Hierarchical Classification of Emotional Speech

Zhongzhe Xiao[1], Emmanuel Dellandrea[1], Weibei Dou[2], Liming Chen[1]
[1]LIRIS Laboratory (UMR 5205), Ecole Centrale de Lyon, Department of Mathematic and Computer Science, 36 avenue Guy de Collongue, 69134 Ecully Cedex, France
Email: {zhongzhe.xiao, emmanuel.dellandrea, liming.chen}@ec-lyon.fr
[2]Department of Electronic Engineering, Tsinghua University, Beijing, 100084, P.R.China
Email: douwb@mail.tsinghua.edu.cn

***Abstract*—Speech emotion as anger, boredom, fear, gladness, *etc*. is high semantic information and its automatic analysis may have many applications such as smart human-computer interactions or multimedia indexing. Main difficulties for an efficient speech emotion classification reside in complex emotional class borders leading to necessity of appropriate audio feature selection. While current work in the literature only relies on classical frequency and energy based features and make use of a global classifier with a identical feature set for different emotion classes, we propose in this paper some new harmonic and Zipf based features for better emotion class characterization and a hierarchical classification scheme as we discovered that different emotional classes need different feature set for a better discrimination. Experimented on Berlin dataset [11] with 68 features, our emotion classifier reaches a classification rate of 76.22% and up to 79.47% when a first gender classification is applied, whereas current works in the literature usually display, as far as we know, a classification rate from 55% to 70%.**

***Index Terms*— emotional speech, harmonic feature, Zipf feature, hierarchical classification**

I. INTRODUCTION

The emotional elements play an important role in the vocal expressions. The automatic classification of emotional speech has several potential applications, such as human-computer interactions, automatic searching in films and TV programs, emotion analysis, *etc*.

As a pattern recognition problem, the main difficulties reside first in the determination of an appropriate discriminant feature set and then the selection of an efficient classifier. For an efficient classification of emotional speech, the most important aspects that we need to studier concern definition of emotion classes, the source of the speech samples for learning, the features adopted in classification and the type of classifiers.

A. Related works

There is not a universal definition of speech emotions in the literature. The number of emotion classes considered varies from 3 classes to more classes allowing a more detailed classification [2] – [6]. All these work can be compared according to several criteria, including the number and type of emotional classes, learning and classifier complexities and classification accuracy on a significant dataset.

Some researchers adopted 3 classes of emotional states to make a clear distinction, especially for some special purposes. In [2], Polzin and Waibel dealt with emotion-sensitive human-computer interfaces. The speech segments were chosen from English movies. They modeled the speech segments with verbal and non-verbal information; the

latter includes prosody features and spectral features. The classification according to 3 classes as sad, anger and neutral are presented and discussed. An accuracy up to 64% was achieved in their work on a significant dataset. According to their experiments, this classification accuracy is quite close to human classification accuracy. One of the originality of 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 they only consider three emotion classes which are furthermore limited to negative emotions.

Slaney and Mcroberts also studied three emotion classes problem in [3] but with another context considering 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. 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 a rate up to 67% correct, and the male utterances were classified correctly with a rate of 57%. Their experiment tends to show that their emotion classification is independent of language 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 emotion classification. However, their three emotion classes are quite specific and very different from the ones usually considered in the literature and in most of applications.

Gender information is also considered by Dimitrios et al [5], [6] with more emotion classes. In their work, 500 speech segments from DES (Danish Emotional Speech) database are used. Speech was expressed in 5 emotional classes, such as anger, happiness, neutral, sadness and surprise. A classical feature set of 87 statistical features of pitch, spectrum and energy was tested, and the feature selection method SFS (Sequential Forward Selection) was used. In [5], 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 [6], 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.

Prior to the work of Dimitrios et al, McGilloway et al [4] also studied 5 emotion classification problem with the 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 is around 55% with a Gaussian SVM when 90% of data were used as training data and 10% as testing data.

These works in considering five emotion types are among pioneer tentative on more realistic emotional speech classification. Their experiments show that classical pitch, frequency and energy based features are quite useful for emotion classification. However, their best classification accuracy rate around 60% tends to suggest that these audio features are not enough for further improve the classification results.

B. Our approach

As we can see, current works in the literature usually make use of frequency and energy based features [4, 5, 6]. Unfortunately, our preliminary experiments shows that they are not enough efficient for discriminating emotion classes, especially for anger and gladness. Moreover, a global classifier with a same feature set generally is applied [2-6] while a detailed analysis on a set of audio emotional samples shows that a same set of audio features can not discriminate efficiently all the emotional classes at the same time and different emotion classes may need different features for a better classification.

In our work, we propose to discriminate emotional aspect of a speech into finer emotional classes and consider 6 emotion classes, including anger, boredom, fear, gladness, neutral and sadness. Further to classical frequency and energy based features, we propose to make use of additional features allowing taking into account other information

contained in speech signals: harmonic features which are perceptual features, and Zipf features which characterize the inner structure of signals. Moreover, as our preliminary experiments evidenced that different emotions may be better characterized by different features, we propose in our work, instead of a single global classifier, a hierarchical classification scheme, combining several classifiers and gender discrimination, to make more accurate the classification.

Experimented on Berlin dataset with 68 features, our emotion classifier reaches a classification rate of 76.22% and up to 79.47% when a first gender classification is applied, whereas current works in the literature usually display, as far as we know, a classification rate from 55% to 70%.

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

II. FEATURE SET

A feature set of 68 acoustic features are used in our work to present the emotions for classification. These features are divided into 4 groups: frequency features, energy features, harmonic features and Zipf features. Among these 4 groups, the first 2 groups are classical features which are often used in the literature, and the latter 2 groups are new features proposed in our work. In this section, we mainly introduce the 2 groups of new features: harmonic features and Zipf features.

A. Harmonic features

The classical features according to the pitch and energy are often used in speech analyzing. Further to these types of features which can be ordered on a single scale, we felt in our preliminary experiments that we needed some additional perceptual features according to the harmonics, which describe the timbre patterns and show the energy pattern as a function of the frequency.

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 can not be presented 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.[7]

In our work, a description of sub-band amplitude modulation of the signal is proposed to present the harmonic distributions. By experiments, we found that the emotions can still be clearly recognized by human ears when only the first 15 harmonics of the speech signal are kept. So the first 15 harmonics are considered in extracting the harmonic features.

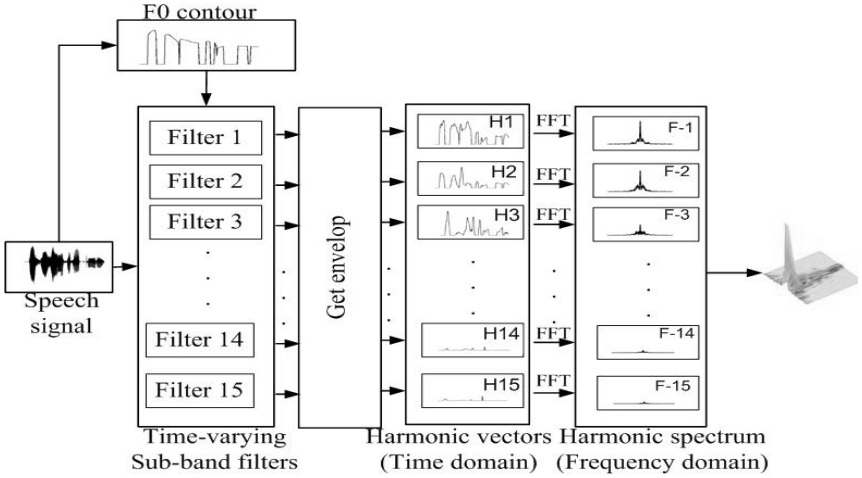

Fig. 1 Harmonic analysis of a speech signal

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 ith sub-band filter at a certain time is ith multiples of the fundamental frequency (ith harmonic) at that time, and the bandwidth is half of 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 (H1, H2, H3…in Fig. 1 and Fig. 3 (a)). That is to say, we use the sum of the 15 amplitude modulated signals using the harmonics as carriers to present the speech signal. Then FFT is applied to the harmonic signals to get the spectrum of the modulation signals (F-1, F-2, F-3…in Fig. 1 and Fig. 3 (b)). We combined these 15 spectrums together into a 3-D harmonic space, as shown in Fig. 3 (c).

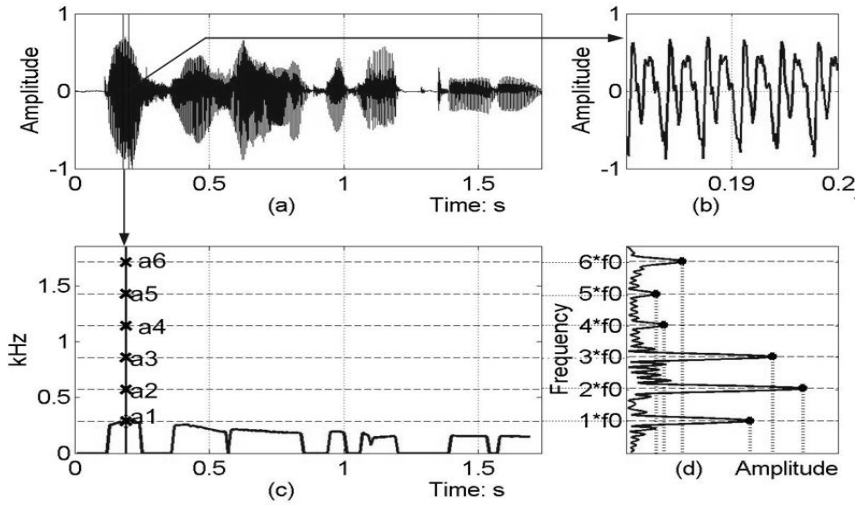

Fig. 2 Calculation process of the harmonic features: (a) waveform in time domain, (b) Zoom out of (a) during 20ms, (c) F0 contour of (a), a1 – a6 are the frequency points of 1 to 6 multiples of the fundamental frequency at the selected time point (d) spectrum of selected time point, the amplitude at 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. 2.

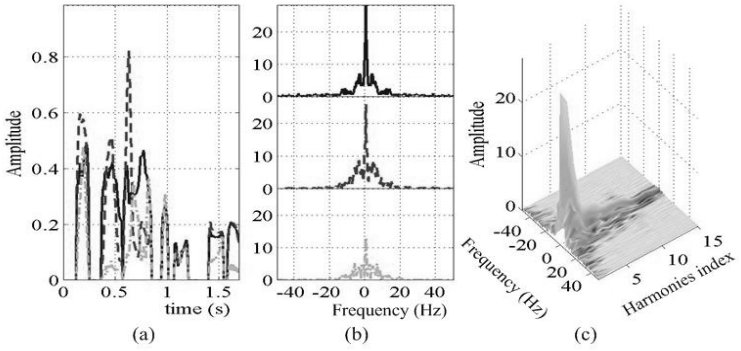

Fig. 3 The amplitude of the harmonic vectors in time domain and their spectrums (a) amplitude of the first 3 harmonic vectors, (b) the spectrums of the first 3 vectors, (c) 15 spectrums combined in 3-D harmonic space

The 3 axes in the 3-D harmonic space are amplitude, frequency and harmonics index Fig. 3 (c). In these 3 axes, both the frequency axis and the harmonics index axis present 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. The values in the side part of this space are relatively low.

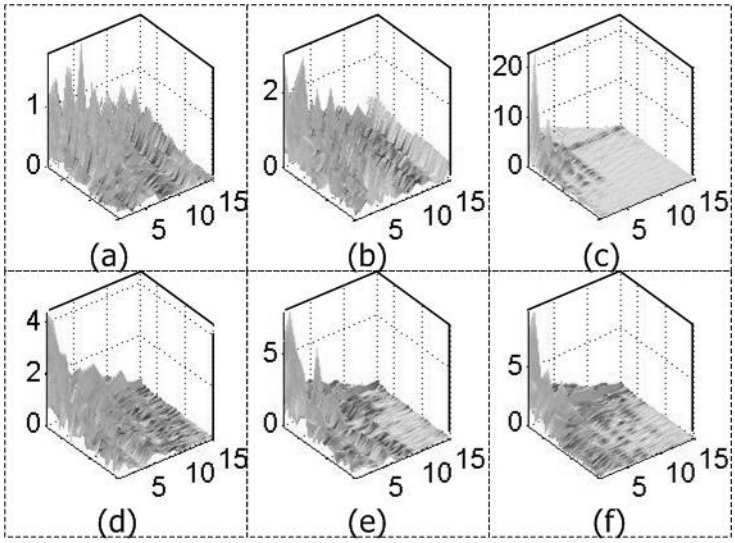

Fig. 4 3-D harmonic space for the 6 emotions from a same sentence: (a) anger, (b) fear, (c) sadness, (d) gladness, (e) neutral, (f) boredom

As the spectrum is symmetric due to FFT properties, we only keep the positive frequency part. Fig. 4 shows the 3-D harmonic space of examples of the 6 emotions from speech samples with a same sentence. The axes in Fig. 4 are the same as in Fig. 3(c). This harmonic space shows obvious difference among the emotions. For example, the emotion 'anger' has relative low main peak and many small peaks in the side parts, while the 'sadness' and the 'boredom' have high main peaks but are quite flat in the side part.

From the difference in the harmonic space among the emotions, we divide the harmonic space into 4 areas as shown in Fig. 5. The ridge, which shows the low frequency part in the frequency domain, is selected as area 1; the other part is divided into 3 areas according to the index of harmonics. 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 latter 3 areas present relatively low frequency part, middle frequency part and high frequency part according to the harmonics index respectively. The mean value, variance value of each area and the value ratios between the areas are used as features to be selected.

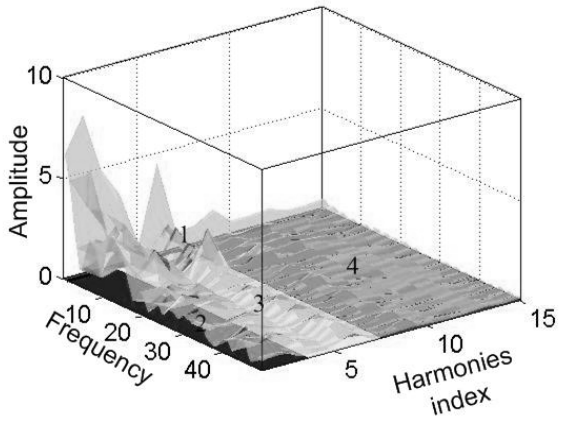

Fig. 5 4 areas for FFT result of 3-D harmonic space

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. Zipf law is an empirical law proposed by G. K. Zipf [8]. 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:

$$f(p) = \alpha r(p)^{-\beta} \qquad \text{(Eq. 1)}$$

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:

$I(f) = \delta f^{-\gamma}$ (Eq. 2)

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. 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 [9]. Three types of coding as temporal coding, frequencial coding and time-scale coding were proposed in [9], in order to bring to the front different informations contained in signals. 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. 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 [10]. The F0 and the formants are only computed through the vowels periods. 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. 6 (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. 6 (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. 6 (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.

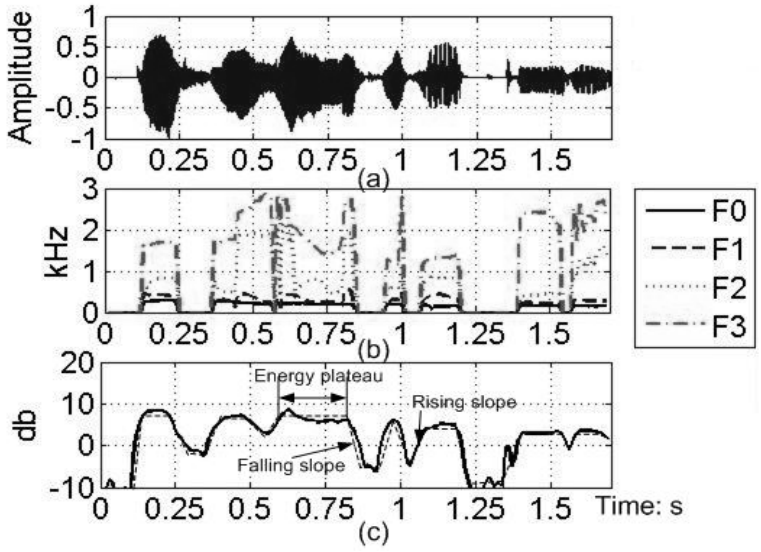

Fig. 6 Basic acoustic features of a speech signal: (a) waveform, (b) fundamental frequency F0 and the first 3 formants (F1, F2, F3), (c) energy contour

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

III. CLASSIFICATION OF EMOTIONAL SPEECH

A hierarchical classification method dealing with the emotional classification in several steps is proposed in this section. A two-stage classification according to the emotion types and a gender classification are combined together to decrease the perturbations between the different emotions. The feature selection is done at each step of the classifier.

A. Feature selection

The number of features used in the classification has to be reduced to simplify the computation procedure and to decrease the interference among the features to get the best performance in classification. Also, as the hierarchical classification is used in our work, different feature subsets are needed for the sub-classifiers.

The feature selection methods can be characterized, by the dependence to the classifiers, into 2 main categories: filter approaches and wrapper approaches. Filter methods normally evaluate the statistical performance of the features over the data without considering the proper classifiers. 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 the accuracy rate.

Filter methods are often efficient, but the performance may be relatively low. For example, the PCA (principal component analysis) is too sensitive to the outliers in the data. Among the wrapper methods, there are some algorithms which have very high performance, for example, the GA (Genetic Algorithm) methods. They make use of the

principles of Darwinian natural selection and biologically inspired operations [12]. Their principle is to maintain during several generations a constant-size population of individuals characterized by their chromosome. Chromosomes are represented by l-long gene strings so that each gene is associated with a feature of the original parameter space. The value of the gene indicates the inclusion (if its value is 1) or the exclusion (if its value is 0) of the associated feature in the new space. The evolution of the population is governed by three rules: selection, crossover, and mutation. GA are very efficient by their major drawback is that it is too computational costly. To make a good compromise between speed and performance, the SFS algorithm is used in our work for the feature selection of each step [13].

SFS begins with an empty subset. 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. For each feature,

$$F_n = \frac{F_{n0} - \min(F_{n0})}{\max(F_{n0}) - \min(F_{n0})} \qquad \text{(Eq.3)}$$

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. Two-stage classification of emotional speech

While different classifiers such as Bayes classifiers, Gaussian mixture model etc. are adopted in the literature, global classifiers are most of time used over all the emotional classes with the same feature subset.

Unfortunately, traditional method with a global classifier leads to unnecessary confusions. A feature with good discriminability to a certain pair of emotional classes may be a feature with high confusion to another pair of emotional classes. Thus the classification performance can be improved by combining several classifiers with their own feature subsets.

On the other hand, some emotional classes, such as anger and gladness, have some similarities with certain features. Thus, we can combine some emotional classes to reduce the total number of the classes as a preparation to simplify the problem. This discovery has led us to propose a two-stage classification of emotional speech with a tree structure [14], as presented in Fig. 7. As we can see on the figure, speech signal is first divided into two intermediate emotional classes: active and not active. Further, speech samples labeled as active class are categorized as terminal emotional classes, i.e. anger and gladness classes. It is much the same for speech signals labeled as not active class. They are first categorized as median and passive classes, and then as fear, neutral, sadness and boredom.

Neural networks have been chosen for their abilities of discriminating non linear data and generalization. In our work, we make use of BP (Back Propagation) neural networks with 2 hidden layers. The transfer function is the log-sigmoid 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.

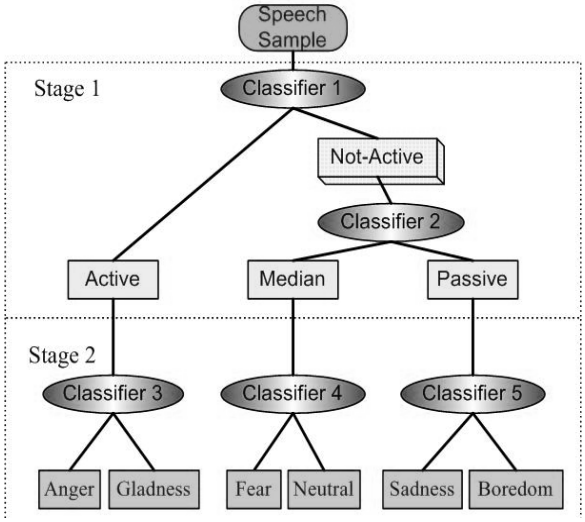

Fig. 7 Two-stage classification of emotional speech

1) Stage 1: 3-state classification

To simplify the classification, a three-state description of emotions is proposed as the first stage [15]. According to the sense of activity, the emotions in speech can be divided into 3 types, or 3 states: active state (including anger and happy), median state (including fear and neutral), and passive state (including sadness and boredom).

By the process of pre-classification, which is a one-step classifier into the 3 states, we found that the active state has a more distinct characteristic with median and passive states. In order to get a better separation between the median state and the passive state, 2 steps are adapted in this stage. In the first step, the active state is separated from the median and the passive states (classifier 1 in Fig. 7); and in the second step, the median state and the passive state are separated (classifier 2 in Fig. 7).

2) Stage 2: Further classification into 6 emotional classes

Although the three-state emotions description of emotional speech is simple and clear, it has its own drawbacks. To view this description in the opinion of the popular dimensional emotions [16], the three-state description has only one dimension, that is, activity dimension, or power dimension. In this way, it is not enough to present the characteristics of the emotions, and some obvious differences between emotions of the same state can be erased by mistake. Therefore, a further classification is needed in the second stage.

Similar classifiers as those proposed in stage 1 are used in this stage. According to Fig. 7, classifier 3 is used for the active state, separating the “anger” and the “gladness”, classifier 4 is used for the median state, separating the “fear” and the “neutral”, and classifier 5 is used for the passive state, separating the “sadness” and the “boredom”.

C. Hierarchical classification of emotional speech

Except the perturbations caused by too many emotional classes, we also found that gender difference in the acoustic features also influences the emotion classification. A new stage as gender classification can be used to allow different models being used for the speech samples according to the gender.

Our hierarchical (three-stage) classification scheme is presented in Fig. 8. This scheme includes two parts: the first part is the male/female discrimination, and the second part is the two-stage classifier as presented in previous section.

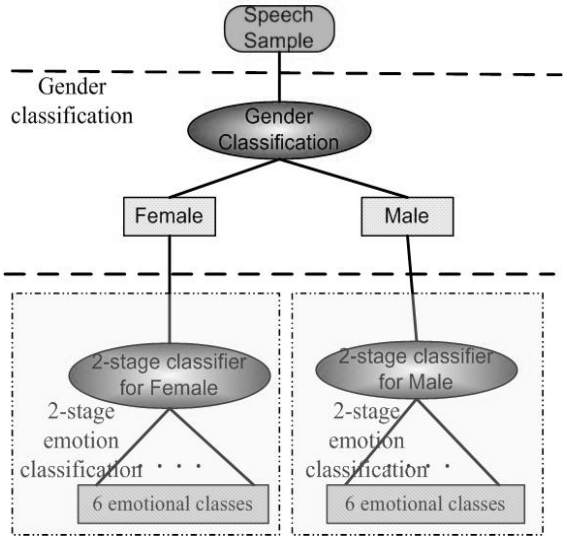

Fig. 8 Hierarchical classification with gender information

In the second part, the two two-stage classifiers have the same structure (as shown in Fig. 7), and work with different parameter set according to the genders that we present in the following section.

IV. EXPERIMENTS AND RESULTS

In order to validate our work, we carried out experiments on the Berlin dataset. In the following, we discuss first the quality of database and describe Berlin dataset. Then, our experimental results are presented.

A. Database

The soundness of any work for automatic classification should be validated on significant dataset. Unfortunately, it is quite difficult to collect labeled emotional speech samples. Generally, there are 3 major categories of emotional speech samples. They are natural vocal expression, induced emotional expression, and simulated emotional expression [1]. 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 of getting speech samples is the simulated emotional expression, that is, to ask 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 DES (Danish Emotional Speech) database and Berlin database. According to [3], even the babies have the ability to recognize the emotions in the speech; we assume that the emotions can be recognized whatever the languages.

The database used in this paper is Berlin emotional speech database. It is developed by Professor Sendlmeier and his fellows in Department of Communication Science, Institute for Speech and Communication, Berlin Technical University [11]. 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 female speech samples, the male speech samples and the combination of all the samples are tested separately. The influence of gender information to the classification result is also discussed in this paper.

B. Experimental results

In our experiment, the data in each case were 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, the type is omitted in training and testing.

1) The two-stage classification for each gender

For the two-stage classification (Fig. 7), the speech samples from the two genders are trained and tested separately. The mixed samples of the both genders are also tested as the third case. The neural networks are used as sub-classifiers 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 I.

The probability distributions of the most frequently selected features (features 65, 24 and 4, ordered according to the frequency selected) are plotted in Fig. 9 to Fig. 11. The gray lines and the black lines present the former and latter classes in the subplots respectively. The solid lines present the result for mixed genders, the dashed lines present the result for female samples, and the point-dashed lines present the result for male samples. We can notice that feature 65 shows very high discriminability in stage 1 (separating the 3 states), but is quite confusable in stage 2; feature 24 shows good discriminability except for classifier 3 "anger" vs. "gladness"; feature 4 shows high discriminability in every sub-classifiers, but also shows obvious difference between the two genders.

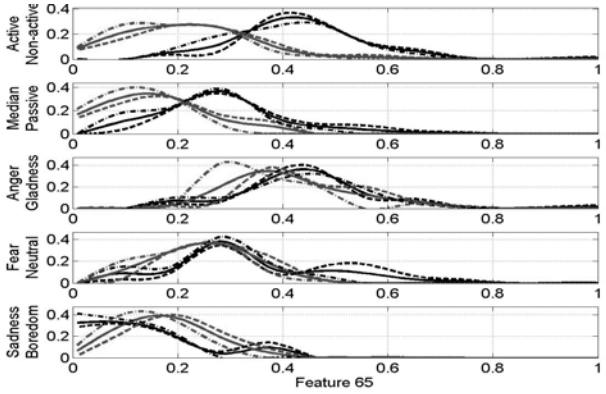

Fig. 9 Pdfs of normalized features 65 for the sub-classifiers in the two-stage classification

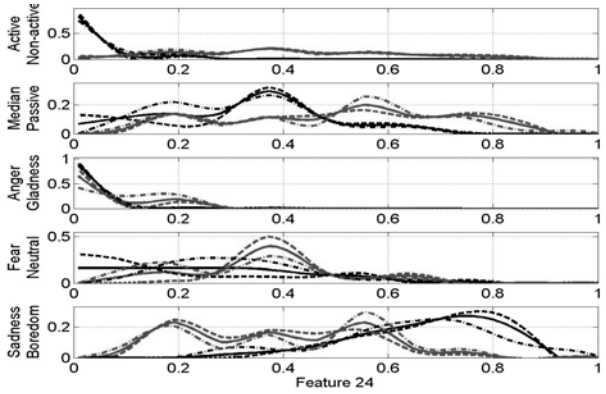

Fig. 10 Pdfs of normalized features 24 for the sub-classifiers

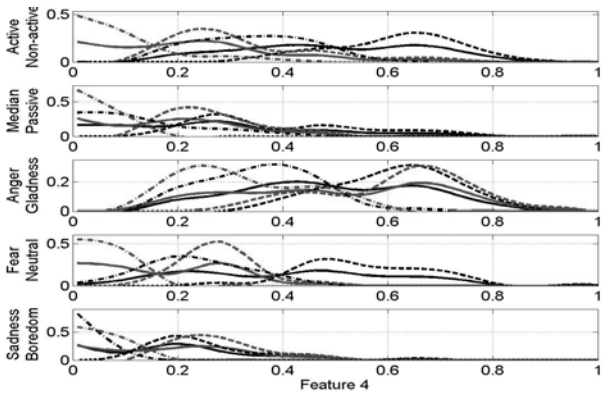

Fig. 11 Pdfs of normalized features 4 for the sub-classifiers

From Table I, we can see that features from feature group 1 (frequency features, see section 4) and feature group 2 (energy features) present normal performance in the 5 sub-classifiers, while feature group 1 is more efficient in classifier 3 ("anger" vs. "gladness") and classifier 5 ("sadness" vs. "boredom"); features from feature group 3 (harmonic features) are selected most frequently in all the 5 sub-classifiers, especially dominate the feature subsets for classifier 2 ("median" and "passive"). Although there are only 2 features in feature group 4 (Zipf features), they show great importance in the feature subset for classifier 1 ("active" vs. "non-active").

These results prove that it is important to select features according to the type of classifiers considered.

The confusion matrixes for the two-stage classification for the two genders and the mixed case are listed in Table II.

The weighted average recognition rate according to the number of speech samples for female samples and male samples is 81.75%, which is and 5.53% higher than the result for mixed speech samples (76.22%). From Table II, we can see that the mixing of the gender cause more misjudgment for the emotion "fear" than for the other emotions.

2) The gender classification

A similar classifier using neural network with SFS feature selection is used for the gender classification. The selected feature subset contains 15 features: 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.

3) The final result for hierarchical classification

The confusion matrix of the hierarchical classification is listed in Table III. The recognition rate (79.47%) is 3.25 higher than the result for the mixed samples with only the two-stage classification (76.22%).

4) Comparison with single global classifier

In order to verify the improvement of the hierarchical scheme, we also made experiments using the traditional global classifier with the same SFS and neural network in the following cases: the two single genders respectively, the mixed data with / without gender classification. The confusion matrixes of these four cases are listed in Table IV. Comparing with Table II and Table III, the recognition rates in the global classifiers are lower than the hierarchical scheme by up to 3% which shows that it is important to decrease a complex global classification problem into smaller classification problems. The overall recognition rates for all the cases are listed in Table V. For both global classifier and classifier with the two-stage classifier, the recognition results for the mixed samples are lower than the weighted average result of the 2 genders, and with the gender classification, the degradation can be reduced. And for both cases with/without gender classification, the classification with two-stage classifier shows better performance than global classifier.

V. CONCLUSIONS AND FUTURE WORK

We studied 68 features from 4 feature groups for emotional speech classification, and a hierarchical classification with gender information has been proposed to improve the recognition rate, combining a gender classification stage and a two-stage emotional classification according to the model of each gender. From these experimental results, we may draw the following lessons:

First, with the architecture of the hierarchical classification by using several two-class classifiers, the disturbance between the classes can be decreased and the recognition rate can be increased.

Secondly, the different groups of features show different importance in the different steps of the two-stage classification. For the two-stage classification, the feature group 3 (harmonic features) has higher discriminability than the other 3 groups, while the feature groups 1 (frequency features) and 2 (energy features) are more important for stage 2, and the feature groups 3 (harmonic features) and 4 (Zipf features) are more important for stage 1.

Thirdly, the gender classification before the 2-stage classifiers can also improve the recognition rate for certain emotions, especially, for the most confusable emotion type for the mixed samples, "fear".

In our future work, we envisage to further validate our approach in considering other datasets and assess the generality of our work go by considering also music signals with the same type of classification system as we have shown that it is particularly efficient in the case of speech signals for multi-classes problems. Besides, for the time being the classifier architecture is fired empirically. To be more adoptive to other classification problems, the architecture should be derived automatically by using for example confusion matrixes.

REFERENCES

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

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

[3]. M Slaney, G Mcroberts, 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.

[4]. S McGilloway, R Cowie, Douglas-Cowie, E., Gielen, C.C.A.M., Westerdijk, M.J.D., & Stroeve, S.H. 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.

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

[6]. D. Ververidis and C. Kotropoulos, Automatic speech classification to five emotional states based on gender information, Proceedings of 12th European Signal Processing Conference, pp.341–344, September 2004, Austria.

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

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

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

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

[11]. Sendlmeier et al., Berlin emotional speech database, available online at http://www.expressive-speech.net/

[12]. http://www.genetic-programming.com/

[13]. Clay Spence, Paul Sajda, 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.

[14]. Zhongzhe Xiao, Emmanuel Dellandrea, Weibei Dou, Liming Chen, 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.

[15]. Zhongzhe Xiao, Emanuel Dellandréa, Weibei Dou, Liming Chen., 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.

[16]. Cécile Pereira, Dimensions of emotional meaning in speech, Proceedings of the ISCA workshop on Speech and Emotion pp. 25-28, 2000, Newcastle, Northern Ireland.

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

		Selected feature subset (Ordered by the sequence of selection)	Recognition rate
Active vs. non-active	Female samples	67^4, 65^3, 25^2, 61^3, 26^2, 51^3, 21^2, 53^3, 28^2	93.95%
	Male samples	24^2, 4^1, 9^1, 19^1, 52^3, 51^3, 17^1, 65^3, 67^4, 12^1	95.18%
	All samples	56^3, 68^4, 25^2, 1^1, 14^1, 26^2, 28^2, 29^2, 42^2, 5^1, 65^3, 27^2	93.10%
Median vs. Passive	Female samples	65^3, 4^1, 27^2, 26^2, 57^3, 53^3, 66^3, 28^2, 56^3, 51^3, 1^1, 24^2	88.52%
	Male samples	66^3, 67^4, 9^1, 56^3, 61^3, 54^3, 53^3, 5^1, 21^2, 26^2, 57^3	91.91%
	All samples	66^3, 28^2, 27^2, 57^3, 65^3, 53^3, 26^2, 32^2	88.26%
Anger vs. Gladness	Female samples	6^1, 7^1, 64^3, 4^1, 53^3, 32^2, 57^3, 24^2	83.55%
	Male samples	24^2, 33^2, 65^3, 9^1, 39^2, 60^3, 28^2, 2^1, 14^1, 18^1	88.93%
	All samples	68^4, 31^2, 18^1, 9^1, 13^1, 53^3, 56^3, 58^3, 65^3, 34^2	83.98%
Fear vs. Neutral	Female samples	4^1, 52^3, 37^2, 9^1, 48^2	94.64%
	Male samples	64^3, 60^3, 37^2, 53^3, 57^3, 44^2, 51^3	96.72%
	All samples	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	87.82%
Sadness vs. Boredom	Female samples	5^1, 67^4, 8^1, 24^2, 19^2, 9^1, 48^2, 16^1, 2^1, 46^2, 65^3, 55^3, 13^1, 56^3	96.75%
	Male samples	59^3, 50^2, 20^1, 22^2, 62^3, 48^2, 60^3, 58^3	95.10%
	All samples	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	92.98%

Table II. Confusion matrix of the 2-stage classification

	Predicted / Actual	Anger	Gladness	Fear	Neutral	Sadness	Boredom
Female samples	Anger	83.43%	12.09%	2.09%	1.94%	0.00%	0.45%
	Gladness	17.75%	69.00%	7.25%	2.00%	0.00%	4.00%
	Fear	7.24%	10.00%	73.45%	4.14%	2.41%	2.76%
	Neutral	0.50%	3.00%	3.50%	75.75%	1.00%	16.25%
	Sadness	0.00%	0.00%	0.86%	5.14%	91.43%	2.57%
	Boredom	1.33%	0.44%	0.22%	12.89%	2.00%	83.11%
Male samples	Anger	89.33%	7.33%	1.67%	1.00%	0.00%	0.67%
	Gladness	17.50%	65.00%	12.50%	3.33%	0.00%	1.67%
	Fear	0.38%	8.46%	78.85%	3. 85%	5.00%	3.46%
	Neutral	0.79%	1.05%	2.11%	83.68%	2.89%	9.47%
	Sadness	0.00%	0.00%	0.00%	2.35%	92.94%	4.71%
	Boredom	0.29%	0.88%	0.00%	5.88%	4.12%	88.82%
Mixed samples	Anger	85.35%	9.45%	2.20%	2.68%	0.00%	0.31%
	Gladness	23.91%	61.88%	9.22%	4.69%	0.00%	0.31%
	Fear	11.27%	11.27%	55.27%	14.00%	3.64%	4.55%
	Neutral	0.90%	1.15%	5.64%	78.33%	2.44%	11.54%
	Sadness	0.77%	1.15%	0.96%	6.15%	82.31%	8.65%
	Boredom	0.76%	0.13%	1.27%	12.15%	4.05%	81.65%

Table III. Confusion matrix of the hierarchical classification

Predicted / Actual	Anger	Gladness	Fear	Neutral	Sadness	Boredom
Anger	85.35%	10.63%	1.73%	1.73%	0.00%	0.55%
Gladness	20.62%	63.28%	10.78%	2.50%	0.00%	2.81%
Fear	4.91%	9.64%	74.18%	4.18%	3.64%	3.45%
Neutral	0.64%	2.95%	3.97%	77.56%	2.05%	12.82%

Sadness	0.00%	0.00%	3.46%	6.92%	86.35%	3.27%
Boredom	1.65%	1.01%	0.25%	9.62%	3.29%	84.18%

Table IV.Confusion matrix of the global classifier

	Predicted / Actual	Anger	Gladness	Fear	Neutral	Sadness	Boredom
Female samples	Anger	85.37%	10.75%	2.84%	0.60%	0.00%	0.45%
	Gladness	26.75%	65.25%	6.50%	1.00%	0.00%	0.50%
	Fear	10.69%	6.55%	68.62%	7.24%	1.38%	4.48%
	Neutral	4.25%	1.50%	1.50%	79.25%	3.75%	9.50%
	Sadness	0.00%	0.86%	0.86%	5.71%	84.86%	6.29%
	Boredom	2.89%	1.11%	3.11%	15.56%	6.00%	70.89%
Male samples	Anger	91.50%	7.33%	1.00%	0.17%	0.00%	0.00%
	Gladness	22.50%	67.92%	5.83%	2.50%	0.00%	0.00%
	Fear	6.54%	7.69%	77.69%	3.85%	0.38%	3.46%
	Neutral	2.63%	1.05%	2.89%	80.53%	4.21%	8.68%
	Sadness	0.00%	0.00%	1.76%	10.00%	82.35%	5.29%
	Boredom	0.00%	0.29%	2.35%	12.94%	6.47%	77.35%
Mixed samples without gender information	Anger	88.19%	8.43%	2.13%	1.02%	0.00%	0.08%
	Gladness	25.78%	63.28%	6.56%	1.41%	0.47%	0.78%
	Fear	12.91%	8.91%	54.36%	8.91%	2.73%	11.09%
	Neutral	4.49%	1.28%	3.85%	68.72%	3.85%	17.82%
	Sadness	0.00%	0.77%	0.77%	8.65%	83.46%	5.77%
	Boredom	1.27%	0.63%	3.16%	12.66%	3.04%	78.99%
Mixed samples with gender information	Anger	87.80%	9.61%	2.05%	0. 31%	0.00%	0.24%
	Gladness	27.19%	62.97%	6.88%	1.72%	0.16%	0.78%
	Fear	9.09%	7.09%	71.09%	6.36%	1.45%	4.18%
	Neutral	3.46%	1.54%	4.36%	76.92%	4.10%	9.49%
	Sadness	0.77%	1.92%	2.69%	6.35%	81.35%	5.00%
	Boredom	1.77%	1.39%	3.54%	14.56%	6.20%	72.03%

Table V. Recognition rates for each case of emotional classification

	Male	Female	Average of the 2 genders	Mixed	Mixed with gender classification
Global	81.56%	76.76%	78.86%	75.12%	76.95%
2-stage	84.17%	79.88%	81.75%	76.22%	--
3-stage (Hierarchical)	--	--	--	--	79.47%