A Dimensional Emotion Model Driven Multi-stage Classification of Emotional Speech

Zhongzhe Xiao, Emmanuel Dellandrea, Weibei Dou, Liming Chen

Abstract— This paper deals with speech emotion analysis within the context of increasing awareness on wide application potential of affective computing. Unlike the most of works in the literature which mainly rely on classical frequency and energy based features along with a single global classifier for emotion recognition, we propose in this paper some new harmonic and Zipf based features for better speech emotion characterization in terms of timbre, rhythm and prosody and a dimensional emotion model driven multi-stage classification scheme for a better emotional class discrimination. Experimented on Berlin dataset [1] with 68 features and six emotion states, our approach shows its effectiveness, displaying a 68.60% classification rate and reaching a 71.52% classification rate when a gender classification is first applied. Using DES dataset having five emotion states, our approach achieves an 81% recognition rate when the best performance in the literature to our knowledge is 66% on the same dataset [2].

Index Terms— emotional speech, harmonic feature, Zipf feature, dimensional emotion model, Multi-stage classification

I. INTRODUCTION

Studies suggest that only 10% of human life is completely unemotional while the rest involves emotion of some sort [3]. As a major part of emotion-oriented computing or affective computing [4], automatic emotional speech recognition has potentially wide applications. For instance, based on automatic speech emotion recognition, one can imagine a smart system routing automatically angry customers in a call-center to a human operator, or a powerful search engine delivering speakers in a multimedia collection that discuss a certain topic in a certain emotional state. Another application of emotional speech recognition concerns the development of personal robots either for educational purpose [5] or for pure entertainment [6]. From the scientific point of view, automatic speech emotion analysis is also a challenging problem because of the semantic gap between low level speech signal and

highly semantic (and even subjective in this case) information.

However, machine recognition of speech emotion is feasible only if there is sound emotion taxonomy model and reliable acoustic correlates of emotions in human speech. In the following subsections, we first discuss emotion taxonomies and acoustic correlates, then we overview the related work to further introduce our approach.

A. Emotion taxonomy

The theoretical model of emotions is the first problem raised in the research for emotion classification. According to different psychological theories of emotion, the emotion domain could be cut into different qualitative states or dimensions by different ways. The two traditional theories that have most strongly shaped past research in this area are discrete and dimensional emotion theory [7].

According to discrete theories, there exist a small number, between 9 and 14, of basic or fundamental emotions that are characterized by very specific response patterns in physiology as well as in facial and vocal expression [7]. The term "big six" has gained attention in the tradition of the discrete description of emotions and it implies the existence of a fundamental set of six basic emotions while there is no agreement on which six ones should be. The terms including happiness, sadness, fear, anger, neutral and surprise are often used in the research with this theory. The discrete description of the emotions is the most direct way than other descriptions to discuss emotional clues conveyed by audio signals. Using such a discrete emotion model, it is more likely to distinguish an emotion from the given kinds than to recognize emotions in the whole emotional space.

In the dimensional theories of emotion [8] [9] [10], emotional states are often mapped in a two or three-dimensional space. The two major dimensions consist of the valence dimension (pleasant–unpleasant, agreeable–disagreeable) and an activity dimension (active–passive) [8]. If a third dimension is used, it often represents either power or control. Usually, several discrete emotion terms are mapped into the dimensional space according to their relationships to the dimensions.

For example, some of the dimensional opinions of the emotions characterize the emotional states in arousal and appraisal components [9]. Intense emotions are accompanied by increased levels of physiological arousal. An example of arousal vs. appraisal plane of emotions is shown in Fig. 1. In this

example, arousal values range from very passive to very active, and appraisal values range from very negative to very positive.

Fig. 1 Example of emotions in arousal vs. appraisal plane [9].

There exist some other more elaborated emotion models as well, for instance componential models of emotion [11] which don't limit the description of emotions to two or three basic dimensions as compared to dimensional theories and also permit to model distinctions between members of the same emotion family.

In practice, it is useful to associate discrete model with the dimensional one by mapping the discrete emotional states into dimensional spaces as illustrated in Fig. 1. However, most of the current machine learning algorithms [12] only consider classification problems of a finite number of clearly labeled classes. Machine recognition of speech emotions is mostly based on discrete emotional model whereas the kinds of emotional states and their number of emotional states are typically application dependant.

B. Acoustic correlates of emotions in the acoustic characteristics

Are there reliable acoustic correlates of emotions in speech signal making feasible machine based emotion recognition? As we know, apart from the words, human being expresses emotion through modulation of facial expression [13] and modulation of voice intonation [14]. There are some reliable correlates of emotion in the acoustic characteristics of the signal: speech emotion is question of prosody and expressed by the modulation of the voice intonation parameterized by features such as tonality, intensity, rhythm.

Emotions are considered as cognitive or physical by different theories, and can be discriminated by distinct physical signatures [4]. Several researchers have studied the acoustic correlates of

emotion/affect in the acoustic features of speech signals [14] [15]. According to [14], there exists considerable evidence of specific vocal expression patterns for different emotions. Emotion may produce changes in respiration, phonation and articulation, which in turn affect the acoustic features of the signal [16]. There are also much evidence points to the existence of phylogenetic continuity in the acoustic patterns of vocal affect expression [17]. However, there is currently little systematic knowledge on the details of acoustic patterns that describe the specific emotions in human vocal expressions. Typical acoustic features which are considered as strongly involved in emotional speech signal include the following: 1) The level, range and contour shape of the fundamental frequency (F0), which reflect the frequency of the vibration of the speech signal and is perceived as pitch; 2) the level of vocal energy, which is perceived as intensity of voice, and the distribution of the energy in the frequency spectrum, which affects the voice quality; 3) the formants, which affects the articulation; 4) speech rate. For example, several emotional states such as anger and happiness (or joy) are considered as with high arousal levels [14], they are characterized by a tense voice with faster speech rate, high F0, and broad pitch range, which are caused by the arousal of sympathetic nervous system with increasing of heart rate and blood pressure, which are accompanied with dry mouth and occasional muscle tremors; yet sadness (or quiet sorrow) and boredom are similar with slower speech rate, lower energy, lower pitch, reduced pitch range and variability for both emotions, which are caused by the arousal of parasympathetic nervous system with decreasing of heart rate and blood pressure and increasing of salivation [14] [15] [18].

Emotion recognition can be language and culture independent: acoustical correlates of basic emotions across different cultures are quite common due to the universal physiological effects of the emotions. Abelin and Allwood investigated in [19] utterances spoken be a native Swedish speaker to be recognized by persons native of 4 different languages as Swedish, English, Finnish and Spanish. Close recognition patterns were obtained by people speaking different languages, showing that the inner characters of vocal emotions can be universal and culture independent. Tickle also proved this point by asking Japanese listeners to recognize emotions expressed by Japanese or American people using meaningless utterance without semantic information [20]. The best recognition score by human was about 60%. Similar result was obtained by Burkhardt and Sendlmeier using semantically neutral but

meaningful sentences [15].

These studies thus show considerable possibilities to achieve machine recognition of vocal emotions. On the other hand, as the human recognition of vocal emotion, with roughly 60% recognition rate, is far from accurate, we probably cannot expect perfect machine recognition. This rather average recognition rate of vocal emotion by human mainly comes from similar physiological correlates between certain emotional states, and thus similarities in acoustic correlates. While human beings make use of all contextual information, such as expression, gesture, *etc*. for resolving ambiguity in actual situations, machine based emotion recognition using only vocal modality should focus on a few kinds of basic emotional states to achieve reasonable performance.

C. Related works

Along with increasing awareness of wide application potential from affective computing [4], there exist active research activities on automatic speech emotion recognition in the literature. According to underlying applications, the number of emotion classes considered varies from 3 classes to more classes allowing a more detailed emotion description [2] [21] – [25]. All these works can be compared according to several criteria, including the number and type of emotional classes for the application under consideration, acoustic features, learning and classifier complexities and classification accuracy.

In [21], Polzin and Waibel dealt with emotion-sensitive human-computer interfaces. The speech segments were chosen from English movies. Only three negative emotion classes, namely sad, anger and neutral, are considered. They modeled speech segments with verbal and non-verbal information: the former includes emotion-specific word information by computing the probability of a certain word given the previous word and the speaker's expressed emotion, while the latter includes prosody features and spectral features. Prosody features include mean and variance of fundamental frequency and the jitter information presented by small perturbations in the contour on the fundamental frequency, and mean and variance of the intensity and tremor information presented by small perturbations in the intensity contour. The spectral features include cepstral coefficients derived from a 30 dimensional mel scale filterbank. The verbal features, prosody features and spectral features were evaluated separately in their work. Accuracy up to 64% was achieved on a significant dataset from English movies containing

more than 1000 segments for each of the three emotional states. According to their experiments, this classification accuracy is quite close to human classification accuracy. One of originalities in this work is preliminary separation of speech signals into verbal signal and non verbal signal. Specific feature set is then applied to each group for emotion classification. The major drawback is that the verbal information only works with language dependent problems and doesn't reflect acoustic characters of vocal emotions. Among the non-verbal features, the pitch, intensity, and cepstral coefficients information were used to describe prosody, spectral characteristics of vocal expression; but the prosody features only contained simple features related to fundamental frequency and intensity contour. The other features such as features related to the formants, energy distribution in the spectrum and the other higher level features concerning the whole structure of emotional speech signals is absent in their feature set.

Slaney and Mcroberts also studied three emotion classes problem in [22] but within another context. They considered the 3 attitudes as approval, attention bids, and prohibition from adults talking to their infants aged of about 10 months. They made use of simple acoustic features, including several statistics measures related to the pitch and MFCC as measures of the formant information, and also timbre cepstral coefficients. 500 utterances were collected from 12 parents talking to their infants. A multidimensional Gaussian mixture model discriminator was used to perform the classification. The female utterances were classified at an accuracy rate up to 67%, and the male utterances at 57% accuracy rate. Their experiment also tends to show that their emotion classification is language independent as their dataset is formed by sentences whose emotion was understood by infants who do not speak yet. Their work also suggests that gender information impacts on emotion classification. However, their three emotion classes are quite specific and very different from the ones usually considered in the literature, thus cannot be used as reference directly for other applications. As the main object of Slaney's work was to prove that it is possible to build machines that sense the "emotional state" of a user. The emotion sensitive features were not the key point of this research. Only simple acoustic features were used in their experiment, and relationships between the features and emotions in terms of prosody, arousal or rhythm were not discussed in details.

Gender information is also considered by Ververidis *et al* [23] [24] [2] with more emotion classes.

In their work, 500 speech segments from DES (Danish Emotional Speech) database are used. Speech is expressed in 5 emotional classes, namely anger, happiness, neutral, sadness and surprise. A feature set of 87 statistical features of pitch, spectrum and energy was tested, using the feature selection method SFS (Sequential Forward Selection). In [23], a correct classification rate of 54% was achieved when all data were used for training and testing with a Bayes classifier using the 5 best features: mean value of rising slopes of energy, maximum range of pitch, interquartile range of rising slopes of pitch, median duration of plateaus at minima of pitch and the maximum value of the second formant. When considering gender information in [24], correct classification rates of 61.1% and 57.1% were obtained for male and female subjects respectively with a Bayes classifier with Gaussian Pdfs (Probability density functions) using 10 features. The best result in their work is obtained by a GMM for male samples at 66% classification rate in [2].

Prior to the work of Ververidis *et al*, McGilloway *et al* [25] also studied 5 emotion classification problem with speech data recorded from 40 volunteers describing the emotion types as afraid, happy, neutral, sad and anger. They already made use of 32 classical pitch, frequency and energy based features selected from 375 speech measures. The accuracy was around 55% with a Gaussian SVM when 90% of data were used as training data and 10% as testing data. An extension of this work was carried out by P.Y.Oudeyer within the framework of personal robot communication [26]. He considered 4 emotional classes as joy/pleasure, sorrow/sadness/grief, normal/neutral, and anger in a cartoon-like speech. Using similar features as applied by McGilloway *et al*. and making a large-scale data mining experiment with several algorithms such as neural network, decision tree, classification by regression, SVM, naïve bayes, and Adaboost on WEKA platform [27], P.Y. Oudeyer displayed an extremely high success rate up to 95.7%. However, a direct comparison of this result with the others is quite difficult as the dataset in their experiments seems not to be highly accorded with the natural speech emotions but exaggerated ones as depicted in the cartoon situation. Moreover, emotion recognition is speaker dependant as the robotic pet basically only needs to understand his master's humor.

The feature sets used in experiments by McGilloway *et al* [25], Ververidis [24], and Oudeyer [27] were basically spectral, pitch and energy (intensity) based features and thus similar to each other. The spectral features include low frequency energy (energy below 250Hz) and the formants information.

The pitch features mainly concern the properties of pitch contour, including the statistical values of the pitch value, the duration and value at the plateaus of the pitch contour, and the rising and falling slopes of the pitch contour. Similar statistical values of the energy contour as the ones with the pitch contour were used as energy features. Their experiments show that classical pitch, frequency and energy based features, while partially capturing voice timber, intensity and rhythm, are quite useful for emotion classification. However, these features are likely to mostly reflect nonspecific physiological arousal, and the existence of emotion-specific acoustic profiles may have been obscured [14]. They are thus not enough for capturing speech intonation, because tonality is not only question of pitch and formants patterns and prosody needs to be better captured. Moreover, except the low frequency energy, all the other features are derived from frame based short-term features. Long-term features enabling a better characterization of vocal tonality and rhythms in emotional expression are missing. In addition, all these works rely on a global one step classifier using a same feature set for all the emotional states while studies on emotion taxonomy suggests that some discrete emotions are very closed each other on the dimensional emotion space and there is confusion of emotion class borders as evidenced in [14] which states that acoustic correlates between fear & surprise or between boredom & sadness are not very clear, thus making very hard an accurate emotion classification by a single step global classifier.

D. Our approach

In this work, as our primary motivation is multimedia indexing for enabling content-based retrieval, some rough and basic emotion classes are investigated here. However, our approach is rather general and can be applied for various discreet emotional states which are, as we have seen previously, mostly application dependent. While we fully develop and illustrate our approach using the following "big six" emotion classes from Berlin dataset, namely anger, boredom, fear, happiness, neutral and sadness, we also show the effectiveness of our approach on DES dataset having some different five emotion classes.

Unlike the most works in the literature, our contributions for vocal emotion recognition are twofold: first, as a complement to classical frequency and energy based features which only partially capture the emotion-specific acoustic profiles, we propose some additional features in order to

characterize other information conveyed by speech signals: harmonic features which are perceptual features capturing more comprehensive information of the spectral and timbre structure of vocal signals than basic pitch and formants patterns, and Zipf features which characterize the inner structure of signals, particularly rhythmic and prosodic aspects of vocal expressions ; Second, as a single global classifier using a same feature set is not suitable for discriminating emotion classes having similar acoustic correlates, especially for emotional states close to each other in the dimensional emotion space, we propose here a multi-stage classification scheme driven by the dimensional emotion models which hierarchically combines several binary classifiers. At each stage, a binary class classifier makes use of a different set of the most discriminant features and distinguishes emotional states according to different emotional dimensions. Finally, an automatic gender classifier is also used for a more accurate classification.

Experimented with 68 features on Berlin dataset considering six emotional states, our emotion classifier reaches a classification accuracy rate of 68.60% and up to 71.52% when a first gender classification is applied. On DES dataset with five emotion classes, our approach displays an 81% classification accuracy rate. So far as we know, current works in the literature display a best classification rate up to 66% on the same DES dataset.

The remainder of this paper is organized as follows. Section II defines our feature set, especially the new harmonic and Zipf features. Our multi-stage classification scheme is then introduced in section III. The experiments and the results are discussed in section IV. Finally, we conclude our work in section V.

II. ACOUSTIC FEATURES OF EMOTIONAL SPEECH

As our study on acoustic correlates and related works highlighted, popular frequency and energy based features only partially capture the voice tonality, intensity and prosody of an emotional speech. In complement to these two groups of classical features also used in our work, we introduce in this section two new feature groups, namely harmonic features for a better description of voice timber pattern, and Zipf features for a better rhythm and prosody characterization.

A. Harmonic features

Timbre has been defined by Plomp (1970) as "… attribute of sensation in terms of which a listener can judge that two steady complex tones having the same loudness, pitch and duration are dissimilar." It is multidimensional and cannot be represented on a single scale. An approach to describe the timbre pattern is to look at the overall distribution of spectral energy, in another word, the energy distribution of the harmonics [28].

In our work, a description of sub-band amplitude modulations of the signal is proposed to represent the harmonic distributions. The first 15 harmonics are considered in extracting the harmonic features.

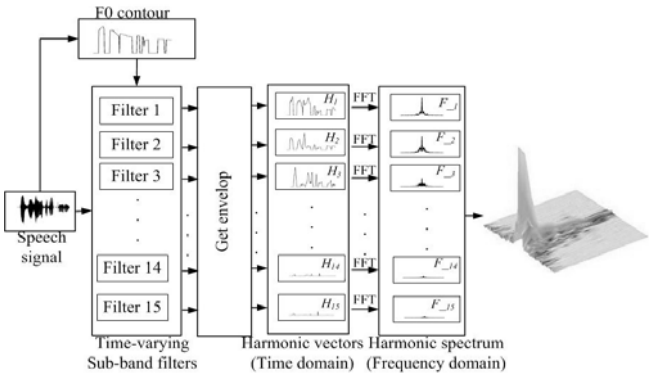

Fig. 2 Harmonic analysis of a speech signal

The extraction process works as follows. First, the speech signal is put into a time-varying sub-band filter bank with 15 filters. The properties of the sub-band filters are determined by the F0 contour, which is derived in section II-C. The center frequency for the i^{th} sub-band filter at a time instant is i^{th} multiples of the fundamental frequency (i^{th} harmonic) at that time, and the bandwidth is the fundamental frequency. The sub-band signals after the filters can be seen as narrowband amplitude modulation signals with time-varying carriers, where the carriers are the center frequency of the sub-band filters mentioned before, and the modulation signals are the envelops of the filtered signals. We call these modulation signals as harmonic vectors (H_1, H_2, H_3…in Fig. 2 and Fig. 4 (a)). That is to say, we use the sum of the 15 amplitude modulated signals using the harmonics as carriers to represent the speech signal as

$$X(n) \approx \sum_{i=1}^{15} H_i(n) * e^{j2\pi i f_0(n) n} \qquad \text{(Eq. 1)}$$

where $X(n)$ is the original speech signal, $H_i(n)$ corresponds to the i[th] harmonic vector in time domain, and $f_0(n)$ is the fundamental frequency.

As the harmonic vectors H_i are in time domain and do not present typical patterns in the timber structure, the amplitudes of spectrums of the harmonic vectors on the whole range of a speech segment are thus used to represent the voice timber pattern:

$$F\_i = FFT(H_i(n)) \qquad \text{(Eq. 2)}$$

The spectrums are shown in Fig. 2 and Fig. 4 (b) (F_1, F_2, F_3…). These 15 spectrums are combined together into a 3-D harmonic space, as shown in Fig. 4 (c).

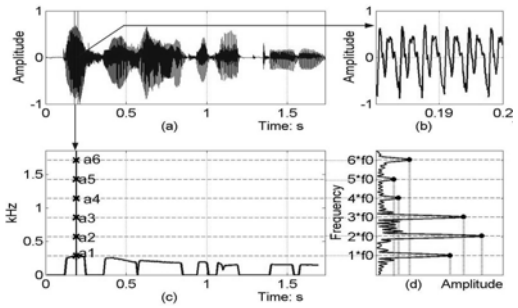

Fig. 3 Calculation process of the harmonic features: (a) waveform in time domain, (b) Zoom out of (a) during 20ms, (c) F0 contour of (a), a1 – a6 are the frequency points of 1 to 6 multiples of the fundamental frequency at the selected time point (d) spectrum of selected time point, the amplitude at a1, a2, a3, a4, a5 and a6

In order to simplify the calculation, we derive the amplitudes at the integer multiples of the F0 contour from the short time spectrum over the same windows as computing the F0 to form the harmonic vectors instead of passing the filter bank, as shown in Fig. 3. As the F0 is derived in our work based on frames of 20ms with 10 ms' overlap (see section II-C), we derive the amplitudes of the 15 harmonic points from the short time spectrum of each frame to approximate the harmonic vectors. Thus, the harmonic vectors in time domain obtained in this way are with sampling frequency of 100Hz, and the frequency axis in the 3-D space ranges between ±50Hz (Fig. 4 (c)).

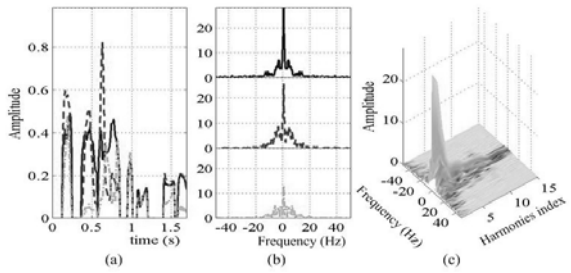

Fig. 4 The amplitude of the harmonic vectors in time domain and their spectrums (a) amplitude contour of the first 3 harmonic vectors in time domain, the dark solid line, dark dashed line and grey dashed line show the first 3 harmonic vectors respectively, (b) the spectrums of the first 3 vectors, (c) 15 spectrums combined in 3-D harmonic space

The 3 axes in the 3-D harmonic space are amplitude, frequency and harmonics index Fig. 4 (c). In these 3 axes, both the frequency axis and the harmonics index axis are in the frequency domain. The harmonics index axis shows the relative frequency according to the fundamental frequency contour, and the frequency axis shows the spectrum distribution of the harmonic vectors. Normally, this space has a main peak at the frequency center of the spectrum of the 1st or the 2nd harmonic vector, and a ridge in the center of the frequency axis, which corresponds to the peak in the spectral center of the harmonic features. The values in the side part of this space are relatively low.

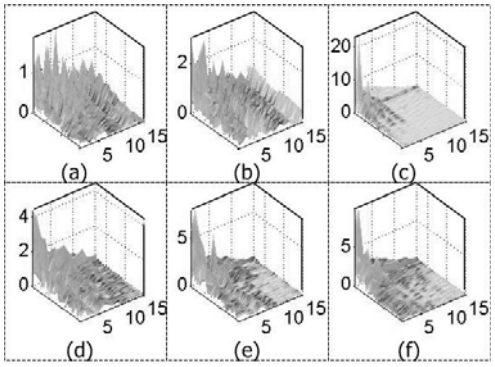

Fig. 5 3-D harmonic space for the 6 emotions from a same sentence: (a) anger, (b) fear, (c) sadness, (d) happiness, (e) neutral, (f) boredom

As the spectrum is symmetric due to FFT properties, we only keep the positive frequency part. Fig. 5 shows the 3-D harmonic space of examples of the 6 emotions from speech samples with a same sentence. The axes in Fig. 5 are the same as in Fig. 4(c). This harmonic space shows obvious difference among the emotions. For example, the emotion ‘anger’ and ‘happiness’ have relative low main peak

and many small peaks in the side parts, and the difference between the harmonic vectors with higher indexes and lower indexes are relatively low, while the ‘sadness’ and the ‘boredom’ have high main peaks but are quite flat in the side part, and the difference between the harmonic vectors with higher indexes and lower indexes are relatively high, the ‘fear and ‘neutral’ have properties between the previous two cases.

In our work, the properties of such a 3-D harmonic space are extracted as features for classification. From the difference in the harmonic space among the emotions, we divide the harmonic space into 4 areas as shown in Fig. 6. The ridge, which shows the low frequency part (lower than 5Hz) in the frequency domain, is selected as area 1; the other part (ranging from 5Hz to 50Hz according to the frequency axis) is divided into 3 areas according to the index of harmonics. Referring to the definition of octaves in the music, these 3 areas are divided with double frequency range to their previous area according to the harmonic index axis. Thus, the area 2 contains the 1st to 3rd harmonic vectors, the area 3 contains the 4th to the 7th harmonic vectors, and the area 4 contains the 8th to the 15th harmonic vectors. The mean value, variance value of each area and the value ratios between the areas are used as features to be selected.

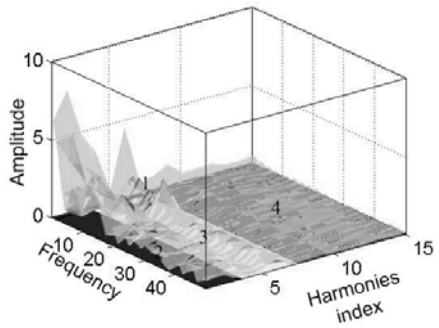

Fig. 6 4 areas for FFT result of 3-D harmonic space

List of harmonic features:

51 – 63. Mean, maximum, variance and normalized variance of the 4 areas

64 – 66. The ratio of mean values of areas 2 ~ 4 to area 1

B. Zipf features

Features derived from an analysis according to Zipf laws are presented in this group to better capture the prosody property of a speech signal.

Zipf law is an empirical law proposed by G. K. Zipf [29]. It says that the frequency *f(p)* of an event *p* and its rank *r(p)* with respect to the frequency (from the most to the least frequent) are linked by a power law:

$$f(p) = \alpha r(p)^{-\beta} \qquad \text{(Eq. 3)}$$

Where α and β are real numbers.

The relation becomes linear when the logarithm of *f*(*p*) and of *r*(*p*) are considered. So, this relation is generally represented in a log-log graph, called Zipf curve. The shape of this curve is related to the structure of the signal. As it is not always well approximated by a straight line, we approximate its corresponding function by a polynomial.

Since the approximation is realized on logarithmic values, the distribution of points is not homogeneous along the graph. So we also compute the polynomial approximation on the resampled curve. It differs from Zipf graph as the distance between consecutive points is constant. In each case, the relative weight associated with most frequent words and with less frequent ones differs.

The Inverse Zipf law corresponds to the study of the event frequency distributions in signals. Zipf has also found a power law which holds only for low frequency events: the number of distinct events *I*(*f*) of apparition frequency f is given by:

$$I(f) = \delta f^{\gamma} \qquad \text{(Eq. 4)}$$

Where δ and γ are real numbers.

Zipf law thus characterizes some structural properties of an informational sequence and is widely used in the compression domain. The most famous application of Zipf law is statistical linguistic. For example, in [30], Zipf law has been evaluated to discriminate natural and artificial language texts; Havlin proved that [31] that the authors can be characterized by the distance between Zipf plots associated with the text of books with shorter distance between the books written by the same author than by different authors.

In order to capture these structural properties from a speech signal, the audio signals are first coded into text-like data, and features linked to Zipf and Inverse Zipf approaches are computed, enabling a characterization of the statistical distribution of patterns in signals [32]. Prosodic information, in particular rhythmic features can be represented by Zipf patterns. Three types of coding as temporal

coding, frequency coding and time-scale coding were proposed in [32], in order to bring to the front different information contained in signals.

For example, the coding principle denoted as TC1 in [32] is to enable to build up a sequence of patterns based on the coding of the original audio signal (Fig. 7). First, three letters – U for Up, F for Flat, and D for Down – are used as a symbolic representation for the signal sample values. The letter U is used when a positive difference between the magnitude values of two successive samples of the audio signal occurs. The letter F is used when the difference is close to zero; and the letter D is used when the difference is negative. Then the letters are assembled by three character long sequences with totally 3^3=27 different possible patterns. Each of them can be associated with a letter of the alphabet; and indicates the local evolution of the temporal signal on three consecutive samples. Adjacent patterns are obtained by shifting the analysis window one step to the right. A sequence of patterns is finally obtained from the audio signal. The pattern sequence can then be formed into words with given length. In the example of Fig. 7, the words length is set to 5.

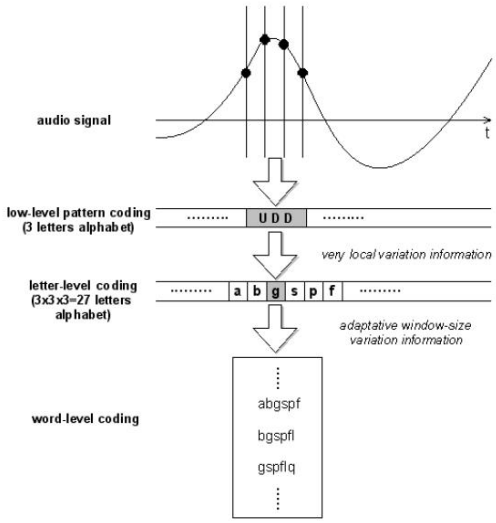

Fig. 7 Description of TC1 coding [32]

From Zipf studies of theses codings, several features are extracted. In this work, 2 features are selected according to their discriminability for the emotions that we consider.

List of Zipf features:

67. Entropy feature of Inverse Zipf of frequency coding

68. Resampled polynomial estimation Zipf feature of UFD (Up – Flat - Down) coding

C. Others – frequency features and energy features

We also considered the classical frequency features and energy features as they partially capture voice tonality and intensity and have shown their efficiency from the overview on the related works in the literature. The frequency features include the statistics of fundamental frequency F0 and the first 3 formants; and the energy features include the statistical features of the energy contour.

The range of F0 is assumed between 60 Hz and 450 Hz for sonant. The F0 and the formants are computed over windows of 20 ms with overlaps of 10ms because the speech signal can be assumed stationary in this time scale and the statistical properties of the F0 and the formants over the length of the speech segments are used as features. The F0 is computed by autocorrelation method, and the formants are computed by solving the roots of the LPC (Linear Predict Coding) polynomial [33]. The F0 and the formants are only computed through the vowels periods, which are segmented by short time energy (STE) and zero crossing rate (ZCR) of signal [34] [35]. For the consonants, the F0 and the formants are assumed as 0, and are not considered in the statistics. See F0 and the formants in Fig. 8 (b).

The energy values in the energy contour are also calculated over windows of 20 ms with overlaps of 10ms as the F0 and the formants. See the solid line in Fig. 8 (c). The edge points of the plateaus of the energy contours are defined as the points at 3 db to the peak points. The energy plateaus and the slopes are obtained by approximating the energy contour with straight lines, see the dashed line in Fig. 8 (c). The examples of energy plateaus and the rising and falling slopes are marked in the figure. The first and last slopes of energy contour of each speech segment are ignored to avoid error values.

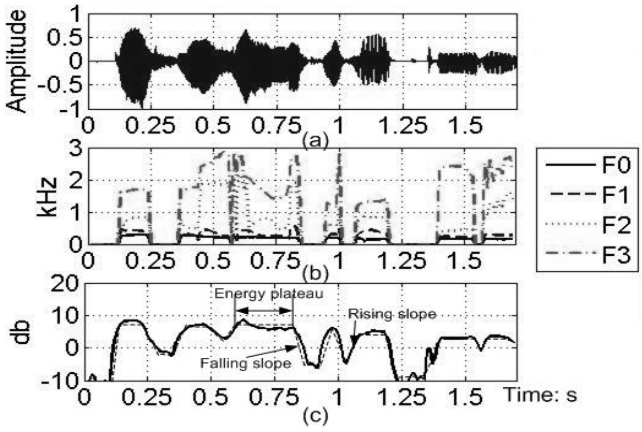

Fig. 8 Basic acoustic features of a speech signal: (a) waveform, (b) fundamental frequency F0 and the first 3 formants (F1, F2, and F3), (c) energy contour

List of frequency features:

1 - 5. Mean, maximum, minimum, median value and the variance of F0

6 – 20. Mean, maximum, minimum, median value and the variance of the first 3 formants

List of energy features:

21 – 23. Mean, maximum, minimum value of energy

24. Energy ratio of the signal below 250 Hz

25 – 28. Mean, maximum, median and variance of energy plateaus duration

29 – 32. Mean, maximum, median value and variance of the values of energy plateaus

33 – 36, 42 – 45. Mean, maximum, median and variance gradient of rising and falling slopes of energy contour

37 – 40, 46 – 49. Mean, maximum, median and variance duration of rising and falling slopes of energy contour

41, 50 Number of rising and falling slopes of energy contour per second

Subsequently, frequency-based features and energy-based features are respectively referred as group 1 and group 2 features while the newly defined harmonic feature set and Zipf one respectively referred as group 3 feature set and group 4 feature set.

III. Hierarchical Classification of emotional speech

Fuzzy neighborhood relationship between some emotional states, for instance between sadness and boredom. As evidenced by studies on acoustic correlates, leads to unnecessary confusion between emotion states when a single global classifier is applied using the same set of features. In this section, we propose a dimensional emotion model guided multi-stage classification method dealing with the emotional classification in several stages. The basic idea here is that emotional states can first be categorized into some broad and rough emotional classes according to the dimensional emotion model in one of the dimensions, such as arousal dimension, and then each broad emotional class can then be further classified into final emotional states according to other dimensions, such as appraisal dimension.

At each classification step, a set of the most relevant features is selected by the SFS feature selection scheme. In doing so, our hierarchical classification scheme enables the use of different relevant feature set for better discriminating emotional states at each stage. Moerover, a gender classifier is also defined which tops our multi-stage emotion classification to further decrease the perturbations between different emotion classes.

A. Feature selection

As the list of audio features introduced in section II is quite important which may lead to the well known phenomenon of "the Curse of Dimensionality" [36], a feature selection is performed as a preprocessing step for each classifier in our hierarchy of classifiers to simplify the computation procedure and to decrease the interference among the features.

There exist two main approaches of feature selection methods according to their dependency to a classifier or not: filter one or wrapper one. Filter methods normally evaluate the statistical performance of the features over the data without considering the underlying classifier. The irrelevant features are filtered out before the classification process. In wrapper methods, the good subsets are selected by using the induction algorithm itself. The criterion of the selection is the optimization of classification accuracy rate.

Filter methods are often fast in the feature selection process, but the resulting classification performance may be relatively low. For example, the PCA (principal component analysis) is too sensitive to data outliers. In our work, we thus made use of a wrapper method, namely SFS algorithm [37] which is a reasonable compromise between speed and performance.

SFS begins with an empty subset of features. The new subset S_k with k features is obtained by adding a single new feature to the subset S_{k-1} which performs the best among the subsets with k-1 features. The correct classification rate achieved by the selected feature subset is used as the selection criterion. The selection process stops when the correct classification rate begins to decrease.

All the features are normalized before the SFS by (Eq.5):

$$F_n = \frac{F_{n0} - \min(F_{n0})}{\max(F_{n0}) - \min(F_{n0})} \qquad \text{(Eq.5)}$$

Where F_{n0} is the original value of feature n, and F_n is the normalized value of feature n, which is

used in the SFS and classification.

B. Dimensional emotion model driven hierarchical classification of emotional speech

As our study on emotion taxonomy and acoustic correlates highlighted, some emotional states can have similar acoustic correlates. Thus a relevant feature with good discrimination to a certain pair of emotional classes may be a feature with high confusion to another pair of emotional classes. Moreover, coming back to our study on emotion taxonomy in section I-A, the relationship between discrete emotion models and dimensional ones reveals that some emotional classes have some similarities with certain features according to their position in the dimensional distribution. Clearly, a hierarchical emotion classification scheme is needed.

In our work, emotion classes come from two public datasets (Berlin dataset [1] and DES dataset [41], see section IV-A). Referring to these discrete emotion states in arousal vs. appraisal plane (Fig. 1, [9]), they can also be mapped into a 2-D emotional space as in Fig. 9 : anger and happiness stand in very active position, and sadness and boredom stand in very negative position according to the arousal dimension, etc.. We thus propose a hierarchical dimensional emotion model driven classification scheme which combines at its early stage, according to neighborhood relationship in arousal or appraisal dimension, some close emotional classes into intermediate broad classes, reducing the number of the classes at each stage to simplify the overall classification complexity.

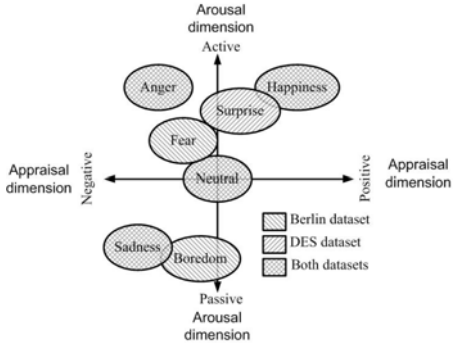

Fig. 9 The emotions in the dimensional space

Fig. 10 illustrates such a hierarchical classification scheme with two stages [39], called subsequently *Dimensional Emotion Classifier* (DEC), applied on emotion classes from Berlin dataset. As we can see from the figure, speech signal is first divided into two intermediate emotional classes according to arousal dimension: active one including anger and happiness, and non active one including

the rest of emotional states. Further, speech samples labeled as active class are categorized into terminal emotional classes, i.e. anger and happiness classes, according this time to appraisal dimension. It is much the same for speech signals labeled as *non active* class. They are first categorized as median and passive classes according to arousal dimension, and then as fear and neutral, sadness and boredom according to appraisal dimension.

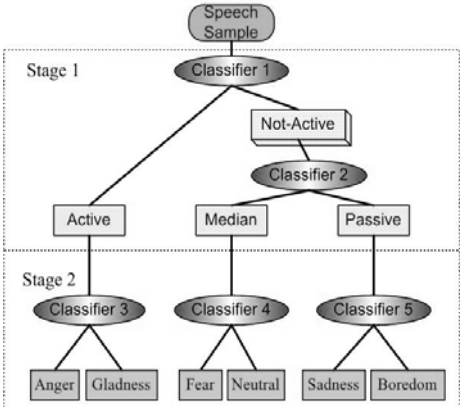

Fig. 10 Dimensional Emotion Classifier (DEC) on Berlin dataset: a Two-stage hierarchical classification scheme of emotional speech driven by the dimensional emotion model

Any machine learning algorithms may be used for the classifiers in such a multi-stage classification scheme. In our work, neural networks have been chosen for their abilities of discriminating non linear data and generalization. We made use of BP (Back Propagation) neural networks with 2 hidden layers, 15 neurons in each layer, and the log-sigmoid function as transfer function. For each network, the inputs are the feature subset, and there is only one output node separating 2 classes by a threshold of 0.5.

1) Stage 1: classification in arousal dimension

Emotional states are first classified according to arousal dimension in two steps into three states, namely active, median and passive state [40]. In the first step, the active state is separated from the median and the passive states (classifier 1 in Fig. 10); and in the second step, the median state and the passive state are further separated (classifier 2 in Fig. 10).

2) Stage 2: classification in appraisal dimension

The first stage of classification in arousal dimension achieves an emotional classification into three rough states (Fig. 9). For each of these three rough emotional states, we further proceed to achieve an

appraisal dimension-based classification to obtain final emotional classes.

Similar classifiers as those proposed in stage 1 are used at this stage. According to Fig. 10, classifier 3 is used for the active state, separating the “anger” from the “happiness”, classifier 4 is used for the median state, separating the “fear” from the “neutral”, and classifier 5 is used for the passive state, separating the “sadness” from the “boredom”.

C. An automatic gender detection based hierarchical classification of emotional speech

The related works in the literature prove that gender difference in the acoustic features also influences the emotion recognition [22] [24]. We thus extend our previous dimensional emotion model driven hierarchical classifier (DEC) by a gender classification to allow different models being used for the speech samples according to the gender. Fig. 11 illustrates the final classification scheme, subsequently labeled Automatic Gender Recognition based DEC, which tops a gender classifier on two DEC schemes as defined in the previous section.

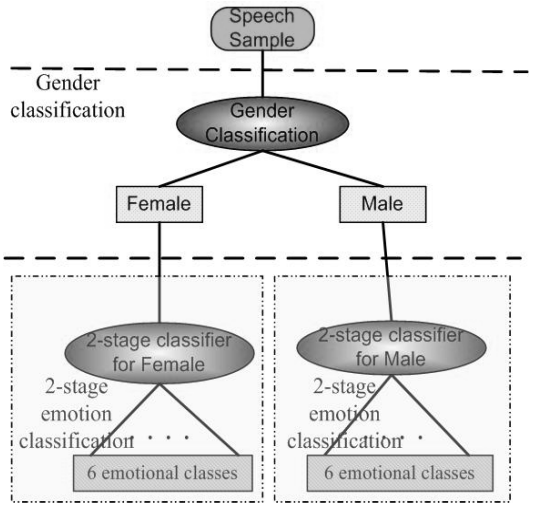

Fig. 11 Gender-Based DEC: a gender classification tops two DECs according to the detected gender

As we can see from this figure, the two two-stage Dimensional Emotion Classifiers (DEC) have the same structure (as shown in Fig. 10), but work with different feature set according to the underlying gender information delivered by the gender classifier.

Any gender classifier might be used. In our work, we build a gender classifier similar as the one defined in our previous work [38] and make use of a neural network with SFS feature selection. The selected feature subset contains 15 features from the whole feature set (see section 2): 19, 55, 1, 58, 44,

28, 59, 63, 14, 16, 4, 5, 8, 11, and 64 (ordered by the sequence of selection). The average recall rate with this feature subset is 94.65% using 10 groups of cross validation on Berlin dataset introduced below.

IV. EXPERIMENTAL RESULTS

The effectiveness of our approach is experimented both on the Berlin dataset and DES dataset. Recall that frequency-based features and energy-based features as introduced in section II rare respectively referred as group 1 and group 2 features while the newly defined harmonic feature set and Zipf one respectively referred as group 3 feature set and group 4 feature set. In the following, we first describe quality of database and introduce Berlin and DES datasets. Then, our experimental results are presented and discussed.

A. Emotional speech datasets

Generally, there are 3 major categories of emotional speech samples. They are natural vocal expression, induced emotional expression, and simulated emotional expression [7]. Natural vocal expression is recorded during naturally occurring emotional states of various sorts. Induced emotions are caused by using psychoactive drugs or some particular circumstances, such as in some kind of games or by using inducing words to get the speech sample of desired emotion. The third category for getting emotional speech samples is the simulated emotional expression which consists of asking actors to produce vocal expressions of certain emotions. In this way, the content and the emotions are given, and the process can be controlled to get more typical expressions. In the literature, the most preferred way of getting emotional speech samples is the third one, the most common used databases being Berlin database [1] and DES (Danish Emotional Speech) database [41].

Berlin emotional speech database is developed by Professor Sendlmeier and his fellows in Department of Communication Science, Institute for Speech and Communication, Berlin Technical University [1]. This database contains speech samples from 5 actors and 5 actresses, 10 different sentences of 7 kinds of emotions: anger, boredom, disgust, fear, happiness, sadness and neutral. There are totally 493 speech samples in this database, in which 286 speech samples are of female voice and 207 samples are of male voice. The length of the speech samples varies from 3 seconds to 8 seconds, and the sampling rate is 16 kHz.

The DES dataset is recorded by Center for Person Komunikation (CPK), Aalborg University, Denmark, as a part of the VAESS project (Voices, Attitudes and Emotions in Speech Synthesis). The sound files were recorded in mono with 16-bit PCM under sample rate of 20 Khz. Four actors were employed for the recording of the DES, two males and two females. Five emotions are considered in the DES: neutral, surprise, happiness, sadness, and anger.

In our work, these two datasets are both used for experimental evaluation of our approach. As there are more emotional types and more actors in the Berlin dataset than the DES dataset, full scale experiments are driven using the Berlin dataset and the preliminary results were reported in the research report [42].

B. Experimental results on Berlin dataset

In our experiments, the data in each case is divided into 10 groups randomly for cross validation and the average of these 10 results is adapted as final result. In each time of experiment, 50% of the samples are used as training set and the other 50% samples are used as testing set. As there are only 8 samples of "disgust" in the male samples, which is much less than the other types, and the acoustic feature for this emotion is inconsistent [14], this type is omitted in training and testing. The influence of gender information on the emotion classification accuracy is also highlighted. For each classification scheme, three experimental settings, using respectively only the female speech samples, the male speech samples and the combination of all the samples (mixed samples), are evaluated and compared.

1) Harmonic and Zipf features vs frequency and energy based features

This first experiment aims at studying the contributions of our harmonic and Zipf features for improvement of emotion classification accuracy when they are used in complement to classic frequency and energy based features. For this experiment, no innovation is brought in classification scheme and we only make use of several well known global classifiers all using the same feature set. Two sets of experimental results are thus produced. The first one contains results produced by the global classifiers when only classic frequency and energy based features are used. The second set of experimental results is obtained when the previous classic frequency and energy based features are extended to also include harmony and Zipf features.

The experiments are carried out on TANAGRA platform [43]. Five types of classifiers are tested: Multi-layer Perception (Neural Network, marked as MP in the following text), C4.5, Linear Discriminant Analysis (LDA), K-NN, and Naive Bayes (NB). Each classifier is tested with several parameter configurations, and only the best results are kept. The correct classification rates are listed in Table I.

Table I. Best recognition rates with one-step global classifiers

	Frequency and energy feature set (FES)			All features (FES+Harmonic +Zipf features)		
	Female	Male	Mixed	Female	Male	Mixed
MP	60.38±2.26	57.91±2.56	60.38±2.26	65.73±2.85	64.45±2.47	64.47±1.93
C4.5	54.27±1.80	53.90±3.93	52.04±2.21	55.46±2.7	58.60±3.70	53.16±1.52
LDA	61.03±1.89	57.09±1.73	59.09±1.25	60.92±2.56	51.16±3.05	64.71±1.64
K-NN	58.24±2.63	53.56±2.89	56.34±1.38	60.14±2.37	60.92±2.71	60.89±1.69
NB	60.70±1.85	56.61±2.26	58.16±1.48	62.67±1.45	62.12±2.47	62.07±1.75
Best	61.03	57.91	60.38	65.73	64.45	64.71

The confusion matrixes with the highest recognition rates are listed in Table II and Table III. As we can see from these tables, the additional features that we have proposed help to improve by at least 4 points the performance achieved by all the global classifiers fed by frequency and energy features, the best amelioration being obtained on male emotional samples with a performance gain of 6 points. The next experiment will precisely show the relevance of our harmonic and Zipf features in the classification process.

Table II. Confusion matrix of the global classifier with frequency and energy features with TANAGRA (%)

	Predicted Actual	Ang.	Hap.	Fea.	Neu.	Sad.	Bor.
Female	Ang.	67.55	23.67	6.24	1.39	0.00	1.15
	Hap.	35.56	45.60	12.50	3.70	0.00	2.64
	Fea.	14.87	23.85	37.18	12.31	4.87	6.92
	Neu.	0.20	1.22	3.47	61.43	3.27	30.41
	Sad.	0.00	0.00	0.27	8.02	86.96	4.76

	Predicted Actual	Ang.	Hap.	Fea.	Neu.	Sad.	Bor.
	Bor.	2.00	1.85	6.62	30.92	8.15	50.46
Male	Ang.	82.67	8.80	6.80	0.67	0.53	0.53
	Hap.	36.18	39.12	20.00	3.53	0.00	1.18
	Fea.	12.82	10.26	55.38	11.54	6.92	3.08
	Neu.	1.52	3.26	5.00	48.91	10.87	30.43
	Sad.	0.20	0.82	2.86	10.61	61.22	24.29
	Bor.	1.84	1.43	1.84	27.96	26.73	40.20
Mixed	Ang.	71.82	21.71	4.73	0.46	0.00	1.27
	Hap.	43.31	41.02	8.63	3.52	0.35	3.17
	Fea.	13.85	25.90	38.72	6.92	7.44	7.18
	Neu.	1.84	1.22	3.27	57.14	6.73	29.8
	Sad.	0.00	0.41	1.63	5.84	80.16	11.96
	Bor.	2.62	2.46	3.54	24.00	12.31	55.08

Table III. Confusion matrix of the global classifier with all features (FES+harmonic+Zipf features)

(%)

	Predicted Actual	Ang.	Hap.	Fea.	Neu.	Sad.	Bor.
Female	Ang.	73.44	21.71	2.66	0.46	0.12	1.62
	Hap.	38.03	50.53	6.51	1.94	2.11	0.88
	Fea.	12.56	23.59	41.79	5.9	10.51	5.64
	Neu.	1.02	1.02	0.61	60.00	6.12	31.22
	Sad.	0.00	0.14	0.68	5.30	86.68	7.20
	Bor.	1.69	1.23	2.00	22.62	8.77	63.69
Male	Ang.	84.93	9.33	5.33	0.27	0.00	0.13
	Hap.	30.88	46.18	17.65	2.94	0.88	1.47
	Fea.	11.03	18.72	55.38	7.44	6.15	1.28
	Neu.	3.26	0.65	3.26	58.04	7.39	27.39

	Sad.	0.00	1.63	3.06	4.29	73.27	17.76
	Bor.	1.22	0.20	1.22	22.65	24.49	50.20
Mixed	Ang.	74.57	18.40	5.19	1.05	0.00	0.80
	Hap.	38.81	44.76	11.14	1.76	1.76	1.76
	Fea.	10.53	15.4	56.48	5.52	8.99	3.08
	Neu.	0.95	1.48	2.63	62.7	5.69	26.55
	Sad.	0.00	0.49	2.04	4.89	79.97	12.62
	Bor.	1.50	1.14	2.73	23.66	14.95	56.02

2) The two-stage Dimensional Emotion model driven Classification (DEC)

The second experiment aims at highlighting contributions on performance improvement from the innovation that we have proposed on classification scheme, namely DEC scheme as represented in (Fig. 10). Recall that all the sub-classifiers in DEC are neural networks and the SFS is applied for each sub-classifier for each gender. The selected feature subsets and the recognition rates for the sub-classifiers are listed in Table IV where the superscript indicates the feature group number which a selected feature comes from (see section II).

Table IV. Selected features and recognition rates for the sub-classifiers (The groups of the features are marked with superscripts)

		Selected feature subset (Ordered by the sequence of selection)	Recognition rate (%)
Active vs. non-active	Female	67^4, 65^3, 25^2, 61^3, 26^2, 51^3, 21^2, 53^3, 28^2	91.13±1.46
	Male	24^2, 4^1, 9^1, 19^1, 52^3, 51^3, 17^1, 65^3, 67^4, 12^1	92.32±2.21
	Mixed	56^3, 68^4, 25^2, 1^1, 14^1, 26^2, 28^2, 29^2, 42^2, 5^1, 65^3, 27^2	90.31±4.59
Median vs. Passive	Female	65^3, 4^1, 27^2, 26^2, 57^3, 53^3, 66^3, 28^2, 56^3, 51^3, 1^1, 24^2	84.98±0.78
	Male	66^3, 67^4, 9^1, 56^3, 61^3, 54^3, 53^3, 5^1, 21^2, 26^2, 57^3	88.23±3.03
	Mixed	66^3, 28^2, 27^2, 57^3, 65^3, 53^3, 26^2, 32^2	84.73±0.14
Anger vs. Happiness	Female	6^1, 7^1, 64^3, 4^1, 53^3, 32^2, 57^3, 24^2	80.21±3.43
	Male	24^2, 33^2, 65^3, 9^1, 39^2, 60^3, 28^2, 2^1, 14^1, 18^1	85.37±6.25
	Mixed	68^4, 31^2, 18^1, 9^1, 13^1, 53^3, 56^3, 58^3, 65^3, 34^2	80.62±7.58
Fear vs. Neutral	Female	4^1, 52^3, 37^2, 9^1, 48^2	90.85±1.02
	Male	64^3, 60^3, 37^2, 53^3, 57^3, 44^2, 51^3	92.85±0.80
	Mixed	4^1, 47^2, 37^2, 44^2, 49^2, 13^1, 60^3, 46^2, 50^2, 42^2, 54^3, 38^2, 56^3	84.31±5.43
Sadness	Female	5^1, 67^4, 8^1, 24^2, 19^2, 9^1, 48^2, 16^1, 2^1, 46^2, 65^3, 55^3,	92.88±0.93

vs. Boredom		13^1, 56^3	
	Male	59^3, 50^2, 20^1, 22^2, 62^3, 48^2, 60^3, 58^3	91.30±0.29
	Mixed	5^1, 9^1, 11^2, 66^3, 13^1, 30^2, 50^2, 57^3, 41^2, 8^1, 24^2, 54^3, 16^1	89.26±2.47

From Table IV, we can see that frequency features (group 1) and energy features (group 2) deliver standard performance for the five sub-classifiers. While group 1 with frequency features is more efficient in classifier 3 ("anger" vs. "happiness") and classifier 5 ("sadness" vs. "boredom"), harmonic features (group 3) are selected most frequently in all the five sub-classifiers, and especially dominate the feature subsets for classifier 2 ("median" and "passive"). For example, feature 65 (the ratio of mean values of areas 3 to area 1 in harmonic space) shows very high discriminality in stage 1 – arousal classification (separating the 3 states), but less efficient in stage 2 – appraisal classification. Although there are only two features in feature group 4 (Zipf features), they show great importance in the feature subset for classifier 1 ("active" vs. "non-active"), which confirms our assumption that the Zipf features have high ability in describing the prosody patterns.

DEC achieves a classification accuracy rate of 71.89%±2.97% in cross-validation for female samples, and 75.75%±3.15% for male samples, and 68.60%±3.36% for mixed samples. The mean confusion matrixes from DEC scheme for the two genders and the mixed case in cross-validation are listed in Table V.

Table V. Mean confusion matrix achieved by DEC (%)

	Predicted / Actual	anger	Happiness.	Fear.	Neutral.	Sadness.	Boredom.
Female	Ang.	83.43	13.76	3.76	3.61	1.67	2.12
	Hap.	19.13	69.00	8.63	3.38	1.38	5.38
	Fea.	8.71	11.47	**73.45**	5.61	3.88	4.23
	Neu.	2.01	4.51	5.01	75.75	2.51	17.76
	Sad.	1.83	1.83	2.69	6.97	91.43	4.40
	Bor.	2.99	2.10	1.88	14.55	3.66	83.11
Male	Ang.	89.33	9.12	3.46	2.79	1.79	2.46
	Hap.	18.80	65.00	13.80	4.63	1.30	2.97

	Fea.	1.96	10.04	**78.85**	5.43	6.58	5.04
	Neu.	2.46	2.72	3.78	83.68	4.56	11.14
	Sad.	1.86	1.86	1.86	4.21	92.94	6.57
	Bor.	2.07	2.66	1.78	7.66	5.90	88.82
Mixed	Ang.	85.35	11.16	3.91	4.39	1.71	2.02
	Hap.	25.15	61.88	10.46	5.93	1.24	1.55
	Fea.	12.38	12.38	**55.27**	15.11	4.75	5.66
	Neu.	2.47	2.72	7.21	78.33	4.01	13.11
	Sad.	2.42	2.80	2.61	7.80	82.31	10.30
	Bor.	2.39	1.76	2.90	13.78	5.68	81.65

The weighted average recognition rate according to the number of speech samples for female samples and male samples is 73.58%, which is 4.78% higher than the result for mixed speech samples (68.60%). From Table V, we can see that the mixing of the gender cause more misjudgment for the emotion "fear" than for the other emotions.

3) Automatic Gender Recognition-based DEC

The third experiment makes use of automatic gender detection on the top of a DEC scheme as introduced in section III-C. The confusion matrix of the multi-stage classification is listed in Table VI. The automatic gender recognition DEC achieves a recognition rate of 71.52%±3.85% which is 2.92% higher than the result from simple DEC (68.60%±3.36%).

Table VI. Confusion matrix of automatic gender recognition based DEC (%)

Predicted Actual	Anger.	Happiness.	Fear.	Neutral.	Sadness.	Boredom.
Anger.	**85.35**	12.34	3.44	3.44	1.71	2.26
Happiness.	21.89	**63.28**	12.05	3.77	1.27	4.08
Fear.	6.39	11.12	**74.18**	5.66	5.12	4.93
Neutral.	2.19	4.50	5.52	**77.56**	3.60	14.37
Sadness.	1.73	1.73	5.19	8.65	**86.35**	5.00
Boredom.	3.33	2.69	1.93	11.30	4.97	**84.18**

4) Synthesis and Discussion

Table VII summarizes the overall performances achieved by the different classification scheme through the previous three experiments. For both global classifier and DEC scheme, the recognition results for the mixed samples are lower than the weighted average result of the 2 genders. The use of an automatic gender recognition classifier can reduce such degradation. As we can see from the synthesis table, when harmonic and Zipf features sets are used in complement to frequency and energy features, single global classifier achieves at least an accuracy gain of 4 points. We further improves the previous classification accuracy when our multi-stage DEC scheme is used, leading to a 71.52% accuracy classification rate with an automatic gender recognition engine on the top of DEC schemes.

Table VII. Synthesis of recognition rates by the four classifiers (%)

	Male	Female	Average of the 2 genders	Mixed	Mixed with gender info
Global with Grp 1 (frequency based features)+ Grp.2 (energy based feature)	57.91±2.56	61.03±1.89	59.55	60.38±2.26	--
Global (Grp.1+2+Harmonic+Zipf features)	64.45±2.47	65.73±2.85	65.12	64.71±1.64	--
DEC scheme	75.75±3.15	71.89±2.97	73.58	68.60±3.36	--
Automatic Gender recognition based DEC	--	--	--	--	71.52±3.85

From these experimental results, we can draw the following lessons:

First, our hierarchical classification scheme (DEC) combining several two-class classifiers according to dimensional emotion model helps to decrease disturbance between neighbor emotion classes and results in an increased recognition rate.

Secondly, the four groups of features show their importance at the different stage in our DEC scheme, thus confirming our intuition for a hierarchical classification scheme. Indeed, feature group 3 (harmonic features), while characterizing the high level timbre structure of speech signals and selected by SFS at every classification stage, displays higher discriminality than the other 3 feature groups. For the DEC scheme, the feature groups 1 (frequency based features) and 2 (energy based features) seem to be more important for stage 2 in appraisal dimension, and our newly proposed features, feature groups 3 with harmonic features and 4 with Zipf features, appear to be more important for stage 1 in arousal

dimension. The ability of different groups of features to discriminate the emotional states in different dimensions in the emotional space shows the possibility to develop automatic classification systems for emotional speech even if the number and types of emotional states change in the applications.

Thirdly, these experimental results confirm the conclusion from several works in the literature stating that there exist much difference between the two genders in the way of expressing their emotions, and an automatic gender discrimination before the 2-stage DEC scheme in our case has helped to improve the recognition rate for some emotions, especially for "fear" - the most confused emotion state for the mixed samples.

C. Experimental results on DES dataset

Encouraged by the previous results on Berlin dataset, we further evaluate the effectiveness of our new features and our multi-stage dimensional emotion model driven classification approach on DES dataset. Recall that there only exist five emotion states in DES dataset which are Anger, Happiness, Neutral, Sadness, and Surprise. Using first arousal dimension and then appraisal dimension in dimensional emotion model as we did for our previous six emotion classification problem, we derived the following hierarchical classification scheme as illustrated in Fig. 12 which splits first all the emotion states, according to arousal dimension, into two broad emotion classes gathering Anger, Happiness and Surprise on one hand, and Neutral and Sadness on the other hand. These broad emotion classes are further divided through three other classifiers to attain the final emotion states.

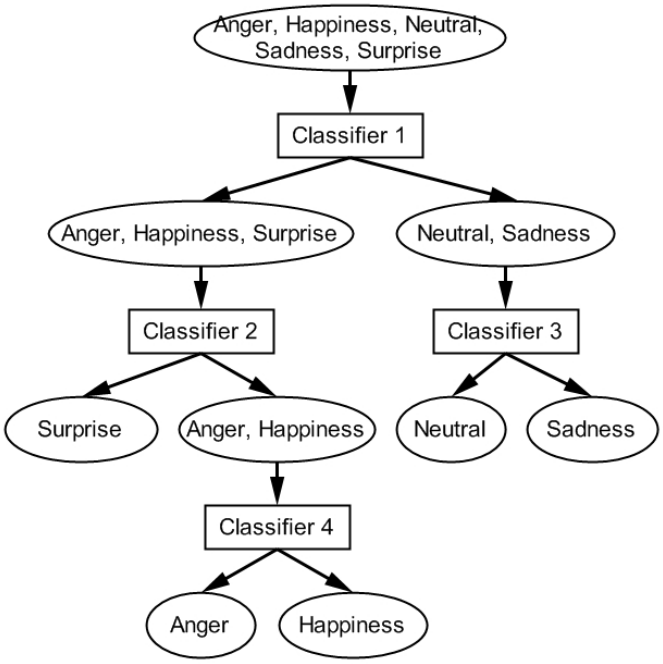

Fig. 12 DEC on DES dataset

In order to compare this result with the work of Ververidis *et al*, the same ratio between training and testing set as 90% and 10% with a cross-validation is applied in this experiment. Table VIII summarizes the accuracy rates, and Table IX gives the confusion matrix of such an evaluation. As we can see, average classification accuracy rates of 81% are achieved in our work. For comparison, the best performance in the literature to our knowledge on the same dataset is 66% classification accuracy rate for only male samples by Ververidis *et al* [2].

Table VIII. Accuracy rates on DES dataset (%)

Female	Male	Mixed
85.14±2.02	87.02±1.44	81.22±1.27

Table IX. Confusion matrix on DES dataset (%)

		Anger	Happiness	Neutral	Sadness	Surprise
Female	Anger	**76.86**	14.71	2.94	1.37	4.12
	Happiness	9.22	**86.08**	0	1.18	3.53
	Neutral	1.37	2.55	**85.88**	8.43	1.76
	Sadness	0	0.96	8.46	**89.04**	1.54
	Surprise	4.81	4.81	1.67	1.11	**87.59**
Male	Anger	**84.51**	5.49	2.16	2.35	5.49

	Happiness	4.63	**85.37**	3.15	0.37	6.48
	Neutral	4.91	3.27	**87.09**	3.64	1.09
	Sadness	0.37	0.74	6.85	**90.93**	1.11
	Surprise	5.9	6.56	0.49	0	**87.05**
Mixed	Anger	73.43	13.14	2.84	3.63	6.96
	Happiness	6.86	**80.67**	1.62	1.62	9.24
	Neutral	3.68	3.49	**81.89**	8.87	2.08
	Sadness	0.38	0.94	8.21	**88.77**	1.7
	Surprise	7.22	7.83	1.39	2.52	**81.04**

V. CONCLUDING REMARKS

In this work, we have proposed, in complement to classic frequency and energy based features, two new feature groups, namely harmonic and Zipf features, for a better characterization of emotional speech in terms of timbre, prosody and rhythm. Moreover, dealing with fuzzy neighborhood of some discreet emotion states having similar acoustic correlates, we have also proposed a hierarchical classification scheme (DEC scheme) using alternatively arousal and appraisal dimension from a dimensional emotion model. Experiments carried out on Berlin dataset show first that our newly proposed Harmonic and Zipf feature groups help to improve emotion recognition rate when used in complement to classic frequency and energy based features, and second, that our DEC scheme further improves the classification accuracy. The effectiveness of our approach has also been validated on another public dataset, DES dataset.

However, there still exist several issues which need to be addressed in a future work.

First, as there is no common agreement on the number and types of discrete emotions, the types of emotions considered in practice are usually application or dataset dependent. Our DEC scheme relies on intuitive mapping of the discreet emotion states into the dimensional emotion model. In this work, this intuitive mapping was thus made manually. An automatic mapping scheme is clearly needed especially when the number of emotions increases and their types vary. We are currently investigating this problem with some preliminary results [44].

Second, as the emotions are very subjective and the emotion borders between closed emotions in the dimensional space are usually not very clear, judgment on emotional state conveyed by an utterance may be between some emotional states or even multiple according to person. Thus ambiguous or multiple judgments also need to be addressed.

As another future research direction, we envisage to further validate our approach in considering other datasets and assess the generality of our work by considering also music signals using similar classification system that we have built for speech signals.

REFERENCES

[1]. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B., A Database of German Emotional Speech Proceedings Interspeech 2005, Lissabon, Portugal

[2]. Ververidis, D. Kotropoulos, C, Emotional Speech Classification Using Gaussian Mixture Models and the Sequential Floating Forward Selection Algorithm, IEEE International Conference on Multimedia and Expo, 2005. ICME 2005, p1500- 1503

[3]. http://emotion-research.net

[4]. Picard, R., "Affective Computing", MIT Press, 1997

[5]. Druin A., Hendler J., Robots for Kids: exploring new technologies for learning, Morgan Kauffman, Los Altos, CA, 2000

[6]. Kusahara M., "The art of creating subjective reality: an analysis of Japanese digital pets", in: Boudreau E. (Ed), in Artificial Life 7 Workshop Proceedings, pp.141-144

[7]. Scherer, K. R., Vocal communication of emotion: A review of research paradigms, Speech Communication 40, pp. 227-256, 2002

[8]. Scherer, K.R., Johnstone, T., Klasmeyer, G., Banziger, T., Can automatic speaker verification be improved by training the algorithms on emotional speech? In: Proc.ICSLP2000, Beijing, China, 2000.

[9]. Wieczorkowska, A., Synak, P., Lewis, R., Ras, Z., W., Extracting Emotions from Music Data, Proceedings of 15th International Symposium, ISMIS 2005, Saratoga Springs, NY, USA, May 25-28, 2005., p456-465

[10]. Pereira, C., Dimensions of emotional meaning in speech, Proceedings of the ISCA workshop on

Speech and Emotion pp. 25-28, 2000, Newcastle, Northern Ireland.

[11]. Scherer, K.R., Schorr, A., Johnstone, T., Appraisal Processes in Emotion: Theory, Methods, Research.Oxford University Press, New York and Oxford, 2001.

[12]. Bishop, C. M., Pattern recognition and machine learning, Ed. Springer, 2006

[13]. Ekman P., "Emotions in the human face", Cambridge University Press, 1982

[14]. Banse, R., Sherer, K.R., 1996. Acoustic profiles in vocal emotion expression, Journal of Personality and Social Psychology 70 (3), 614–636.

[15]. Burkhardt, F., Sendlmeier, W., 2000. Verification of acoustical correlates of emotional speech using formant-synthesis, In: Proceedings of the ISCA Workshop on Speech and Emotion.

[16]. Scherer, K. R., Vocal correlates of emotion, in A. Manstead & H. Wagner (Eds.), Handbook of psychophysiology: emotion and social behavious (pp.165-197). London: Wiley, 1989

[17]. Scherer, K.R., A. Kappas, 1988: Primate vocal expression of affective state, in D. Todt, P. Goedeking, & D. Symmes (Eds.), Primate vocal communication (pp. 171-194). Berlin: Springer

[18]. Breazeal, C., 2001. Designing Social Robots. MIT Press, Cambridge, MA.

[19]. Abelin, A., Allwood, J., 2000. Cross-linguistic interpretation of emotional prosody. In: Proceedings of the ISCA Workshop on Speech and Emotion.

[20]. Tickle, A., 2000. Englishand Japanese speaker's emotion vocalizations and recognition: a comparison highlighting vowel quality. ISCA Workshop on Speech and Emotion, Belfast, 2000.

[21]. Polzin, T., Waibel, A., Emotion-Sensitive Human-Computer Interfaces, Proceedings of the ISCA workshop on Speech and Emotion, pp. 201~206, 2000, Newcastle, Northern Ireland.

[22]. Slaney, M., Mcroberts, G., Baby Ears: A Recognition System for Affective Vocalizations, Proceedings of the 1998 International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 12-15, 1998, Seattle, WA.

[23]. Ververidis, D. and Kotropoulos, C., Pitas, I., Automatic emotional speech classification, Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004), pp. 593 – 596, 2004, Montreal, Canada.

[24]. Ververidis, D., Kotropoulos, C., Automatic speech classification to five emotional states based on gender information, Proceedings of 12th European Signal Processing Conference, pp.341–344,

September 2004, Austria.

[25]. McGilloway, S., Cowie, R., Cowie, E. D., Gielen, S., Westerdijk, M., Stroeve, S. Approaching automatic recognition of emotion from voice: a rough benchmark, Proceedings of the ISCA workshop on Speech and Emotion, pp. 207-212, 2000, Newcastle, Northern Ireland.

[26]. Oudeyer P. Y., The production and recognition of emotions in speech: features and algorithms, International Journal of Human-Computer Studies, v.59 n.1-2, p.157-183, July 2003

[27]. Witten, I.H., & Frank, E., Data Mining: Practical machine learning tools and techniques with Java implementations, Morgan Kaufmann, San Francisco, CA, USA, 2000.

[28]. Brian C.J. Moore, An Introduction to the Psychology of hearing, Academic Press, 1997

[29]. Zipf, G. K., Human Behavior and the Principle of Least Effort. Addison-Wesley Press, 1949.

[30]. Cohen, A., Mantegna, R. N., Havlin, S., Numerical analysis of word frequencies in artificial and natural language texts, Fractals, vol. 5, no. 1, pp. 95–104, 1997.

[31]. Havlin, S., The distance between Zipf Plots, Physica A216, pp. 148–150, 1995.

[32]. Dellandrea, E., Makris, P., Vincent, N., Zipf Analysis of Audio Signals, Fractals, World Scientific Publishing Company, vol. 12(1), p. 73-85, 2004.

[33]. PRAAT, a system for doing phonetics by computer. Glot International 5(9/10), 341-345, 2001

[34]. Atal, B., Rabiner, L., A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition, IEEE Transactions on ASSP, Vol-24, Issue 3 , Jun 1976, Pages: 201 - 212.

[35].Childers, D. G., Hand, M., Larar, J. M., Silent and Voiced/Unvoied/ Mixed Excitation(Four-Way), Classification of Speech, IEEE Transaction on ASSP, Vol-37, No-11, pp. 1771-74, Nov 1989.

[36]. Bellman, R., Adaptive Contrôle Processes : Aguided tou, Princeton University Press, 1961

[37]. Spence, C., Sajda, P.,The role of feature selection in building pattern recognizers for computer-aided diagnosis, Proceedings of SPIE -- Volume 3338, Medical Imaging 1998: Image Processing, Kenneth M. Hanson, Editor, pp. 1434-1441, June 1998.

[38]. Harb, H., Chen, L., “voice-based Gender Identification in multimedia applications”, Journal of intelligent information systems, J. Intell. Inf. Syst. 24, vol 24(2), 2005, pp.179-198

[39]. Xiao, Z., Dellandrea, E., Dou, W., Chen, L., Two-stage Classification of Emotional

Speech, International Conference on Digital Telecommunications (ICDT'06), p. 32-37, August 29 - 31, 2006, Cap Esterel, Côte d'Azur, France.

[40]. Xiao, Z., Dellandrea, E., Dou, W., Chen, L., Features extraction and selection in emotional speech, International Conference on Advanced Video and Signal based Surveillance (AVSS 2005). p. 411-416. September 2005, Como, Italy.

[41]. Engberg, I. S., Hansen, A. V., Documentation of the Danish Emotional Speech Database DES, Aalborg September 1996

[42].Xiao, Z., Dellandréa, E., Dou, W., Chen, L., Hierarchical Classification of Emotional Speech, research report RR-LIRIS-2007-006, LIRIS UMR 5205 CNRS, 2007

[43]. Rakotomalala, R., « TANAGRA : un logiciel gratuit pour l'enseignement et la recherche", in Actes de EGC'2005, RNTI-E-3, vol. 2, pp.697-702, 2005

[44]. Xiao, Z., Dellandrea, E., Dou, W., Chen, L., Automatic Hierarchical Classification of Emotional Speech, accepted by MIPR 2007, Taichung, Taiwan, R.O.C., December 10-12, 2007