Speech Emotion Recognition Using Fourier Parameters

Kunxia Wang, Ning An, *Senior Member, IEEE*, Bing Nan Li, *Senior Member, IEEE*, Yanyong Zhang, *Member, IEEE*, Lian Li, *Member, IEEE*

Abstract—Recently attention has been paid on harmony features for speech emotion recognition. It is found in our study that the first- and second-order differences of harmony features also play an important role in speech emotion recognition. Therefore, we propose a new Fourier parameter model by using the perceptual content of voice quality, the first- and second-order differences for speaker-independent speech emotion recognition. Experiment results show that the proposed Fourier parameter (FP) features are effective in identifying various emotion states in speech signals. They improve the recognition rates over the methods using Mel Frequency Cepstral Coefficient (MFCC) features by 16.2 points, 6.8 points and 16.6 points on the German database (EMODB), the Chinese language database (CASIA) and the Chinese elderly emotion database (EESDB). In particular, if combining FP with MFCC, the recognition rates can be further improved by 17.5 points, 10 points and 10.5 points on the aforementioned databases, respectively.

Index Terms—Fourier parameter model, speaker-independent, speech emotion recognition, affective computing

1 Introduction

Speech emotion recognition, defined as extracting the emotional states of a speaker from his or her speech, is attracting more and more attention. It is believed that speech emotion recognition can improve the performance of speech recognition systems [1], and thus is very helpful for criminal investigation, intelligent assistance [2], surveillance and detection of potentially hazardous events [3], and health care systems as well [4]. Speech emotion recognition is particularly useful in man-machine interaction [1],[6].

In order 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 made a lot of studies [6]-[12]. First, it is found that continuous features including pitch-related features, formants features, energy-related features, timing features deliver important emotional cues [7][11][31]. In addition to time-dependent acoustic features, various spectral features such as linear predictor coefficients (LPC) [32], linear predictor cepstral coefficients (LPCC) [33] and mel-frequency cepstral coefficients (MFCC) [45] play a significant role in speech emotion recognition. Bou-Ghazale et al. [33] explored that the features based on cepstral analysis, such as LPCC and MFCC, outperform the linear features LPC in detecting speech emotions. Third, the Teager-energy-operator (TEO), introduced by Teager [35] and Kaiser [36], can be used to detect stress in speech [37]. There are also other TEO-based features proposed for detecting neutral versus stressed speech [38]. Although the abovementioned features turn out to be useful for recognizing specific emotions, there is yet no sufficiently effective feature to describe complicated emotional states [13].

It has been demonstrated that voice quality features are related to speech emotions [14],[15],[39],[40],[42],[54]. According to an extensive study by Cowie [11], 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 deciding voice quality terms. The first one depends on the fact that speech signals can be modelled as the output of vocal tract filter excited by a glottal source signal [32]; hence voice quality can be measured by removing the filtering effect of the vocal tract and by measuring the parameters of the glottal signal [41]. However, the glottal signal has to be estimated by exploiting the characteristics of the source signal and the vocal tract filter because neither of them is known [1]. In the second approach, voice quality is represented by the parameters estimated from speech signals. In [39], 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 65.5 points when using MFCC features only. The classification accuracy was improved to 68.1 points when MFCC was combined with jitter, 68.5 points when MFCC was combined with shimmer and 69.1 points when MFCC was combined with both of them. In [54], the voice quality parameters were estimated by

- *K.X. Wang is with the School of Computer and Information, Hefei Univercity of Technology, and works in the Department of Electronic Engineering, Anhui Univercity of Architechture, Hefei, China. E-mail: kxwang@ ahjzu.edu.cn.*
- *N. An and L. Li are with the School of Computer and Information, Hefei Univercity of Technology, Hefei, China. E-mail: ning.g.an@acm.org, llian@hfut.edu.cn*
- *B.N. Li is with the Department of Biomedical Engineering, Hefei Univercity of Technology, Hefei, China. E-mail: bingoon@ieee.org*
- *Y.Y.Zhang is with WINLAB of Rutgers University, North Brunswick, NJ, USA,E-mail: yyzhang@winlab.rutgers.edu*

spectral gradients of the vocal tract compensated speech signal, and were applied in [40], [42] for classifying utterances from the Berlin emotional database [18] to improve speaker-independent emotion classification. To the best of our knowledge, Yang and Lugger [15] 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 70.9 points when using voice quality features and the standard features.

In spite of these contributions, further study about how voice quality in delivering emotions is need. Acoustic interpretation explains that 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 fisrt- and second-order differences, for speaker-independent speech emotion recognition. Both Bayesian classification and SVM (Support Vector Machine) are evaluated.

The main contributions of this paper for speaker-independent emotion recognition are summarized as follows: 1) proposing a new FP model using FP features and their one- and second-order differences for speech emotion recognition; 2) proposing to further improve speaker-independent speech emotion recognition by combining FP and MFCC features; 3) carrying out extensive validations on three speech databases in two languages. This paper is organized as follows. Section 2 presents the Fourier parameter model based on Fourier series. Section 3 details the FP features for speech emotion analysis and the evaluations on a German database and a Chinese database. Section 4 discusses the experiment results of speaker-independent speech emotion recognition. Concluding remarks are drawn in section 5.

2 FOURIER PARAMETER MODEL OF SPEECH

Fourier series [44] 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 [17]. A speech signal $x(m)$ that is divided into l frames can be represented by a combination of a FP model as in (1):

$$x(m)=\sum_{k=1}^{M} H_k^{\,l}(m)(\cos(2\pi\frac{f_k^l}{F_s}m) \ +\varphi_k^l)\ , \qquad (1)$$

where F_s is the sampling frequency of $x(m)$, $H_k^{\,l}$ and φ_k^l are the amplitude and phrase of the k^{th} harmonic's sine component, l is the index of frame, 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 that sampling the Fourier transform of a discrete-time signal at N discrete-frequencies, which correspond to the integer multiples of the foundamental sampling interval 2π/N. For a finite duration discrete-time signal $x(m)$ of length N samples, DFT is defined as (2), where $H(k)$ is FPs from k=0 ... N-1.

$$H(k)\ =\sum_{m=0}^{N-1} x(m)e^{-j\frac{2\pi}{N}mk} \ \text{k=0,1,2,...,N-1}\ . \qquad (2)$$

3 FOURIER PARAMETER FEATURES FOR SPEECH EMOTION ANALYSIS

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

3.1 Emotion Databases

Three databases are considered: a German emotional corpus (EMODB) [18], a Chinese emotional database (CASIA) [46] and a Chinese elderly emotional speech database (EESDB) [55], which are summarized as follows.

EMODB was collected by the Institute of Communication Science of the Technical University of Berlin. It has been exploited by many researchers as a standard dataset for studying speech emotion recognition. EMODB comprises 10 sentences which cover 7 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 9600 wave files that represent different emotional states: happiness, sadness, anger, surprise, fear, and neutrality. Four actors (2 females and 2 males) simulated this set of emotions, and produced 400 utterances in 6 classes of different emotions.

EESDB database includes 7 classes of emotions (angry, disgust, fear, happy, neutral, sadness and surprise). The sources of this database came from a part of Chinese's TV statements presented by 11 elderly people over 60 years old (5 females and 6 males).

In the first step, two speech emotion databases, EMODB and CASIA, are employed to validate the method for extracting FP features.

3.2. Fourier Parameter Features

Harmonics include frequency, amplitude and phase. It has been reported that harmonic frequency features are effective for speech emotion recognition [15]. In this study, we also make use of harmonic amplitude and phase features. For every frame, FP is estimated by Fourier analysis. As shown in (2), 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. In other words, a new speech feature vector H_k may be evaluated for all frames in the speech signal from 1 to l (number of frames). Fig.1 shows the averaged H_3 among various emotions of one person. It is observed that amplitudes vary with different classes of emotions. We further figure out the mean of each phase of speech with different emotions, but the difference is trivial.

3.3 Global Fourier Parameter Features

It has been reported [1] that global features are superior in terms of classification accuracy and computational efficiency. Therefore, the mean, maximum, minimum, median and standard deviation of the amplitudes of the first 20 Fourier parameters are calculated as in [1],[7],[15],[47],[48].

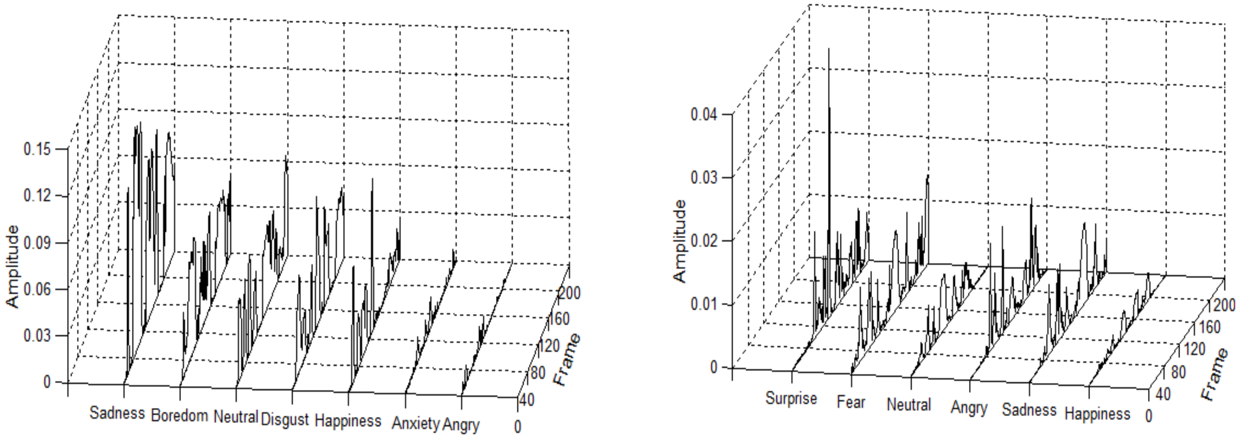
Fig. 2(a) shows that the means of H_1 to H_{20} are different with regard to 7 emotions. The average values of H_3 for sadness, boredom and neutral are higher than disgust, happy and anxiety. The peak of every emotion is at the lower harmonics. For example, the peaks of happy and angry emotions are obtained at the 6th harmonic; the peaks of neutral, boring, anxious and disgusted emotions are obtained at the 4th harmonic. It is also observed that the variation of the low-order FPs for every emotion is large, while that of the high-order FPs is relatively smooth. The amplitudes of happy and angry emotions are below those of neutrality before the former 10 harmonics. Similar results have been observed by Ramamohan and Dandapat [19] that angy and happy emotions have higher values of energy compared to those with neutral emotion.

Fig. 2(b) shows the means of 6 emotions from CASIA. The amplitudes of anger and surprise are higher than those of other emotions, while that of happy is lower. The variation of happiness is relatively smooth, while those of anger and surprise are obvious. It also shows that the peaks of happiness, surprise and anger lie at the 8th harmonic, and the peaks of neutrality, sadness and fear at the 6th 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 that from 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 [25].

4 Speaker-Independent Recognition

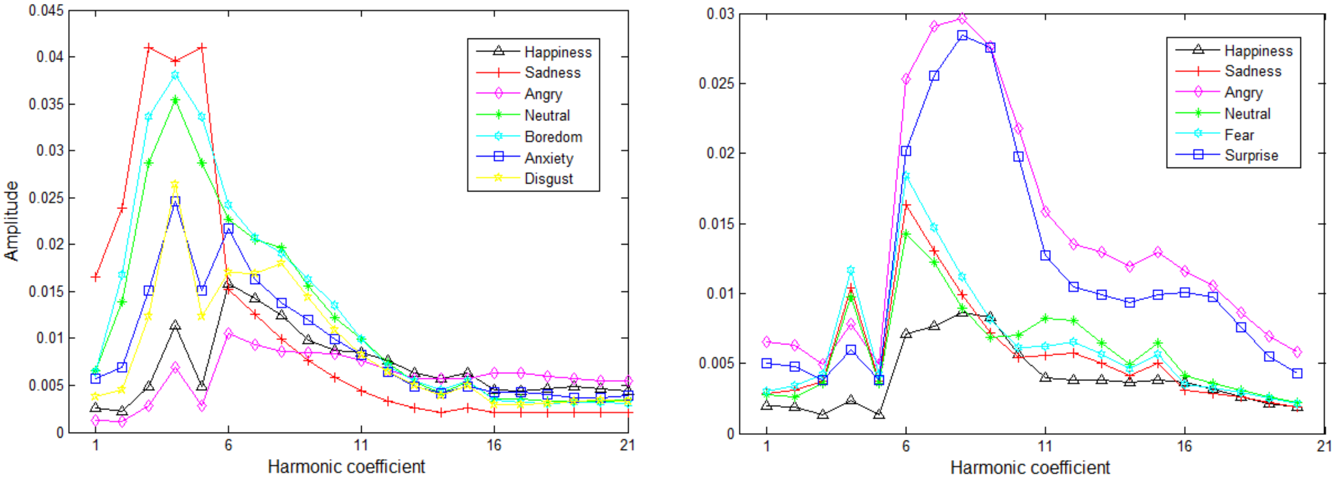
Speaker-independent emotion recognition is one of the latest challenges in the field of speech emotion recognition. It is able to cope with unknown speakers, and thus has better generalization than those speaker-dependent approaches [4]. Till now there have been quite a few studies reported on speaker-independent emotion recognition [4],[15],[16],[49]. Yang and Lugger [15] proposed a set of harmony features derived from the pitch contour, and employed them for speaker-independent recognition of 6 classes of emotions: happiness, boredom, neutrality, sadness, anger, and anxiety. Ruvolo et al. [16] made use of the hierarchical aggregation of features in order to com-



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



(a) German

(b) Chinese

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

bine short-, medium- and long-scale features. They employed MFCC and LPCC for speaker-independent experiments. Bitouk et al. [49] 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 relative duration of voiced segments for speaker-independent experiments. Kotti and Savvas Paternò [4] extracted 2327 features in total for speaker-independent recognition that were related to the statistics of pitch, formants, energy contours, as well as spectrum, cepstrum, autocorrelation, voice quality, jitter, shimmer and others.

4.1. Feature Extraction

Both MFCC and FP features are extracted for speaker-independent emotion recognition. Continuous features [1] such as Fundamental Frequency (F0) [7], Energy and Zero-Crossing Rate (ZCR) are also extracted.

4.1.1. MFCC features

MFCC was first introduced and applied to speech recognition in [45]. It has been popularly used for speech emotion recognition [34],[40],[25]. By considering the feelings 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-dimentional MFCC feature vector in total.

4.1.2. Fourier parameter features

We extracted a set of FP features from speech signals as described in Sections 3.2-3.3. Here the first 120 harmonic coefficients were extracted. The dynamic features were extracted in that temporal derivative features may improve the performance of emotion recognition [20]. In other words, the FP feature vector comprised amplitude (H), 1-order difference (ΔH) and 2-order difference ($\Delta\Delta H$). Their minimum, maximum, mean, median and standard deviation were also computed. There were a total of 1800 features for speaker-independent speech emotion recognition.

4.1.3 Continuous features

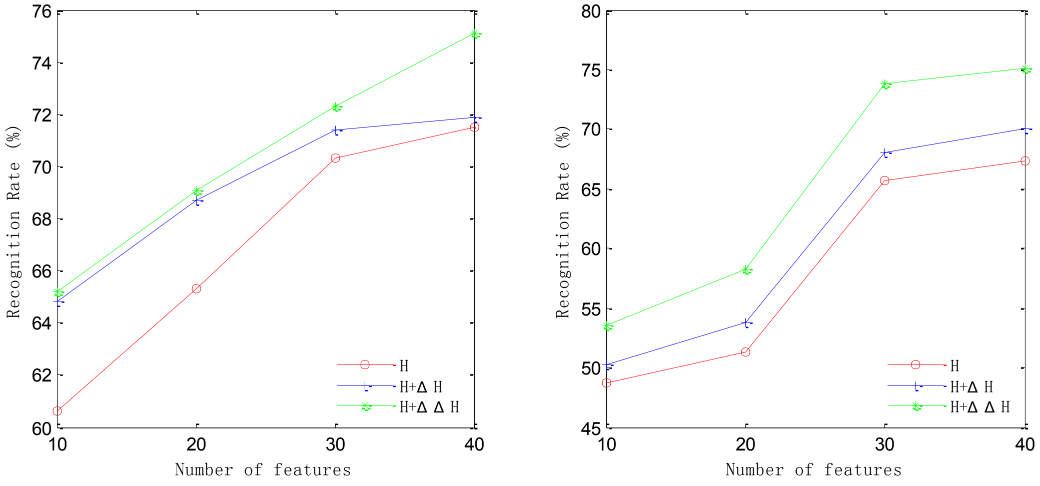
Continuous features are important in delivering emotional cues of speakers [7],[9],[11], and thus have been widely used in speech emotion recognition [1],[7],[22]. F0 or pitch is a prosodic feature, which provides the tonal and rhythmic properties of the speech. Energy refers to the intensity of speech 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, ZCR has been found better than other common feature sets including Formant and LPCC [57]. 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.

4.2. Feature Normalization

Normalization is an important aspect for a robust emotion recognition system [30]. The goal is to eliminate speaker and recording variability while keeping the effectiveness of emotional discrimination [52]. In particular, it could compensate for speaker variability. Here z-score normalization [43] 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 devia

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2392101, IEEE Transactions on Affective Computing



(a) (b)

Fig.3. Result of six class emotion recognition using H , $H+\Delta H$ and $H+\Delta\Delta H$ (a) EMODB database (b) CASIA database The x-axis represents 1 to n number Fourier parameter

tion, $std(H^s)$, were first derived out. Then, the normalized feature was estimated by following (3).

$$\widehat{H^s} = \frac{H^s - E(H^s)}{std(H^s)} \qquad (3)$$

4.3. Support Vector Machine Classification

With respect to emotional speech recognition, many classifiers including Gaussian Mixture Model (GMM), Artificial Neutral Networks (ANN) [22], Hidden Markov Model (HMM) [23] and Support Vector Machine (SVM) [21],[24],[25] 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 [26], including bioinformatics, text and facial expression recognition [27]. It was also used for speech emotion recognition [4],[28-29],[51] and outperformed other well-known classifiers [1].

There are two different families of solutions aiming to extend SVM for multiclass problems [53]. The first one follows the strategy of "one-versus-all", while the other one takes the strategy of "one-versus-one". We selected the second method by using LIBSVM [59] in that it is more convenient in practice [53]. FP features were fed as inputs to the SVM classifier with Gaussian radial basis function kernel, where the controlling parameters have been evaluated for c ∈ (0,10) and γ ∈ (0,1) .

4.4. Experiment Results

We firstly used first 40 FP features for speech emotion recognition. As shown in Fig. 3, the recognition rate increased with increment of 10 FP features. Moreover, the recognition rates futher increased when the first- and second-order differences were incorporated.

The third- and forth-order differences were also evaluated, but their contributions were not so effective. The same protocol was conducted by using the phase features and their differences. The recognition rate was as low as roughly 55 points. It suggests that the phase feature is not efficient for speech emotion recognition, which has also been demonstrated in [19].

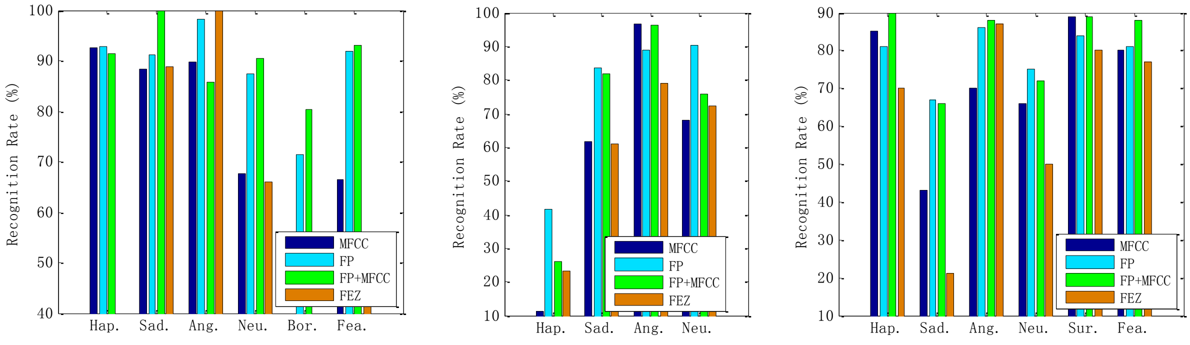
The method of Sequential Floating Forward Search (SFFS) [58] was then used to reduce the inputting features and improve the recognition rate. Table 1 shows the confusion matrix of SVM classification with 120 FP features on EMODB.

In [4], a total number of 2327 features were extracted for speech emotion recognition. The average accuracy in [4] was 83.3 points, 89.7 points for happiness, 90.5 points for neutrality, 87.7 points for anxiety, 90.1 points for anger, 88.6 points for sadness and 89.3 points for boredom. In contrast, the approach presented in this paper achieved the recognition rates for happiness (92.92 points), anger (98.29 points), sadness (91.21 points) and anxiety (91.92). In other words, the proposed FP and FP+MFCC features are able to improve the recognition rate about 5.6 points and 6.8 points than the result of [4].

In [15], harmony features were proposed for speaker-independent emotion recognition by using a Bayesian classifier on EMODB. The rates of emotion recognition were 52.7 points for happiness, 84.8 points for boredom, 52.9 points for neutrality, 87.6 points for sadness, 86.1 points for anger and 76.9 points 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 points, where 87.59 points for happiness, 54.36 points for boredom, 84.31 points for neutrality, 89.60 points for sadness, 93.62 points for anger and 67.60 points for anxiety. In other words, the proposed FP features are able to improve the recognition rate about 6.01 points than the best result of [15].

Tables 2 and 3 report the confusion matrices of 120 FP features on CASIA and EESDB, respectively. The recognition rate on EESDB is below than that of the other two databases. The main reason might be due to that the emo tions expressed by the elderly are usually more difficult to identify [56].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2392101, IEEE Transactions on Affective Computing



(a) German EMODB database (b) Chinese EESDB database (c) Chinese CASIA database

Fig.4. Comparisons of the recognition rates by FP, MFCC, FP+MFCC and F0+ENERGY+ZCR (FEZ)

According to Tables 1 to 3, it seems that the rates of emotion recognition vary between German and Chinese. It is reasonable in that different countries have different cultures, and the way by which they express their emotion is also different [25].

Fig. 4 shows the results of using MFCC, FP, MFCC+FP and F0+ENERGY+ZCR (FEZ) features on the three databases. In general, the FP features themselves achieved higher average recognition rates than MFCC and FEZ, in particular on the EMODB database. If 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 usally led to worse performance. In other words, although continuous features deliver important emotional cues of speakers [7],[11],[31], FEZ features did not demonstrate better performance in speaker-independent emotion recognition. We also combined FP with FEZ features, but it had little impact on recognition accuracy.

Table 4 shows 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 87.5 points, the proposed FP features outperformed 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 combining FP and MFCC features to roughly 17.5 points, 10 points and 10.5 points on the aforementioned databases, respectively.

5 Conclusion

MFCC was widely emploited for speech emotion recognition over decades. 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 and EESDB. 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 further 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.

TABLE 1
CONFUSION MATRIX OF EMOTION RECOGNITION
USING 120 FP FEATURES ON THE GERMAN DATABASE (%)

	Happ.	Bored.	Neutr.	Sad.	Angry	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
Angry		1.71			98.29	
Anxi.		3.08	1.25	2.08	1.67	91.92

TABLE 2
CONFUSION MATRIX OF EMOTION RECOGNITION
USING 120 FP FEATURES ON THE CASIA DATABASE (%)

	Happ.	Surpri.	Neutr.	Sadn.	Anger	Fear
Happ.	81		2	4	11	2
Surpri..		84		3	10	3
Neutr.			75	16		9
Sadn.	11	8	2	67	6	6
Anger				6	86	8
Fear.		3	4	5	7	81

TABLE 3
CONFUSION MATRIX OF EMOTION RECOGNITION
USING 120 FP FEATURES ON THE EESDB 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

Happ.= happiness, Surpri.=surprise, Neutr.=neutrality, Sadn.=sadness, Anxi.=anxiety, Bored.=boredom

TABLE 4
THE BEST FEATURE AMONG MFCC, FP, FP+MFCC AND FEZ ON THREE DATABASES

	EMODB	CASIA	EESDB
Happ.	FP	FP+MFCC	FP
Sadn.	FP+MFCC	FP	FP
Anger	FEZ	FP+MFCC	MFCC
Neutr.	FP+MFCC	FP	FP
Bored	FP+MFCC		
Surpri.		FP+MFCC	
Fear	FP+MFCC	FP+MFCC	

ACKNOWLEDGMENT

The authors would like to thank all anonymous reviewers for their critical comments. This work was supported by the 111 Project under grant B14025, by National Natural Science Foundation of China under grants 61203312, 61271123, 61370219, 61305064, and by Anhui Provincial Natural Science Foundation under grant KJ2014A047. Any correspondence should be made to N. An (ning.g.an@acm.org) or B.N. Li (bingoon@ieee.org).

REFERENCES

[1] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition* vol.44, no.3, pp.572-587, 2011.

[2] M. Wolfgang, et al, "Challenges in speech-based human–computer interfaces," *International Journal of Speech Technology* vol. 10, no. 2-3, pp: 109-119, 2007.

[3] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "An adaptive framework for acoustic monitoring of potential hazards," *EURASIP Journal on Audio, Speech, and Music Processing*, no.13, 2009.

[4] M. Kotti and F. Paternò, "Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema," *International Journal of Speech Technology*, pp.131-150, 2012.

[5] N.Mavridis, M. S. Katsaiti, S. Naef , A. Falasi, A. Nuaimi, H. Araifi, and A. Kitbi, "Opinions and attitudes toward humanoid robots in the Middle East,"*AI & SOCIETY*, vol. 27, no. 4, pp. 517-534, 2012.

[6] R.A. Calvo and S. D',Mello, "Affect detection: an interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affective Computing,* vol. 1, no. 1, pp. 18-37, Jan.-Jun. 2010.

[7] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol.17, no.4, pp.582-596, 2009.

[8] Y.J. Yuan, P. H. Zhao, and Q. Zhou, "Research of speaker recognition based on combination of LPCC and MFCC," in Proceedings of *IEEE ICIS*, vol. 3, 2010.

[9] T. Kinnunen, and H.Z. Li, "An overview of text independent speaker recognition: from features to supervectors," *Speech Communication,* vol.52, pp. 12-40, 2010.

[10] M. Sheikhan, D. Gharavian, and F. Ashoftedl, "Using DTW neural–based MFCC warping to improve emotional speech recognition," *Neural Computing and Application,* pp.1765-1773, 2011.

[11] R.Cowie, D.Cowie, E. Tsapatsoulis, N.Votsis, G. Kollias, S. W. Fellenz, and J. G.Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine,* vol.18, pp. 32-80, 2001.

[12] C. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Process,* pp. 293–303, 2005.

[13] E.Messina, G. Arosio, and F. Archetti, "Audio-based emotion recognition in judicial domain: a multilayer support vector machines approach,"*Machine Learning and Data Mining in Pattern Recognition*, pp. 594-602, 2009.

[14] C. Gobl and A.Ni Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, pp.189–212, 2003.

[15] B.Yang and M. Lugger, "Emotion recognition from speech signals using new harmony features," *Signal Processing*, pp. 1415–1423, 2010.

[16] P. Ruvolo, I. Fasel, and J. R. Movellan, "A learning approach to hierarchical feature selection and aggregation for audio classification," *Pattern Recognition Letters*, pp.1535–1542, 2010.

[17] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech Signal Process*.,vol. 34, no. 4, pp. 744–754, Aug. 1986.

[18] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B.Weiss, "A database of german emotional speech," in: *Proceedings of INTERSPEECH'05*, pp. 1517–1520, 2005.

[19] S.Ramamohan and S.Dandapat "Sinusoidal model-based analysis and classification of stressed speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp.737-746, 2006.

[20] B.A. Hanson and T. H. Applebaum, "Robust speaker-independent word recognition using static, dynamic and acceleration features: Experiments with Lombard and noisy speech," in: *Proceedings of ICASSP,* pp. 857-860, 1990.

[21] M.Y. You, C.Chen, J.J. Bu, J. Liu, and J.H. Tao, "Emotion recognition from noisy speech," in: *Proceedings of IEEE ICME*, pp.1653-1656, Jul. 2006.

[22] P.H. David, V. Bogdan, B. Ronald, and W.Andreas, "The performance of the speaking rate parameter in emotion recognition from speech," in: *Proceedings of IEEE ICME*, 2012.

[23] W. Johannes, V. Thurid, and A. Elisabeth, "A systematic comparison of different HMM designs for emotion recognition from acted and spontaneous speech," *Lecture Notes in Computer Science,* vol. 4738, pp.114-125, 2007.

[24] V.N. Vapnik, *Statistical Learning Theory*, New York, Wiley, 1998.

[25] K. Norhaslinda, W. Abdul, and Q. Chai, "Cultural dependency analysis for understanding speech emotion," *Expert Systems with Applications,* pp. 5115–5133, 2012.

[26] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Knowledge Discovery Data Mining*, vol. 2, pp. 121–167, 1998.

[27] M. Hayat and M. Bennamoun, "An automatic framework for textured 3D video-based facial expression recognition," *IEEE Trans. Affective Computing*, vol. PP, no.99, pp.1-37, 2014.

[28] S. Chandrakala and C. C. Sekhar, "Combination of generative models and SVM based classifier for speech emotion recognition," *Proceedings of IEEE International Joint Conference on Neural Networks (IJCNN),* pp.1374-1379, 2009.

[29] 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," *Proceedings of ICASSP,* vol. 1, pp. 577–580, 2004.

[30] O. Küstner, R. Tato, T. Kemp, and B. Meffert, "Towards real life applications in emotion recognition," *Affective Dialogue Systems (ADS'05),* E. Andre, L. Dybkaer, W. Minker, and P. Heisterkamp, eds., pp. 25-35, Springer Verlag, May 2004.

[31] E. Vayrynen, J. Kortelainen, and T. Seppanen, "Classifier-based learning of nonlinear feature manifold for visualization of emotional speech prosody", *IEEE Trans. Affective Computing* , vol.4, no.1, pp.47-56, Jan.-Mar. 2013

[32] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*, first ed., Pearson Education, 1978.

[33] B.S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Am.*, vol. 55, no. 6, pp.1304–1312, 1974.

[34] S.E. Bou-Ghazale, J. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 429–442, 2000.

[35] H. Teager, "Some observations on oral air flow during phonation," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 28, no. 5, pp. 599–601, 1990.

[36] L. Kaiser, "Communication of affects by single vowels," *Synthese*, vol.14, no. 4, pp. 300–319, 1962.

[37] D. Caims and J. Hansen, "Nonlinear analysis and detection of speech under stressed conditions," *J. Acoust. Soc. Am.*, 96, pp. 3392–3400, 1994.

[38] G. Zhou, J. Hansen, and J. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 201–216, 2001.

[39] X. Li, J. Tao, M.T. Johnson, J. Soltis, A. Savage, K.M. Leong, and J.D. Newman, "Stress and emotion classification using jitter and shimmer features," in: *Proceedings of ICASSP*, vol. 4, pp. IV-1081–IV-1084, 2007.

[40] M. Lugger and B. Yang, "Combining classifiers with diverse feature sets for robust speaker independent emotion recognition," in: *Proceedings of EUSIPCO*, 2009.

[41] R. Sun, E. Moore, and J.F. Torres, "Investigating glottal parameters for differentiating emotional categories with similar prosodics," *Proceedings of* ICASSP, pp. 4509–4512, 2009.

[42] M. Lugger and B. Yang, "Psychological motivated multi-stage emotion classification exploiting voice quality features," in: F. Mihelic and J. Zibert (Eds.), *Speech Recognition,* In-Tech, 2008.

[43] 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," *in: Proceedings of the ICASSP*, vol. 1, pp. 577–580, 2004.

[44] J. S. Walker, "Fourier series," in: *Encyclopedia of Physical Science and Technology*, Academic Press, 2001.

[45] S. Davis and P.Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Audio Speech Language Process.*, vol. 28, pp. 357 – 366, 1980.

[46] J.H. Tao, F.Z. Liu, M. Zhang, H.B. Jia, "Design of speech corpus for Mandarin text to speech," in: *Proceedings of The Blizzard Challenge 2008 Workshop*, pp. 1, 2008.

[47] M. Grimm, K. Kroschel, E. Mower, and S.Narayanan, "Primitives based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, pp. 787 – 800, 2007.

[48] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette, "Feartype emotion recognition for future audio-based surveillance systems," *Speech Communication*, vol. 50, pp. 487 – 503, 2008.

[49] D. Bitouk, R. Verma, and A. Nenkova, "Class-level spectral features for emotion recognition," *Speech Communication*, vol. 52, no. 7–8, pp. 613–625, 2010.

[50] H.L.F. Helmholtz, *On the Sensations of Tone as a Physiological Basis for the Theory of Music*, second ed., Dover Publications, New York, 1877.

[51] H. Boukricha, I.Wachsmuth, M. N.Carminati, and P. Knoeferle, "Stress detection from audio on multiple window analysis size in a public speaking task," in: *Proceedings of ACII 2013*.

[52] C.Busso, A.Metallinou, "Iterative Feature Normalization Scheme for Automatic Emotion Detection from Speech", *IEEE Trans. Affective Computing,*, vol.4,no4,pp.386-397,2013

[53] C.W. Hsu and C.J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.,* vol. 13, no. 2, pp. 415–425, Mar. 2002.

[54] M. Lugger, B. Yang, W. Wokurek, "Robust estimation of voice quality parameters under realworld disturbances," *Proceedings of ICASSP*, vol.1, pp.14-19, May 2006.

[55] K.X. Wang, Q.L. Zhang, and S.Y. Liao, "A database of elderly emotional speech," *Proceedings of International Symposium on Signal Processing Biomedical Engineering, and Informatics (SPBEI),* pp.549-553, 2014.

[56] I.S. Engberg and A.V. Hansen, *Documentation of the Danish Emotional Speech Database (DES)*, Aalborg University, Denmark, 1996.

[57] K.X. Wang, N. An, and L. Li, "Emotional Speech Recognition Using a Novel Feature Set," *Journal of Computational Information Systems*, vol. 9, pp. 1-8, 2013.

[58] P. Pudil, F. Ferri, J. Novovicova, and J. Kittler, "Floating search method for feature selection with nonmonotonic criterion functions," *Pattern Recognition*, vol.2, pp. 279-283, 1994.

[59] C.C. Chang and C.J. Lin, *LIBSVM: a Library for Support Vector Machines*. Technical report, Department of Computer Science, National Taiwan University, 2009.

Kunxia Wang received her B.S. and M.S. degrees in computer science and technology from Anhui Univercity of Science and Technology, Huainan, China, in 2002 and 2005, respectively. She is currently working toward the PhD degree at Hefei Univercity of Technology, Hefei, China. She has been an Associate Professor at Anhui University of Architechure, Hefei, China. Her research interests include speech signal processing, machine learning and affective computing.

Ning An (corresponding author) received his B.S. and M.S. degrees in computer science from Lanzhou University, Lanzhou, China, in 1993 and 1996, respectively, and his Ph.D. in computer science and technology from Pennsylvania State University, PN, USA in 2002. He is a Full Professor at the School of Computer and Information, Hefei Univercity of Technology, Hefei, China. His research interests are data management, the internet of things, perceived health. His current research includes affective computing, human-machine interface and machine learning methods.

Bing Nan Li (corresponding author) received his B.E. degree in biomedical engineering from Southeast University, Nanjing, China, and completed his first Ph.D. in electronics engineering from University of Macau in 2009 and his second Ph.D. degree in bioengineering from National University of Singapore in 2011, respectively. He is now a Full Professor at the Department of Biomedical Engineering, Hefei Univercity of Technology, Hefei, China. His research interests are healthonics, medical imaging and computing.

Yanyong Zhang received her B.S. degree in computer science from University of Science and Technology of China in 1997, and obtained her Ph.D. degree in computer science and engineering from Penn State University in 2002. She has 10 years of research experience in the field of distributed computing and performance evaluation. She is an Associate Professor at the Department of Electrical and Computer Engineering, Rutgers University, NJ, USA.

Lian Li is a Full Professor at the School of Computer and Information, Hefei Univercity of Technology, Hefei, China. His research interests are computing mathematics, grid computing and machine learning.