFCCs, jitter, shimmer are among the computed features. However, in our case feature computation is exhaustive and it results a total number of 2327 extracted features.
An emotion recognizer that may operate jointly with an automatic speech recognizer is examined by Pittermann. The feature vector comprises of MFCCs (along with their first- and second-order differences), intensity, and three formants, along with pitch and pitch statistics, namely minimum, mean, maximum, deviation and range. No feature selection technique is applied, while the HMMs are employed as classifiers to a speaker-dependent protocol. In our approach that applies feature selection and a speaker-independent protocol.. However, in both cases compute the first- and second- differences of the features in order to capture their temporal evolution.

3) Fourier Parameter Features:
Harmonics include frequency, amplitude and phase. Harmonic frequency features are effective for speech emotion recognition. Here make use of harmonic amplitude and phase features. For every frame, FP is estimated by Fourier analysis. $(H_k)^l$ is the l^{th} frame's FP. The i^{th} FP amplitude is H_i. It then leads to the average values of H_i. A new speech feature vector H_k may be evaluated for all frames in the speech signal from 1 to l (number of frames).



(a) Speech signals uttered in German with seven emotion types, sadness, boredom, neutral, disgust, happiness, anxiety/fear and angry

(b) Speech signals uttered in Chinese with six emotion types, surprise, fear, neutral, angry, sadness and happiness

Fig. 1The mean of H_3 for speech signals with different emotions from the German and Chinese databases.

Fig.1 shows the averaged H_3 among various emotions for one person. It is observed that amplitudes vary with different classes of emotions. We also discern the mean of each phase of speech with different emotions, but the difference is trivial.

4) Global Fourier Parameter Features:
It has been reported that global features are superior in terms of classification accuracy and computational efficiency. Therefore, the mean, maximum, minimum,

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

medianand standard deviations of the amplitudes of the first 20 Fourier parameters are calculated.

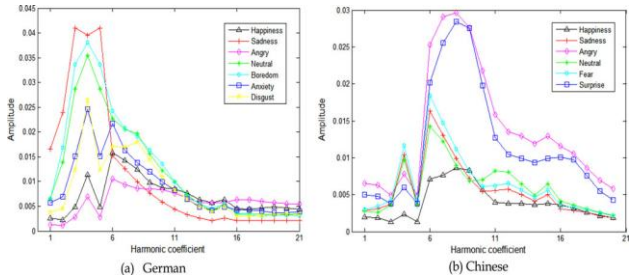


Fig. 2 The means of H_1 to H_{20} with different emotions

Fig. 2a shows that the means of H_1 to H_{20} are different with regard to seven emotions. The average values of H_3 for sadness, boredom and neutral are higher than disgust, happy and anxiety. The peak of every emotion is at the lower harmonics. For examplethe peaks of happy and angry emotions are obtained at the sixth harmonic; the peaks of neutral, boring, anxious and disgusted emotions are obtained at the fourth harmonic. It is also observed that the variation of the low-order FPs for every emotion is large, while the high-order FPs is relatively smooth. The amplitudes of happy and angry emotions are below those of neutrality before the former 10 harmonics. Angry and happy emotions have higher values of energy compared to those with neutral emotion.

Fig.2b shows the means of six emotions from CASIA. The amplitudes of anger and surprise are higher than those of other emotions, while happy is lower. The variation of happiness is relatively smooth, while anger and surprise are obvious. It also shows that the peaks of happiness, surprise and anger lie at the eighth harmonic, and the peaks of neutrality, sadness and fear lie at the sixth harmonic. It is noteworthy that among these speech databases, the same emotions between the German speech database and the Chinese speech database may have different FP features. The angry emotion from the Chinese speech database is higher than the other emotions, while the German database is lower. Moreover, the happy emotion in both databases is low. The reason is that different countries have different cultures so that the ways in which they convey and perceive emotions are different.

C. SPEAKER-INDEPENDENT RECOGNITION

Speaker-independent emotion recognition is one of the latest challenges in the field of speech emotion recognition. It is able to cope with unknown speakers and thus has better generalization than those speaker-dependent approaches. Until now, there have been quite a few studies reported on speaker-independent emotion recognition. Yang and Lugger proposed a set of harmony features derived from the pitch contour and employed them for the speaker-independent recognition of six classes of emotions: happiness, boredom, neutrality, sadness, anger, and anxiety. Ruvolo et al made use of the hierarchical aggregation of featuresto combine short-, medium- and long scale features. They employed MFCC and LPCC for speaker-independent experiments. Bitouk et al. defined three classes of phonemes in the utterance, namely, stressed vowels, unstressed vowels and consonants, and further calculated the statistics of fundamental frequency, first formant, voice intensity, jitter, shimmer and the relative duration of voiced segments for speaker-independent experiments. Kotti and Patern-o extracted 2,327 features in total for speaker-independent recognition that were related to the statistics of pitch, formants, and energy contours as well as spectrum, cepstrum, autocorrelation, voice quality, jitter, shimmer and others.

1) FEATURE EXTRACTION:

Both MFCC and FP features are extracted for speaker-independent emotion recognition. Continuous features such as fundamental frequency (F0), energy and zero-crossing rate (ZCR) are also extracted.

MFCC Features

MFCC was first introduced and applied to speech recognition. It has been popularly used for speech emotion recognition. By considering the reaction of human ears to different frequencies, the Mel frequency is determined according to the characteristics of human audition. In this study, MFCC features were extracted for comparison with the proposed FP features. For emotion recognition, MFCC features usually include mean, maximum, minimum, median, and standard deviation. All speech signals were first filtered by a high pass filter with a pre-emphasis coefficient of 0.97. The first 13 MFCCs and the associated delta- and double-delta MFCCs were extracted to form a 39-dimensional feature vector. Its mean, maximum, minimum, median and standard deviation were further derived out, which led to a 195-dimensional MFCC feature vector in total.

Fourier Parameter Features

Here, the first 120 harmonic coefficients were extracted. The dynamic features were extracted so that temporal derivative features may improve the performance of emotion recognition. In other words, the FP feature vector is comprised of amplitude (H), first-order difference (DH) and second-order difference (DDH). Their minimum, maximum, mean, median and standard deviation was also computed. There were a total of 1,800 features for speaker-independent speech emotion recognition.

Continuous Features

Continuous features are important in delivering the emotional cues of speakers and thus have been widely used in speech emotion recognition. F0 or pitch is a prosodic feature, which provides the tonal and rhythmic properties of the speech. Energy refers to the intensity of the speech signal and reflects the pause and where the accent of the voice signal is. ZCR reflects the time when adjacent samples of a voice signal are going to change the symbol. In our earlier studies, the feature set with F0, energy, and ZCR has been better than other common feature sets including formant and LPCC. In this study, the minimum, maximum, mean, median, and standard deviation of F0, energy and ZCR were also calculated for comparison with the proposed FP features.

DOI 10.17148/IJARCCE.2016.51110

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

2) FEATURE NORMALIZATION:
Normalization is an important aspect for a robust emotion recognition system. The goal is to eliminate speaker and recording variability while keeping the effectiveness of emotional discrimination. In particular, it could compensate for speaker variability. Here, z-score normalization was adopted for feature normalization.

For a given FP feature H from a speech signal of a speaker s, its mean value, $E(H^s)$, and its standard deviation, $std(H^s)$, were first derived out. Then, the normalized feature was estimated by

$$\bar{(H^s)}=[H^s-E(H^s)]/std(H^s) \quad (3)$$

3) SUPPORT VECTOR MACHINE CLASSIFICATION:
With respect to emotional speech recognition, many classifiers including the Gaussian mixture model (GMM), artificial neutral networks (ANN), hidden Markov model and support vector machine have been studied more than once. SVM makes use of convex quadratic optimization that is advantageous in making a globally optimal solution. SVM has demonstrated good performance on several classical problems of pattern recognition, including bioinformatics, text recognition and facial expression recognition. It was also used for speech emotion recognition and outperformed other well-known classifiers. There are two different families of solutions aiming to extend SVM for multiclass problems. The first solution follows the strategy of "one-versus-all", while the second solution follows the strategy of "one-versus-one". We selected the second method by using LIBSVM because it is more convenient in practice. FP features were fed as inputs to the SVM classifier with the Gaussian radial basis function kernel, where the controlling parameters have been evaluated for c ∈(0, 10) and g ∈ (0, 1).

4) EXPERIMENT RESULT

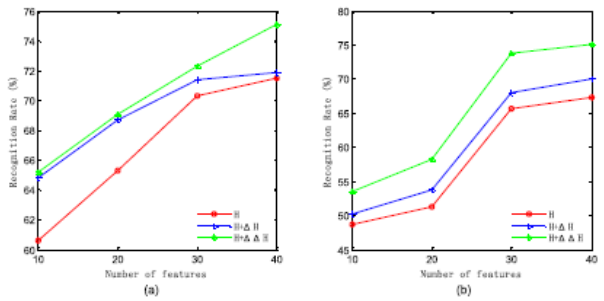


Fig. 4Result of six- class emotion recognition using H,H+ΔH and H+ΔΔH

First used 40 FP features for speech emotion recognition. As shown in Fig. 4, the recognition rate increased with increments of 10 FP features. Moreover, the recognition rates increased when the first- and second-order differences were incorporated. The third- and fourth-order differences were also evaluated, but their contributions were not as effective. The same protocol was used with the phase features and their differences. The method of sequential floating forward search (SFFS) was then used to reduce the inputting features and to improve the recognition rate.

TABLE I Confusion matrix of emotion recognition using 120 FP features on the German database

	Happ.	Bored	Neutr.	Sadn.	Anger	Anxi.
Happ.	92.92				1.25	5.83
Bored.	1.11	71.48	10.22	1.25	12.44	3.5
Neutr.		2.54	87.46	4.44		5.56
Sadn.				91.21		8.79
Anger		1.71			98.29	
Anxi.		3.08	1.25	2.08	1.67	91.92

Table II shows the confusion matrix of SVM classification with 120 FP features on EMODB. A total of 2,327 features were extracted for speech emotion recognition. The average accuracy was 83.3, 89.7 percent for happiness, 90.5 percent for neutrality, 87.7 percent for anxiety, 90.1 percent for anger, 88.6 percent for sadness and 89.3 percent for boredom. In contrast, the approach presented here achieved recognition rates for happiness (92.92 percent), anger (98.29 percent), sadness (91.21 percent) and anxiety (91.92 percent). In other words, FP and FP + MFCC features improve the recognition rate at approximately 5.6 and 6.8 points. Harmony features were proposed for speaker-independent emotion recognition by using a Bayesian classifier on EMODB. The rates of emotion recognition were 52.7 percent for happiness, 84.8 percent for boredom, 52.9 percent for neutrality,87.6percent for sadness, 86.1 percent for anger and 76.9 percent for anxiety. We also developed a Bayesian classifier with Gaussian class-conditional likelihood on EMODB. By using the same 120 FP features, the average accuracy was 79.51 percent, with 87.59 percent for happiness, 54.36 percent for boredom, 84.31 percent for neutrality, 89.60 percent for sadness, 93.62 percent for anger and 67.60 percent for anxiety. The FP features are able to improve the recognition rate at approximately 6.01 points.

Tables II and III report the confusion matrices of 120 FP features on CASIA and EESDB, respectively.

TABLE II Confusion matrix of emotion recognition using 120 FP features on the CASIA database

	EMODB	CASIA	EESDB
Happiness	FP	FP+MFCC	FP
Sadness	FP+MFCC	FP	FP
Anger	FEZ	FP+MFCC	MFCC
Neutrality	FP+MFCC	FP	FP
Bored	FP+MFCC		
Surprise		FP+MFCC	
Fear	FP+MFCC	FP+MFCC	

The recognition rate on EESDB is below that of the other two databases. The main reason might be because the emotions expressed by the elderly are usually more difficult to identify.

DOI 10.17148/IJARCCE.2016.51110

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

TABLE III Confusion matrix of emotion recognition using 120 FP features on the EESDB database

	Happ.	Surpri	Neutra.	Sadn.	Anger	Fear
Happ.	81		2	4	11	2
Surpri		84		3	10	3
Neutr.			75	16		5.56
Sadn	11	8	2	67	6	6
Anger				6	86	8
Fear		3	4	5	7	81

Happ.=Happiness, Surpri.= Surprise, Neutr.= Neutrality, Sadn.= Sadness, Anxi.= Anxiety, Bored.= Boredom

According to Tables I to III, it seems that the rates of emotion recognition vary between German and Chinese. It is reasonable that different countries have different cultures, and the way in which they express their emotion is also different. In general, the FP features themselves achieved higher average recognition rates than MFCC and FEZ, particularly on the EMODB database. When combining the FP and MFCC features (MFCC+FP), it was able to further improve the performance of speech emotion recognition. On the contrary, the FEZ features usually led to worse performance. In other words, although continuous features deliver the important emotional cues of speakers, FEZ features did not demonstrate better performance in speaker-independent emotion recognition.

Table IVshows the optimal combination of features by using FP, MFCC, FP+MFCC and FEZ for different classes of emotions on the three databases. With an average rate of recognition at 87.5 percent, the proposed FP features out-performed the others in all cases. In summary, compared with MFCC features, the proposed FP features improved speaker-independent emotion recognition by 16.2 points on the German database, 6.8 points on the CASIA database and 16.6 points on the EESDB database. The performance could be further enhanced by approximately 17.5 points, 10 and 10.5 points by combining the FP and MFCC features on the afore-mentioned databases.

TABLE IV Best features among MFCC, FP, FP+MFCC and FEZ on three database

	Happ.	Sadn.	Anger	Neutr.
Happ.	41.5	28.9	14.7	14.9
Sadn.	2	83.6	9.4	5
Anger	2.6	8.6	88.8	
Neutr.	4.2	2.9	2.8	90.1

III.PROPOSED SYSTEM

A. NEURAL NETWORKS

Here used a two-layer back propagation neural network architecture with a 8, 10 or 14 element input vector, 10 or 20 nodes in the hidden sigmoid layer and five nodes in the output linear layer. The number of inputs corresponds to the number of features and the number of outputs corresponds to the number of emotional categories. To train and test our algorithms we used the data sets s70, s80 and s90. These sets were randomly split into training (70% of utterances) and test (30%) subsets. We created several neural network classifiers trained with different initial weight matrices. This approach applied to the s70 data set and the 8-feature set gave the average accuracy of about 65% with the following distribution for emotional categories: normal state is 55-65%, happiness is 60-70%, anger is 60-80%, sadness is 60-70%, and fear is 25-50%.

B. ENSEMBLES OF NEURAL NETWORK CLASSIFIERS

An ensemble consists of an odd number of neural network classifiers, which have been trained on different subsets of the training set using the bootstrap aggregation and the cross-validated committee's techniques. The ensemble makes decision based on the majority voting principle. We used ensemble sizes from 7 to 15. Figure 2 shows the average accuracy of recognition for ensembles of 15 neural networks, the s70 data set, all three sets of features, and both neural network architectures (10 and 20 neurons in the hidden layer). We can see that the accuracy for happiness stays the same (~65%) for thedifferent sets of features and architectures. The accuracy for fear is relatively low (35-53%). The accuracy for anger starts at 73% for the 8-feature set and increases to 81% the 14-feature set. The accuracy for sadness varies from 73% to 83% and achieves its maximum for the 10- feature set. The average total accuracy is about 70%.

A wavelet is a wave-like oscillation with amplitude that begins at zero, increases, and then decreases back to zero. It can typically be visualized as a "brief oscillation" like one might see recorded by a seismograph or heart monitor. Generally, wavelets are purposefully crafted to have specific properties that make them useful for signal processing. Wavelets can be combined, using a "reverse, shift, multiply and integrate" technique called convolution, with portions of a known signal to extract information from the unknown signal. For example, a wavelet could be created to have a frequency of Middle C and a short duration of roughly a 32nd note. If this wavelet was to be convolved with a signal created from the recording of a song, then the resulting signal would be useful for determining when the Middle C note was being played in the song. Mathematically, the wavelet will correlate with the signal if the unknown signal contains information of similar frequency. This concept of correlation is at the core of many practical applications of wavelet theory.

The main difference between wavelet and Fourier transform is that wavelets are localized in both time and frequency whereas the standard Fourier transform is only localized in frequency. The classical Fourier transform of a function allows you to make a measurement with 0 bandwidth: the evaluation f ^ (k) tells us precisely the size of the component of frequency k . But by doing so you lose all control on spatial duration: you do not know when in time the signal is sounded. This is the limiting case of the Uncertainty Principle: absolute precision on frequency and zero control on temporal spread.

DOI 10.17148/IJARCCE.2016.51110

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

IV. CONCLUSION

In previous studies, different features were employed for speech emotion recognition. In this paper, we proposed a new FP model to extract salient features from emotional speech signals and validated it on three well-known databases including EMODB, CASIA andEESDB. It is observed that different emotions did lead to different FPs. Furthermore, FP features were evaluated for speaker-independent emotion recognition by using SVM and a Bayesian classifier. The study showed that FP features are effective in characterizing and recognizing emotions in speech signals. Moreover, it is possible to improve the performance of emotion recognition by combining FP and MFCC features. These results establish that the proposed FP model is helpful for speaker-independent speech emotion recognition. Neural network classifier can be used to improve the classification of different emotions

ACKNOWLEDGMENT

I sincerely thank to all those who helped me in completing this task.

REFERENCES

[1]. Busso, S. Marioor-yad, S. Narayanan and A. Metallinou, “Iterative feature normalization scheme for automatic emotion detection from speech,” IEEE Trans. Affective Comput., vol. 4, no. 4, pp. 386–397, Oct.-Dec. 2013

[2] P. H. David, V. Bogdan, B. Ronald, and W. Andreas, “The performance of the speaking rate parameter in emotion recognition from speech,” in Proc. IEEE Int. Conf. Multimedia Expo Workshops, 2012, pp. 296–301

[3] M. Kotti and F. Patern_o, “Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema,” Int. J. Speech Technol., vol. 15, pp. 131–150, 2012.

[4] B. Yang and M. Lugger, “Emotion recognition from speech signals using New Harmony features,” Signal Process, vol. 90, pp. 1415–1423, and 2010

DOI 10.17148/IJARCCE.2016.51110