ANNEE: 2008

THESE

POUR OBTENIR LE GRADE DE

DOCTEUR DE L'ECOLE CENTRALE DE LYON

DISCIPLINE : INFORMATIQUE

Présentée et soutenue publiquement par

Zhongzhe XIAO

Recognition of Emotions in Audio Signals

DIRECTEUR DE THESE: **Liming CHEN**

Ecole Doctorale Informatique et Information pour la Société (EDIIS)

JURY

Mme. ANDRE-OBRECHT Régine	Professeur, Université Paul Sabatier	Rapporteur
M. BESACIER Laurent	Maître de Conférences (HDR), Université Joseph Fourier	Rapporteur
M. HATON Jean-Paul	Professeur, Université Henri Poincaré	Examinateur
M. CHEN Liming	Professeur, Ecole Centrale de Lyon	Directeur de thèse
M. DELLANDREA Emmanuel	Maître de Conférences, Ecole Centrale de Lyon	Co-directeur
Mme. DOU Weibei	Professeur, Université de Tsinghua	Co-directrice

Numéro d'ordre : 2008-02

Abstract

This Ph.D thesis work is dedicated to automatic emotion/mood recognition in audio signals. Indeed, audio emotion is high semantic information and its automatic analysis may have many applications such as smart human-computer interactions or multimedia indexing. The purpose of this thesis is thus to investigate machine-based audio emotion analysis solutions for both speech and music signals.

Our work makes use of a discrete emotional model combined with the dimensional one and relies upon existing studies on acoustics correlates of emotional speech and music mood. The key contributions are the following. First, we have proposed, in complement to popular frequency-based and energy-based features, some new audio features, namely harmonic and Zipf features, to better characterize timbre and prosodic properties of emotional speech. Second, as there exists very few emotional resources either for speech or music for machine learning as compared to audio features that one can extract, an evidence theory-based feature selection scheme named Embedded Sequential Forward Selection (ESFS) is proposed to deal with the classic "curse of dimensionality" problem and thus over-fitting. Third, using a manually built dimensional emotion model-based hierarchical classifier to deal with fuzzy borders of emotional states, we demonstrated that a hierarchical classification scheme performs better than single global classifier mostly used in the literature. Furthermore, as there does not exist any universal agreement on basic emotion definition and as emotional states are typically application dependent, we also proposed a ESFS-based algorithm for automatically building a hierarchical classification scheme (HCS) which is best adapted to a specific set of application dependent emotional states. The HCS divides a complex classification problem into simpler and smaller problems by combining several binary sub-classifiers in the structure of a binary tree in several stages, and gives the result as the type of emotional states of the audio samples. Finally, to deal with the subjective nature of emotions, we also proposed an evidence theory-based ambiguous classifier allowing multiple emotions labeling as human often does.

The effectiveness of all these recognition techniques was evaluated on Berlin and DES datasets for emotional speech recognition and on a music mood dataset that we collected in our laboratory as there exist no public dataset so far.

Keywords: audio signal, emotion classification, music mood analysis, audio features, feature selection, hierarchical classification, ambiguous classification, evidence theory.

Résumé

Les travaux de recherche réalisés dans le cadre de cette thèse de doctorat portent sur la reconnaissance automatique de l'émotion et de l'humeur au sein de signaux sonores. En effet, l'émotion portée par les signaux audio constitue une information sémantique particulièrement importante dont l'analyse automatique offre de nombreuses possibilités en termes d'applications, telles que les interactions homme-machine intelligentes et l'indexation multimédia. L'objectif de cette thèse est ainsi d'étudier des solutions informatique d'analyse de l'émotion audio tant pour la parole que pour les signaux musicaux.

Nous utilisons dans notre travail un modèle émotionnel discret combiné à un modèle dimensionnel, en nous appuyant sur des études existantes sur les corrélations entre les propriétés acoustiques et l'émotion dans la parole ainsi que l'humeur dans les signaux de musique. Les principales contributions de nos travaux sont les suivantes. Tout d'abord, nous avons proposé, en complément des caractéristiques audio basées sur les propriétés fréquentielles et d'énergie, de nouvelles caractéristiques harmoniques et Zipf, afin d'améliorer la caractérisation des propriétés des signaux de parole en terme de timbre et de prosodie. Deuxièmement, dans la mesure où très peu de ressources pour l'étude de l'émotion dans la parole et dans la musique sont disponibles par rapport au nombre important de caractéristiques audio qu'il est envisageable d'extraire, une méthode de sélection de caractéristiques nomméeESFS, basée sur la théorie de l'évidence est proposée afin de simplifier le modèle de classification et d'en améliorer les performances. De plus, nous avons montré que l'utilisation d'un classifieur hiérarchique basé sur un modèle dimensionnel de l'émotion, permet d'obtenir de meilleurs résultats de classification qu'un unique classifieur global, souvent utilisé dans la littérature. Par ailleurs, puisqu'il n'existe pas d'accord universel sur la définition des émotions de base, et parce que les états émotionnels considérés sont très dépendant des applications, nous avons également proposé un algorithme basés sur ESFS et permettant de construire automatiquement un classifieur hiérarchique adapté à un ensemble spécifique d'états émotionnels dans le cadre d'une application particulière. Cette classification hiérarchique procède en divisant un problème de classification complexe en un ensemble de problèmes plus petits et plus simples grâce à la combinaison d'un ensemble de sous-classifieurs binaires organisés sous forme d'un arbre binaire. Enfin, les émotions étant par nature des notions subjectives, nous avons également proposé un classifieur ambigu, basé sur la théorie de l'évidence, permettant l'association d'un signal audio à de multiples émotions, comme le font souvent les êtres humains.

L’efficacité de ces techniques de reconnaissance a été évaluée sur les ensembles de données Berlin et DES pour la reconnaissance de l’émotion dans la parole, et sur un ensemble de données construit au sein de notre laboratoire pour l’humeur dans les signaux de musique, dans la mesure où il n’existe pour le moment aucun jeu de données public.

Mots clés: signal audio, classification de l’emotion, analyse de l’humeur dans la musique, caractéristiques audio, sélection de caractéristiques, classification hiérarchique, classification ambiguë, théorie de l’évidence.

Acknowledgements

I would like to express my gratitude here to the many people who were helping me during my thesis work since 2004.

First, I wish to thank my supervisor Prof. Liming Chen for accepting me in his group for my thesis and supporting me during the whole thesis work.

I am so grateful to Prof. Weibei Dou, professor of Department of Electronic Engineering at Tsinghua University, for being my co-supervisor, and the initiator of my connection to France, and for all her valuable help. I would like to thank Dr. Emmanuel Dellandrea, my co-supervisor, for his patience and priceless advices during these years.

I would like to thank Dr. Aliaksandr Paradzinets for sharing his ideas and all his fruitful discussions in the thesis work. I also want to thank all my friends in the laboratory: Alain Pujol, Karima Ouji, Kun Peng, Chu Duc Nguyen, Yan Liu, Huanzhang Fu, and Xi Zhao. I thank the MI department members, Christian and Colette Vial, the secretaries Francoise Chatelin and Isabelle San-Jose, and all the others.

At the end, I want to thank my family, who are the most important people for me in this world: my parents Yongsheng Xiao and Yuqi Qiao, my husband Gengzhao Xu, and my sister Yingzhe Xiao, for their love and supporting.

Table of Contents

Chapter 1
Introduction

1.1 Research topic

Studies suggest that only 10% of human life is completely unemotional while the rest involves emotion of some sort [Emo]. A problem of interest in the emotions lays in the emotions of audio signals, including emotional speech and music mood. Indeed, the automatic recognition of emotional audio signals has great potential of applications, such as human-computer interactions (routing angry customers in a call-center automatically to a human operator [Dev05], or Embodied Conversational Agent (ECA)), automatic searching in films and TV programs, searching for speakers in a multimedia collection that discuss a certain topic in a certain emotional state, or selecting music works under the given mood in a music collection.

In this work, we focused on the problem of automatic classification and recognition of audio emotions with fast and accurate solutions mainly for emotional speech, but also for music mood, considering simulated speech expressions and classical music.

1.2 Problems and Objective

The interest in expressive speech can be traced back to the early Greek and Roman manuals on rhetoric that were the basis of the later theory of emotional appeal in western philosophy [Sch02]. In the 19th century, a new interest in the expression of emotion in face and voice was motivated by the emergence of modern evolutionary biology, particularly due to Darwin's research in 1872 on how animals and humans express and signal to others their emotions. Systematic research on the emotions started in the 1960s when psychiatrists renewed their interest in diagnosing affective states via vocal expression. Emotion psychologists, linguists, phoneticians, engineers

and phoneticians also took part into the research of audio emotions from various aspects from the 1970s and then, the automatic detection of the emotions began to come into interest in the last few years.

Although many research works dealing with the notion of emotion have been made, there is still no universal agreement on the basic definition of emotions. The two traditional theories on emotions are the discrete and the dimensional emotion theories. For the discrete emotion theory, researchers propose different numbers and different types of emotions. The term "big six" gained attention implying the existence of a fundamental set of six basic emotions while there does not seem to be any agreement on which six these should be [Sch02]. When the terms of the emotions are applied to the music, they should be modified to fit the longer lasting emotion states as moods. In the dimensional emotion approach, different emotional states are mapped into a two or three-dimensional space [Per00] [Sch00c] [Sch02] [Tha89]. The two major dimensions consist in the valence dimension (or appraisal dimension, pleasant – unpleasant) and the activity dimension (or arousal dimension, or energy dimension, active – passive). A joint description of the emotion definition combining the two traditional theories is proposed in our work in which the discrete emotion states are distributed in a dimensional space.

In the scope of emotion recognition in audio signals, the problems following the emotion taxonomy elaboration concern the selection of acoustic features presenting the emotion or mood aspects and the classification algorithms. Further to the effective features commonly used in speech recognition and other work on music analysis it is necessary for the recognition task to find new features which have the ability to represent the emotional characteristics. According to the different needs in the classification and recognition problems for the emotional speech and music mood, and the situation that there is no agreement with the definition of emotion types, we have developed an automatic hierarchical classification algorithm that can adapt itself to different classification problems with different number and types of emotions. Due to the uncertainty of the emotions or moods as a subjective judgment of human beings and the complex border between the emotion states, we also considered a *preliminary attempt* to the *ambiguous* recognition.

Thus the objective of our work can be summarized as to implement accurate and reliable emotion classification and recognition systems applying the two

traditional theories of emotions at the same time with effective feature sets, ambiguous emotions management and having the ability to automatically adapt themselves to the different recognition problems. The research results raised in our work can also be applied as a part of audio indexing systems.

1.3 Our Approach

Although emotions in audio signals are high-level semantic notions, they can be represented by proper features extracted from the signals. It is proved that there exist reliable acoustic correlates of emotions in vocal signals [Ekm82] [Ban96] [Pic97] [Bur00] [Sch89] [Sch88]. Generally, global classifiers with the same feature set are applied in the literature for classification problems of emotions in audio [Pol00] [McG00] [[Oud03] [Sla98] [Ver04a] [Ver04b] [Ver05a] [Ver05b]. This type of classifier, which can be based on different algorithms, such as neural networks, are designed to manage all the classes in a single classifier which leads generally to a complex classifier having difficulties to generalize. According to the correlates of emotions with the acoustic features [Ban96] [Bre01] [Bur00] and our analysis on a set of audio emotional samples, some different emotions such as anger and happiness, or sadness and boredom may have similar features caused by similar physiological arousals. Therefore the same set of audio features cannot discriminate efficiently all the emotional classes at the same time. This is the reason why we have developed a multi-stage classification system which is based on a hierarchy of classifiers, each one being specialized in the classification of one particular emotion and using its own appropriate feature set. Moreover, as the number and the types of the discrete emotional states are typically application dependant, we also propose an automatic scheme which derives an optimal hierarchy of classifiers best fitting the discrete emotional states under consideration. As speech emotion or music mood are highly subjective and often fuzzy, a classification method allowing the management of ambiguous emotions, based on the evidence theory, has also been elaborated in order to meet the demand of real applications, although it is only a first study in the ambiguous recognition and needs to be greatly improve in the future. The influences of the different fusion operators and the duration of the audio segments to the recognition are also considered in our work.

Therefore, our approach consists in two ways of classification of the audio emotions: a hierarchical classification and a classification with ambiguity

management, under the consideration of accuracy and speed. Our work is mainly focused on the classification of speech emotions. However, we have also applied the two approaches on the classification of music mood.

1.4 Contributions

In order to illustrate the contributions of this work, we present first the feature sets describing the emotional states, and then we discuss the different solutions of the automatic recognition and classification of both emotional speech and music mood.

New features: Appropriate feature sets adapted to the classification problem are essential for an efficient classification. Some new acoustic features are proposed in our work designed to give a better description of the emotions in audio signals. The feature sets are thus made up of the new proposed features and some traditional audio features. For the recognition of emotional speech, new harmonic features based on a sub-band amplitude modulation of the signal and features derived from an analysis according to Zipf laws are proposed. For the description of the music mood, the features derived for the spectrum of the sub bands according to the octaves and the musical features according to the tempo and the tonality are adopted. Traditional acoustic features concerning the pitch, frequency and energy are also used in the feature sets.

Feature selection scheme: Because the features have different discriminatory power for the classification, the selection of features plays an important role to optimize an automatic classification solution. An embedded feature selection method namely ESFS (Embedded Sequential Forward Selection) is applied in this work by using the concept of the evidence masses from the evidence theory. Several t-norm operators are studied in this part in order to find out the best features with different parameters. As the discriminatory power of the features can vary for different emotional states, the classification problem can be cut into several sub-problems with fewer classes which could be more efficiently solved to get a better overall correct classification rate. The feature selection processing is thus applied to each of the sub-problems.

Hierarchical classification of audio signals: In order to diminish the perturbations between the different emotions or mood, hierarchical classifications are proposed. While a multi-stage classification scheme on six discrete emotional states is

first manually elaborated within multimedia indexing framework, we further propose a Hierarchical Classification Scheme (HCS) that automatically generates the structures of the hierarchical tree with classifiers for the sub-problems having their own parameters of operators in feature selection. Each layer in the hierarchical classification fits well the dimensional theory of the emotions, while the final classification results fit well the discrete theory of the emotions. The major advantage of such a scheme is that the number and types of discrete emotional states can be now application dependant.

Classification with ambiguity management of speech emotion and music mood: As the emotions are subjective judgments of human beings, the border between the difference emotions are usually ambiguous. With the consideration of a limited number of types of emotions, a certain emotional state can be between some pre-defined emotional states, while a too large number of types can lead to an insolvable classification problem. In this case, a classification with the management of ambiguous emotions is necessary. Thus, we have developed a method based on the evidence theory, which performs the combination of the sub-problems classifiers – Ambiguous Classification Scheme (ACS). The possibilities of each of the emotional states are given as the recognition result.

For the classification of the emotional speech, a gender classification step is added before the emotion classification for both hierarchical approach and ambiguity managed approach, as we have shown that the gender influences the recognition.

Influence of segment duration to the music mood: Since the mood is usually changeable in classical music works, it is necessary to choose an appropriate duration of music clips for the detection of the mood. The database used in our work is cut into several versions with the segment duration from 4 seconds to 32 seconds. Two kinds of tests are made in studying the influence of segment duration. First, both the learning set and the test set are taken from the data with fixed segment duration to compare the performance with different segment durations to find out the best duration. Second, the data with certain segment duration are used as the learning set, and the model derived from these data is tested with other versions of dataset to see the compatibility of the models for different durations.

Music mood tracking: The mood usually changes in a whole piece of classical music work. These changes can be tracked by dividing the music into several

independent segments, each of which containing a constant mood, and by detecting the mood type in each segment respectively. A simple criterion of cutting segments of constant mood is derived from the result of the influence of the duration to the music mood.

Fig. 1-1 illustrates the key points implemented in our work, including the feature extraction, feature selection, the classification and recognition of the emotion/mood for speech and music signals. The system is first developed upon emotional speech signals, and then applied into the classification of music mood. The music mood tracking is also implemented as an extension of the music mood recognition.

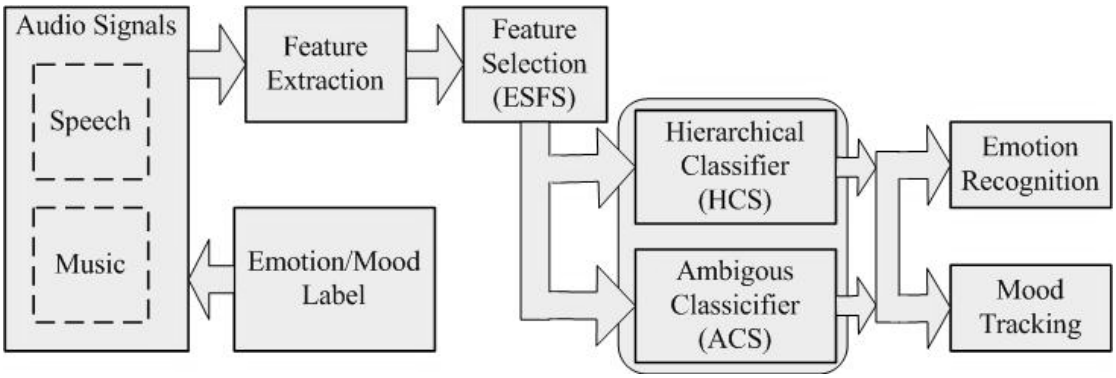

Fig. 1-1 Recognition of emotion from audio signals

Taken the emotion/mood label apart of our work, there remains two main parts of contributions in our emotion/mood classification system: on the one hand the feature extraction and selection and on the other hand the classification algorithms. For the two types of audio signal sources, different feature sets are extracted based on the properties of signals; the feature selection and the classification algorithms can be adapted to both of the two types of audio signal sources. The feature selection scheme (ESFS) based on the evidence theory in our work also serves as the sub-classifier itself, and it is the basis of this system.

1.5 Outline of the dissertation

This dissertation is organized as follows:

Chapter 2 introduces the background about the emotion recognition of speech signals, including the taxonomy and acoustic correlates. The joint of the two traditional theories is proposed as the basis of the emotion classification in our work.

In Chapter 3, we make a brief review of the state of the art on the classification of vocal emotions. The emotional speech corpuses and the related works are introduced. We also make a synthesis of the commonly used features in the literature, mostly frequency-based features and energy-based features. It outlines the state of the art of this field and points out the goal of this thesis: to provide *automatic* emotion recognition systems of audio signals which can *adapt* the change of the emotion/mood definitions with the possibility of *ambiguous* recognition.

Our early work in this thesis with an empirical built dimensional emotion model driven multi-stage classification method (DEC) based on neural network is introduced in Chapter 4. In this chapter, we first proposed several new acoustic features - harmonic features which are perceptual features containing more comprehensive information of the spectral and timbre structure of vocal signals than basic pitch and formants patterns, and Zipf features which characterize the inherent structure of signals, particularly the rhythm and prosody aspects of the vocal expressions. The manually built empirical multi-stage classifier contains a first gender classification stage and two 2-stage classifiers for each gender, and different empirical DEC were built considering Berlin dataset and DES dataset respectively.

Chapter 5 presents improvements of this first classification system according to two aspects: a new feature selection scheme based on the belief masses and an automatic approach of hierarchical classification scheme. The first part of this chapter introduces a feature selection scheme namely ESFS based on the belief masses which is inspired by the evidence theory, several t-norm operators being used to make the combination of the features. The second part of this chapter addresses the development a hierarchical classification scheme (HCS) of emotional states, which can reduce the confusions between the similar emotional states and adapt to different classification problems. The automatic HCS is experimented on Berlin and DES datasets.

An ambiguous approach (ACS) based on the evidence theory which manages the subjective characteristic of the emotions is proposed in Chapter 6 as a first step in the automatic ambiguous recognition of emotions to simulate the human manner in the emotions judgment as close as possible. This approach is also experimented using the two datasets of emotional speech: Berlin and DES datasets.

The two classification approaches are applied on the automatic classification of music mood in Chapter 7. A specific feature set for the recognition of music mood is proposed, and the definition of four mood states in the two-dimensional space according to Thayer's model is adopted. The algorithms are experimented on the four versions of our dataset with different segment durations for the music mood. Besides the mood classification algorithm, the influence of the duration of the music segments and the music mood tracking are also discussed.

Chapter 8 summarizes the thesis results and the dissertation contributions. Finally further research suggestions are given.

Chapter 2

Emotion Taxonomy and Acoustic Correlates of Emotions in Human Speech

Machine recognition of speech emotion is feasible only if there is a model of sound emotion taxonomy as well as reliable acoustic correlates of emotions in human speech.

The theoretical model of emotions is the first problem raised in the research of the classification of emotions. 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 theories [Sch02].

Researchers in the discrete theories propose that there exists 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 expressions [Sch02]. The term “big six” has gained attention in the tradition of the discrete description of emotions. It implies the existence of a fundamental set of six basic emotions. However, there is no agreement on which these six should be. The terms including happiness, sadness, fear, anger, neutral and surprise are often used in the research field with this theory. The discrete description of emotions is the most direct way and is much clearer than other descriptions to discuss the emotional clues conveyed in audio signals. However, with a limited number of emotions considered in the discrete emotion theory, the differences between the different emotions in the

same emotional family cannot be easily taken into consideration. Using this way for emotion description, it is more likely to distinguish an emotion from the given kinds than to recognize it in the whole emotional space.

In the dimensional theories of emotion, the emotional states are often mapped into a two or three-dimensional space. The two major dimensions consist of a valence dimension (pleasant–unpleasant, agreeable–disagreeable, also presented as appraisal dimension) and an activity dimension (active–passive, also presented as energy dimension or arousal dimension) [Sch00b]. If a third dimension is used, it often represents either power or control. According to Youngstrom, a third dimension does emerge as the control for the emotions; for example, if someone felt relatively in control of a threatening situation, he or she might respond with anger, without control, the reaction might be fear [Grif]. 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 [Wie05a]. Intense emotions are accompanied by increased levels of physiological arousal. An example of arousal vs. appraisal plane of emotions is shown in Fig. 2-1. In this example, arousal values range from very passive to very active, and appraisal values range from very negative to very positive.

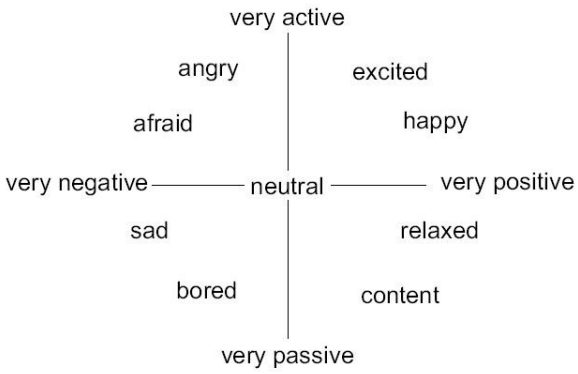

Fig. 2-1 Example of emotions in arousal vs. appraisal plane [Wie05a].

Some researchers proposed a dimensional and hierarchical structure of affects like the Tellegen-Watson-Clark emotion model [Tel99]. This three-level hierarchy (Fig. 2-2) incorporates in one structure a general bipolar happiness versus unhappiness dimension, the relatively independent PA (positive affect) and NA (negative affect) dimensions at the level below it, and discrete emotions at the base.

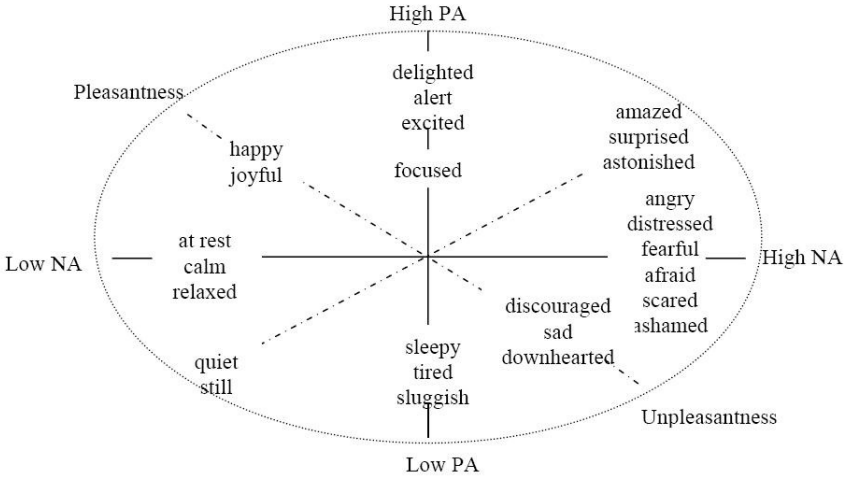

Fig. 2-2 Elements of the Tellegen-Watson-Clark emotion model [Yan04] [Tel99]. Dotted lines are top-level dimensions. The Positive Affect (PA) and Negative Affect (NA) dimensions shown as solid lines form the middle of the hierarchy, and provide heuristics needed to discern the specific discrete emotion words based on function. Discrete emotions that are close to an axis are highly correlated with that dimension

The taxonomies of emotional speech are discussed in the following subsections.

2.1 Taxonomy of emotional speech

For the theoretical study of the taxonomy of emotional speech, precise models are the componential models of emotion, which are often based on the appraisal theory [Sch01]. These models emphasize the variability of different emotional states, do not limit the description of emotions to the two or three basic dimensions comparing to the dimensional theories and they do not restrict the attention to a limited numbers of supposedly basic emotions comparing to the discrete theories. Moreover, these models allow modeling the distinctions between members of the same emotion family. For example, for “anger”, there are explosive rage, controlled anger, cold anger, *etc.*; and for “happiness”, there are difference between exuberant joy and quiet happiness. In many cases, a mild state of a certain emotion in speech may have completely different affect in acoustic features with the intense state of the same emotion family. The most important advantage of componential models is that these approaches provide a solid basis for theoretical elaborations of the mechanisms

that are supposed to underlie the emotion–voice relationship and permit to generate very concrete hypotheses that can be tested empirically.

Another model proposed in [Cow00] looks at the various ways in which category labels may be used to describe emotional states. Several terms similar to the discrete emotions and the dimensional emotions are proposed. The first term as the basic emotion categories suggests obtaining a list of the primary emotions as an appropriate starting point for search whose aim is to study the speech patterns associated with the basic emotions. The "big six" is also included as a potential list of the basic emotion categories. The second term as the second order emotion categories suggests that everyday language contains an abundance of emotion-related categories. A list by Plutchik [Plu80] contains 142 words describing emotional states, and another list by Whissell [Whi89] contains 107 words, which cover a great range of emotional states with very few of which could be regarded as basic. The second order emotions are less important but are more natural emotional states. The basic emotion categories correspond to the traditional discrete emotion theories, and the second order emotions have similarities in some degree with the componential models.

From practical point of view, the discrete definition of emotional speech is the most commonly used with different number and types of emotions according to the application under consideration. Since too many types of emotions can lead to too complex automatic recognition problems, the number of emotional states considered in the literature usually varies from 3 to 6 [Pol00] [Sla98] [Ver04a] [Ver04b] [Oud03].

In this thesis work, the primary aim of speech emotion recognition lies in the multimedia searching and indexing of movies or TV series. Discrete emotional states such as anger, boredom, happiness, fear, neutral, sadness and surprise appear most frequently and are most typical in such programs [Nie00]. We thus first investigate these emotional states that are mapped into a 2-dimensional space during the classification process.

2.2 Acoustic correlates of emotions in the acoustic characteristics

Apart from the words, human beings express emotions through modulations of facial expressions [Ekm82] and modulations of the voice intonation [Ban96]. There

are some reliable correlates of emotion in the acoustic characteristics of the signal: speech emotion is question of prosody and is expressed by the modulation of the voice intonation parameterized by features such as tonality, intensity and rhythm.

Emotions are considered as cognitive or physical by different theories, and can be discriminated by distinct physical signatures [Pic97]. Several researchers have studied the acoustic correlates of emotion/affect in the acoustic features of speech signals [Ban96] [Bur00]. According to [Ban96], 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 [Sch89]. There are also much evidence points of the existence of phylogenetic continuity in the acoustic patterns of vocal affect expression [Sch88], while currently there is little systematic knowledge about the details of the acoustic patterns that describe the specific emotions in human vocal expressions. Typical acoustic features which are considered as strongly involved in this process include the following: 1) The level, range and contour shape of the fundamental frequency (F0), which reflect the frequency of the speech signal vibration 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 affect the articulation; 4) the speech rate. For example, several emotional states such as anger, fear, and happiness (or joy) are considered as with high arousal levels [Ban96]. 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 an increase of the heart rate and blood pressure, being 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 a decrease of the heart rate and blood pressure and an increase of salivation [Bre01]. Picard summarizes the effects of vocal emotion by the arousal of sympathetic and parasympathetic nervous systems as primarily in the frequency and timing, and secondly in the loudness and enunciation [Pic97]. There is a special emotional state as disgust with inconsistent character according to Banse *et al*. [Ban96]. Indeed, it represents an increase in mean F0 in disgust and unpleasant films while a decrease in mean F0 in actor simulation [Pit93], because from an evolutionary perspective, disgust is an emotion needs to be

communicated only over relatively short distances and thus can be expressed very well visually, but usually bad communicated in the voice [Joh04].

Emotion recognition can be language and culture independent. The acoustical correlates of basic emotions across different cultures are quite common due to the universal physiological effects of the emotions. According to Darwin, *"Even monkeys express strong feelings in different tones — anger and impatience by low, fear and pain by high notes"* [Dar1871]. Abelin and Allwood considered in [Abe00] utterances spoken by a native Swedish speaker to be recognized by persons as native speaker of 4 different languages as Swedish, English, Finnish and Spanish. The carrier phrase in Abelin's work was composed by several terms of food (translated as salted herring, mashed potatoes and pancakes) since many different emotions can be hold towards food, and thus the sentence was neutral in respect to different emotions. Close recognition patterns were obtained by people speaking different languages, which shows that *the inherent characters of vocal emotions can be universal and culture independent*. Three dimensions as lust - non lust, active – passive, and secure – insecure were evaluated for the emotions in their work: happiness, anger and surprise comprise high activity in common, while happiness and surprise are with high lust and anger is with low lust as difference between them; sadness, fear and shyness comprise low lust, and sadness and shyness are with low activity, fear and shyness are with low security. The common characters of acoustic features of the emotions based on the dimensional analysis with F0 variance, intensity, and speech duration (between silence periods) were also discussed in Abelin's work. Happiness, fear, shyness and sadness are quite even with F0 variance, and surprise, anger, and dominance have strongly varying F0; for the intensity, anger, surprise, disgust, and dominance have the highest value, and sadness and shyness are weakest; the longest duration occurs with happiness, disgust, and surprise, and shyness and sadness have relatively longer pauses between utterances.

Tickle also proved this relative language and culture independence of emotion recognition by asking Japanese listeners to decide the emotions expressed by Japanese or American people using meaningless utterance without semantic information [Tic00]. The best recognition score by human is about 60%. Similar result was obtained by Burkhardt and Sendlmeier [Bur00] using utterance as a set of syllables containing a set of phonemes with assigned prosody descriptors similar to the MBROLA-format [Dut96]. Listeners were asked to assign the stimuli to nine types of

emotions as neutral, hot anger, cold anger, happiness, joy, crying despair, quite sorrow, fear and boredom. The prototypes for these emotions include faster speech rate for hot/cold anger, joy, and fear, slower speech rate for happiness, crying despair, quiet sorrow, and boredom, raised pitch and broader pitch range for hot anger, joy, and fear, lowered pitch for cold anger, quiet sorrow and boredom. Most confusion in human testing occurs between hot anger and cold anger, happiness and joy.

Quang, Besacier, and Castelli investigated cross lingual experiments on automatic question detection, which can be considered as a special emotional state that is also influenced by the vocal intonation [Qua07]. Prosody features mainly concerning the F0 and lexical features are used in their work with two languages of French and Vietnamese.

These studies suggest thus that some reliable acoustic correlates of emotions in the acoustic features of vocal signals offer the possibility to achieve machine recognition of vocal emotions. On the other hand, as emotion recognition by humans, with roughly 60% recognition rate is not so accurate, we probably cannot expect perfect machine recognition. This relatively low recognition rate by humans can mainly be explained by similar physiological properties for certain emotional states that lead to the similarity of acoustic features. While human beings can make use of all the contextual information (speech, gesture, facial expression, *etc*.) for resolving ambiguity, an automatic emotion recognition only based on vocal signal should focus on a few kinds of basic emotional states with reasonable efficiency.

2.3 Conclusion

Emotions in audio signals can be represented either by a discrete description or by a dimensional description. In synthesized consideration of the two descriptions, we propose to map the discrete emotional/mood states into a dimensional emotion space as a joint of the two traditional theories of emotions to better fit with the acoustic correlates of emotions. Because considering too many discrete classes can lead to a too difficult classification problem, and thus low performance results, the number of classes considered in our work is currently limited to the most representative ones for a given application, with the upper limit of six.

Chapter 3

Automatic Classification of Vocal Emotions: A State of the Art

As a major part of emotion-oriented computing or affective computing [Pic97], 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 [Dru00] or for pure entertainment [Kus01]. 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. The aim of our work with this subject resides in the multimedia indexing, such as automatic search and recommending of movies/TV programs under certain emotional themes.

The state of the art concerning speech emotion including the data resources of the emotional audio signals and related works is introduced in this chapter.

3.1 Resources of emotional speech

As a pattern classification problem, the dataset of the emotional audio signals is essential for building models of recognition. Most of existing audio corpuses concern topics such as speech recognition, speaker identification, music genre classification, *etc*. Specially designed datasets should be built to fit the problem of emotion recognition according to the characteristics of the emotional audio signals.

The possible resources and existing datasets of emotional speech are introduced respectively in the following subsections.

3.1.1 Emotional speech dataset taxonomy

Speech corpus for the studies of emotions needs to represent the different types of emotion by unique patterns or configurations of acoustic clues in order to communicate reliably the underlying speaker's emotional states. There have been a relatively large number of studies dealing with the way to obtain emotional speech corpuses, which can be generally classified into three major categories as natural vocal expression, induced emotional expression, and simulated emotional expression [Sch02].

Natural vocal expression is recorded during naturally occurring emotional states of various sorts. It has very high ecological validity, which measures the degree to which the characteristics are actually correlated with the underlying speaker state (according to Brunswik's terminology [Bru56]). Moreover, it reflects more faithfully the acoustic features and other clues of emotions in the speech. The natural vocal expression can be obtained from real life or from some kind of TV program, such as talk shows or interaction game shows. In spite of the high ecological validity, the natural vocal expression has its severe disadvantages. Indeed the segments with obvious emotional clues in natural voice samples are usually suffering from bad recording quality, which leads to difficulties to determine the precise nature of the underlying emotion [Sch02]. Although it is a concerned character of vocal emotion, the training samples with emotional states that are not clear enough may disturb the training of the recognition algorithm by obscuring the acoustic correlates between the audio features and emotional states.

Induced emotions are caused by using psychoactive drugs or some particular circumstances, such as stress induction via difficult tasks to be completed under time pressure, the presentation of emotion inducing films or slides, or imagery methods [Sch02]. Getting speech samples in this way is always favored by psychologists, but this method cannot ensure to get the desired emotional speech, because people do not always have the same reaction to the same stimulation.

The third way to get speech samples is the simulated or portrayed emotional expression. It consists in asking actors, including ordinary people and professional

actors, to produce vocal expressions of certain emotions. These samples use given content and produce given emotions. The emotions in portrayed speech have more typical expressions than induced emotion, sometimes even more intense than natural emotions. The problem is that, in portrayed emotions, some obvious clues may be over emphasized, while some more subtle clues may be ignored, and cannot reflect all the clues in the emotional vocal faithfully. Furthermore, some considerations indicate that the portrayed emotions may inflect something of culture background of the speaker more than when they are occurring under natural conditions. However, it is argued that all the speeches to the public have more or less sense of acting. As long as the portrayed emotions can be recognized by listeners, they reflect at least part of the emotional patterns. The portrayed emotional expression is preferred by some of the researchers [Sch02], because the simulated vocal portrayal of emotions usually yields much more intense, highly controlled, prototypical expressions than induced states or even natural emotions.

3.1.2 Emotional speech datasets

There exist several portrayed emotional expression databases in different languages. According to [Sch00a], vocal emotion expressions may be at least in large part driven by universal psychobiological mechanisms since judges from different cultures, speaking different languages, recognize the expressed emotions with much better accuracy than chance. This point can also be supported by [Sla98] who showed that even the small babies who do not speak yet could also recognize the emotional clues from the speech of the adults. In the work of Zhu [Zhu07], a language-independent machine recognition of human emotion in speech is also implemented with a corpus of emotional speech from various subjects and different languages for developing and testing the feasibility of the system, which proved the work of [Abe00] [Tic00] [Bur00] with human testing. Therefore, the languages used in the emotional speech databases do not have much influence on the study of vocal emotions.

Several databases were built by different research groups.

Pao *et al* built a mandarin emotional speech database with five emotional states (anger, happiness, sadness, boredom and neutral) acted by 18 males and 16 females with 20 different sentences, resulting in 3400 emotional speech sentences.

After a three-pass procedure of listening test, the speech samples whose emotional content is hard to identify were deleted, and 839 sentences were remained in this database. The final results in listening test in their work vary from 73.22% to 89.56% for different emotions. Their aim is to help hearing-impaired people to improve the naturalness of their emotion expression [Pao04].

A cartoon-like Japanese emotional speech database was built by Sony for training pet robots [Oud03]. Six professional speakers who worked on many radio/TV commercials, movies and animations were asked to pronounce short sentences or phrases over four emotional classes: joy/pleasure, sorrow/sadness/grief, anger, and normal/neutral, imagining themselves speaking to a pet robot. 200 examples per speaker and per emotion are collected which make 4800 samples in total in this database.

A speech corpus containing utterances for testing an effective language-independent emotion recognition system was built by Zhu in [Zhu07] concerning six classes: happiness, sadness, anger, fear, surprise and disgust. Subjects from different language backgrounds of English and Chinese were selected for recording over 500 utterances with a sampling rate of 22050 Hz.

Berlin emotional speech database, developed by Professor Sendlmeier and his fellows in Department of Communication Science, Institute for Speech and Communication, Berlin Technical University, was recorded in 1997 and 1999 by 5 actors and 5 actresses, pronouncing 10 sentences in German over 7 emotional classes: anger, boredom, disgust, anxiety/fear, happiness, sadness, and neutral [Sen] [Bur05]. More than 500 speech samples were recorded. The length of the speech samples varies from 3 seconds to 8 seconds, and the sampling rate is 16 kHz.

Another easily accessible and well-annotated database is the DES (Danish Emotional Speech) database [Eng96]. The speech samples are expressed by four professional actors, two male and two female in five emotional states such as anger, happiness, neutral, sadness, and surprise. This database contains around 10 short sentences and word and 2 paragraphs per actor and per emotion.

These five existing datasets on emotional speech contain from four to seven emotional classes, which generally include anger, boredom, disgust, happiness (joy, pleasure), fear (anxiety), neutral (normal), sadness (sorrow, grief), and surprise. Different languages such as Chinese (Mandarin), English, Japanese, German, and

Danish are recorded. The largest dataset is from Sony [Oud03] that contains 4800 speech samples, while the aim of developing this dataset was for training of pet robots with cartoon-like emotional speech, and cannot easily fit the other problems in recognition of vocal emotions. The scales of the other five datasets are almost the same (around 800 samples in the dataset of Pao [Pao04], around 500 samples in the other three). According to the availability of the databases and the possibility to compare our results with those of other works, the latter two databases (Berlin database and DES database) are used in our work. The language independence character of the emotional speech is also investigated using these two datasets with different languages (German and Danish).

3.2 Related works in emotional speech classification

Along with increasing awareness of wide application potential from affective computing [Pic97], 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 [Pol00] [Sla98] [McG00] [Oud03] [Ver04a] [Ver04b]. Every works are applying a discrete emotional model for the comfort of automatic recognition. 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.

Some researchers adopted three or four classes of emotional states to make a clear distinction, especially for some special purposes.

In [Pol00], Polzin and Waibel dealt with emotion-sensitive human-computer interfaces. Their corpus includes speech segments from English movies. Only three negative emotion classes, namely sad, anger and neutral, are considered. They modeled the 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 or 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 filter bank. The verbal features, prosody features and spectral features were evaluated separately in their work. For human judgment, an overall accuracy of about 70% was obtained by listening test, about 55% using only textual representation. In their automatic classification, accuracy up to 60.4% was achieved with prosodic information and 63.9% with spectral information. According to their experiments, this classification accuracy is quite close to human classification accuracy. One of the originality of this work is the preliminary separation of speech signals into verbal signal and non-verbal signal. A 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 cannot reflect the acoustic characters of vocal emotions. Among the non-verbal features, the pitch, intensity, and cepstral coefficients information were considered to present 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, the energy distribution in the spectrum and the other higher level features concerning the whole structure of emotional speech signals were absent in their feature set.

Slaney and Mcroberts also studied a classification problem with three classes of emotions in [Sla98] but within another context considering the three attitudes as approval, attention bids, and prohibition from adults talking to their infants aged of about 10 months. 500 utterances were collected from 12 parents talking to their infants. Each utterance was first split into three segments as the first, middle, and final third of the sound, and the features are calculated on each segment separately. They made use of simple acoustic features, including several statistics measures related to the pitch and MFCC as measures of the formant information, as well as timbre cepstral coefficients. 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 also 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 influences 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, thus cannot be used for reference directly in other applications. The main goal 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, thus only simple acoustic features were used in their experiment, and the relations 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* [Ver04a] [Ver04b] [Ver05a] [Ver05b] with more emotion classes. In their work, 500 speech segments from DES (Danish Emotional Speech) database are used [Eng96]. Speech is expressed in five 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 [Ver04a], a correct classification rate of 54% was achieved when all data were used for training and testing. A Bayes classifier has been used with the five 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 [Ver04b], 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 ten features. The GMMs (Gaussian Mixture Models) were emphasized in [Ver05a], and short-term features they used were better explained in [Ver05b]. Their best result was obtained by a GMM for male samples at 66% classification rate in [Ver05b].

Prior to the work of Ververidis *et al*, McGilloway *et al* [McG00] also studied a five emotion classification problem with the speech data recorded from 40 volunteers describing the emotion types as afraid, happy, neutral, sad and anger, based on a system called ASSESS (Automatic Statistical Summary of Elementary Speech Structures). They already made use of 32 classical features based on pitch, frequency and energy, selected from 375 speech measures. The accuracy they reached 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 [Oud03]. He considered four 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 networks, decision trees, classification by regression, SVM, naïve Bayes, and Adaboost on WEKA platform [Wit05], 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.

Pao *et al*. also considered five emotions such as anger, happiness, sadness, boredom and neutral aiming at helping hearing-impaired people to improve the naturalness of their emotion expression [Pao04]. Since the key point concerned in their work is the construction and testing of the speech dataset, the features are not thoroughly discussed. MFCC features were evaluated in their system using K-NN classifier with 70% of speech data for training and 30% of data for testing. An emotion radar chart with multi-axes was applied to evaluate the emotions in order to keep the ambiguous information, as shown in Fig. 3-1. The distances of testing data to each category are measured to plot the radar chart with an M-KNN (M stands for the different emotions). The dashed lines in Fig. 3-2 show examples of sentences with anger emotion, (a) is a sentence close to anger but a little ambiguous, and (b) is an non ambiguous emotion as anger.

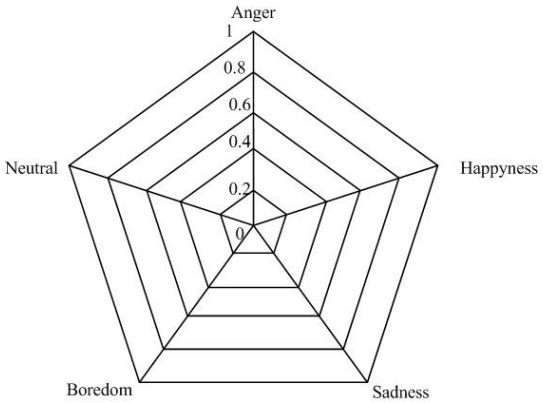

Fig. 3-1 Emotion radar chart [Pao04]

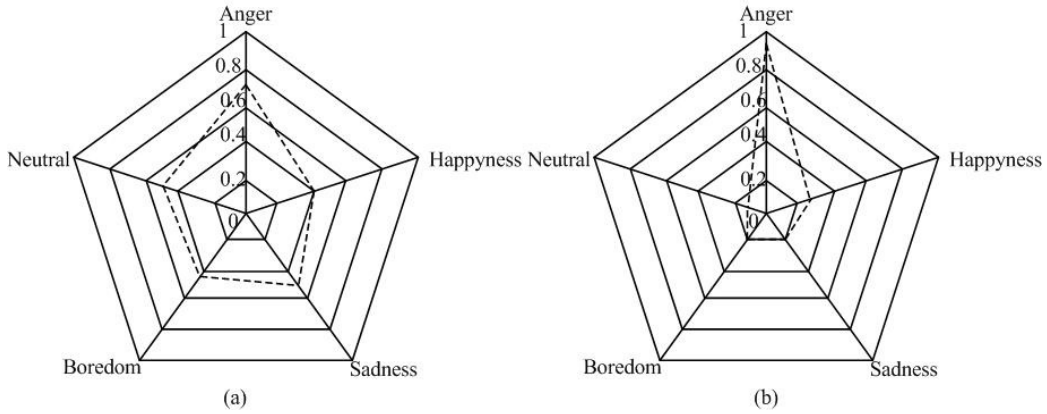

Fig. 3-2 Sentences with anger emotion in the radar chart [Pao04]

Six emotions as anger, disgust, fear, joy, sadness, and surprise were investigated by Nwe *et al* in [Nwe03] on an emotional corpus containing two languages of Burmese and Mandarin with 720 utterance from twelve actors (three females and three males of native Burmese speakers, and three females and three males of native Mandarin speakers). LFPC (log frequency power coefficients) features were tested with HMM (Hidden Markov Models) classifiers. The average classification accuracy for the test on the 12 speakers reached 78.1% for the six emotions. However, as their experiments were taken on each speaker separately, it is hard to compare their result with other works, because their results are highly speaker dependent.

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

The feature sets used by McGilloway [McG00], Ververidis [Ver05a] and Oudeyer [Oud03] were basically spectral features, pitch features and energy (intensity) 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 applied 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. In particular, the fundamental frequency movements were analyzed by Paeschke [Pae00] *et al* to present the prosodic characteristics of emotional speech. The range, declination, and duration and steepness of accents were discussed in their work, and confirmed that there is reliable discrimination between emotions with low and high arousal with sad, boredom, and neutral on the one hand, and fear, happiness, and anger on the other hand.

However, these features are likely to mostly reflect nonspecific physiological arousal, and the existence of emotion-specific acoustic profiles may have been obscured [Ban96]. 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 suggest that some discrete emotions are very closed to each other on the dimensional emotion space and there is confusion of emotion class borders as evidenced in [Ban96] 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.

In the next section, we summarize traditional features such as spectral, pitch and energy based features.

3.3 Synthesis of commonly used frequency and energy based features

In the related works mentioned in the previous subsection, the most commonly used features include features concerning the fundamental frequency, formants, intensity, MFCC, *etc* [McG00] [Oud03] [Ver04a] [Ver04b]. Referring to the work of McGilloway [McG00] and Ververidis [Ver05a], the statistical values such as mean, maximum, minimum, and variance of the original acoustic characteristics of the speech signal can be considered. We group these traditional features into frequency

features, energy features, and MFCC features. The frequency features include the statistics of fundamental frequency F0 and the first three formants; the energy features include the statistical features of the energy contour, and the MFCC features include the statistical of the first 24 MFCC coefficients. They are further applied in our work in complement to harmony and Zipf based features, as we will describe in the next chapter.

3.3.1 Frequency-based features

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 [Pra01]. The F0 and the formants are only computed through the vowels periods. For the consonants, the F0 and the formants are assumed as zero, and are not considered in the statistics. See F0 and the formants in Fig. 3-3 (b).

3.3.2 Energy-based features

The energy distribution over the spectrum, especially the ratio of low frequency energy (set to below 250Hz in our work, referring to the work of [McG00]).

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, and presented in db. See the solid line in Fig. 3-3 (c). The edge points of the plateaus of the energy contours are defined as the points at three db lower than the peak points. The valley of the energy contours are obtained with similar method as the points at three db higher than the local lowest points. The energy plateaus and the slopes are obtained by approximating the energy contour with straight lines (see the dashed line in Fig. 3-3 (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.

The durations of the energy plateaus show approximately the lengths of the vowels, and the durations of the energy valleys show approximately the lengths of the silence periods in the utterances. The slops in the energy contours present the intonation. Thus, the prosodic characteristics can be extracted by the analysis of the energy contours.

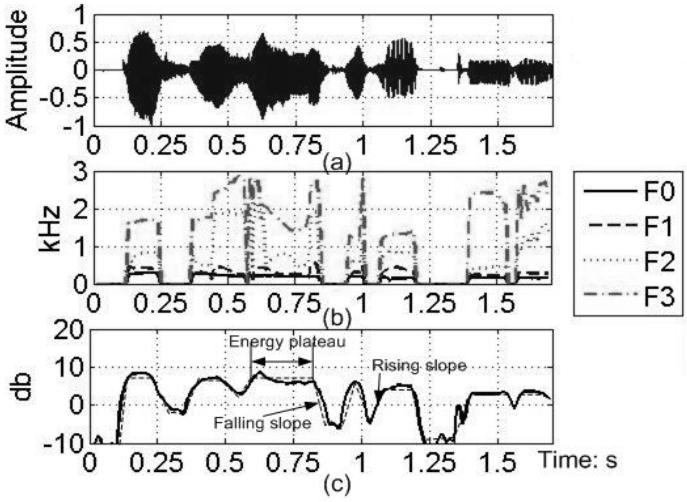

Fig. 3-3 Basic acoustic features of a speech signal: (a) waveform, (b) fundamental frequency F0 and the first 3 formants (F1, F2, and F3), (c) energy contour

3.3.3 MFCC features

Mel-Frequency Cepstrum Coefficients are the most known from the family of cepstral characteristics. These are cepstral coefficients obtained from a spectrum filtered by Mel scale [Ste37]. The Mel scale is a scale inspired by the characteristics of human perception, which is different from the normal cepstrum, in such a way that the frequency bands are positioned logarithmically allowing approximating the human auditory system's response more closely than the linearly spaced frequency bands obtained directly from the FFT or DCT. It affirms that high frequencies are caught by human's ear with less precision in comparison to low frequencies.

The MFCC are obtained as follows:

1. Divide signal into frames.
2. Compute the FFT to obtain the amplitude spectrum.
3. Take the logarithm to map the spectrum onto Mel scale.
4. Apply the Mel filter.

5. Take the Discrete Cosine Transform of Mel log-amplitudes.

The MFCC are widely used in speech modeling and recognition. In our work, MFCC vectors with 24 coefficients are considered.

3.3.4 Summary

We summarize the commonly used traditional features, and list them as below.

List of frequency features:

- Mean, maximum, minimum, median value and the variance of F0
- Mean, maximum, minimum, median value and the variance of the first 3 formants

List of energy features:

- Mean, maximum, minimum value of energy
- Energy ratio of the signal below 250 Hz
- Mean, maximum, median and variance of energy plateaus duration
- Mean, maximum, median value and variance of the values of energy plateaus
- Mean, maximum, median and variance gradient of rising and falling slopes of energy contour
- Mean, maximum, median and variance duration of rising and falling slopes of energy contour
- Number of rising and falling slopes of energy contour per second

List of MFCC features:

- Mean, maximum, minimum and variance of each of the 24 first MFCC coefficients

3.4 Discussion

Several conclusions can be drawn from this short state of the art on vocal emotions including emotional speech resources and related works in automatic classification of emotional.

First, different definitions of emotions are used by different researchers in their work. Usually, three to six emotional types with the discrete description are considered. Moreover, the relations between the emotional types are not always clear with the discrete terms associated with the emotions.

Secondly, traditional speech features related to pitch, formants, and energy were commonly used in previous works. As these features were originally used in speech recognition or speaker identification, some of the specific profiles of emotion clues in vocal expressions, such as timbre and prosody characters, are missing and special new features are needed in practice.

The third point concerns classifiers. Indeed, approaches mentioned in the state of the art make use of global classifiers that perform in one-step the classification of all the emotions considered in the classification. The similarities between close emotions in the dimensional emotion space may be ignored and the differences between the emotions may also be obscured.

Finally, as the emotion is a rather subjective topic in audio signal processing and even human beings cannot always tell precisely the underlying emotional states in vocal sentences [Tic00] [Bur00] [Pol00], the ambiguous aspect need to be investigated, while only very few work considered the possibility of the ambiguous vocal emotions [Pao04]. Indeed, the emotion contained in a certain segment of audio signal may be not defined as one type of emotion but it can be between some emotion types or be a combination of several emotions, especially between the emotions located close to it in the dimensional emotion space.

We propose to address the previous issues respectively in Chapter 4, Chapter 5, and Chapter 6.

For the emotion definitions, we propose to map the discrete emotions into 2-dimensional space with an arousal dimension and an appraisal dimension in order to make clear the relationship and differences between the emotional states. Moreover, this process allows to keep the correlates between the acoustic features and emotions relatively stable according to the two dimensions even with the change of the terms describing the emotions.

Furthermore, we propose new harmonic features correlated to the timbre structure and Zipf features representing the prosody information in order to

complement the traditional feature set, which was not rich enough to carry the information expressed in audio signals for the different type of emotions.

For the last two points, we have proposed two approaches of classification.

The first one is a multi-stage (hierarchical) classifier with the structure of a binary tree with each step corresponding to a dimension in the emotional space to maximize the ability of discrimination of the speech features for the emotions. We first developed an empirical multi-stage classifier for six specific emotional states. Then an automatic generation method of hierarchical classifiers that can adapt to different emotion definition has been developed. The nodes in the binary tree are generated automatically according to the training data. Single judgment of emotional states can be made in this approach.

The second one is an ambiguous recognition that allows multiple judgments. This method, based on the evidence theory, makes the fusion of several simple classifiers present the recognition result as the belief mass of each of the emotional states. Due to the lack of universal agreement on the emotion definitions, it is also important for both of the approaches that the generation of the classifiers can adapt to the change of the number and types of the emotional/mood states.

Chapter 4

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

In this chapter, we develop our first approach for emotional speech classification. We first define the problem and our approach. Then, we introduce the two major contributions of this first approach for dealing with emotional speech classification: on the one hand two new groups of features for better emotion characterization in complement to the classic frequency and energy based features, and on the other hand, a dimensional emotion model-driven hierarchical classifier (DEC) instead of a single global classifier mostly found in the literature. Finally, we show effectiveness of this first approach with experimental results on two different public emotion datasets.

4.1 Problem statement and our approach

Our primary goal for vocal emotion classification was motivated by multimedia indexing for enabling content-based retrieval. The kind of application might be a powerful search engine delivering speakers in a multimedia collection that discuss a certain topic in a certain emotional state. We thus investigate some rough and basic emotion classes here. However, the number and types of discreet emotion states may be application dependant as we have previously seen in Chapter 3 on related work. We thus want our approach to be general, speaker independent, and possibly language independent so that it can be applied for various discrete emotional states independent in different application context. In the following, 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 and its generalization capabilities on DES dataset having some different five emotion classes.

Unlike most of the most in the literature, our contributions for vocal emotion recognition in this first approach 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 inherent 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 that hierarchically combines several binary classifiers. At each stage, a binary class classifier makes use of a different set of the most discriminative features and distinguishes emotional states according to different emotional dimensions. Finally, an automatic gender classifier is also used for a more accurate classification.

Experimented 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. As far as we know, current works in the literature display a best classification rate up to 66% on the same DES dataset [Ver05b].

4.2 Harmonic and Zipf features

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. Tonality is perceived with the timbre structures, which can be partly presented by position and amplitude of pitch and formants, and

the energy distribution in the spectrum. Intensity is mainly related to the level of signal energy. Prosody, which comes from information including the rhythm, stress, and intonation of speech and other elements of the language that may not be encoded by grammar, reflecting the emotional state of a speaker [Bar06], is related to high-level features that cannot be simply presented by basic speech features [Aud05]. In many previous works on the emotional speech, the features considered for the classification of emotional states normally limited to the classical acoustic features in speech signal processing, such as the features related to the pitch, the formants and the energy, which are common features used widely in all the problems related to speech signal processing. Some of these features have been proved to be effective for the classification of speech emotional states in [McG00] [Oud03] [Ver04a], especially in representing intensity and tonality in some degree. These fundamental speech features are adopted into our feature set, but they are not enough for the recognition of vocal emotions because the timbre and prosody structures of speech signal need to be better described with other features in order to better present the emotions.

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 and Zipf features. Harmonic features are derived from a of sub-band amplitude modulation of the signal [Xia07a] [Xia07b] for a better description of voice timber pattern whereas Zipf features are derived by mapping the speech signal into letters and words for a better rhythm and prosody characterization.

4.2.1 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 [Plo70] 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 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 [Moo97].

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 when extracting the harmonic features.

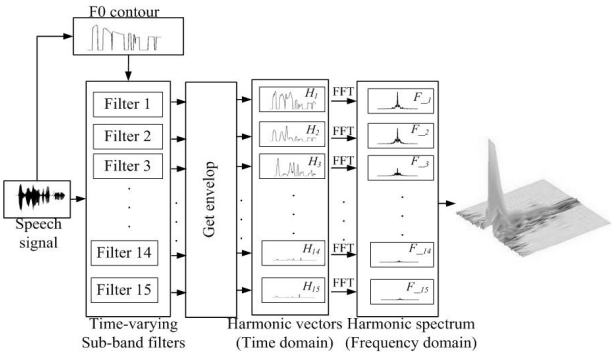

Fig. 4-1 Harmonic analysis of a speech signal

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. The center frequency for the i^{th} sub-band filter at a certain time is i^{th} multiples of the fundamental frequency (i^{th} 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. 4-1 and Fig. 4-3 (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 as

$$X(n) \approx \sum_{i=1}^{15} H_i(n) * e^{j2\pi i f_0(n) n} \qquad (4.1)$$

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*, which is set to 2 seconds in our experiments, are thus used to represent the voice timbre pattern:

$$F_{\_i} = FFT(H_i(n)) \qquad (4.2)$$

The spectrums are shown in Fig. 4-1 and Fig. 4-3 (b) ($F_{\_1}$, $F_{\_2}$, $F_{\_3}$…). These 15 spectrums are combined together into a 3-D harmonic space, as shown in Fig. 4-3 (c).

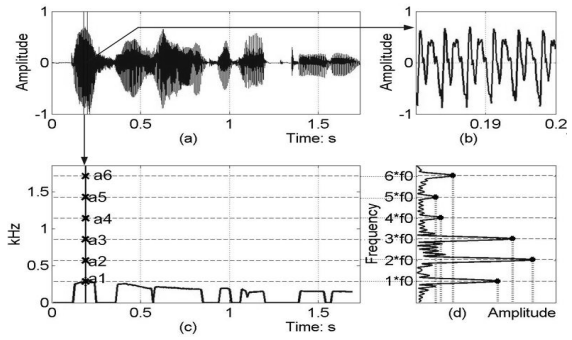

Fig. 4-2 Calculation process of the harmonic features: (a) waveform in time domain, (b) Zoom out of (a) during 20ms, (c) F0 contour of (a), a1 – a6 are the frequency points of 1 to 6 multiples of the fundamental frequency at the selected time point (d) spectrum of selected time point, the amplitude at 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. 4-2. As the F0 is derived in our work based on frames of 20ms with 10 ms' overlap (see section 3.3), 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. 4-3 (c)).

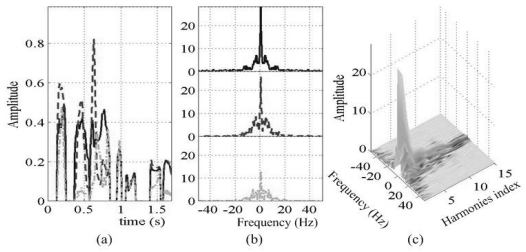

Fig. 4-3 The amplitude of the harmonic vectors in time domain and their spectrums (a) amplitude of the first 3 harmonic vectors, (b) the spectrums of the first 3 vectors, (c) 15 spectrums combined in 3-D harmonic space

The three axes in the 3-D harmonic space are amplitude, frequency and harmonics index Fig. 4-3 (c). In these three 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 vectors, and a ridge in the center of the frequency axis. The values in the side part of this space are relatively low.

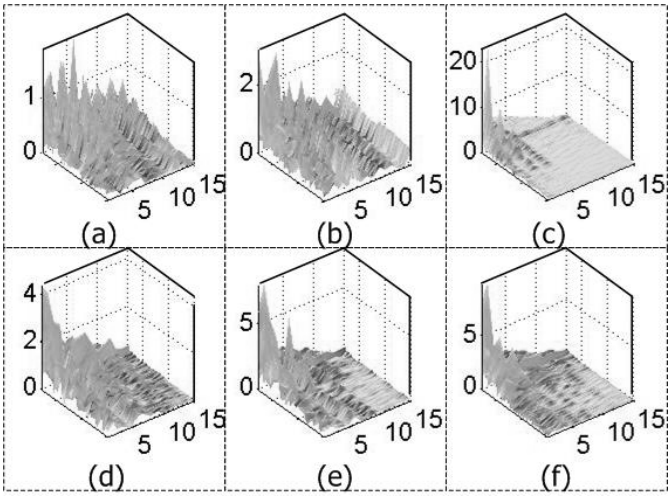

Fig. 4-4 3-D harmonic space for the 6 emotions from a same sentence: (a) anger, (b) fear, (c) sadness, (d) happiness, (e) neutral, (f) boredom

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

'anger' and 'happiness' emotions have relatively 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. On the other hand, 'sadness' and '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. '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. 4-5. 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.

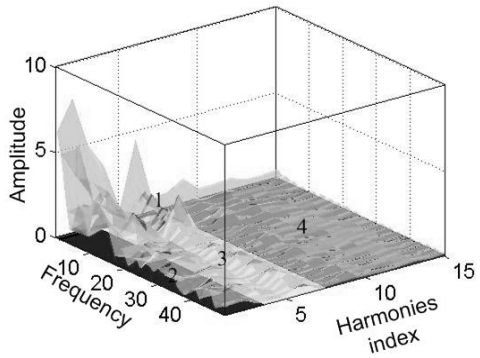

Fig. 4-5 4 areas for FFT result of 3-D harmonic space

We summarize the list of harmonic features as follows:

- Mean, maximum, variance and normalized variance of the 4 areas
- The ratio of mean values of areas 2 ~ 4 to area 1

4.2.2 Zipf features

Features derived from an analysis according to Zipf laws are presented in this group to better capture the prosody property [Bar06] of a speech signal by analyzing the structure of vocal signal in both time domain and frequency domain.

Zipf law is an empirical law proposed by G. K. Zipf [Zip49]. 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:

$$f(p) = \alpha r(p)^{-\beta} \qquad (4.3)$$

where α and β are real numbers.

The relation becomes linear when the logarithms 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. Therefore, we also compute the polynomial approximation on the resampled curve. It differs from Zipf graph as the distance between consecutive points is constant

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:

$$I(f) = \delta f^{\gamma} \qquad (4.4)$$

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 [Coh97], Zipf law has been evaluated to discriminate natural and artificial language texts; Havlin proved that [Hav95] 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 [Del04]. Three types of coding as temporal coding, frequency coding and time-scale coding were proposed in [Del04], in order to bring to the front different information contained in signals.

For example, the coding principle denoted as TC1 in [Del04] is to allow to build up a sequence of patterns based on the coding of the original audio signal. Three levels of coding are used to map the signal into text. In the first step as a low-level letter coding, 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 in the second step as letter-level coding, the letters U, F, and D are assembled by three character long sequences with totally 3^3=27 different possible patterns into a new 27 letters alphabet. 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. In the third step, word level coding, the pattern sequence can then be formed into words with given length. In the example of Fig. 4-6, the words length is set to five letters.

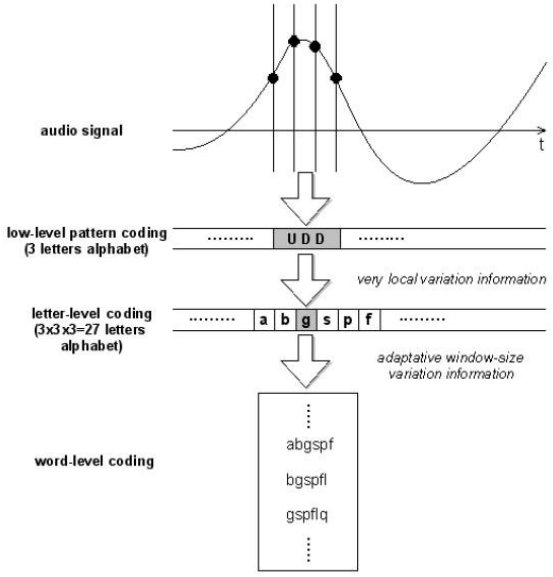

Fig. 4-6 Description of TC1 coding [Del04]

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

We summarize the list of Zipf features as follows:

- Entropy feature of Inverse Zipf of frequency coding
- Resampled polynomial estimation Zipf feature of UFD (Up – Flat - Down) coding

4.2.3 Conclusion

Two groups of new proposed features are extracted for presenting the speech emotions. When the traditional features including frequency, energy, and MFCC based features are considered together (section 3.3), there are totally 226 features that can be used for the classification of emotional states in speech.

4.3 Hierarchical Classification of emotional speech

Fuzzy neighborhood relationships between some emotional states, for instance between sadness and boredom, as evidenced by studies on acoustic correlates in Chapter 2, lead 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 (DEC) 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. Moreover, a gender classifier is also defined as the first step of our multi-stage emotion classification to further decrease the perturbations between different emotion classes.

4.3.1 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 2.1, 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 [Bur05] and DES dataset [Eng96], see section 3.1.1). Referring to these discrete emotion states in arousal vs. appraisal plane (Fig. 2-1, [Wie05a]), they can also be mapped into a 2-D emotional space as in Fig. 4-7: 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.

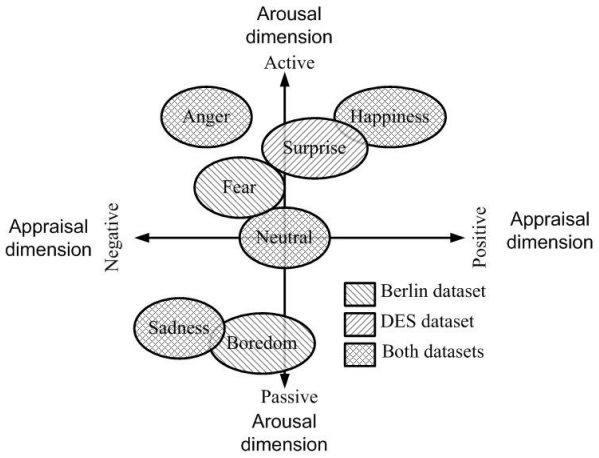

Fig. 4-7 Emotions mapped in the dimensional space

Fig. 4-8 illustrates such a hierarchical classification scheme with two stages [Xia06], 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.

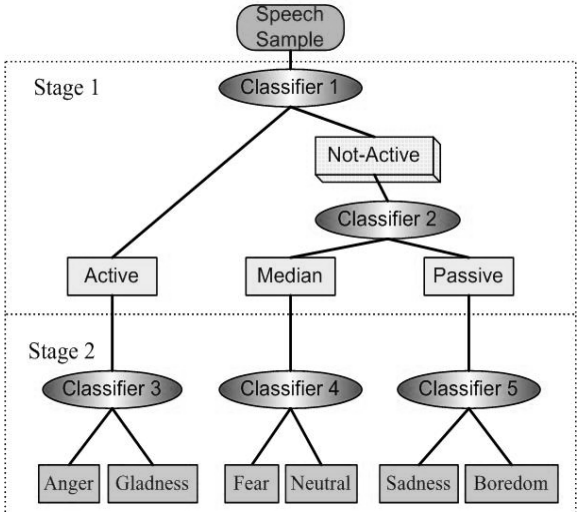

Fig. 4-8 Dimensional Emotion Classifier (DEC) on Berlin dataset: a two-stage hierarchical classification scheme of emotional speech driven by the dimensional emotion model

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 to discriminate non-linear data and as well as their generalization properties. We made use of BP (Back Propagation) neural networks with two 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 two classes by a threshold of 0.5.

4.3.1.1 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 states [Xia05]. In the first step, the active state (including anger and happiness) is separated from the median one (including fear and neutral) and the passive states (including sadness and boredom). This is performed by classifier 1 represented in Fig. 4-8. In the second step, the median state and the passive state are further separated (classifier 2 in Fig. 4-8).

4.3.1.2 Stage 2: classification in appraisal dimension

The first stage of classification in arousal dimension achieves an emotional classification into three rough states (Fig. 4-7). 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. 4-8, 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".

4.3.2 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 [Sla98] [Ver04b] [Ver05b] [Xia07a]. 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. 4-9 illustrates the final classification scheme, subsequently called Automatic Gender Recognition based DEC, which tops a gender classifier on two DEC schemes as defined in the previous section.

The approaches on gender classification have been studied in-depth in [Har03a] [Har04] [Har05b]. Our goal in this study is only to prove that the gender recognition can improve the overall classification rate for the emotional speech but not to make the best result for the gender classification that is already a solved problem. Thus, the same scheme in the sub-classifiers in emotion classification is also applied in gender classification in order to simplify the classification system, while the other algorithms special for gender discrimination can also replace this part to further improve the overall performance.

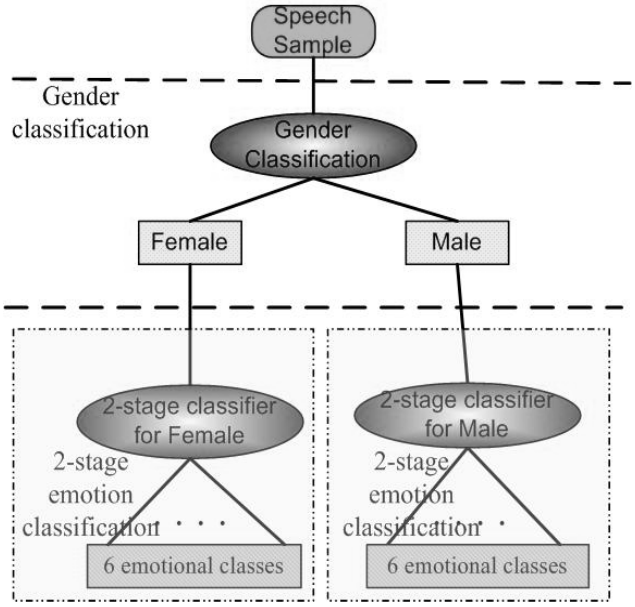

Fig. 4-9 Gender-Based DEC: a gender classification tops two DECs according to the detected gender

As we can see from this figure, the two two-stage Dimensional Emotion Classifiers (DEC) have the same structure (as shown in Fig. 4-8), but work with different feature sets according to the underlying gender information delivered by the gender classifier.

Any gender classifier might be used. We have chosen to build a gender classifier similar as the one defined in our previous work [Har05b] 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 3.3 and section 4.2): 41, 10, 23, 14, 69, 53, 15, 19, 36, 38, 26, 27, 30, 33, and 20 (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.

4.4 Experiments and results

The effectiveness of our approach is experimented on both Berlin dataset and DES datasets. In the following, we first introduce Berlin and DES datasets. Then, our experimental results are presented and discussed.

4.4.1 Introduction to the datasets

Two datasets available to public are tested in our work, namely Berlin dataset and DES dataset (section 3.1.1). These two datasets will be introduced in more details in the following subsections.

4.4.1.1 Berlin dataset

The 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 [Sen] [Bur05]. This database contains speech samples from five actors and five actresses, ten different sentences in German of seven kinds of emotions: anger, boredom, disgust, fear, happiness, sadness and neutral. The texts in the dataset are listed in Table. 4-1. There are totally 535 speech samples in this database, in which 302 speech samples are of female voice and 233 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.

Table. 4-1 Text in Berlin dataset [Sen]

code	text (in German)	English translation
a01	Der Lappen liegt auf dem Eisschrank.	The tablecloth is lying on the fridge.
a02	Das will sie am Mittwoch abgeben.	She will hand it in on Wednesday.
a04	Heute abend könnte ich es ihm sagen.	Tonight I could tell him.
a05	Das schwarze Stück Papier befindet sich da oben neben dem Holzstück.	The black sheet of paper is located up there besides the piece of timber.
a07	In sieben Stunden wird es soweit sein.	In seven hours it will be.
b01	Was sind denn das für Tüten, die da unter dem Tisch stehen?	What about the bags standing there under the table?
b02	Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter.	They just carried it upstairs and now they are going down again.
b03	An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht.	Currently at the weekends I always went home and saw Agnes.
b09	Ich will das eben wegbringen und dann mit Karl was trinken gehen.	I will just discard this and then go for a drink with Karl.
b10	Die wird auf dem Platz sein, wo wir sie immer hinlegen.	It will be in the place where we always store it.

The samples with the emotional state “disgust” are ignored in our experiment due to its inconsistency in acoustic features (see section 2.2). The other samples are

cut into segments of 2 seconds. The tail parts of the samples longer than 1.5 seconds are also kept. The numbers of segments obtained in the final experiment in this dataset of each of the emotional states are listed in Table. 4-2.

Table. 4-2 Numbers of segments for the emotional states in Berlin dataset

	Anger	Happiness	Fear	Neutral	Sadness	Boredom
Female	87	57	39	49	74	65
Male	75	34	39	46	49	49

4.4.1.2 DES dataset

The DES dataset is recorded for Center for PersonKomunikation (CPK), Aalborg University, Denmark, as a part of the VAESS project (Voices, Attitudes and Emotions in Speech Synthesis) [Eng96]. 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 as listed in Table. 4-3.

Table. 4-3 Gender and age for the four actors used in collecting DES [Eng96]

Initials	Gender	Age
DHC	Female	34
KLA	Female	52
JZB	Male	38
HO	Male	52

Five emotions are considered in this dataset: Neutral, Surprise, Happiness, Sadness, and Anger. For each emotion, each actor recorded following segments:

- 2 single words
- 9 sentences
- 2 passages of fluent speech

The single words and sentences are taken as test set in our experiment. As the passages are much longer than the words and the sentences, and the emotions contained in these passages are not as typical as in the words and sentences, they are ignored in our test. The texts and translation are listed in Table. 4-4.

Table. 4-4 Words and sentences in DES [Eng96]

Words	Text	Translation
1	Ja.	Yes.
2	Nej.	No.
Sentences		
1	Du er en sød dreng.	You are a nice boy.
2	Jeg er ikke sulten.	I am not hungry.
3	Jeg ved det heller ikke.	I don't know either.
4	Hvad er det?	What is this?
5	Hvor er du?	Where are you?
6	Hvor skal du hen?	Where are you going?
7	Kom med mig!	Come with me!
8	Kommer du her igen?	Is it you again?
9	Jeg synes vi mangler nogle, som er lidt længere.	I think we need some that are a little longer.

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 [Xia07a].

4.4.2 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 considered 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 eight samples of “disgust” in the male samples, which is much less than the other types and the acoustic feature for this emotion is inconsistent [Ban96], 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.

4.4.2.1 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 [Rak05]. 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. 4-5.

Table. 4-5 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 matrices with the highest recognition rates are listed in Table. 4-6 and Table. 4-7. 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. 4-6 Confusion matrix of the global classifier with frequency and energy features with TANAGRA (%)

	Predicted / Actual	Anger	Happiness	Fear	Neutral	Sadness	Boredom
Female	Anger	67.55	23.67	6.24	1.39	0.00	1.15
	Happiness	35.56	45.60	12.50	3.70	0.00	2.64
	Fear	14.87	23.85	37.18	12.31	4.87	6.92
	Neutral	0.20	1.22	3.47	61.43	3.27	30.41
	Sadness	0.00	0.00	0.27	8.02	86.96	4.76
	Boredom	2.00	1.85	6.62	30.92	8.15	50.46
Male	Anger	82.67	8.80	6.80	0.67	0.53	0.53
	Happiness	36.18	39.12	20.00	3.53	0.00	1.18
	Fear	12.82	10.26	55.38	11.54	6.92	3.08
	Neutral	1.52	3.26	5.00	48.91	10.87	30.43
	Sadness	0.20	0.82	2.86	10.61	61.22	24.29
	Boredom	1.84	1.43	1.84	27.96	26.73	40.20
Mixed	Anger	71.82	21.71	4.73	0.46	0.00	1.27
	Happiness	43.31	41.02	8.63	3.52	0.35	3.17
	Fear	13.85	25.90	38.72	6.92	7.44	7.18
	Neutral	1.84	1.22	3.27	57.14	6.73	29.8
	Sadness	0.00	0.41	1.63	5.84	80.16	11.96
	Boredom	2.62	2.46	3.54	24.00	12.31	55.08

Table. 4-7 Confusion matrix of the global classifier with all features (FES+harmonic+Zipf features) (%)

	Predicted / Actual	Anger	Happiness	Fear	Neutral	Sadness	Boredom
Female	Anger	73.44	21.71	2.66	0.46	0.12	1.62
	Happiness	38.03	50.53	6.51	1.94	2.11	0.88
	Fear	12.56	23.59	41.79	5.9	10.51	5.64
	Neutral	1.02	1.02	0.61	60.00	6.12	31.22
	Sadness	0.00	0.14	0.68	5.30	86.68	7.20
	Boredom	1.69	1.23	2.00	22.62	8.77	63.69
Male	Anger	84.93	9.33	5.33	0.27	0.00	0.13
	Happiness	30.88	46.18	17.65	2.94	0.88	1.47
	Fear	11.03	18.72	55.38	7.44	6.15	1.28

	Neutral	3.26	0.65	3.26	58.04	7.39	27.39
	Sadness	0.00	1.63	3.06	4.29	73.27	17.76
	Boredom	1.22	0.20	1.22	22.65	24.49	50.20
Mixed	Anger	74.57	18.40	5.19	1.05	0.00	0.80
	Happiness	38.81	44.76	11.14	1.76	1.76	1.76
	Fear	10.53	15.4	56.48	5.52	8.99	3.08
	Neutral	0.95	1.48	2.63	62.7	5.69	26.55
	Sadness	0.00	0.49	2.04	4.89	79.97	12.62
	Boredom	1.50	1.14	2.73	23.66	14.95	56.02

4.4.2.2 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. 4-8). 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. 4-8 where the superscript indicates the feature group number which a selected feature comes from.

Table. 4-8 Selected features and recognition rates for the sub-classifiers (The groups of the features are marked with superscript, "F" stands for frequency features, "E" stands for energy features, "H" stands for harmonic features, and "Z" stands for Zipf features)

		Selected feature subset (Ordered by the sequence of selection)	Recognition rate (%)
Active vs. non-active	Female	225^{Z}, 21^{H}, 50^{E}, 16^{H}, 51^{E}, 4^{H}, 43^{E}, 6^{H}, 53^{E}	91.13
	Male	46^{E}, 26^{F}, 31^{F}, 36^{F}, 5^{H}, 4^{H}, 34^{F}, 21^{H}, 225^{Z}, 29^{F}	92.32
	Mixed	10^{H}, 226^{Z}, 50^{E}, 23^{F}, 36^{F}, 51^{E}, 53^{E}, 54^{E}, 67^{E}, 27^{F}, 21^{H}, 52^{E}	90.31
Median vs. Passive	Female	21^{H}, 26^{F}, 52^{E}, 51^{E}, 11^{H}, 6^{H}, 22^{H}, 28^{F}, 10^{H}, 4^{H}, 23^{F}, 46^{E}	84.98
	Male	22^{H}, 225^{Z}, 31^{F}, 10^{H}, 16^{H}, 7^{H}, 6^{H}, 27^{F}, 43^{E}, 51^{E}, 11^{H}	88.23
	Mixed	22^{H}, 53^{E}, 52^{E}, 11^{H}, 21^{H}, 6^{H}, 51^{E}, 57^{E}	84.73

Anger vs. Happiness	Female	28^{F}, 29^{F}, 20^{H}, 26^{F}, 6^{H}, 57^{E}, 11^{H}, 46^{E}	80.21
	Male	46^{E}, 58^{E}, 21^{H}, 31^{F}, 64^{E}, 15^{H}, 53^{E}, 24^{F}, 36^{F}, 40^{F}	85.37
	Mixed	226^{H}, 56^{E}, 40^{F}, 31^{F}, 35^{F}, 6^{H}, 10^{H}, 12^{H}, 21^{H}, 60^{E}	80.62
Fear vs. Neutral	Female	26^{F}, 5^{H}, 63^{E}, 31^{F}, 48^{E}	90.85
	Male	20^{H}, 15^{H}, 37^{F}, 6^{H}, 11^{H}, 69^{E}, 4^{H}	92.85
	Mixed	26^{F}, 72^{E}, 63^{E}, 69^{E}, 74^{E}, 35^{F}, 15^{H}, 46^{E}, 75^{E}, 67^{E}, 7^{H}, 64^{E}, 10^{H}	84.31
Sadness vs. Boredom	Female	27^{F}, 225^{Z}, 30^{F}, 46^{E}, 41^{F}, 31^{F}, 73^{E}, 38^{F}, 24^{F}, 46^{E}, 21^{H}, 8^{H}, 35^{F}, 10^{H}	92.88
	Male	14^{H}, 75^{E}, 42^{F}, 44^{E}, 18^{H}, 73^{E}, 15^{H}, 12^{H}	91.30
	Mixed	27^{F}, 31^{F}, 33^{F}, 22^{H}, 35^{F}, 55^{E}, 75^{E}, 11^{H}, 66^{E}, 30^{F}, 46^{E}, 7^{H}, 38^{F}	89.26

From Table. 4-8, we can see that frequency features and energy features deliver standard performance for the five sub-classifiers. While frequency features is more efficient in classifier 3 ("anger" vs. "happiness") and classifier 5 ("sadness" vs. "boredom"), harmonic features 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 21 (the ratio of mean values of areas 3 to area 1 in harmonic space) shows very high discriminative power in stage 1 – arousal classification (separating the 3 states), but less efficient in stage 2 – appraisal classification. Although there are only two 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. 4-9.

Table. 4-9 Mean confusion matrix achieved by DEC (%)

	Predicted / Actual	Anger	Happiness	Fear	Neutral	Sadness	Boredom
Female	Anger	83.43	13.76	3.76	3.61	1.67	2.12
	Happiness	19.13	69.00	8.63	3.38	1.38	5.38
	Fear	8.71	11.47	**73.45**	5.61	3.88	4.23

	Neutral	2.01	4.51	5.01	75.75	2.51	17.76
	Sadness	1.83	1.83	2.69	6.97	91.43	4.40
	Boredom	2.99	2.10	1.88	14.55	3.66	83.11
Male	Anger	89.33	9.12	3.46	2.79	1.79	2.46
	Happiness	18.80	65.00	13.80	4.63	1.30	2.97
	Fear	1.96	10.04	**78.85**	5.43	6.58	5.04
	Neutral	2.46	2.72	3.78	83.68	4.56	11.14
	Sadness	1.86	1.86	1.86	4.21	92.94	6.57
	Boredom	2.07	2.66	1.78	7.66	5.90	88.82
Mixed	Anger	85.35	11.16	3.91	4.39	1.71	2.02
	Happiness	25.15	61.88	10.46	5.93	1.24	1.55
	Fear	12.38	12.38	**55.27**	15.11	4.75	5.66
	Neutral	2.47	2.72	7.21	78.33	4.01	13.11
	Sadness	2.42	2.80	2.61	7.80	82.31	10.30
	Boredom	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 and 4.78% higher than the result for mixed speech samples (68.60%). From Table. 4-9, we can see that the mixing of the gender cause more misjudgment for the emotion “fear” than for the other emotions.

4.4.2.3 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 4.3.2. The confusion matrix of the multi-stage classification is listed in Table. 4-10. 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. 4-10 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

4.4.2.4 Synthesis and Discussion

Table 4-11 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 two 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 4-11. Synthesis of recognition rates by the four classifiers (%). F: frequency-based features, E: energy-based feature, H: harmonic features, Z: Zipf features

	Male	Female	Average of the 2 genders	Mixed	Mixed with gender info
Global: F+E	57.91±2.56	61.03±1.89	59.55	60.38±2.26	--
Global: F+E+H+Z	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 discriminative power 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.

4.4.3 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 classification problem with six emotions, we derived the following hierarchical classification scheme as illustrated in Fig. 4-10. This process 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 obtain the final emotion states.

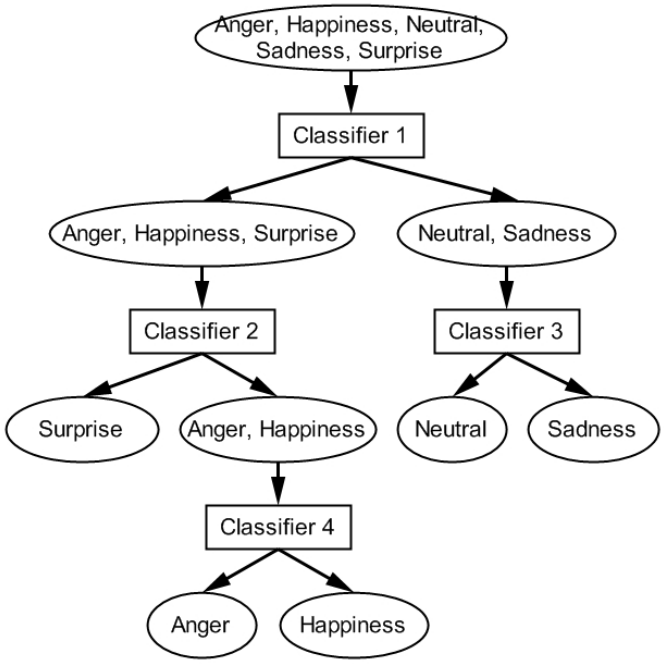

Fig. 4-10 DEC on DES dataset

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 4-12 summarizes the accuracy rates, and Table 4-13 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* [Ver05b].

Table 4-12. Accuracy rates on DES dataset (%)

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

Table 4-13. Confusion matrix on DES dataset (%)

	Predicted / Actual	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**

4.5 Conclusion

In this chapter, 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 that we propose to deal with in the following chapters.

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 the intuitive mapping of the discreet emotion states into the dimensional emotion model. In this work, this intuitive mapping was thus made manually and empirically. An automatic mapping scheme is clearly needed especially when the number of emotions increases and their types vary.

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. We thus address the issue of ambiguous or multiple judgments in Chapter 6 where we develop an automatic algorithm by representing probability of the emotional states using the belief masses according to the evidence theory.

But first of all, we describe in the next chapter a new embedded feature selection scheme, called ESFS for Embedded Sequential Forward Selection, that we have developed. It relies on both the evidence theory and SFS. This scheme is indeed the fundamental preprocessing step used for an efficient learning process proposed in Chapter 5 and Chapter 6.

Chapter 5

An Automatically Multi-stage Classification of Emotional Speech: HCS

The purpose of this chapter is to address the first major issue that we have identified in Chapter 4, namely manual and empiric nature of dimensional model driven hierarchical classifier of emotion analysis. To this end, we propose an evidence theory based feature selection scheme for an automatic approach of hierarchical classification scheme. This chapter is organized as follows. We first state the problem and introduce our approach. Then, we define ESFS, an Embedded evidence theory based Feature Selection Scheme we have developed that is used as the basis for automatic derivation of speech emotion classifier hierarchy HCS (Hierarchical Classification Scheme), which is then described. Finally, experimental results obtained by our classifier on two datasets are presented.

5.1 The problem and our approach

In the previous chapter, we showed that a hierarchical classification approach could greatly improve the classification rate as compared to a global approach. The DEC scheme we have developed and presented in the previous chapter implements this concept. However, the classification hierarchy was built manually using an empirical mapping of a discrete emotion model into the dimensional one. Consequently, its generalization to a new vocal emotion classification problem inevitably leads to a manual work of adaptation.

In order to tackle this difficulty, we propose in this chapter an automatic scheme for building this classifier hierarchy which is represented by a binary tree whose root is the union of all emotion classes, leaves are single emotion classes and nodes are subsets containing several emotion classes obtained by a sub_classifier. Each of these sub_classifiers is based on a feature combination relying on the evidence theory that is a powerful method for merging several information sources. This approach allows to easily represent classifiers characterized by their mass function, which is the combination of the information given by an appropriate feature subset, each sub-classifier having its own one.

As the feature selection process is crucial to ensure good classification efficiency and is at the heart of our automatic approach for building the hierarchical classification scheme, we also introduce a new embedded feature selection method based on SFS. This method makes use of the evidence theory in order to better select and merge the features for improving the classification accuracy.

5.2 ESFS: a new feature selection method based on SFS and the evidence theory

Feature subset selection is an important subject when training classifier in Machine Learning (ML) problems. Too many input features in a ML problem may lead to the so-called "curse of dimensionality" [Koj00], which describes the problem of exponential sample complexity with the increasing number of features. Practical ML algorithms are known to degrade in prediction accuracy when faced with many features that are not necessary.

The redundant and irrelevant features often yield the opposite of intended as slowed execution, less understandable results, and much reduced accuracy [Hal97]. The objective of feature selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and gaining a deeper insight into the underlying processes that generated the data.

In our case of emotional speech analysis, we have been using Berlin and DES datasets for learning and testing. Whereas more than 226 features were generated in order to characterize the various aspects of vocal emotion, Berlin dataset only contains 500 speech samples over seven emotional classes and DES dataset has the

same order of magnitude. Clearly, a feature selection scheme is needed in combination with a supervised classifier.

5.2.1 Related work

A feature selection method thus aims at finding the most relevant features. There exist considerable works in the literature on the question. Interesting overviews include [Koh97] [Guy03] [Sae07]. However, the relevance notion is not perfectly defined and may depend on the feature selection method. One of these definitions [Blu97] is to consider that a feature f is relevant if it is incremental useful to a learning algorithm L with respect to a feature subset S: the accuracy that L produces an hypothesis using the feature set $f \cup S$ is higher than the accuracy achieved only using S. In the case of classification problems, the accuracy can be the correct classification rate.

Feature selection methods can be split into three main categories according to the dependence to the classifiers: filter approaches, wrapper approaches and embedded approaches [Koj00] [Seb04]. 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 [Hal97]. In wrapper methods, the good feature subsets are selected by using the induction algorithm itself. The criterion of the selection is the optimization of the accuracy rate [Koh97]. In embedded methods, the selection stage and the classification stage are not separated. The selection of an optimal feature subset is included into the classifier construction [Blu97].

Filter methods include Relief method [Kir92] [Aro04], Focus algorithm [Alm91], *etc.* [Blu97]. Their main advantage is their low computational complexity that makes them very fast. Their main drawback is that they are not optimized to be used with a particular classifier as they are completely independent of the classification stage.

Wrapper methods on the contrary evaluate feature subsets with the classification algorithm in order to measure their efficiency according to the correct classification rate [Koh97]. Thus, feature subsets are generated thanks to some search strategy, and the feature subset that leads to the best correct classification rate is kept. Among algorithms having very high performance, Genetic Algorithm (GA) methods

are widely used. However, wrapper methods can be computationally burdensome and can make the problems with a very large feature set intractable [Kor04], since for a feature set with n features, n! subsets have to be evaluated in order to find out the best subset. Fast algorithms are thus needed to be applied in the wrappers. The SFS method (Sequential Forward Selection) and SBS method (Sequential Backward Selection) are two fast wrapper methods that are quite popular. The SFS starts from an empty feature subset, and the subsets are constructed sequentially by adding an additional feature at each iteration. The subset S_k with k features is constructed by adding to the subset S_{k-1} with k-1 features a single feature which gives the optimal performance for the new subset. The number of subsets which have to be evaluated is reduced to n*(n+1)/2 instead of n!. The SBS works in the opposite way and start from the whole feature set. A single feature is removed from the subsets at each iteration during the search of the best subset. Modifications of these methods such as so-called floating sequential selection methods (SFFS, Sequential Forward Floating Selection, and SBFS, Sequential Backward Floating Selection) have been proposed [Pud94]. In SFFS, a backward step is added at each iteration in the forward selection. Although sequential selection methods gain in computational complexity, they remain sub-optimal and the selected subsets do not necessary include the optimal one [Spe98].

In embedded feature selection methods, similarly to wrapper methods, the feature selection is linked to the classification stage, this link being in this case much stronger as the feature selection in embedded methods is included into the classifier construction. Recursive partitioning methods for decision trees such as ID3, C4.5 and CART [Tso90] are examples of such method. Indeed, a search through the space of decision trees at each stage is performed, using an evaluation function to select the attribute that has the best ability to discriminate among the classes. The process of partition the training data based on this attribute is repeated on each subset extending the tree downward until no further discrimination is possible. Embedded methods offers the same advantages as wrapper methods concerning the interaction between the feature selection and the classification and moreover present a better computational complexity since the selection of features is directly included in the classifier construction during training process.

In our work, we introduce a new embedded feature selection method we have developed and called ESFS. It is inspired from the wrapper method SFS since it relies

on the simple principle to add incrementally most relevant features, and makes use of mass functions that are introduced from the evidence theory. This process allows to merge elegantly feature information in an embedded way, leading to a lower computational cost than original SFS.

A brief introduction of the evidence theory is proposed in next subsection, followed by the presentation of feature selection scheme.

5.2.2 Introduction to the evidence theory

In our feature selection scheme, the term "belief mass" from the evidence theory is introduced into the processing of features.

Dempster and Shafer wanted in the 1970's to calculate a general uncertainty level from the Bayesian theory. They developed the concept of "uncertainty mapping" to measure the uncertainty between a lower limit and an upper limit [Dem67] [Dem68]. Similar to the probabilities in the Bayesian theory, they presented a combination rule of the belief masses (or mass function) m().

The evidence theory was completed and presented by Shafer in [Sha76]. It relies on the definition of a set of n hypothesis Ω which have to be exclusive and exhaustive. In this theory, the reasoning concerns the frame of discernment 2^{Ω} which is the set composed of the 2^n subsets of Ω [Sha90] [Sha92] [Fio04] [Les03]. In order to express the degree of confidence we have in a source of information for an event A of 2^{Ω}, we associate to it an elementary mass of evidence $m(A)$.

The elementary mass function or belief mass that represents the chance of being a true statement is defined as:

$$m: 2^{\Omega} \rightarrow [0, 1] \qquad (5.1)$$

which satisfies

$$m(\Phi) = 0 \text{ and } \sum_{A \subseteq 2^{\Omega}} m(A) = 1 \qquad (5.2)$$

The belief function is defined if it satisfies $Bel(\Phi)=0$ and $Bel(\Omega)=1$ and for any collection $A_1 \ldots A_n$ of subsets of Ω

$$Bel(A_1 \cup ... \cup A_n) \geq \sum_{\substack{I \subseteq \{1...n\} \\ I \neq \phi}} (-1)^{|I|+1} Bel(\bigcap_{i \in I} A_i) \tag{5.3}$$

The belief function shows the *lower* bound on the chances, and it corresponds to the mass function with the following formulae

$$\begin{aligned} Bel(A) &= \sum_{B \subseteq A} m(B) \text{ for all A} \subset \Omega \\ m(A) &= \sum_{B \subseteq A} (-1)^{|A-B|} Bel(B) \end{aligned} \tag{5.4}$$

where $|X|$ means the number of elements in the subset.

The doubt function is defined as

$$Dou(A) = Bel(\neg A) \tag{5.5}$$

and the *upper* probability function is defined as

$$P^*(A) = 1 - Dou(A) \tag{5.6}$$

The true belief in *A* should be between *Bel*(*A*) and $P^*(A)$.

The Dempster's combination rule can combine two or more *independent* sets of mass assignments by using an orthogonal sum. For the case of two mass functions, let m_1 and m_2 be mass functions on the same frame 2^{Ω}, the orthogonal sum is defined as $m = m_1 \oplus m_2$, to be $m(\phi) = 0$, and

$$\begin{aligned} m(A) &= K \sum_{X \cap Y = A} m_1(X) \bullet m_2(Y) \\ K &= \frac{1}{1 - \sum_{X \cap Y = \phi} m_1(X) \bullet m_2(Y)} \end{aligned} \tag{5.7}$$

For the case with more than two mass functions, let $m = m_1 \oplus ... \oplus m_n$, it satisfies $m(\phi) = 0$ and

$$m(A) = K \sum_{\bigcap A_i = A} \prod_{1 \le i \le n} m_i(A_i)$$
$$K = \frac{1}{1 - \sum_{\bigcap A_i = \phi} \prod_{1 \le i \le n} m_i(A_i)} \qquad (5.8)$$

The evidence theory is applied in two aspects in this thesis. First, the mass function is used to represent the features in the feature selection and classification ESFS (section 5.2.3). Second, in the ambiguous approach of automatic classification, the Dempster's combination rule is applied for combining the sub-classifiers (Chapter 6).

For the first aspect of the application of the evidence theory in the ESFS, the mass functions of the acoustic features are obtained from the probability densities computed using the training samples. In this particular case, the probabilities and the Bayesian method may also be applied in our feature selection scheme. However, thanks to the evidence theory, ESFS scheme could be applied on other classification problems even if the prior probability densities cannot be obtained by the training data. The mass functions could then be built from other information contained in signals. For the second aspect of the application of the evidence theory, in the ambiguous classification, the Bayesian method cannot be used instead of the evidence theory for the following reasons. First, the classes considered in the sub-classifiers do not necessary contain all the emotional classes, while the Bayesian method needs a complete set of hypotheses [Wiki]. Second, in the fusion step of the ambiguous approach, the Dempster's combination rule is applied, and the concept of conflict K (equation (5.8)) does not exist in the Bayesian method. The details of the ESFS and the ACS are introduced in section 5.2.3 and Chapter 6.

5.2.3 ESFS scheme

An exhaustive search of the best subset of features leads to the exploration of a space of 2^n subsets, which is unconceivable in practice. Thus, we have to consider a heuristic method for the feature selection. We have chosen to base our approach on the SFS method as it presents interesting properties of efficiency and simplicity. For this classifier dependent sub-optimal selection method, we have provided in this work two innovations. First, the range of subsets to be evaluated in the forward process is extended to multiple subsets for each size, and the feature set is reduced according to

a certain threshold before the selection in order to decrease the computational burden caused by the extension of the subsets in the evaluation. Second, since the SFS is a classifier dependent method, the concept of belief masses that comes from the evidence theory is introduced to consider the audio feature as a classifier that leads to an embedded feature selection method.

5.2.3.1 The principle

A heuristic feature selection algorithm can be characterized by its stance on four basic issues that determine the nature of the heuristic search process [Blu97]. First, one must determine the starting point in the space of feature subsets, which influences the direction of search and operators used to generate successor states. The second decision involves the organization of the search. As an exhaustive search in a space of 2^n feature subsets is impractical, one needs to rely on a more realistic approach such as greedy methods to traverse the space. At each point of the search, one considers local changes to the current state of the features, selects one and iterates. The third issue concerns the strategy used to evaluate alternative subsets of features. Finally, one must decide on some criterion for halting the search. In the following, we bring our answers to the previous four questions.

The SFS algorithm 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. In the original algorithm of SFS, there are totally n*(n+1)/2 subsets which need to be evaluated and the optimal subset may be missing in the searching.

In order to avoid departure too far from the optimal performance, we proposed an improvement of the original SFS method by extending the subsets to be evaluated. In each step of forward selection, instead of keeping only one subset for each size of subsets, a threshold is set according to the compromise between the performance and the computational burden (which is decided from the performance from experiments with a small amount of data in our work) and all the subsets with the performance above the threshold are kept to enter the evaluation in the next step. Since remaining multiple subsets in each step may lead to heavy computational burden, only the features selected in the first step (subsets with single feature), thus having the best

abilities to discriminate among classes that occur in the training data, are used in the evaluation in posterior steps.

As the features are added to the potential subsets one by one in the SFS process, the forward process of creating a feature subset with size k can be seen as a combination between two elements: a subset with size k-1 and a single feature. Thus, if we consider each subset as a feature itself, the process of creating a new feature subset can be interpreted as generating a new feature from two features.

A wrapper feature selection scheme such as the SFS needs to specify a classifier in order to evaluate improvement of classification accuracy as feature selection criterion. In our case, the classifier used in this feature selection method is simply based on the belief masses (5.1) and (5.2) of the features which are modeled from the distribution of the features for each class obtained from the training data. The belief masses of samples in the testing sets are calculated with the model of the belief masses. The class with the highest belief mass is then taken as the output of the classification. This classifier is repeated for every subset in evaluation for searching the best feature subset.

For summary, our embedded SFS using belief masses includes the following steps:

- Computation of the belief masses of the single features from the training set.
- Evaluation and ordering of the single features to decide the initial set for potential features.
- Combination of features for the generation of the feature subsets, making use of combination operators.
- Selection of the best feature subset.

These steps are detailed in next subsection.

5.2.3.2 Feature selection procedure

The feature selection procedure is introduced in this section with four steps. This procedure is illustrated in Fig. 5-1.

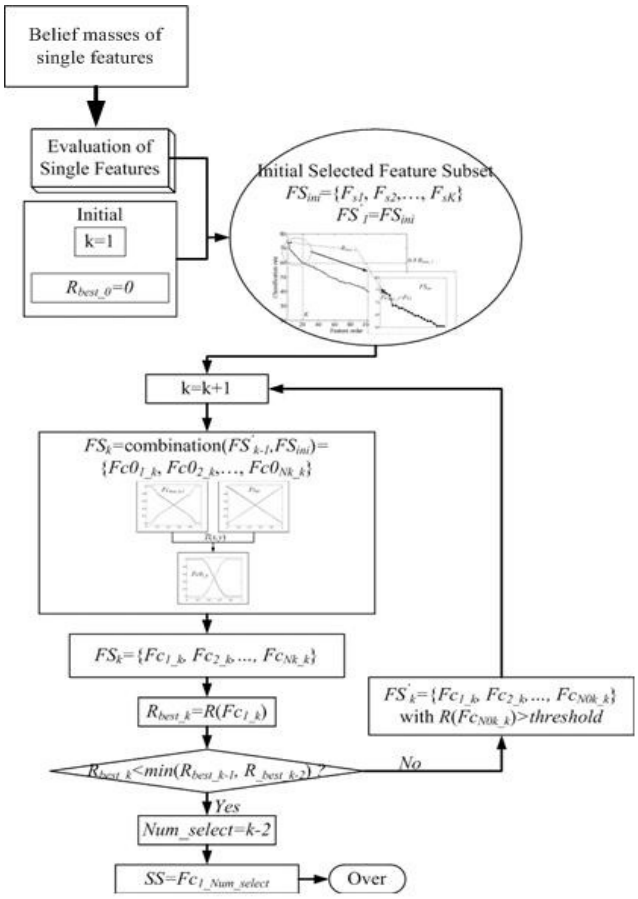

Fig. 5-1 Procedure of feature selection

Step 1: Calculation of the belief masses of the single features

Before the feature selection starts, all features are normalized into [0, 1]. For each feature,

$$Fea_n = \frac{Fea_{n0} - \min(Fea_{n0})}{\max(Fea_{n0}) - \min(Fea_{n0})} \qquad (5.9)$$

where Fea_{n0} is the set of original value of the n^{th} feature, and Fea_n is the normalized value of the n^{th} feature.

By definition of the belief masses, the mass can be obtained by different ways that can represent the chance for a statement to be true. In our work, the PDFs (probability density functions) of the features of the training data are used for calculating the masses of the single features.

The curves of PDFs of the features are obtained by applying polynomial interpolation to the statistics of the distribution of the feature values from the training data.

Taking the case of a 2-class classifier as example, the classes are defined as subset A and subset A^C. First, the probability densities of the features in each of the 2 subsets are estimated from the training samples by the statistics of the values of the features in each class. We define the probability density of the k^{th} feature Fea_k in subset A as $Pr_k(A, f_k)$ and the probability density in subset A^C as $Pr_k(A^C, f_k)$, where the f_k is the value of the feature Fea_k (Fig. 5-3 (a)). According to the probability densities, the masses of feature Fea_k on these 2 subsets can be defined to meet the requirement in (5.2) as

$$m_k(A, f_k) = \frac{\Pr_k(A, f_k)}{\Pr_k(A, f_k) + \Pr_k(A^C, f_k)}$$
$$m_k(A^C, f_k) = \frac{\Pr_k(A^C, f_k)}{\Pr_k(A, f_k) + \Pr_k(A^C, f_k)} \qquad (5.10)$$

where at any possible value of the k^{th} feature f_k, $m_k(A, f_k)+m_k(A^C, f_k)=1$.

In the case of N classes, the classes are defined as $A_1, A_2,\ldots,A_N$. The masses of feature F_k of the i^{th} class A_i can be obtained as

$$m_k(A_i, f_k) = \frac{\Pr_k(A_i, f_k)}{\sum_{n=1}^{N} \Pr_k(A_n, f_k)} \qquad (5.11)$$

which satisfies

$$\sum_{i=1}^{N} m_k(A_i, f_k) = 1 .$$

We take a simple example with discrete features with the following problem of "and" function. In this example we have four samples to be classified into two classes with two features. The value of the features and the ground truth of the output for the 4 samples are listed in Table. 5-1.

Table. 5-1 Features and classes in the example of "and" function

Samples	S_1	S_2	S_3	S_4
Feature 1	0	0	1	1

Feature 2	0	1	0	1
Output	0	0	0	1
Class	A	A	A	A^C

In this two class problem, $\Omega=\{A, A^C\}$, where class A corresponds to the subset of samples with output value of 0, and class A^C corresponds to the subset of samples with output value of 1. We first estimate the probability density of the two features as

$$\Pr_1(A, f_1)=\begin{cases}\frac{2}{3}, & , f_1=0\\ \frac{1}{3}, & , f_1=1\end{cases},\quad \Pr_1(A^C, f_1)=\begin{cases}0, & , f_1=0\\ 1, & , f_1=1\end{cases} \tag{5.12}$$

$$\Pr_2(A, f_2)=\begin{cases}\frac{2}{3}, & , f_2=0\\ \frac{1}{3}, & , f_2=1\end{cases},\quad \Pr_2(A^C, f_2)=\begin{cases}0, & , f_2=0\\ 1, & , f_2=1\end{cases} \tag{5.13}$$

Then we calculate the mass function of the two features:

$$m_1(A, f_1)=\frac{\Pr_1(A, f_1)}{\Pr_1(A, f_1)+\Pr_1(A^C, f_1)}=\begin{cases}\frac{2/3}{2/3+0}=1, & , f_1=0\\ \frac{1/3}{1/3+1}=\frac{1}{4}, & , f_1=1\end{cases}$$

$$m_1(A^c, f_1)=\frac{\Pr_1(A^c, f_1)}{\Pr_1(A, f_1)+\Pr_1(A^C, f_1)}=\begin{cases}\frac{0}{2/3+0}=0, & , f_1=0\\ \frac{1}{1/3+1}=\frac{3}{4}, & , f_1=1\end{cases} \tag{5.14}$$

$$m_2(A, f_2)=\frac{\Pr_1(A, f_2)}{\Pr_1(A, f_2)+\Pr_1(A^C, f_2)}=\begin{cases}\frac{2/3}{2/3+0}=1, & , f_2=0\\ \frac{1/3}{1/3+1}=\frac{1}{4}, & , f_2=1\end{cases}$$

$$m_2(A^c, f_2)=\frac{\Pr_1(A^c, f_2)}{\Pr_1(A, f_2)+\Pr_1(A^C, f_2)}=\begin{cases}\frac{0}{2/3+0}=0, & , f_2=0\\ \frac{1}{1/3+1}=\frac{3}{4}, & , f_2=1\end{cases} \tag{5.15}$$

These mass functions satisfy

$$m_1(A, f_1) + m_1(A^C, f_1) = \begin{cases} 1+0=1, & , f_1 = 0 \\ \frac{1}{4}+\frac{3}{4}=1, & , f_1 = 1 \end{cases}$$
$$m_2(A, f_2) + m_2(A^C, f_2) = \begin{cases} 1+0=1, & , f_2 = 0 \\ \frac{1}{4}+\frac{3}{4}=1, & , f_2 = 1 \end{cases} \quad (5.16)$$

The two features are actually symmetrical to each other in this example. The probability density and the masses of the two features are illustrated in Fig. 5-2.

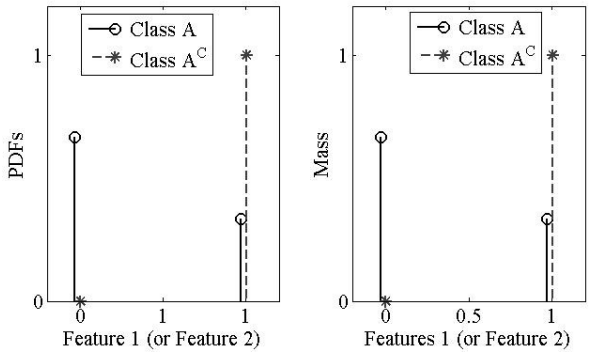

Fig. 5-2 The probabilities and masses of the two features in the example

An example of the pdfs and masses of a single feature in our application is shown in Fig. 5-3. The feature considered in this figure is the mean value of the fundamental frequency derived from the Berlin dataset with female utterances. The two classes considered are A_2={happiness}, A_2^C={anger, fear, neutral, sadness, boredom}.

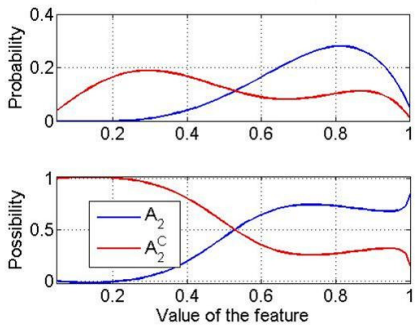

Fig. 5-3 Example of probability and evidence masses for single feature,
(a) probability of the feature (b) belief masses of the feature

Step 2: evaluation of the single features and selection of the initial set for potential features

When the distribution model of the belief masses of the single features on the classes in the classification have been extracted from the training data, the single features are evaluated by passing the distribution model derived from the training data. For each sample, its belief mass value can be extracted for each data sample from its mass function with the features. The samples are assigned to the class that has the highest belief mass and thus performances of correct classification rates can be obtained.

Come back to our simple example with "and" function. The belief masses for the four samples to the two classes according to each of the two features and the correct classification rates of the single features in this example are listed in Table. 5-2. In this case, the correct classification rates for both of the features are 75%.

Table. 5-2 Classification of single features in the example of "and" function

Samples			S_1	S_2	S_3	S_4	Classification rate
Ground truth			A	A	A	A^C	
Feature 1	Feature value		0	0	1	1	75%
	Mass	A	1	1	¼	¼	
		A^C	0	0	¾	¾	
	Predicted class		A	A	A^C	A^C	
Feature 2	Feature value		0	1	0	1	75%
	Mass	A	1	¼	1	¼	
		A^C	0	¾	0	¾	
	Predicted class		A	A^C	A	A^C	

With this method, the single features can then be ordered by the correct classification rate with the mass function and thus the best features can be selected.

The features are ordered in descending order according to the correct classification rates $R_{single}(F_k)$ as $\{F_{s1}, F_{s2},\ldots, F_{sN}\}$, where N means the total number of features in the whole feature set.

In order to reduce the computational burden in the feature selection, an initial feature set FS_{ini} is constructed with the best K features in the re-ordered feature set according to a certain threshold in classification rates as

$$FS_{ini}=\{F_{s1}, F_{s2},\ldots, F_{sK}\} \qquad (5.17)$$

The threshold of the classification rates is decided according to the best classification rate as:

$$R_{single}(F_{s\_K}) \geq thres\_1 * R_{best\_1} \qquad (5.18)$$

Where $R_{best\_1}=R_{single}(F_{s\_1})$. In our work on emotion analysis, the threshold value $thres\_1$ is set to 0.8 according to a balance between the overall performance and the calculation time by experiments. This threshold may vary with different problems, and around 30 features are kept in our applications above the threshold of 0.8.

Only the features selected in the set FS_{ini} will attend in the latter steps of feature selection process. The elements (features) in FS_{ini} are seen as subsets with size 1 at the same time.

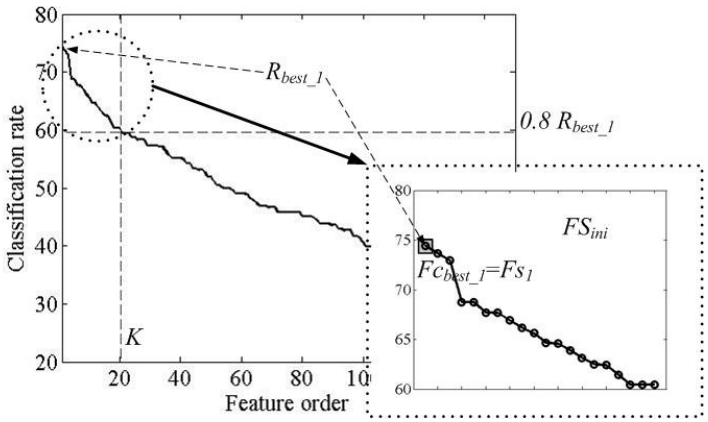

Fig. 5-4 An example of ordered single features in FS_{ini}

Step 3: Combination of features for the generation of the feature subsets

For the iterations with subsets with size k (k≥2), the generation of a subset is converted into a new feature by using an operator of combination from two original features, and the subsets are selected according to a threshold similar to the case with single features for each size of subsets.

We note the set of all the feature subsets in the evaluation with size k as FS_k and the set of the selected subsets with size k as FS'_k. Thus, FS_1 equals to the original whole feature set, and $FS'_1=FS_{ini}$. From k=2, the set of the feature subsets FS_k is noted as:

$$FS_k=Combine(FS'_{k-1}, FS_{ini})=\{Fc0_{1\_k}, Fc0_{2\_k},\ldots, Fc0_{Nk\_k}\} \qquad (5.19)$$

where the function "*Combine*" means to generate new features by combining features from each of the two sets FS'_{k-1} and FS_{ini} with all the possible combinations except the case in which the element from FS_{ini} appears in the original features during the generation process of the element from FS'_{k-1}; $Fc0_{n\_k}$ represents the generated new features; and Nk is the number of elements in the set FS_k.

The creation of a new feature from two features is implemented by combining the contribution of the belief masses of the two features, making use of an operator of combination. The combining process works as follows.

Assume that N classes are considered in the classifier. For the i^{th} class A_i, the pre-processed mass m^* for the new feature $Fc0_{t\_k}$, which is generated with Fc_{x_k-1} from FS'_{k-1} and Fs_y from FS_{ini}, $Fc0_{t\_k}$= *Combine* (Fc_{x_k-1}, Fs_y), is calculated as

$$m^*(A_i, fc0_{t\_k}) = T(m(A_i, fc_{x\_k-1}), m(A_i, fs_y)) \tag{5.20}$$

where the f_x is the value of the feature F_x, and $T(x,y)$ is an operator of combination. The commonly used existing operators for two elements, the triangle-norms, are selected in our work to combine the features, which will be explained in details in next subsection. The sum of m^*s may not be 1 according to different operators. In order to meet the definition of belief masses, the m^*s can then be normalized as the masses for the new feature:

$$m(A_i, fc0_{t\_k}) = \frac{m^*(A_i, fc0_{t\_k})}{\sum_{n=1}^{N} m^*(A_n, fc0_{t\_k})} \tag{5.21}$$

The performance of the combined new feature may be better than both two features in the combination, as illustrated in Fig. 5-5. However, the combined new feature may even performance worse than any of the two original features, which will be eliminated in the selection.

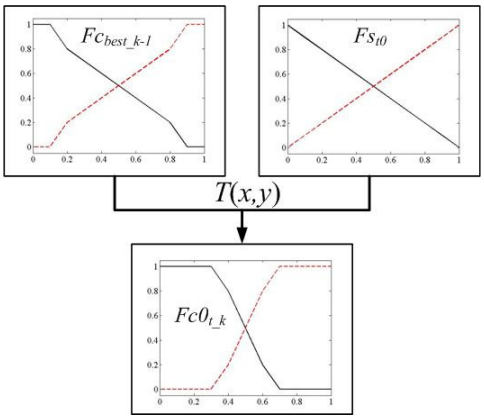

Fig. 5-5 Example of masses of a combined new feature in the case of 2 classes

Taking again the simple example of the "and" function in step one and applying the Lukasiewicz operator in the combination of the two features: $T(x,y)=\max(x+y-1, 0)$. The belief masses of the four samples with the combination of the two features are listed in Table. 5-3. The four samples can be perfectly separated into the two classes with the combination of the two features.

Table. 5-3 Classification with combination of the two features in the example of "and" function (with Lukasiewicz operator in the combination)

Samples			S_1	S_2	S_3	S_4	Classification rate
Ground truth			A	A	A	A^C	
Feature 1	Feature value		0	0	1	1	75%
	Mass	A	1	1	¼	¼	
		A^C	0	0	¾	¾	
Feature 2	Feature value		0	1	0	1	75%
	Mass	A	1	¼	1	¼	
		A^C	0	¾	0	¾	
Combination of the two features	$T(m_1,m_2)=$ $\max(m_1+m_2-1, 0)$	A	1	¼	¼	0	-
		A^C	0	0	0	½	-
	$m_{1\&2}$	A	1	1	1	0	-
		A^C	0	0	0	1	-
	Predicted class		A	A	A	A^C	100%

The correct classification rates of the combined new features can be obtained with the belief masses by assigning the class with the highest belief mass to the data samples, and the combined new features can then be ordered in descending order according to the correct classification rates as with the single features:

$$FS_k=\{Fc0_{1\_k}, Fc0_{2\_k},\ldots, Fc0_{Nk\_k}\}=\{Fc_{1\_k}, Fc_{2\_k},\ldots, Fc_{Nk\_k}\} \quad (5.22)$$

The best feature with size k is noted as $Fc_{best\_k}=Fc_{1\_k}$, and the recognition rate of feature Fc_{best_k} is recorded as R_{best_k}.

Similar to the selection of FS_{ini} in the evaluation of the single features, a threshold is set to select a certain number of subsets with size *k* to take part to the next step of forward selection. The set of the subsets remained is noted as

$$FS'_k=\{Fc_{1\_k}, Fc_{2\_k},\ldots, Fc_{N0k\_k}\} \quad (5.23)$$

which satisfies $R(Fc_{N0k\_k})\geq thres\_k * R_{best\_k}$. In order to simplify the selection, the threshold value $thres\_k$ is set in our work to the same value as 0.8 in every step without any adaptation to each step.

Step 4: Stop criterion and the selection of the best feature subset

The stop criterion of ESFS occurs when the best classification rate begins to decrease while increasing the size of the feature subsets. In our work, in order to avoid missing the real peak of the classification performance, the forward selection stops when the classification performance continues to decrease in two steps, $R_{best\_k}<\min(R_{best\_k-1}, R_{best\_k-2})$. The number of the selected features is noted as *Num_select*, and the selected feature subset is

$$SS=Fc_{1\_Num\_select}= Combine(Fc_{x\_Num\_select-1}, Fs_y)=\ldots$$
$$= Combine(Combine\ldots Combine(Fs_p, Fs_q)) \quad (5.24)$$

5.2.3.3 Operators of combination

Aggregation and fusion of different information sources are basic concerns in different knowledge based systems, including decision making, pattern recognition, machine learning, and so on. Different approaches such as evidence theory, possibility theory or fuzzy set theory make use of different fusion techniques, but all these approaches can be based on some numerical aggregation operators. Generally speaking, the aggregation operators are mathematical functions consisting of reducing a set of numbers into a unique representative number [Det00].

Since the combination of the masses of the features in our feature selection scheme amounts to combine two features, we have chose to consider the commonly used existing operators for two elements, the triangle-norms, in order to combine the features. (applied in (Equation (5.20)).

The triangular norm (abbreviated as t-norm) is a kind of binary operation used in the framework of probabilistic metric spaces and in multi-valued logic that was first introduced by Menger [Men42] in order to generalize the triangular inequality of a metric. The current concept of a t-norm and its dual operator (t-conorm) is developed due to Schweizer and Sklar [Sch60] [Sch83]. The t-norms generalize the conjunctive 'AND' operator and the t-conorms generalize the disjunctive 'OR' operator. These properties allow them to be used to define the intersection and union operations [Det00] [Ful96]. Bonissone [Bon85] investigated the properties of these operators with the goal of using them in the development of intelligent systems.

The definitions of a t-norm and a t-conorm are as follows [Det00]:

t-norm: A t-norm is a function T: $[0,1]\times[0,1]\rightarrow[0,1]$, having the following properties

1	$T(x,y)=T(y,x)$	(T1) Commutativity
2	$T(x,y)\leq T(u,v)$, if $x\leq u$ and $y\leq$v	(T2) Monotonicity (increasing)
3	$T(x,(T(y,z))=T(T(x,y),z)$	(T3) Associativity
4	$T(x,1)=x$	(T4) One as a neutral element

A well known property of t-norms is:

$$T(x,y)\leq min(x,y) \qquad (5.25)$$

t-conorm: Formally, a t-conorm is a function S $[0,1]\times[0,1]\rightarrow[0,1]$, having the following properties:

1	$S(x,y)=S(y,x)$	(S1) Commutativity
2	$S(x,y)\leq S(u,v)$, if $x\leq u$ and $y\leq$v	(S2) Monotonicity (increasing)
3	$S(x,(S(y,z))=S(S(x,y),z)$	(S3) Associativity
4	$S(x,0)=x$	(S4) Zero as a neutral element

A well known property of t-conorms is:

$$S(x,y)\geq \mathrm{Max}(x,y) \qquad (5.26)$$

We say that a t-norm and a t-conorm are dual (or associated) if they satisfy the DeMorgan law.

$$1-T(x,y)=S(1-x,1-y) \qquad \text{(the DeMorgan law)}$$

The minimum is the biggest t-norm; and its dual is the smallest t-conorm.

Six parameterized t-norms, namely Lukasiewicz, Hamacher, Yager, Weber-Sugeno, Schweizer & Sklar, and Frank, which are frequently proposed in the literatures [Det00] [Ful96] [Dou06], are tested with different parameters in our approach. They are defined as follows:

1) Lukasiewicz

$$T(x,y)=\max(x+y-1,\ 0) \qquad (5.27)$$

2) Hamacher

$$T(x,y)=\frac{x\cdot y}{\gamma+(1-\gamma)\cdot(x+y-x\cdot y)},\quad \gamma\geq 0 \qquad (5.28)$$

3) Yager

$$T(x,y)=\max\left(1-\left[(1-x)^p+(1-y)^p\right]^{1/p},0\right),\quad p>0 \qquad (5.29)$$

4) Weber-Sugeno

$$T(x,y)=\max\left(\frac{x+y-1+\lambda_T\cdot x\cdot y}{1+\lambda_T},0\right),\quad \lambda_T>-1 \qquad (5.30)$$

5) Schweizer & Sklar

$$T(x,y)=1-\left[(1-x)^q+(1-y)^q-(1-x)^q(1-y)^q\right]^{\frac{1}{q}},\quad q>0 \qquad (5.31)$$

6) Frank

$$T(x,y)=\log_s\left[1+\frac{(s^x-1)\cdot(s^y-1)}{s-1}\right],\quad s>0, s\neq 1 \qquad (5.32)$$

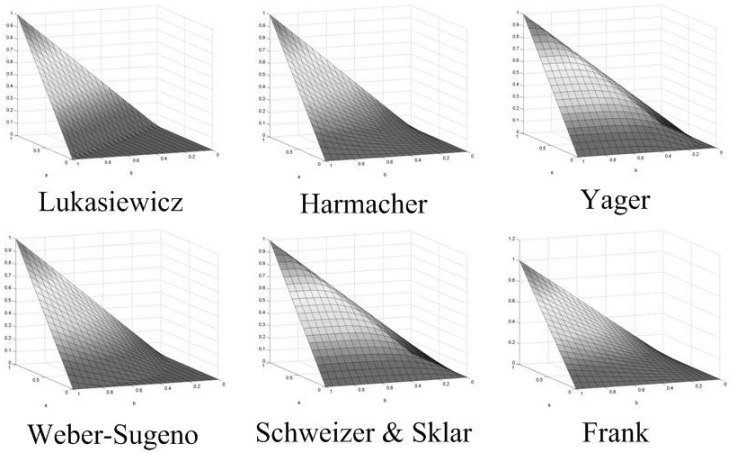

Fig. 5-6 The property curve surfaces of the operators

The property curve surfaces of the operators mentioned above are displayed in Fig. 5-6. The x-axis and y-axis present the two elements, and the z-axis presents the output of the operators. Among these six operators, the Yager, Weber-Sugeno, and Schweizer & Sklar operators have convex curve surfaces, while the Lukasiewicz and Frank operators have flatter or even concave curve surfaces. For each operator, the degree of convex or concave in the curve surfaces is also affected by the parameters. The difference in the shape of the curve surfaces may influence the performance when they are applied in the classification. This influence is discussed in Chapter 5, Chapter 6 and Chapter 7 according to experimental results.

Besides these t-norm operators, the average and the geometric average of the features are also used for the combination of the features:

Average:

$$A(x,y)=(x+y)/2 \qquad (5.33)$$

Geometric average:

$$G\_a(x,y)=\sqrt{x\cdot y} \qquad (5.34)$$

The property of the curve surfaces of average and geometric average are displayed in Fig. 5-8.

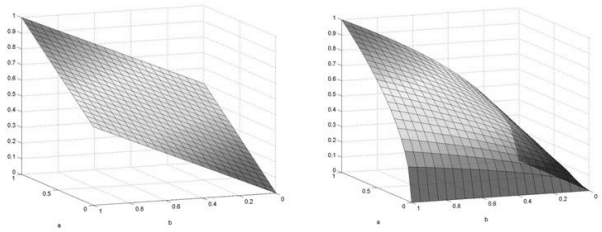

Fig. 5-7 The property curve surfaces of average and geometric average

It should be noticed that since a step of normalization is applied in calculating the masses of the combined new features in equation (5.21), the character "associativity" of the t-norms is not effective in our case. Furthermore, in order to ensure the performance of the final combined new feature, the order of the selected features cannot be moved randomly.

With these operators, the ESFS method can work on the linearly separable classification problems.

5.2.4 Experimental results

Brief experiments are carried out to test the performance of this ESFS.

Three groups of experiments are made with different features on Berlin dataset: one with all the features without selection, the second with features selected with fisher filter method [Nar77], and the third with the best features selected by the ESFS (in the following sections, section 5.4.1 and section 5.4.2).

Five types of one step global 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 experiments are carried out on TANAGRA platform [Rak05] with 10-folds cross-validation. The experimental results are listed in Table. 5-4.

Table. 5-4 Comparison between the result without feature selection and with the features selected by ESFS on Berlin dataset

	Female			Male		
	without selection	Filter selection	Selected by ESFS	without selection	Filter selection	Selected by ESFS
MP	**65.73**±2.85	66.38±2.73	71.03±1.39	**61.78**±2.93	65.75±3.19	66.44±2.50
C4.5	55.46±2.7	56.22±2.95	55.73±3.38	55.75±0.66	54.66±2.32	56.51±3.53

LDA	60.92±2.56	**70.16**±3.14	**74.00**±2.08	51.16±3.05	**70.62**±2.37	**71.97**±1.57
K-NN	60.14±2.37	64.16±3.44	67.41±1.42	57.88±2.85	61.51±2.23	66.23±2.31
NB	62.67±1.45	59.78±1.10	67.41±1.46	56.30±1.12	57.60±0.59	62.81±2.79
Best	65.73	70.16	74.00	61.78	70.62	71.97
ESFS	71.75%±3.10%			73.77%±2.33%		

The features selected by the embedded method ESFS are actually working in a filter way on the several classifiers in this experiment. The result show that for most of the classifiers tested in this experiment, the features selected by ESFS work better than the features selected by fisher filtering criterion. Especially, the features selected by ESFS fit the LDA very well, and classification rate on the LDA with these features is even better than the result from the ESFS itself on female voice samples. This result shows that the ESFS method is able to select the most discrimative features on the problem of classification of emotional speech, and the features selected with this method are more suitable to be used in the linear classifier methods than the non-linear ones.

Table. 5-5 Comparison of classification accuracy between ESFS and other classifiers (%)

ESFS	MP	C4.5	LDA	K-NN	NB
72.80	69.14±2.23	57.33±2.61	60.86±3.16	68.99±2.03	70.80±1.79

We also made experiments on the problem of classification of music mood with four classes (See introduction to the music mood dataset in section 7.5.1). In order to test the ESFS itself without the effects of the structure of the classifiers, global classifiers with one step in the classification of the four classes are applied. The same classifiers on the TANAGRA platform – MP, C4.5, LDA, K-NN, and NB – as used on the Berlin dataset are also tested on the music mood dataset. The result of ESFS on the problem of classification of music mood with global classifier is 72.80%, which is 2% higher than that obtained from the experiments on TANAGRA as 70.80% with Naïve Bayes. Although the superiority of the result of the ESFS is not so obvious, the ESFS still shows at least a little better than the popular used classification schemes, and with lower computational complexity because the feature selection and the classification processes are implemented simultaneously.

5.3 Building Automatically Hierarchical Scheme for Vocal Emotion Classifier - HCS

In Chapter 4, we showed the effectiveness of a hierarchical classifier for vocal emotion analysis, using respectively Berlin and DES datasets. The drawback of our previous approach is that this hierarchical classifier was built on the basis of an intuitive thus manual mapping of a discrete emotional model onto the dimensional one. In the case of Berlin dataset, six emotional states were first divided into 3 rough states according to the active-passive dimension: “Active state” includes anger and happiness, “Median state” includes fear and neutral, and “Passive state” includes sadness and boredom [Xia06]. For the DES dataset, the five emotional states were divided according to the active-passive dimension as “Active state” which includes anger, happiness, and surprise and “Passive state” which includes neutral and sadness.

This intuitive mapping made our hierarchical classifier-based approach rather hard to fit problem specific emotion analysis where emotion states are generally application dependant. The basic idea of the hierarchical framework is based on the separation of the emotional states into proper binary pairs according to the distribution of the emotional states in the dimensional space. The hierarchical classifiers showed good classification performance, and similar hierarchical frameworks are also proposed in different classification problems in the literature [Mei04] [Lu06]. The idea of a hierarchical classification is kept in our new approach, while there are two main weaknesses on the empirically built hierarchical frameworks. First, the positions of the emotional states in the dimensional space are not always regular. They may have quite unbalanced distribution. For example, in our manually built hierarchical classifier for Berlin database, the active dimension is divided into 3 parts and the six emotional states lies more in the negative part in the appraisal dimension. Second, due to the lack of universal agreement on the emotion definitions, the number of types of discrete emotional states and their distribution in the dimensional emotion space may change in different problems. The hierarchical structures of the empirically built frameworks have to be adjusted when the emotional space changes. Thus, our previous framework cannot apply in a straightforward manner to similar problems and a great deal of repeated works is introduced.

In this section, we introduce an automatic way for building the hierarchical classifier as we developed in the previous chapter. More precisely, we describe here a

method, so that the proper binary pairs of the discrete emotional states are automatically separated and selected, leading to a hierarchical classifier represented by a binary tree whose root is the union of all emotion classes, leaves single emotion classes and nodes subsets containing several emotion classes obtained by a sub classifier. Each of these sub classifiers is based on the new embedded feature selection method (ESFS) we have developed in the previous section, allowing to easily represent classifiers characterized by their mass function, which is the combination, thanks to a combination operator, of the mass function of the features belonging to the automatically selected feature subset. The main goal for developing this automatic hierarchical classification scheme (HCS) is to adapt classification problems with different classes to avoid repeated work, while for a given classification problem, the automatic generated classifier does not necessary give better performance than an empirical hierarchical framework.

The generation process of the HCS is shown in Fig. 5-8.

The N discrete emotional states concerned in the classification are first assigned into a set of hypotheses $\Omega=\{E_1, E_2, \ldots, E_N\}$ where E_n stands for the n^{th} emotional state in the set Ω. For example, the set of hypotheses for Berlin database is Ω_{Belin}={Anger, Happiness, Fear, Neutral, Sadness, Boredom}, and the set for DES dataset is Ω_{DES}={Anger, Happiness, Neutral, Surprise, Sadness}.

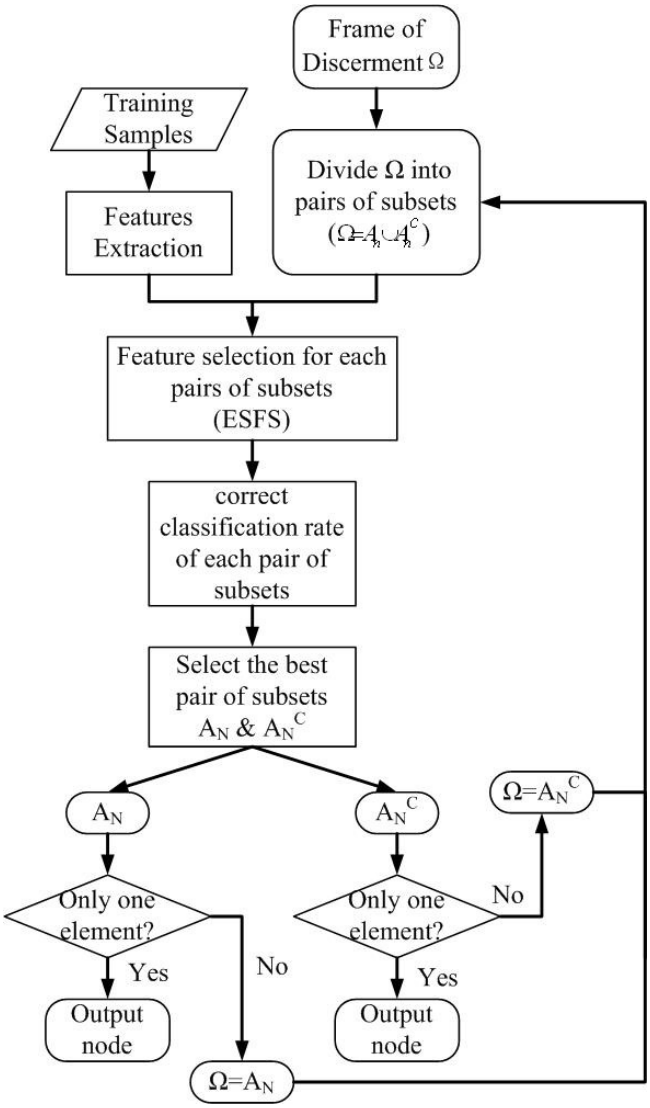

Fig. 5-8 Generation of the hierarchical classifier

The classifier is represented by a binary tree. The initial frame of discernment is set as the root node of the binary tree.

5.3.1 The basic scheme

The fundamental steps for generating the HCS are listed as follows:

1) Step 1:

The hierarchical structure is composed of several binary sub-classifiers. The set Ω is divided into pairs of nonempty subsets exhausting all possible sub-classifiers. The two subsets in each pair are complements to each other, and each subset presents a class in the sub-classifier:

$$\Omega = A_n \cup A_n^C, A_n \neq \phi, A_n^C \neq \phi \qquad (5.35)$$

All the possible pairs of complements are evaluated to decide the optimized case. The pairs are defined to ensure that the number of elements in subset A_n is not larger than the number of elements in A_n^C. For example, in the case with 4 classes, 7 pairs of subsets can be evaluated as listed in Table. 5-6:

Table. 5-6 list of pairs of subsets for 4 classes: E_1, E_2, E_3 and E_4

Index of pairs	A_n	A_n^C
1	E_1	E_2, E_3, E_4
2	E_2	E_1, E_3, E_4
3	E_3	E_1, E_2, E_4
4	E_4	E_1, E_2, E_3
5	E_1, E_2	E_2, E_3
6	E_1, E_3	E_2, E_4
7	E_1, E_4	E_2, E_3

Our feature combination and selection process (ESFS) is applied to each pair of the subsets and the belief masses of the training samples in the subsets can be obtained. All these pairs can then be sorted according to their classification accuracy rates.

2) Step 2

The two subsets in the pair with the highest classification rate (assuming it is the n^{th} pair of subsets) are assigned as the children nodes: A_n as the left child node, and A_n^C as the right child node.

The two children nodes of A_n and A_n^C are processed respectively in the same way. The numbers of elements in the children nodes A_n/A_n^C are counted. Note the subsets A_n or A_n^C as A^*:

If $Size_{A^*}=1$ (only one element in the subset), this node is marked as a leaf node.

If $Size_{A^*}>1$ (the subset can be further classified), the frame of discernment is updated as $\Omega=A^*$, and the construction of the binary tree continues with step 1.

3) Step 3:

When the number of leaf nodes equals to the number of emotional classes, the generation process of the binary tree stops. The information about the binary tree is stored in the model of the classifier.

5.3.2 Practice and Improvement

In practice, we want our HCS resulted from the previous scheme to be as balanced as possible. Indeed, the overall classification accuracy rate of a multi-stage hierarchical classifier is approximately the product of the classification rates at each stage. Assuming the different stages in the classifier have correct classification rates close to each other as R_{stage}, for an n stage classifier, the overall classification rate can be approximated by R_{stage}^{n}. Thus, too many stages may lead to dramatic degrading of the overall classification accuracy rate. In order to reach a classification accuracy as high as possible, one needs to reduce the depth of the tree-based hierarchical classifier so that it is a balanced structure.

In our work, balanced pair of subsets are put forward. For the each pair of subsets A_n & A_n^C, a subset distance is calculated as the difference of the number of elements of the two subsets:

$$D_n = Size_{A_n^C} - Size_{A_n} \tag{5.36}$$

with

$$Size_{A^*} = \left|A^*\right| \tag{5.37}$$

where |X| means the number of elements in the set X.

Because the subsets A_n/A_n^C are defined as A_n with fewer or same number of elements than A_n^C, D_n always satisfies

$$D_n \geq 0 \tag{5.38}$$

when the n^{th} pair of subsets satisfies $D_n \leq 1$, it is defined as a balanced pair of subsets.

If the pair of subsets with the highest classification rate (assuming it is the n_1^{th} pair, and the classification rate is R_{n1}) is a balanced pair, the generation of the binary

tree continues normally; if it is not a balanced pair, it will be compared with the balanced pair with the highest classification rate (assuming it is the n_2^{th} pair, and the classification rate is R_{n2}). As we have only five or six classes in our applications, there should be two or three stages in a balance binary tree. So a threshold of rate difference *thre_diff* is set according to the criterion that $(R_{n1}\text{-}thre\_diff)^2 \geq (R_{n1})^3$, assuming the number of the stages does not exceed three. The approximate values of *thre_diff* related to R_{n1} are listed in Table. 5-7. The highest classification rate R_{n1} in the first stage of the hierarchical structure is normally around 90% in the experiments, so the most commonly selected *thre_diff* is between 4% and 5%. When the number of classes increases in the classification problems, the thresholds should be adjusted according to the number of classes (the computational complexity will also increase exponentially with too many of classes as well with this approach, which limits the application of this automatic approach).

Table. 5-7 Thresholds according to the highest classification rate (%)

R_{n1}	97	95.8	94.8	93.6	92.6	**91.4**	**90.2**	**89**	87.8	86.6
thre_diff	1.5	2	2.5	3	3.5	**4**	**4.5**	**5**	5.5	6

If $R_{n1}\text{-}R_{n2}<thre\_diff$, the binary tree with balanced pair is assumed to have better overall performance in the classification; and the n_2^{th} pair is selected instead of the n_1^{th} pair. While when the unbalanced pair have significant better recognition rate than the balanced one, it will be still selected.

The common structure of the HCS generated with this approach is shown in Fig. 5-9. The grey doubled line illustrates the possible recognition route of an audio sample.

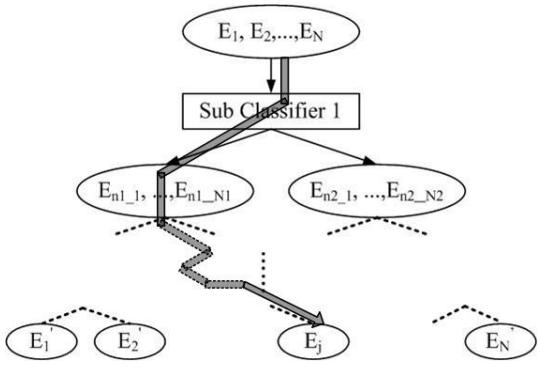

Fig. 5-9 Hierarchical classifier and recognition route

The audio datasets experimented in our works concerns the number of the emotional/mood states from 4 to 6. The typical classifiers with these numbers of states

are shown in Fig. 5-10 with balanced pairs of subsets. Only the states in each node are displayed and the sub-classifiers are omitted in Fig. 5-10.

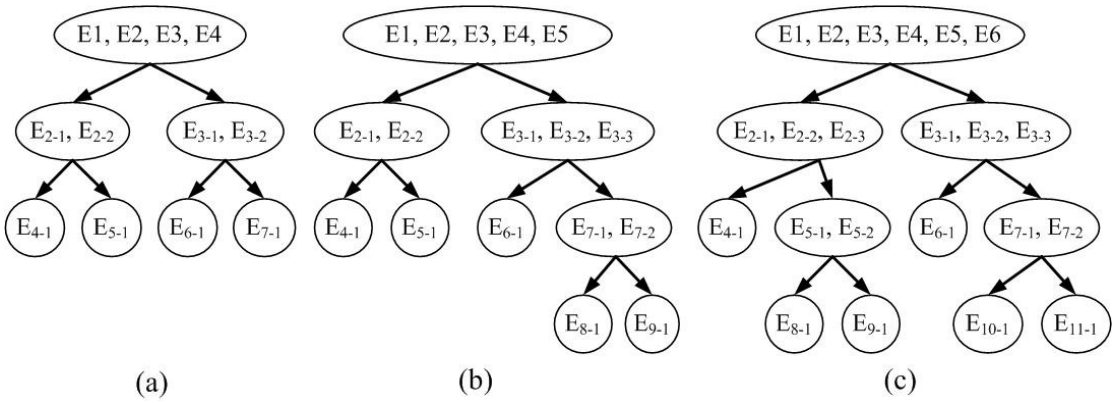

Fig. 5-10 Typical balanced HCS classifiers for 4, 5, 6 classes

5.4 Experimental results

The approach for automatic building of HCS that we have developed so far is benchmarked on Berlin dataset from Berlin Technical University and DES dataset from Aalborg University, Denmark that were used in the previous chapter on empiric hierarchical classifier mapping a discrete emotional model into the dimensional one.

Recall that there are six emotional states that were considered in Berlin dataset (the seventh state – disgust – is ignored due to too few samples) and five emotional states in DES dataset. The emotional states are illustrated in Fig. 4-7. Four emotional states are common in the two dataset: “Anger”, “Happiness”, “Neutral”, and “Sadness”. The two other emotional states in Berlin dataset are “Fear” and “Boredom” while “Surprise” is the last remaining emotional state in DES dataset.

As we have previously seen in Chapter 4, gender difference in the acoustic features also influences the emotion classification. Therefore, a preprocessing step for gender classification as illustrated in Fig. 5-11 can improve the overall accuracy of vocal emotion analysis. While gender classification has been studied in-depth in the literature [Har03a] [Har04] [Har05b] and any gender classifier might be used, we made use of the gender classifier described in Chapter 4 (section 4.3.2).

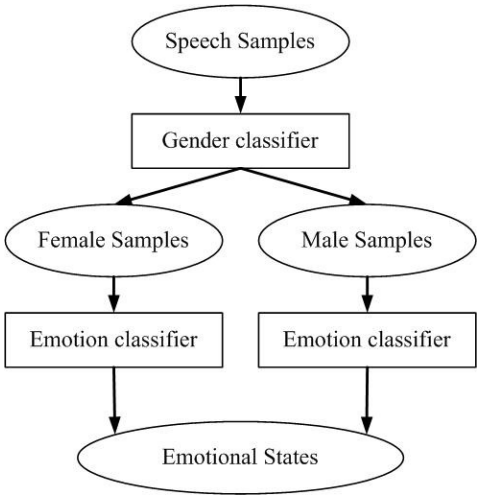

Fig. 5-11 Emotion classifier with gender information

In the following subsections, we present the experimental results on Berlin and DES datasets. In addition, influence of the language is also discussed. All these experiments are taken with hold-out cross-validation [Koh95] with 10 iterations. The structures of the automatic generated hierarchical classifiers and the experimental results are presented in these subsections, then further analyzed and summarized in subsection 5.4.4.

All the previously defined audio features, namely frequency features, energy features, MFCC features, harmonic features and Zipf features are used in the subsequent experiments.

5.4.1 Experiments on Berlin dataset

The HCS generated for the two genders for the six emotional classes of the Berlin dataset are shown in Fig. 5-12. As we can see, these two gender dependant hierarchical classifier trees have similar structures, which differ, however, on intermediate emotional subclasses starting from the first level.

As compared to our previous empirical DEC, the three emotional states as anger, happiness, and fear are separated from the other three emotions in the first stage for both genders. Getting back to the distribution of these emotions in the dimensional space as shown in Fig. 4-7, this division fits the distribution in the arousal dimension as well. Our automatically built hierarchical classifiers also capture the gender difference of expressing emotions. Because the way of expressing the emotions is different for women and men, the following stages are different between

the two genders. For the female speech in the second stage, the active emotional states are further divided again according to the arousal dimension, and the passive emotional states are further divided according to the appraisal dimension. On the contrary, for the male samples, the active emotional states are further divided in the appraisal dimension, and the passive emotional states are further divided according to the arousal dimension.

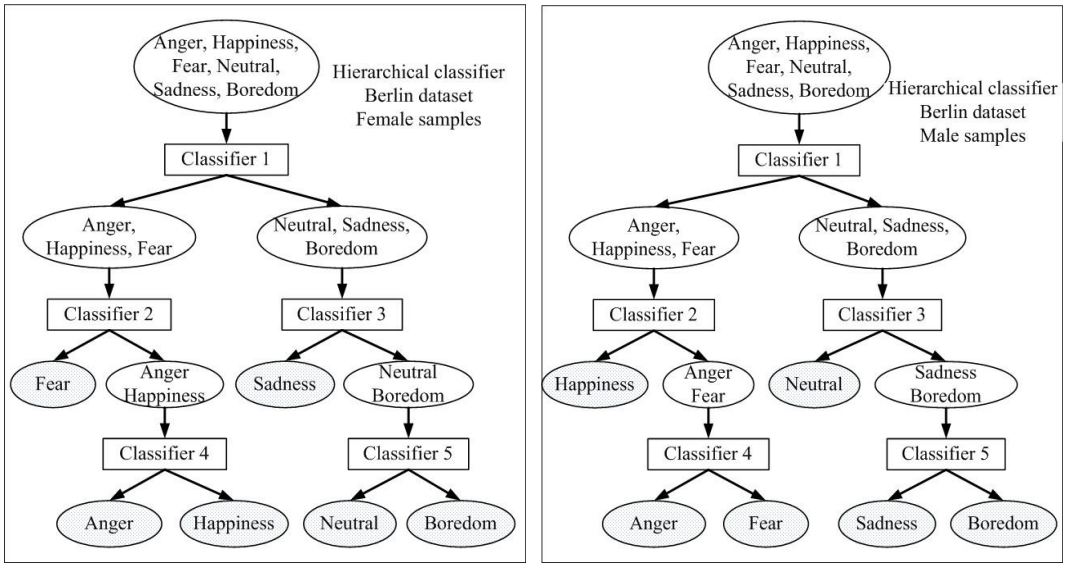

Fig. 5-12 Hierarchical classifier for Berlin dataset

Hold-out cross-validations with 10 iterations are used in our experiments on Berlin dataset. In each of the iterations in the cross-validation, 50% of samples are used as training set, and 50% as test set. The average classification rates and the root mean square errors of the classification rates are calculated with several operators with different parameters. The detailed results are listed in Table. 5-8 (female samples), Table. 5-9 (male samples), and Table. 5-10 (mixed samples, with/without gender classification). The emotional states are indexed in the confusion matrix as: E1=Anger, E2=Happiness, E3=Fear, E4=Neutral, E5=Sadness, E6=Boredom for Berlin dataset.

Table. 5-8 Berlin database, hold-out cross-validation with 10 iterations, female samples in HCS

Operator	Lukasiewicz	Hamacher			
Parameter	-	γ=0.5	γ=3	γ=5	γ=10
Mean rate	67.36±2.04	70.92±2.46	71.37±2.73	70.75±1.82	70.27±2.05
Operator	Yager				Weber-Sugemo

Parameter	p=1	p=3	p=5	p=10	λ_T=-0.5
Mean rate	67.39±2.02	70.78±2.67	69.54±2.46	68.57±2.31	64.39±2.83
Operator	Weber-Sugemo				
Parameter	λ_T=0	λ_T=0.5	λ_T=2	λ_T=4	λ_T=5
Mean rate	64.80±2.31	65.69±2.33	68.95±2.43	71.05±1.99	71.18±2.56
Operator	Schweizer & Sklar				
Parameter	q=0.2	q=0.4	q=0.6	q=0.8	q=1
Mean rate	68.19±1.87	**71.75±3.10**	71.54±2.79	71.54±2.83	71.08±2.50
Operator	Schweizer & Sklar		Frank		
Parameter	q=3	q=5	s=2	s=5	s=8
Mean rate	70.54±2.56	66.47±1.36	53.15±2.28	64.21±3.61	65.12±2.15
Operator	Frank			Average	Geometric Average
Parameter	s=10	s=12	s=15	-	-
Mean rate	67.55±2.49	68.03±1.93	67.93±1.90	65.93±1.17	66.34±1.24

Average confusion matrix for the best parameter (%)

Predicted / Actual	E1	E2	E3	E4	E5	E6
E1	75.06	18.28	5.06	1.03	0.23	0.34
E2	28.95	59.12	7.72	1.58	0.35	2.28
E3	13.59	16.41	55.64	7.95	3.08	3.33
E4	0.82	1.84	4.49	78.98	4.49	9.39
E5	0.00	0.27	2.03	2.57	89.73	5.41
E6	0.62	1.38	4.15	20.00	11.69	62.15

Table. 5-9 Berlin database, hold out cross-validation with 10 iterations, male samples in HCS.

Operator	Lukasiewicz	Hamacher			
Parameter	-	γ=0.5	γ=3	γ=5	γ=10
Mean rate	69.08±2.76	73.25±3.55	73.01±1.43	72.29±2.03	72.23±1.97
Operator	Yager				Weber-Sugemo
Parameter	p=1	p=3	p=5	p=10	λ_T=-0.5
Mean rate	69.08±2.58	73.01±3.43	72.02±2.49	71.58±2.58	68.12±1.93
Operator	Weber-Sugemo				
Parameter	λ_T=0	λ_T=0.5	λ_T=2	λ_T=4	λ_T=5
Mean rate	70.21±2.20	70.65±1.93	70.62±2.62	71.92±2.19	73.29±2.90

Operator	Schweizer & Sklar				
Parameter	q=0.2	q=0.4	q=0.6	q=0.8	q=1
Mean rate	73.08±2.32	73.39±2.34	**73.77**±2.33	73.01±2.81	73.73±2.60
Operator	Schweizer & Sklar		Frank		
Parameter	q=3	q=5	s=2	s=5	s=8
Mean rate	72.77±2.04	71.85±2.39	57.26±3.31	61.30±2.08	66.85±3.12
Operator	Frank			Average	Geometric Average
Parameter	s=10	s=12	s=15	-	-
Mean rate	68.15±1.86	67.33±1.81	68.18±3.71	68.77±2.36	69.25±2.68

Average confusion matrix for the best parameter (%)

Predicted / Actual	E1	E2	E3	E4	E5	E6
E1	86.67	8.67	3.87	0.80	0.00	0.00
E2	25.59	59.12	12.35	2.06	0.29	0.59
E3	6.15	10.51	70.77	8.72	2.56	1.28
E4	1.30	3.70	5.65	84.78	0.87	3.70
E5	0.61	0.41	0.61	8.16	73.06	17.14
E6	1.22	1.63	3.47	18.57	18.16	56.94

As mentioned in Chapter 4, any gender classification scheme on gender classification might be applied in the classification of emotional speech, the same Neural Network based gender classifier as in section 4.3.2 is also used here.

Table. 5-10 Berlin database, hold-out cross-validation with 10 iterations, all samples in HCS.
Rate 1: with gender classification; Rate 2: without gender classification

Operator	Lukasiewicz	Hamacher				Yager
Parameter	-	γ=0.5	γ=3	γ=5	γ=10	p=1
Rate 1 (%)	66.75±2.76	70.79±3.55	68.65±1.43	70.26±2.03	69.19±1.97	64.80±2.58
Rate 2 (%)	55.07±3.10	57.50±2.82	57.22±2.54	56.94±2.22	56.11±2.75	55.11±3.11
Operator	Yager			Weber-Sugemo		
Parameter	p=3	p=5	p=10	λ_T=2	λ_T=4	λ_T=5
Rate 1 (%)	68.80±3.43	68.49±2.49	68.96±2.58	68.68±2.62	69.00±2.19	70.63±2.90
Rate 2 (%)	56.40±2.55	57.30±2.57	56.23±2.16	56.40±2.57	57.35±2.52	57.86±2.76
Operator	Schweizer & Sklar					
Parameter	q=0.4	q=0.6	q=0.8	q=1	q=3	

Rate 1 (%)	69.86±2.34	**71.38±2.33**	70.65±2.81	71.26±2.60	69.65±2.04
Rate 2 (%)	**57.95**±2.87	57.77±2.49	57.86±2.70	57.13±2.07	57.12±2.27

Operator	Frank			Average	Geometric Average
Parameter	s=10	s=12	s=15	-	-
Rate 1 (%)	66.23±1.86	66.02±1.81	65.44±3.71	65.47±2.36	64.59±2.68
Rate 2 (%)	54.72±1.93	54.86±2.03	55.25±2.78	55.05±2.66	54.30±1.39

Best confusion matrix with gender classification		E1	E2	E3	E4	E5	E6
	E1	78.67	13.58	4.88	1.45	0.68	0.73
	E2	26.91	57.67	10.26	2.33	0.88	1.96
	E3	10.1	13.57	61.61	8.62	3.29	2.8
	E4	1.59	3.24	5.47	79.65	3.16	6.89
	E5	0.86	0.9	1.84	5.75	79.18	11.46
	E6	1.46	2.03	4.25	19.2	14.99	58.08

Best confusion matrix without gender classification		E1	E2	E3	E4	E5	E6
	E1	59.26	25.56	8.02	5.06	1.42	0.68
	E2	25.16	49.45	11.76	8.24	2.53	2.86
	E3	10.90	12.18	34.49	21.15	15.13	6.15
	E4	1.47	2.63	5.16	64.63	17.16	8.95
	E5	0.08	0.24	2.20	4.63	83.74	9.11
	E6	0.35	1.58	2.81	22.89	26.84	45.53

Fig. 5-13 shows the best classification rates of the tested operators. The classification rates for the case of all features are obtained by adding a preprocessing of gender classification. The error bars in the figure show the root mean square errors of the classification rates.

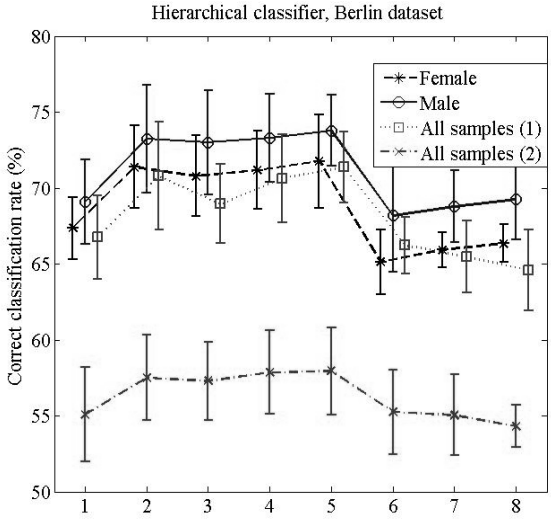

Fig. 5-13 Classification rate with HCS on Berlin dataset, The indexes on the X axis stand for the operators: 1 – Lukasiewicz, 2 – Hamacher, 3 – Yager, 4 - Weber-Sugemo, 5 - Schweizer & Sklar, 6 – Frank, 7 - Average, 8 - Geometric Average. "All samples (1)" refers to the classification on the mixed genders samples with gender classification, and "All samples (2)" refers to the classification on the mixed genders samples without gender classification.

The best result is 71.75%±3.10% for the female samples, 73.77%±2.33% for the male samples, and 71.38%±2.33% for the mixed genders with gender classification and 57.95%±2.87% without gender classification, which are quite closed to the results obtained with the manual DEC scheme. All of the best results are obtained with Schweizer & Sklar operator. From the two curves "All samples (1)" and "All samples (2)", we can see that the gender classification can obviously improve the overall classification performance for the mixed gender samples. The operators Hamacher, Yager, Weber-Sugemo, and Schweizer & Sklar which have properties of convex curve surfaces which perform better among all these results.

The best features are analyzed into three categories as 1) Classifier 1 for the first stage, 2) Classifiers 2 & 4 for the active emotions, and 3) Classifiers 3 & 5 for the passive emotions. Each time a combination operator is used in ESFS for building automatically a classifier hierarchy, the selected features for each sub-classifier in the classification hierarchy may be different. We thus count the number of times where each feature is selected in these sub-classifiers. The selection frequency of a feature gives a hint of its importance for vocal emotion analysis. The indexes of the most

frequently selected features are listed according to the feature groups in Table. 5-11. Up to 5 features are listed for each group, and the features are ordered according to the frequency of selection. At the end of Table. 5-11, the percentages of the features selected in each group for the three categories are listed. The Zipf features are calculated together with the harmonic features because there are only 2 Zipf features in our feature set. The description of the most important features is shown in Annex A.

Table. 5-11 Most frequently selected features in the HCS for Berlin dataset.
H – Harmonic, Z – Zipf, F – Frequency, E - Energy & rhythm, M - MFCC

Feature Groups		H	Z	F	E	M
Classifier 1	Female	3	225,226	23,27	78,80,76,46	81,92,135,107,217
	Male	-	225,226	27,23,28,31,35	46,78	107,89,223,222,224
Classifier 2&4	Female	21,20,15	-	30,31,25,41,35	46,78,64,62,45	92,205,175,174,161
	Male	3,7,15,11	-	31,28,41,33,32	45,80,49,78,46	108,151,150,222,136
Classifier 3&5	Female	11,3,20,19	-	35,24,42,38,28	46,56,44,54,49	162,163,205,192,164
	Male	15,19,3	-	36,25,24	46,49,51,55,56	156,175,99,116,102

Female (%)				Male (%)			
group	Classifier 1	Classifier 2&4	Classifier 3&5	group	Classifier 1	Classifier 2&4	Classifier 3&5
H & Z	1	1	6	H & Z	0	8	6
F	62	42	14	F	54	35	12
E	20	30	32	E	11	21	25
M	18	27	48	M	35	37	57

Although the new features proposed in our work compose only a small percentage of the number of the selected features (8% at the most), they act as a crucial factor in enhancing the classification performance in the classification of the speech emotion, and they are usually among the earliest features in the selection. The harmonic features are not very significant in the first stage of the hierarchical classifier (the arousal dimension), but are essential in the latter stages (more in the appraisal dimension), which capture the timber or tonality in the speech. On the contrary, the Zipf features show great importance in the feature subset for classifier 1 in the arousal dimension, which are rather related to the rhythm.

The traditional features still dominate in the number of selected features. The frequency features make up more percentages in the classifier 1 (arousal dimension) than the other sub-classifiers. The energy and rhythm related features are equally

important to all the sub-classifiers. The ratio of energy below 250 Hz is the most important feature among all the features in the feature set, which means that the emotions with higher arousals tend to focus more energy in the high frequency band and the emotions with lower arousals tend to focus more energy in the low frequency band. The MFCC features are more efficient for the sub-classifiers 2~5 especially for the classifiers 3&5 for the passive emotional states; they can compose around 50% of selected features in the classifiers 3&5.

5.4.2 Experiments on DES dataset

The automatically HCS classifiers as defined in section 5.3 for the two genders for the five emotional classes of the DES dataset are shown in Fig. 5-14.

It appears that the same structure is generated for both of the two genders. Similar to the case of the Berlin dataset, the first stage consists in the arousal dimension (or energy/active dimension) with the separation between neutral & sadness vs. anger & happiness & surprise. For the three active emotions, surprise is separated from anger and happiness in the second stage that is still in the arousal dimension. Anger and happiness are separated in the appraisal dimension in the last stage in the hierarchical framework as for the female samples in Berlin dataset.

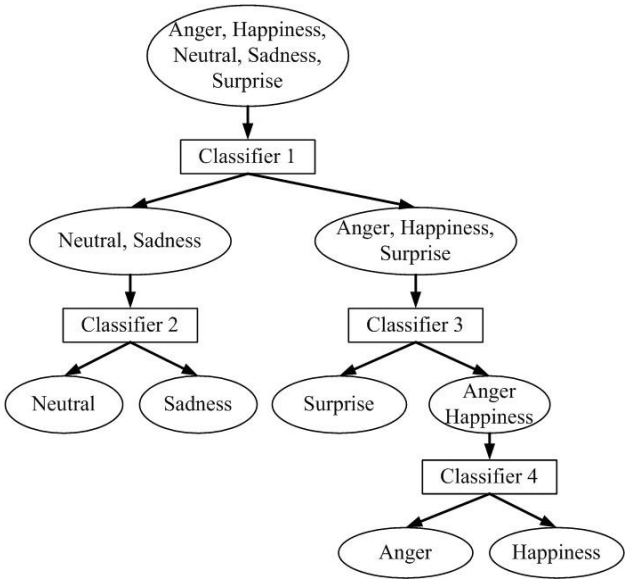

Fig. 5-14 HCS classifier for DES dataset

Hold-out cross-validations with 10 iterations are used in our experiments on DES dataset. In order to compare with previous works [Ver04a] [Ver04b] [Ver05a] [Ver05b], for each iteration of experiments, 90% of segments are used as training set

and 10% are used as testing set. The training set and testing set are selected randomly in each group.

The average classification rates and the root mean square errors of the classification rates are calculated with several operators with different parameters. The detailed results are listed in Table. 5-12 (female samples), Table. 5-13 (male samples), and Table. 5-15 (mixed samples). The emotional states are indexed in the confusion matrix as: E1=Anger, E2=Happiness, E3= Neutral, E5=Sadness, E6=Surprise for DES dataset.

Table. 5-12 DES database, hold-out cross-validations with 10 iterations, female samples in HCS (%)

Operator	Lukasiewicz	Hamacher			
Parameter	-	γ=0.5	γ=3	γ=5	γ=10
Mean rate	72.66±1.96	78.11±1.12	**79.54**±1.95	78.92±1.82	78.61±1.67
Operator	Yager				Weber-Sugemo
Parameter	p=1	p=3	p=5	p=10	λ_T=-0.5
Mean rate	72.7±2.01	78.73±1.70	76.53±1.48	75.37±1.80	74.09±1.89
Operator	Weber-Sugemo				
Parameter	λ_T=0	λ_T=0.5	λ_T=2	λ_T=4	λ_T=5
Mean rate	72.28±2.12	72.93±2.91	76.29±2.41	79.03±2.70	79.03±1.86
Operator	Schweizer & Sklar				
Parameter	q=0.2	q=0.4	q=0.6	q=0.8	q=1
Mean rate	79.42±2.47	78.96±2.18	79.38±1.74	79.31±1.15	79.34±1.14
Operator	Schweizer & Sklar		Frank		
Parameter	q=3	q=5	s=2	s=5	s=8
Mean rate	76.76±1.19	75.91±1.77	53.09±1.59	57.8±3.41	73.09±3.01
Operator	Frank			Average	Geometric Average
Parameter	s=10	s=12	s=15	-	-
Mean rate	72.36±1.86	73.36±1.65	72.66±1.87	70.93±0.60	74.94±1.49
Average confusion matrix for the best parameter (%)					
Predicted / Actual	E1	E2	E3	E4	E5
E1	81.18	7.06	4.71	4.31	2.75
E2	9.8	77.25	2.35	0.98	9.61
E3	6.27	2.55	80.59	4.12	6.47

E4	0.38	1.15	9.42	77.12	11.92
E4	1.11	7.59	3.89	5.93	81.48

Table. 5-13 DES database, hold-out cross-validations with 10 iterations, male samples in HCS (%)

Operator	Lukasiewicz	Hamacher			
Parameter	-	γ=0.5	γ=3	γ=5	γ=10
Mean rate	74.29±1.39	80.33±1.41	81.09±0.98	80.58±0.82	79.93±1.04
Operator	Yager				Weber-Sugemo
Parameter	p=1	p=3	p=5	p=10	λ_T=-0.5
Mean rate	74.33±1.46	80.55±1.51	79.27±1.82	78.73±1.51	68.40±1.68
Operator	Weber-Sugemo				
Parameter	λ_T=0	λ_T=0.5	λ_T=2	λ_T=4	λ_T=5
Mean rate	71.89±1.60	73.96±1.96	77.60±1.03	78.91±1.23	79.82±1.09
Operator	Schweizer & Sklar				
Parameter	q=0.2	q=0.4	q=0.6	q=0.8	q=1
Mean rate	78.98±2.62	81.49±0.79	81.53±0.86	81.71±1.09	**81.96**±1.27
Operator	Schweizer & Sklar		Frank		
Parameter	q=3	q=5	s=2	s=5	s=8
Mean rate	80.04±1.75	76.00±2.23	66.76±1.06	69.56±3.80	69.71±1.98
Operator	Frank			Average	Geometric Average
Parameter	s=10	s=12	s=15	-	-
Mean rate	71.16±1.14	70.91±1.71	71.38±1.44	73.93±1.26	77.75±1.67
Average confusion matrix for the best parameter (%)					
Predicted / Actual	E1	E2	E3	E4	E5
E1	73.14	20.00	0.39	2.16	4.31
E2	2.96	92.59	2.22	2.04	0.19
E3	2.36	11.09	69.45	13.64	3.45
E4	0.56	1.30	5.37	92.04	0.74
E4	13.77	0.66	1.64	1.64	82.30

In order to capture the gender difference in vocal emotion expressions, we also applied a gender classifier before vocal emotion classification as we did for Berlin dataset. While any gender classifier, for instance the ones defined in [Har03a] [Har04] [Har05b], might be used, we also experimented our ESFS using all the audio features

for gender classification on DES dataset. The best accurate rate in gender classification on DES dataset is 94.94% with Schweizer & Sklar operator when q=0.6. The detailed results are listed in Table. 5-14.

Table. 5-14 Rate of gender classification, DES dataset (%)

Operator	Lukasiewicz	Hamacher				Yager
Parameter	-	γ=0.5	γ=3	γ=5	γ=10	p=1
Rate (%)	77.90	94.19	89.70	93.07	93.07	84.64
Operator	Yager			Weber-Sugemo		
Parameter	p=3	p=5	p=10	λ_T=2	λ_T=4	λ_T=5
Rate (%)	87.27	94.57	87.45	86.33	91.76	90.64
Operator	Schweizer & Sklar					
Parameter	q=0.4	q=0.6	q=0.8	q=1	q=3	
Rate (%)	89.51	**94.94**	94.57	93.45	93.45	
Operator	Frank			Average	Geometric Average	
Parameter	s=10	s=12	s=15	-	-	
Rate (%)	88.01	87.08	84.46	88.95	88.39	

Table. 5-15 DES database, hold-out cross-validations with 10 iterations, all samples in HCS (%)
Rate 1: with gender classification; Rate 2: without gender classification

Operator	Lukasiewicz	Hamacher				Yager
Parameter	-	γ=0.5	γ=3	γ=5	γ=10	p=1
Rate 1 (%)	65.69±1.03	75.73±0.92	73.28±0.76	74.74±1.02	74.74±0.83	65.69±1.11
Rate 2 (%)	50.92±1.64	52.47±1.71	53.00±1.98	52.15±2.39	52.30±1.88	50.92±1.82
Operator	Yager			Weber-Sugemo		
Parameter	p=3	p=5	p=10	λ_T=2	λ_T=4	λ_T=5
Rate 1 (%)	72.85±0.60	74.81±0.78	70.67±0.94	69.70±0.94	73.90±1.62	73.58±1.34
Rate 2 (%)	53.67±2.18	50.34±1.01	49.85±1.52	52.17±1.02	52.81±2.36	52.79±2.47
Operator	Schweizer & Sklar					
Parameter	q=0.4	q=0.6	q=0.8	q=1	q=3	
Rate 1 (%)	73.46±0.89	**76.74**±0.83	76.72±0.64	76.16±0.78	74.66±1.10	
Rate 2 (%)	52.58±1.14	**53.75**±1.71	53.37±2.56	52.87±1.77	50.45±2.08	
Operator	Frank			Average	Geometric Average	
Parameter	s=10	s=12	s=15	-	-	
Rate 1 (%)	65.34±1.14	64.85±1.08	63.91±1.44	67.08±1.55	70.26±1.39	
Rate 2 (%)	50.66±1.71	51.82±2.88	50.75±2.35	50.66±1.87	49.64±2.77	

	Predicted / Actual	E1	E2	E3	E4	E5
Best confusion matrix with gender classification	E1	67.75	17.06	3.63	3.04	8.53
	E2	7.24	82.76	3.43	1.33	5.24
	E3	4.25	8.30	75.19	8.30	3.96
	E4	0.66	1.70	8.49	83.77	5.38
	E5	10.96	7.04	2.96	4.87	74.17
	Predicted / Actual	E1	E2	E3	E4	E5
Best confusion matrix without gender classification	E1	47.16	19.12	4.51	2.94	26.27
	E2	14.29	58.57	3.62	1.62	21.90
	E3	14.91	8.87	51.98	3.77	20.47
	E4	11.13	4.25	21.79	45.19	17.64
	E5	7.91	15.65	5.39	6.35	64.70

Fig. 5-15 shows the best classification rates of the tested operators. The classification rates for the case of all speech samples from both genders are obtained by adding a preprocessing of gender classification as we did in experiments on the Berlin dataset. The error bars in the figure show the root mean square errors of the classification rates.

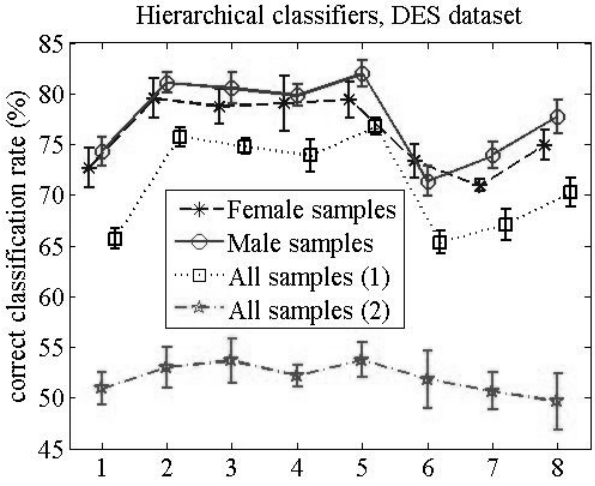

Fig. 5-15 Classification rate with HCS on DES dataset.

The indexes on the X axis stand for the operators: 1 – Lukasiewicz, 2 – Hamacher, 3 – Yager, 4 - Weber-Sugemo, 5 - Schweizer & Sklar, 6 – Frank, 7 - Average, 8 - Geometric Average. "All samples (1)" refers to the classification on the mixed genders samples with gender classification, and

"All samples (2)" refers to the classification on the mixed genders samples without gender classification.

The best result is 79.54%±1.95% for the female samples with Hamacher operator when γ=3, 81.96%±1.27% for the male samples with Schweizer & Sklar operator when q=1, and 76.74%±0.83% for the mixed genders with gender classification and 53.75%±1.71% without gender classification with Schweizer & Sklar operator when q=0.6. Significant improvement of up to 23% is obtained for the mixed genders with the gender classification as a preprocessing step.

Experimented on the same DES dataset with the same 90% data of training and 10% data of testing in cross validation, the best result obtained in the literature by Ververidis *et al.* is 66% for only male samples using a one step GMM (Gaussian Mixture Model) classifier for all the five emotions [Ver05b]. Significant improvement in the correct classification rate is achieved with our automatic hierarchical classifier.

Similar to the case with Berlin dataset, the best features are analyzed into three categories as 1) Classifier 1 for the first stage in the arousal emotions, 2) Classifiers 3 & 4 for the active emotions, and 3) Classifiers 2 for the passive emotions. The numbers of times of the features being selected in these sub-classifiers are also used to evaluate the importance of the features as in the previous section for Berlin dataset. The indexes of the most frequently selected features are listed according to the feature groups in Table. 5-16. Up to 5 features are listed for each group, and the features are ordered according to the frequency of selection. At the end of Table. 5-16, the percentages of the features selected in each group for the three categories are listed. The description of the most important features is shown in Annex A.

Table. 5-16 Most frequently selected features for the hierarchical classifier on DES dataset

Feature Groups		Harmonic	Zipf	Frequency	Energy & rhythm	MFCC
Classifier 1	Female	-	225,226	32,29,27	44,43,69,71	87,216,206,120,110
	Male	15	225,226	27,28,31,23,29	44,74,60,70,61	91,183,161,90,137
Classifier 3&4	Female	15	-	31,27,41,38,29	44,46,73,43,63	124,209,143,135,194
	Male	7,3	-	32,28,27,24,31	78,43,74,64,80	100,197,147,107,141
Classifier 2	Female	-	-	41,27,31,26,23	46,60,61,68,71	123,190,96,191,129
	Male	-	-	31,28,27	46,71,65,49,72	100,187,181,159,155

Female (%)				Male (%)			
group	Classifier 1	Classifier	Classifier 2	group	Classifier 1	Classifier	Classifier 2

		3&4				3&4	
H&Z	0	1	0	H&Z	1	5	0
F	34	26	29	F	46	17	11
E	20	12	23	E	34	7	38
M	46	61	48	M	18	71	51

The relationships between the feature groups and the emotional states presented in the hierarchical classifier on the DES database are similar as on the Berlin dataset. The harmonic features are more important in the latter stages in the hierarchical framework (separation in the active emotional states); the Zipf features which capture rhythms from the speech signals perform better in the first stage (arousal dimension). The ratio of energy below 250 Hz is still the most important feature among all the features in the feature set. The MFCC features are more efficient for the sub-classifier for the passive emotional states in appraisal dimension (classifier 2).

5.4.3 Experiments on the influence of languages on the emotions

Our overview on acoustic correlates in Chapter 2 section 2.2 reveals that human recognition of vocal emotions can be language and culture independent due to the universal physiological effects of the emotions [Abe00] [Bur00] [Tic00]. Is it true for machine-based vocal emotion recognition that makes use of our audio feature set? In order to answer this question, we tested our automatic HCS using Berlin and DES datasets respectively expressed in German and Danish. As we have highlighted, they share four common emotion states, namely: anger, happiness, neutral, and sadness.

Using ESFS and the HCS described in section 5.3, we thus automatically generate a hierarchy of classifiers as depicted in Fig. 5-16 for the four common emotional states. In each round of experiment, one of the datasets is used as the training dataset while the other one as the testing dataset. The female samples and male samples are tested separately in the experiments.

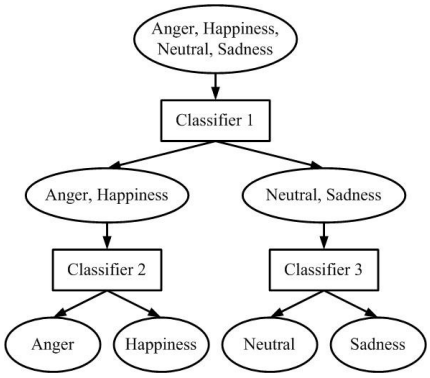

Fig. 5-16 HCS for the four common emotions of Berlin and DES dataset

The results of the correct classification rates for all the four groups of experiments – with alternative training and testing set and female and male samples – are slightly over 50%, which vary from 50.73% to 59.35%. The results are listed in Table. 5-17.

Table. 5-17 Correct classification rates for the four common emotions between the two dataset (%)

	Training – Berlin, Testing - DES		Training – DES, Testing - Berlin	
	Best rate	Operator	Best rate	Operator
Female	50.73	Schweizer & Sklar	52.06	Lukasiewicz
Male	59.35	Lukasiewicz	51.96	Yager

These results tend to approximate the human recognition rate that is roughly 60%. Note that for a classification problem with four classes, the correct classification rate by randomly guessing should around 25% when our experimental results display a classification accuracy rate over 50%.

The confusion matrices in the cross language tests for the four emotions are shown in Table. 5-18. The correct classification rates for most of the cases for the four emotional types are around 50% except an extremely low rate for anger with female samples and a very high rate for sadness with male samples when Berlin dataset is used as training set and DES dataset is used as testing set.

Table. 5-18 Confusion matrix in the cross language tests (%)

	Training – Berlin, Testing – DES					Training – DES, Testing - Berlin			
	Predicted / Actual	Anger	Happiness	Neutral	Sadness	Anger	Happiness	Neutral	Sadness
Female	Anger	**15.69**	19.61	60.78	3.92	**52.87**	39.08	0.00	8.05
	Happiness	0.00	**41.18**	52.94	5.88	35.09	**54.39**	1.75	8.77

	Neutral	9.80	5.88	**74.51**	9.80	26.53	16.33	**40.82**	16.33
	Sadness	0.00	5.77	23.08	**71.15**	14.86	2.70	25.68	**56.76**
Male	Anger	**56.86**	23.53	7.84	11.76	**52.00**	30.67	17.33	0.00
	Happiness	29.63	**48.15**	14.81	7.41	26.47	**55.88**	14.71	2.94
	Neutral	7.27	21.82	**49.09**	21.82	17.39	8.70	**47.83**	26.09
	Sadness	7.41	5.56	3.70	**83.33**	10.20	10.20	26.53	**53.06**

The two datasets are developed by different institutes and the recording conditions for the two datasets are not comparable which should introduce a certain level of heterogeneity in our derived acoustic features in some degree and thus increase the difficulty of the machine based emotion recognition. The results in Table. 5-18 rather confirm to a certain extent the conclusion of previous studies on language independence in vocal emotion recognition. They show that the emotional clues in vocal speech signals can be recognized by an automatic approach in spite of different languages, and that the universal inherent characters of vocal emotions can be grasped by our acoustic feature set for the emotional speech.

5.4.4 Synthesis on the experimental results

The performances of our automatic derived HCS on Berlin dataset and DES dataset were presented in this section.

In Table. 5-19, we compare the performances between the automatically HCSs and the early empirical built hierarchical DECs on Berlin and DES datasets. Almost the same results are obtained by the two kinds of hierarchical classifiers for Berlin dataset (71.52% vs. 71.38%), while the empirical hierarchical DEC classifier for DES dataset performs slightly better than the automatically derived one (81.22% vs. 76.74%).

Table. 5-19 Comparison between the DEC and HCS on the two datasets (%)

		Female samples	Male samples	Mixed without gender information	Mixed with gender information
Berlin dataset	Empirical	71.89±2.97	75.75±3.15	68.60±3.36	71.52±3.85
	Automatic	71.75±3.10	73.77±2.33	57.95±2.87	71.38±2.33
DES dataset	Empirical	85.14±2.02	87.02±1.44	57.34±1.79	81.22±1.27
	Automatic	79.54±1.95	81.96±1.27	53.75±1.71	76.74±0.83

Although the automatically HCS do not offer better performance than the empirical ones, they make the possibility to avoid repeated empiric work when the emotion classification problem changes. From these experiments, we can draw several conclusions on our automatically derived hierarchical classifier scheme as to classifier structures, efficient features, and suitable combination operators.

Firstly, from the affective computing point of view, the automatically generated structures of the HCS classifiers on both datasets have the sub-classifiers in the arousal dimension (energy dimension) as the first stage, and then in the appraisal dimension in the later stages. These classifier structures fit well the joint definition of the emotions with the discrete emotional states mapped into dimensional space. Furthermore, the consistent appearance of the arousal dimension in the first stage shows that it is easier to make discrimination in the arousal dimension than in the appraisal dimension.

Secondly, the best features selected on the two datasets of emotional speech with different number and types of emotional states are consistent with several typical features (harmonic features 3, 7, 15, and 20, frequency features 23, 27, and 31, energy features 44, 46, 78, and 80, MFCC features 81, 89, 92, 100, and 107, and Zipf features 225 and 226, see feature descriptions in Annex A). The harmonic features work better on the appraisal dimension and the Zipf features work better on the arousal dimension. This consistency revealed on these two datasets further confirms that these features reflect the inherent clues of emotions in speech signals.

Compared to the features selected in the empirical built DEC (Table. 4-8, in section 4.4.2.2) and in the automatically derived hierarchical classification scheme (HCS) on Berlin dataset (Table. 5-11, section 5.4.1), the harmonic features seem to be more important in the empirical DEC than in the automatic HCS. The features are selected in the empirical DEC with a wrapper method using SFS whereas they are selected in HCS with ESFS, our embedded feature selection method as introduced in section 5.2. Both of these feature selection methods are dependent on the classification scheme or embedded in the classification scheme. Therefore, it is natural to have different features selected in these two classification schemes, DEC and HCS. However, several features were selected in both DEC and HCS, which rather proves the relevance of these features and the inherent correlates between these features and the emotions. The common selected features include features 23, 27, 28,

31 (frequency features), 46 (energy feature), 225 and 226 (Zipf features) in the classification of arousal dimension, and features 7, 11, 15, 20, 21 (harmonic features), 24, 28, 30, 31, 33, 35, 36, 38, 41, 42 (frequency features), 44, 46, 55, 56, and 64 (energy features).

The universal inherent characters of vocal emotions are also justified by our language independence experiments on the two datasets expressed respectively in German and Danish when one dataset is used in training and the other one for testing. Correct recognition rates of between 50% and 60% are achieved in these cross language experiments.

Thirdly, the different combination operators show different performance in the classification. Most of the best results on hierarchical classifiers are obtained with Schweizer & Sklar operator. The performance curves in Fig. 5-13 and Fig. 5-15 show that the operators Hamacher, Yager, Weber-Sugemo, and Schweizer & Sklar, especially the latter two, have better performance in the classification than the other operators. Referring to the property curve surfaces of the operators in Fig. 5-6, the operators that have properties of *convex* curve surfaces are more suitable in the hierarchical classification of the emotional speech.

5.5 Conclusion

In this chapter, we have introduced a new embedded feature selection scheme ESFS, which is then used as the basis for deriving an automatic method for building a hierarchical classifier HCS. Such a hierarchical classifier is represented by a binary tree whose root is the union of all emotion classes, leaves are single emotion classes and nodes are subsets containing several emotion classes obtained by a sub classifier. Each of these sub classifiers is based on a new embedded feature selection method, ESFS, which allows to easily represent classifiers characterized by their mass function that is the combination of the information given by an appropriate feature subset, each sub classifier having its own one. Benchmarked on Berlin and DES datasets, our automatic approach has shown its effectiveness for vocal emotion analysis, leading to comparable performance as compared to our previous empiric dimensional emotion model driven hierarchical classification scheme (DEC). Furthermore, we also studied the influence of the language for emotion analysis.

Many issues need to be further studied. For instance, from machine learning point of view, our ESFS needs to be compared with other feature selection scheme. Moreover, our automatic approach for building the HCS is rather empiric and its theoretic basis needs to be studied.

So far we have proposed an automatic scheme for automatically building hierarchical vocal emotional classifier which dealt with the problem that vocal emotion analysis is rather application dependent as there exist no universal agreement on the emotion definition. Another issue in machine recognition of vocal emotions is the fuzzy and subjective character of vocal emotion. In the following chapter, we make a preliminary attempt to deal with the subjective properties of emotions with an ambiguous classifier that allows multiple judgments of emotions on utterances.

Chapter 6

Automatic Ambiguous Classification Scheme - ACS

The previous chapter has dealt with the automatic way for building a multi-stage emotion classification problem. We turn in this chapter to another problem of vocal emotion, namely subjective character of human judgment that often leads to a multiple emotions labeling according to person's cultural background. This chapter is organized as follows. We first state the problem, and then describe our evidence theory based solution, an automatic ambiguous classification scheme (ACS) in section 6.2. Experimental results are discussed in section 6.3. Section 6.4 contains some remarks for further work.

6.1 Problem and our approach

The emotions are most of time subjective judgment of persons. Thus, the emotional states contained in a certain segment of speech may be not considered as one definite emotional state. Illustrated by dimensional emotion, basically, there are two problems to be highlighted: First, an emotion state is quite fuzzy and should be continuous, e.g. explosive happiness vs. calm happiness, or hot fury vs. cold anger; second, the judgment from the human being may be multiple as the feeling of emotion is subjective. In this chapter, we tend to make a preliminary approach on the automatic recognition of multiple judgments of the emotions.

In order to justify the possibility of multiple judgments by human, we made a human test on the Berlin dataset. Five human subjects were asked to classify the speech segments from the Berlin dataset according to the expressed emotion. Subjects were asked to mark emotional labels to the speech segments, in the case of difficult in

judging the emotional state, two or three labels are allowed to be marked on the same utterance.

The average confusion matrices in human testing are listed in Table. 6-1. Since the utterances are allowed to be labeled with multiple emotions, the sums of each row in the confusion matrix are possible to exceed 100%.

Table. 6-1 Confusion matrix in human judgment for multi-possibility on Berlin dataset (%)

	Human judge / Original label	Anger	Happiness	Fear	Neutral	Sadness	Boredom
Female	Anger	74.63	14.90	0	4.48	14.90	8.06
	Happiness	12.50	92.50	0	5.00	5.00	15.00
	Fear	10.34	0	96.55	10.34	13.79	13.79
	Neutral	0	0	2.50	95.00	10.00	20.00
	Sadness	0	0	0	2.86	100	8.57
	Boredom	0	0	0	22.22	2.22	93.33
Male	Anger	63.33	16.70	0	1.67	8.35	12.00
	Happiness	8.57	87.50	4.17	20.83	0	8.33
	Fear	7.70	0	96.15	3.85	0	0
	Neutral	0	0	0	100	0	0
	Sadness	0	0	0	5.88	100	0
	Boredom	0	0	0	25.40	17.65	67.65

According to these results, the emotion sadness is almost perfectly recognized. Utterances from every emotion have chances to be judged as neutral, especially for the emotions happiness and boredom. Utterances from anger and happiness have relatively higher chance to be confused between each other, and boredom tends to be misjudged as sadness.

In order to make the automatic recognition of emotions as close as possible to the judgments produced by humans, we also developed an ambiguous classifier that allows multiple labels to emotional speech. Results from the multiple human judgments on Berlin dataset are used as ground truth for our experiments.

In our work, different sub-classifiers concerning all the emotional classes or some of the classes are evaluated and the best few classifiers are selected to form the ambiguous classifier based on the evidence theory of Dempster-Shafer [Sha90] [Sha92] [Fio04]. This theory is chosen for its ability for modeling and quantifying the

assigned belief to facts by giving an order of confidence to these facts [Tel04]. Moreover, the beliefs can be combined by an orthogonal sum with the Dempster's combination rule (See section 5.2.2). The principle of building the sub-classifiers follows the same ESFS scheme as used in the hierarchical approach.

Our automatic ambiguous classifier is only a preliminary attempt on the ambiguous recognition of emotions. Indeed, only the possibility of the multiple judgments is considered in our experiments, while the continuity of the emotions is not yet investigated, and should be discussed in our future work.

6.2 Principle of the ambiguous classifier

In our work for ambiguous classification, several sub-classifiers on simple classification problems with fewer classes are also used to perform the classification. Instead of placing the sub-classifiers hierarchically as in the previous approach (section 5.3), the sub-classifiers process the classification in parallel, and the results from the sub-classifiers are combined thanks to the fusion of the belief masses from each of the emotional states,according to the Dempster's . The aim of this approach is to make possible the multiple judgments for emotions in case of utterances with ambiguous emotions. That is to say, when the emotion states contained in a certain utterance are between some emotional states considered in the problem, all the emotional states with relatively high beliefs can be presented in the judgment.

The generation process of the ambiguous classifier (ACS) is shown in Fig. 6-1.

Likewise in the hierarchical classifier, the N discrete emotional/mood states concerned in the classification are first assigned as set of hypotheses $\Omega=\{E_1, E_2, \ldots, E_N\}$. For example, for the emotion classification on Berlin dataset, Ω_{Belin}={Anger, Happiness, Fear, Neutral, Sadness, Boredom}.

Three fundamental steps are taken in the generation of the ACS. First, the possible sub-classifiers are proposed, then the sub-classifiers are evaluated and several among the best sub-classifiers are chosen according to the classification performance. The selected sub-classifiers then pass a fusion process to get the final beliefs of each emotion. The detailed processes in these three steps are listed as below:

1) Step 1: building of possible sub-classifiers

The set Ω is first divided into different groups of non-empty subsets, and each group can correspond to a sub-classifier. Suppose the n^{th} group of subsets G_n contains N_n subsets A_{n_1},…, $A_{n_{N_n}}$, where the group G_n correspond to the n^{th} sub-classifier, and each subset A_{n_i} presents a class in the n^{th} sub-classifier. Each group satisfies

$$\begin{aligned} &G_n = A_{n_1} \cup ... \cup A_{n_{N_n}} \\ &G_n \subseteq \Omega \\ &A_{n_i} \neq \phi, 1 < i < N_n \\ &A_{n_i} \cap A_{n_j} = \phi, i \neq j \end{aligned} \tag{6.1}$$

The union of the subsets in each group are do not necessary equals the set of hypotheses Ω, as the missing emotion in one group may be derived from the fusion of other sub-classifiers.

For example, we can have the following possible groups to be classified as sub-classifiers on Ω_{Belin}:

$$G_{i_1} = \{Anger, Happiness\} \cup \{Fear, Neutral\} \cup \{Sadness, Boredom\}$$

$$G_{i_2} = \{Anger\} \cup \{Happiness, Fear, Neutral\}$$

$$G_{i_3} = \{Happiness, Fear\} \cup \{Neutral, Sadness\}$$

Taken the group G_{i_1} as example, it corresponds to a classification problem concerning three classes: the first class as Anger & Happiness, the second class as Fear & Neutral, and the third class as Sadness & Boredom. A sub-classifier on these three classes can be built based on this group.

In order to avoid excessive computational load, not all the possible groups of subsets are evaluated. Since the classification problems with fewer classes tend to get better performance under the same situation, the number of subsets (classes) in the groups is limited to $N_n \leq 3$.

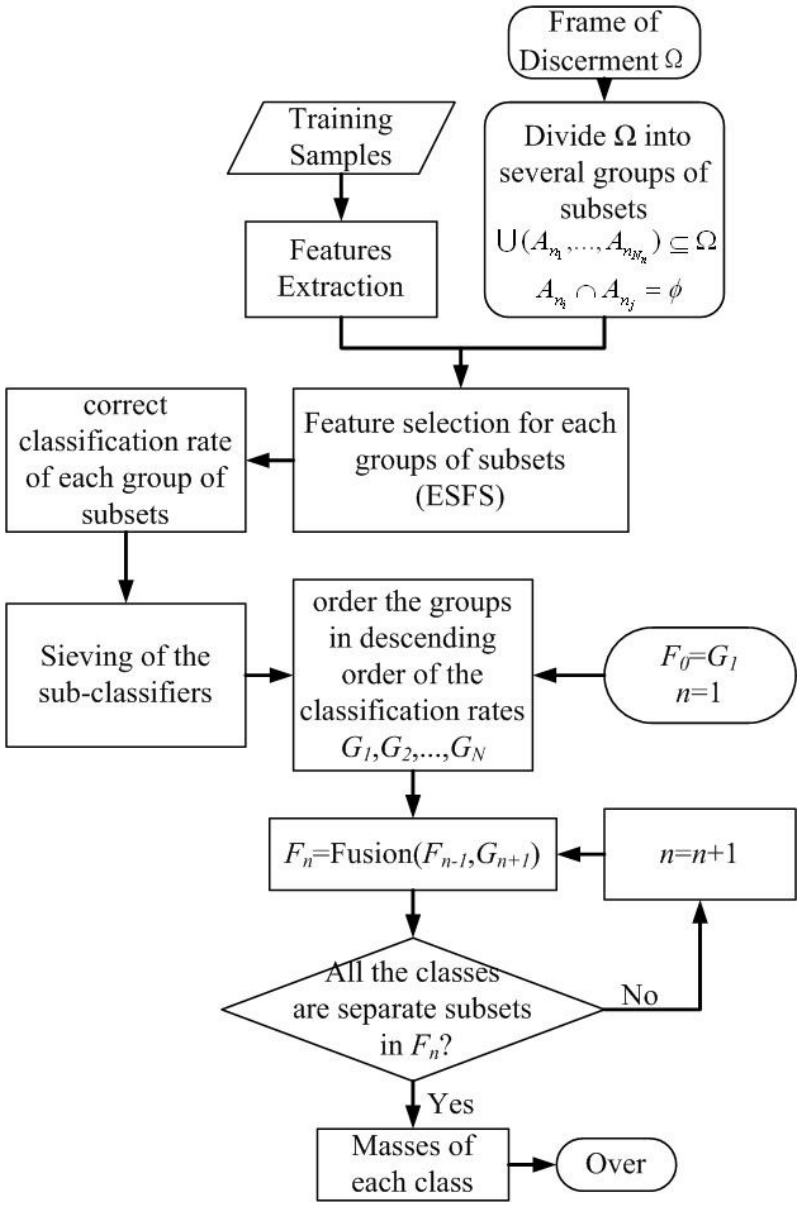

Fig. 6-1 Generation of the ambiguous classifier

The subsets in the same group are classified with the feature combination and selection scheme proposed in section 5.2.3.2. For each group, the belief masses of each audio segment belonging to each of the subsets can be obtained with the selected combined feature with the given operator.

Although the belief masses are given as the output of the sub-classifiers, the judgment of the class whose the utterance belongs to is still needed to get the correct classification rates in order to make evaluations of the performance of the sub-classifiers. If the subset with the highest values is assigned as the

judged class, the correct classification rate R_{0n} can be calculated for the n^{th} groups of subsets G_{0n}.

2) Step 2: Sieving of the sub-classifiers

The groups of subsets are reordered in descending order according to the correct classification rates. From the rules of making groups of subsets (classes) according to equation (6.1), up to several hundred groups of sub-classifiers will be evaluated for a classification problem with six or more classes in the previous step. As the final step will be fusions between the sub-classifiers according to the Dempster's rule of combination, a sieving step to the sub-classifiers is applied here before the fusions.

The Dempster's rule of combination, which is based on an orthogonal sum of the mass functions, requires the mass assignments to be independent to each other. Since the feature selection and classification of sub-classifiers on different groups of emotional subsets (classes) are processed separately, we assume the sub-classifiers based on the groups with distinct classes are approximately independent to each other, because the classes to be discriminated in these sub-classifiers are very different to each other, and thus the features selected in these sub-classifiers are supposed to be different. While some of the groups are rather "similar" to each other, it may lead to similar features selected, and thus some of the groups need to be deleted to meet the independent requirement according to the Dempster's rule.

For the groups considered to be "similar" to each other, only the group with the highest classification rate is kept; and all the other groups are deleted. For example, if 4 groups $G_{0n_1}, G_{0n_2}, G_{0n_3}, G_{0n_4}$ have classification rates as $R_{n_2} > R_{n_4} > R_{n_1} > R_{n_3}$, only the group G_{0n_2} is kept; and the other 3 groups are deleted.

Three criterions of judging "similar" groups are applied as follows:

a) The groups with the identical union of all the subsets.

For example, if the classification rates for the 4 groups

$$G_{0n_1} = \{E_1\} \cup \{E_2\} \cup \{E_3\}$$

$$G_{0n_2} = \{E_1\} \cup \{E_2, E_3\}$$

$$G_{0n_3} = \{E_2\} \cup \{E_1, E_3\}$$

$$G_{0n_4} = \{E_3\} \cup \{E_1, E_2\}$$

The union of all the subsets in these four groups is $\{E_1, E_2, E_3\}$, the four groups are considered as similar.

b) The groups with at least one common subset, and at least one common element in the other subsets.

For example, for the 2 groups

$$G_{0n_1} = \{E_1\} \cup \{E_2, E_3\}$$

$$G_{0n_2} = \{E_1\} \cup \{E_2, E_4\}$$

The subset $\{E_1\}$ is common for the 2 groups and the element E2 is common in the other subsets, the two groups are considered as similar.

c) The groups with all subsets included in the subsets of another group.

For example, for the 2 groups

$$G_{0n_1} = \{E_1, E_2\} \cup \{E_3, E_4\}$$

$$G_{0n_2} = \{E_1\} \cup \{E_4\}$$

The subset $\{E_1\}$ is included in the subset $\{E_1, E_2\}$, and the subset $\{E_4\}$ is included in the subset $\{E_3, E_4\}$, the two groups are considered as similar.

After the step of sieving of the groups, the kept groups are ordered in descending order according to the correct classification rate of the sub-classifiers as $G_1, G_2, \ldots, G_N$.

3) Step 3: Fusion between the sub-classifiers – application of the evidence theory

A step of fusion is applied to combine the sub-classifiers to get final decisions of the classification. The Dempster's rule of combination is applied to make data fusion between the groups of subsets. Each step of fusion is made between two groups of subset from groups G_1 and G_2. New fusion continues between the previous result subsets and the next group G until all the classes are separated in the subsets.

The total number of fusions made K is defined as the fusion depth. The evidence theory and the Dempster's rule of combination are introduced in section 5.2.2. An example of fusion of 2 groups is shown in Fig. 6-2.

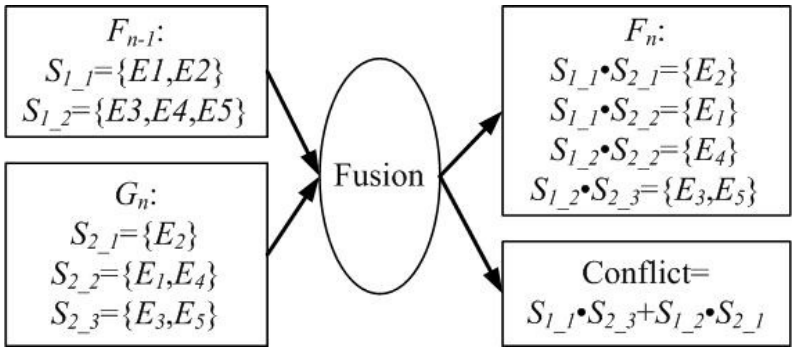

Fig. 6-2 Example of fusion of 2 groups of subsets

Fig. 6-3 shows the ambiguous classifier with fusion depth of K.

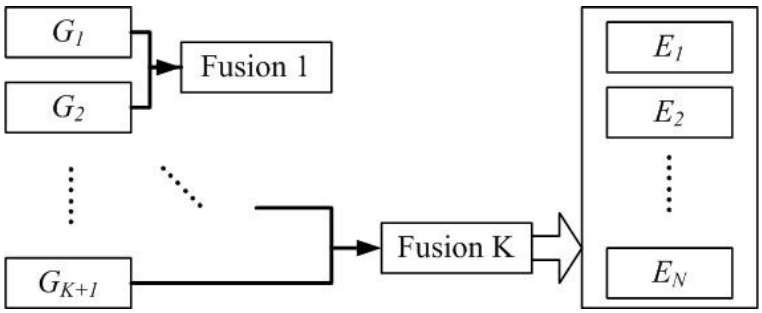

Fig. 6-3 Ambiguous classifier with fusion depth of K

The output of the classifier is the belief masses of the emotional states which satisfies

$$\sum_{i=1}^{N} m(E_i) = 1 \qquad (6.2)$$

To compute the correct classification rate with the belief masses, two ways of evaluation are proposed in this approach. The first way, as the traditional classifiers with definite single judgment of recognized class for each sample, the class with the highest mass is judged as the recognized result. The second way considers multiple possibilities of result. All the classes with masses larger than 0.3 are considered as possible recognition results.

6.3 Experimental results

In this section, the experiments of the ambiguous classification scheme (ACS) are also carried out on Berlin dataset and DES dataset. The structures of the ambiguous classifiers and the experimental results are displayed in subsection 6.3.1 and 6.3.2 respectively. In subsection 6.3.3, we analyzed and summarized the experimental results.

6.3.1 Experiments on Berlin dataset

The approach of the ambiguous classification is first validated on Berlin dataset.

The automatically generated ACS for the six classes in Berlin dataset is shown in Fig. 6-4. The same structure is generated with all the operators and parameters tested in our experiment for each gender with a little difference between the two genders. Five sub-classifiers are selected to form five steps of fusions to get the belief masses of the six emotional states.

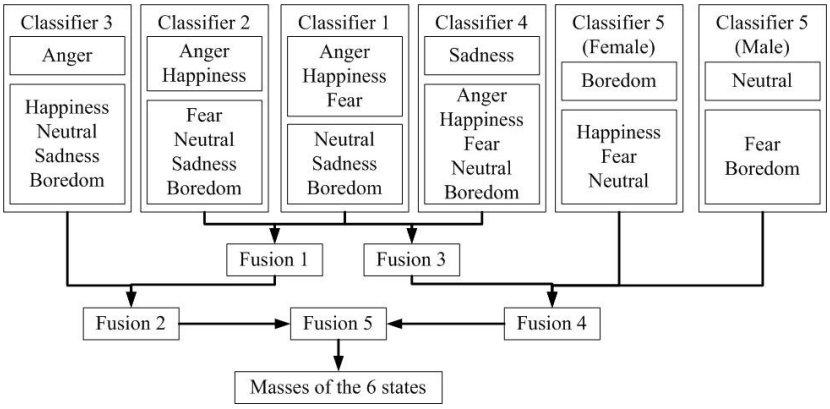

Fig. 6-4 ACS for Berlin dataset

The outputs of the ambiguous classifier are the belief masses of the emotions for each utterance. We proposed two ways of evaluations: traditional single judgment and judgment of multiple possibilities. For the evaluation of single judgment, the utterances are classified as the emotional state with the highest belief mass among all the six emotions; and for the multiple judgment, all the emotional states with higher belief masses than a certain threshold are taken as possible results in the classification. In our experiments, the threshold in the multiple judgments is set to 0.3. For both of the evaluations, the classification results are obtained by the belief masses from the

same classifiers, only the way of applying the belief masses makes the difference between single and multiple judgments.

The average classification rates and the root mean square errors of the classification rates are computed with several operators with different parameters for the two ways of evaluations. The emotion labels obtained in the human testing are taken as the ground truth in the evaluation of multiple judgments. The detailed results are listed in Table. 6-2 (female samples), Table. 6-3 (male samples), Table. 6-4 (mixed samples with single judgment), and Table. 6-5 (mixed samples with multiple judgments).

Table. 6-2 Berlin database, hold-out cross-validation with 10 iterations, female samples in ACS (%). S means the results in classification with single judgment, and M means with multiple judgments

Operator		Lukasiewicz	Hamacher			
Parameter		-	γ=0.5	γ=3	γ=5	γ=10
Mean rate	S	69.54±2.28	69.62±1.74	69.27±3.05	69.27±2.24	70.19±2.07
	M	70.30±2.29	75.71±1.66	74.68±2.49	74.00±2.30	73.32±2.41
Operator		Yager				Weber-Sugemo
Parameter		p=1	p=3	p=5	p=10	λ_T=-0.5
Mean rate	S	69.11±2.86	72.10±2.90	68.33±2.37	66.52±2.36	67.22±1.80
	M	70.22±2.23	**75.81±1.56**	73.81±1.73	72.39±1.75	70.19±1.49
Operator		Weber-Sugemo				
Parameter		λ_T=0	λ_T=0.5	λ_T=2	λ_T=4	λ_T=5
Mean rate	S	67.65±2.59	68.63±2.45	70.84±1.65	71.70±1.89	**72.26±1.82**
	M	70.30±2.29	71.38±2.11	72.97±2.29	74.64±2.40	74.24±2.70
Operator		Schweizer & Sklar				
Parameter		q=0.2	q=0.4	q=0.6	q=0.8	q=1
Mean rate	S	71.67±2.32	69.112.89	70.62±2.86	69.51±3.21	68.71±2.98
	M	74.76±2.17	75.27±2.66	74.76±2.77	75.61±2.44	75.74±1.61
Operator		Schweizer & Sklar		Frank		
Parameter		q=3	q=5	s=2	s=5	s=8
Mean rate		67.98±2.39	68.49±1.76	54.61±1.12	62.94±1.67	65.88±2.31
		74.76±1.66	72.74±1.84	60.13±1.13	66.87±2.53	69.73±1.74
Operator		Frank			Average	Geometric Average

Parameter		s=10	s=12	s=15	-	-
Mean rate	S	67.12±1.47	68.60±2.17	69.92±1.83	59.41±2.62	63.07±2.31
	M	70.58±1.67	68.72±1.40	70.08±2.30	64.74±2.87	68.59±1.93

Average confusion matrix for the best parameter (%)

	Predicted / Actual	E1	E2	E3	E4	E5	E6
Single judgment	E1	90.11	4.83	3.91	1.15	0.00	0.00
	E2	47.02	42.28	6.84	3.86	0.00	0.00
	E3	26.67	10.26	47.95	11.54	1.79	1.79
	E4	0.00	0.00	4.90	72.86	2.04	20.20
	E5	0.00	0.14	2.30	0.41	92.03	5.14
	E6	0.31	0.46	5.69	13.08	14.15	66.31
Multiple judgments		E1	E2	E3	E4	E5	E6
	E1	80.19	15.82	9.89	5.36	1.69	1.23
	E2	38.95	54.04	7.89	6.02	2.69	2.98
	E3	20.43	17.18	52.48	16.92	6.32	5.21
	E4	1.9	2.65	7.48	82.38	9.66	8.64
	E5	1.08	0.72	2.39	2.39	92.07	4.95
	E6	2.41	3.79	11.38	21.95	18.31	58.82

Table. 6-3 Berlin database, hold-out cross-validation with 10 iterations, male samples in ACS with single judgment.

Operator		Lukasiewicz	Hamacher			
Parameter		-	γ=0.5	γ=3	γ=5	γ=10
Mean rate	S	70.58±2.18	71.95±1.89	72.09±2.55	73.08±2.84	73.36±2.87
	M	72.32±2.66	75.31±1.76	75.87±1.86	75.67±2.21	75.06±1.93
Operator		Yager				Weber-Sugemo
Parameter		p=1	p=3	p=5	p=10	λ_T=-0.5
Mean rate	S	70.24±2.40	73.56±1.36	70.10±2.62	67.88±2.21	70.07±1.87
	M	72.24±2.73	76.29±2.47	73.31±2.87	71.64±2.48	72.37±1.68
Operator		Weber-Sugemo				
Parameter		λ_T=0	λ_T=0.5	λ_T=2	λ_T=4	λ_T=5
Mean rate	S	70.62±2.75	70.96±2.41	72.36±1.76	73.39±1.81	**74.62±1.82**
	M	72.32±2.66	72.74±2.40	74.32±1.62	75.26±1.53	75.65±2.03
Operator		Schweizer & Sklar				
Parameter		q=0.2	q=0.4	q=0.6	q=0.8	q=1

Mean rate	S	73.15±2.43	72.33±2.71	72.60±2.01	72.53±2.05	71.71±2.37
	M	75.34±2.27	76.26±2.30	**76.50±1.69**	76.44±1.60	75.48±1.85
Operator		Schweizer & Sklar		Frank		
Parameter		q=3	q=5	s=2	s=5	s=8
Mean rate	S	70.72±2.55	69.11±2.23	57.30±2.98	62.71±2.94	67.54±3.19
	M	74.69±1.65	72.41±2.45	60.67±2.57	67.42±4.27	69.91±2.24
Operator		Frank			Average	Geometric Average
Parameter		s=10	s=12	s=15	-	-
Mean rate	S	67.71±3.04	68.25±2.72	69.59±2.23	57.09±1.96	65.10±2.90
	M	70.42±2.73	69.87±2.75	70.45±2.12	60.25±2.58	66.82±2.12

Average confusion matrix for the best parameter (%)

	Predicted / Actual	E1	E2	E3	E4	E5	E6
Single	E1	91.07	6.00	2.00	0.80	0.00	0.13
	E2	31.18	57.06	7.65	2.94	0.00	1.18
	E3	8.21	13.59	65.90	7.44	2.56	2.31
	E4	0.87	1.09	6.30	81.09	1.74	8.91
	E5	0.00	0.00	1.22	1.84	86.73	10.20
	E6	1.02	1.63	3.67	11.02	32.24	50.41
Multiple	Predicted / Actual	E1	E2	E3	E4	E5	E6
	E1	88.53	6.93	4.36	3.51	0.13	0.58
	E2	27.06	62.65	9.12	9.51	0.29	1.67
	E3	5.56	12.91	72.48	10.77	4.44	3.59
	E4	2.03	3.99	8.70	80.29	5.29	7.54
	E5	1.02	1.29	6.12	5.85	81.22	13.81
	E6	1.50	2.93	10.48	16.80	30.95	49.80

The same neural network based gender classifier as in section 4.3.2 is also used in the gender classification of the ambiguous classification of emotion on Berlin dataset.

Table. 6-4 Berlin database, hold-out cross-validation with 10 iterations, all samples in ACS: single judgment.
Rate 1: with gender classification; Rate 2: without gender classification

Operator	Lukasiewicz	Hamacher				Yager
Parameter	-	γ=0.5	γ=3	γ=5	γ=10	p=1

Rate 1 (%)	68.03±2.19	67.32±1.78	67.93±2.80	68.96±2.57	70.09±2.51	68.23±2.59
Rate 2 (%)	58.34±1.88	57.93±2.93	57.12±2.77	56.86±2.27	57.93±2.37	57.68±2.33
Operator	Yager			Weber-Sugemo		
Parameter	p=3	p=5	p=10	λ_T=2	λ_T=4	λ_T=5
Rate 1 (%)	70.12±2.14	65.51±2.49	63.63±2.24	69.70±1.69	68.97±1.85	**71.83±1.80**
Rate 2 (%)	58.73±3.47	57.71±2.80	56.55±2.94	59.94±2.18	60.80±1.95	61.04±1.94

Operator	Schweizer & Sklar				
Parameter	q=0.4	q=0.6	q=0.8	q=1	q=3
Rate 1 (%)	67.10±2.77	68.10±2.45	68.57±2.65	67.39±2.70	66.23±2.47
Rate 2 (%)	58.01±2.72	58.05±3.38	58.25±3.54	57.33±3.34	57.07±2.71

Operator	Frank			Average	Geometric Average
Parameter	s=10	s=12	s=15	-	-
Rate 1 (%)	64.20±2.28	66.45±2.45	67.74±2.07	55.69±2.25	61.75±2.60
Rate 2 (%)	58.36±1.35	59.25±1.34	59.73±1.53	51.67±3.61	53.50±3.38

	Predicted / Actual	E1	E2	E3	E4	E5	E6
Best confusion matrix with gender classification	E1	87.7	5.85	3.49	1.59	0.65	0.71
	E2	38.22	48.38	7.61	3.92	0.65	1.22
	E3	17.41	12.11	55.35	9.77	2.74	2.62
	E4	1.07	1.18	6.03	74.62	2.47	14.64
	E5	0.65	0.72	2.34	1.73	86.54	8.02
	E6	1.29	1.66	5.15	12.23	22.94	56.73
Best confusion matrix without gender classification	Predicted / Actual	E1	E2	E3	E4	E5	E6
	E1	71.36	15.31	6.11	6.42	0.37	0.43
	E2	32.64	38.57	14.73	12.64	0.66	0.77
	E3	14.87	8.33	31.54	28.97	9.62	6.67
	E4	0.00	0.53	5.47	63.47	11.05	19.47
	E5	0.00	0.08	2.28	0.81	87.72	9.11
	E6	0.18	0.44	3.95	12.11	29.65	53.68

Table. 6-5 Berlin database, hold-out cross-validation with 10 iterations, all samples in ACS: multiple judgments
Rate 1: with gender classification; Rate 2: without gender classification

Operator	Lukasiewicz	Hamacher				Yager
Parameter	-	γ=0.5	γ=3	γ=5	γ=10	p=1
Rate 1 (%)	68.03±2.20	68.43±2.15	68.22±2.90	69.36±2.53	70.50±2.51	68.33±2.69

Rate 2 (%)	58.51±1.85	59.67±2.68	57.89±2.78	57.53±2.33	58.43±2.28	57.81±2.38
Operator	Yager			Weber-Sugemo		
Parameter	p=3	p=5	p=10	λ_T=2	λ_T=4	λ_T=5
Rate 1 (%)	70.94±1.93	66.98±2.72	65.46±2.19	69.70±1.63	69.10±1.85	**72.32±1.83**
Rate 2 (%)	60.03±3.31	59.89±2.68	59.74±2.73	60.29±2.03	61.33±2.02	**61.55±1.82**

Operator	Schweizer & Sklar				
Parameter	q=0.4	q=0.6	q=0.8	q=1	q=3
Rate 1 (%)	67.68±2.73	68.54±2.62	69.56±2.67	68.24±2.78	67.17±2.60
Rate 2 (%)	58.76±2.61	58.90±3.44	59.80±3.54	58.85±3.49	59.28±2.61

Operator	Frank			Average	Geometric Average
Parameter	s=10	s=12	s=15	-	-
Rate 1 (%)	64.22±2.32	66.58±2.44	67.89±1.97	59.42±2.83	63.82±2.84
Rate 2 (%)	58.67±1.38	59.49±1.39	59.94±1.49	57.04±4.05	57.92±3.15

Best confusion matrix with gender classification	Predicted / Actual	E1	E2	E3	E4	E5	E6
	E1	87.99	5.96	6.09	3.55	0.89	1.05
	E2	39.30	47.81	11.82	8.65	1.15	2.13
	E3	22.00	11.96	56.57	11.68	3.68	4.66
	E4	1.88	4.35	7.17	76.56	6.16	11.78
	E5	0.88	2.08	4.13	3.91	86.50	8.35
	E6	1.28	3.70	8.59	15.83	27.60	53.18

Best confusion matrix without gender classification	Predicted / Actual	E1	E2	E3	E4	E5	E6
	E1	71.46	15.56	6.89	7.16	0.53	0.58
	E2	33.00	38.86	15.57	13.63	1.06	1.10
	E3	15.34	8.93	31.97	29.49	10.09	7.22
	E4	0.04	0.84	5.93	63.61	11.75	19.79
	E5	0.03	0.24	2.47	1.54	87.75	9.21
	E6	0.23	0.61	4.04	13.16	30.03	53.86

Fig. 6-5 shows the best classification rates of the tested operators with single judgment, and Fig. 6-6 shows the best classification rates with multiple judgments. In both of the figures, the curve “All samples (1)” refers to the case of classification for the utterances from mixed genders with gender classification, and the curve “All samples (2)” refers to the case of classification for the utterances from mixed genders without gender classification. The error bars in the figures show the root mean square errors of the classification rates. From the difference between the curves “All samples

(1)" and "All samples (2)", the gender classification shows an obvious improvement in the overall classification performance for the mixed gender samples.

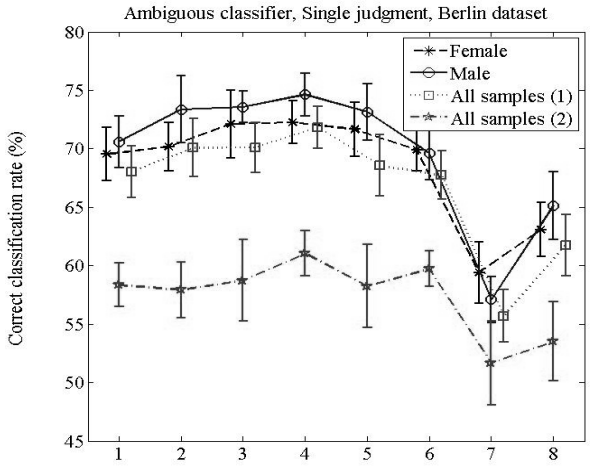

Fig. 6-5 Classification rate with ACS with single judgment for Berlin dataset

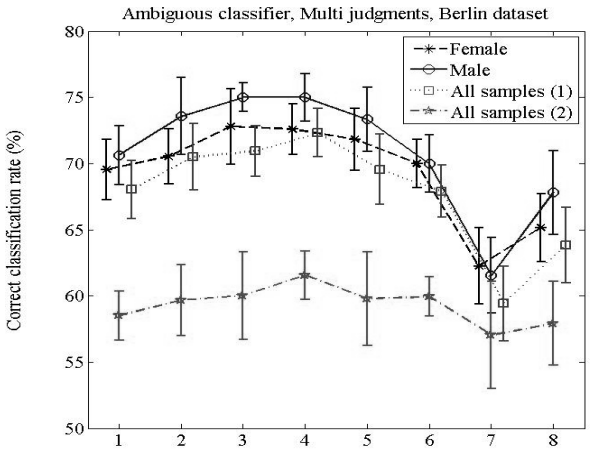

Fig. 6-6 Classification rate with ACS with multiple judgments for Berlin dataset

In the case of single judgment, the best result for female samples is 72.26%±1.82%, 74.62%±1.82% for male samples, 71.83%±1.80% for mixed genders with gender classification and 61.04%±1.94% for mixed genders without gender classification, all these cases with Weber-Sugemo operator when λ_T=5.

In the case of multiple judgments, the best result for female samples is 75.81%±1.56% with Yager operator when p=3, 76.50%±1.69% with Schweizer & Sklar when q=0.6 for male samples, 72.32%±1.83% for mixed genders with gender

classification and 61.55%±1.82% for mixed genders without gender classification with Weber-Sugemo operator when λ_T=5. The multiple labels in human testing on Berlin dataset are taken as ground truth.

Same as the case with the approach of hierarchical classification, the operators with convex curve surfaces of properties give better results than the others.

Distances between the confusion matrices obtained from human testing and automatic ambiguous classification with multiple judgments for the two genders respectively. The values of percentages are taken when computing the distances. The root mean square value of the difference between the matrices is considered to represent the distance:

$$D = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(CF_1(i,j) - CF_2(i,j))^2}{N^2}} \qquad (6.3)$$

where N is the number of emotions, and CF_1 and CF_2 are the two confusion matrices respectively. Meanwhile, the distances that evaluate each emotion are also computed on each line of the confusion matrices as

$$D_i = \sqrt{\frac{\sum_{j=1}^{N}(CF_1(i,j) - CF_2(i,j))^2}{N}} \qquad (6.4)$$

The distances between the results of human testing and automatic ambiguous classification with multiple judgments on Berlin dataset (presented on percentage) are listed in Table. 6-6.

Table. 6-6 Distance between the results of human testing and automatic ambiguous classification

	Whole matrix	E1	E2	E3	E4	E5	E6
Female	14.20	7.65	19.97	20.46	7.35	3.73	16.33
Male	11.50	12.59	13.88	11.63	9.74	9.86	10.72

From the distances listed in Table. 6-6, we can see that the judgments on emotional states neutral and sadness by the automatic approach is closer to human testing than the other emotions.

The best features are analyzed for the two genders respectively. The indexes of the most frequently selected features are listed according to the feature groups in Table. 6-7. Up to five features and the percentages of the features are listed for each group, and the features are ordered according to the frequency of selection. The Zipf features are calculated together with the harmonic features because there are only 2 Zipf features in our feature set. The description of the most important features is shown in Annex A.

Table. 6-7 Most frequently selected features for the ambiguous classifier for Berlin dataset

Feature Groups	Harmonic	Zipf	Frequency	Energy & rhythm	MFCC
Female	3,22,20,7,15	225,226	27,23,28,31,36	46,80,78,55,49	89,223,224,129,99
Male	3,7,20,11,15	225,226	27,23,25,30,35	46,78,80,76,56	92,162,163,81,164

Percentages of number of features in each group		
group	Female	Male
Harmonic & Zipf	2	3
Frequency	35	43
Energy & rhythm	22	28
MFCC	41	26

The best and most frequently selected feature is the 46th feature: ratio of energy below 250 Hz, which refers to the distribution of the energy in the spectrum and thus partially reflects the tonality of the speech signal. The harmonic features that describe the timber structure and Zipf features that better capture the prosody information occupy a small percentage of the best features in numbers while they have high priorities in the sequence of the selected features.

6.3.2 Experiments on DES dataset

The same experiments as the ones on the Berlin dataset are also carried on DES dataset, which is also used in the work in the literature, for example, in the work by Ververidis [Ver04a] [Ver04b] [Ver05a] [Ver05b].

The ambiguous classifier generated for the five classes of DES dataset is shown in Fig. 6-7. The classifiers for the two genders are different. Four sub-classifiers are selected to form four steps of fusions to get the belief masses of the five emotional states.

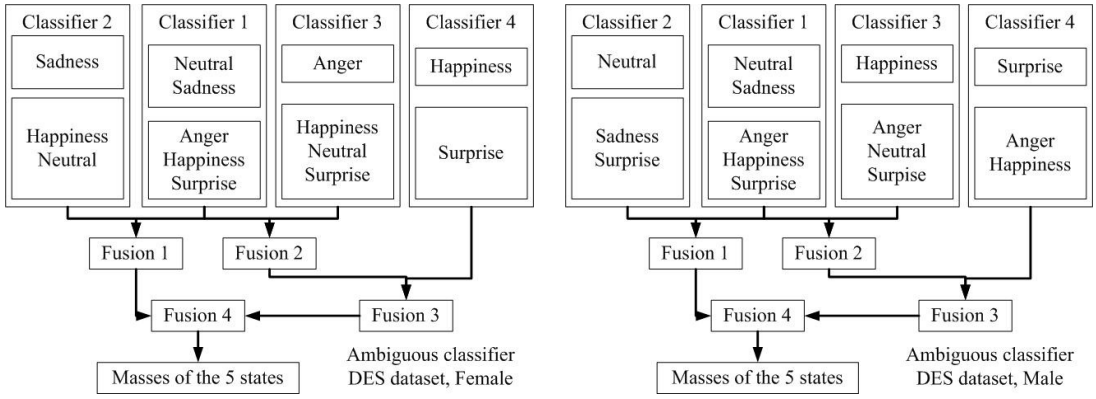

Fig. 6-7 Ambiguous classifier for DES dataset

The evaluation of multiple judgments on the classification for the DES dataset is based on the original emotional labels of the DES dataset as no human testing with multiple labels has been performed on this dataset. The emotion of a certain utterance is judged as correct when the original emotional label of the utterance of the DES dataset appear in the automatically judged emotions.

The average classification rates and the root mean square errors of the classification rates are computed with several operators with different parameters in two ways of evaluations: single judgment and judgment of multiple possibilities. The detailed results are listed in Table. 6-8 (female samples), Table. 6-9 (male samples), Table. 6-10 (mixed samples with single judgment), and Table. 6-11 (mixed samples with multiple judgments).

Table. 6-8 DES database, hold-out cross-validation with 10 iterations, female samples in ACS (%). S means the results in classification with single judgment, and M means with multiple judgments

Operator		Lukasiewicz	Hamacher			
Parameter		-	γ=0.5	γ=3	γ=5	γ=10
Mean rate	S	68.42±2.99	75.37±3.01	78.88±1.83	78.65±1.42	78.53±2.31
	M	68.65±2.78	78.80±1.99	79.77±1.86	79.31±1.57	79.03±2.04
Operator		Yager				Weber-Sugemo
Parameter		p=1	p=3	p=5	p=10	λ_T=-0.5
Mean rate	S	70.62±2.40	76.25±2.29	69.42±2.87	64.71±7.35	73.28±2.38
	M	70.73±2.45	79.61±2.42	74.94±2.58	72.20±6.11	73.6±2.31
Operator		Weber-Sugemo				
Parameter		λ_T=0	λ_T=0.5	λ_T=2	λ_T=4	λ_T=5
Mean	S	73.13±1.40	75.94±1.32	75.06±3.26	**79.54±2.51**	78.84±1.89

rate	M	73.14±1.40	75.98±1.36	75.79±2.70	80.35±2.21	79.92±1.95
Operator		Schweizer & Sklar				
Parameter		q=0.2	q=0.4	q=0.6	q=0.8	q=1
Mean rate	S	78.12±1.61	78.15±2.54	78.92±1.77	78.19±2.11	78.30±1.83
	M	80.46±1.62	79.42±2.44	80.42±1.88	80.46±2.55	**80.85±1.46**
Operator		Schweizer & Sklar		Frank		
Parameter		q=3	q=5	s=2	s=5	s=8
Mean rate	S	71.89±2.75	71.12±2.16	38.42±2.27	47.22±7.51	70.85±3.37
	M	77.34±2.01	72.91±1.91	43.96±2.20	51.26±6.89	71.79±3.36
Operator		Frank			Average	Geometric Average
Parameter		s=10	s=12	s=15	-	-
Mean rate	S	68.65±3.56	71.39±3.34	69.34±4.08	42.66±6.34	60.58±6.55
	M	70.23±3.99	72.24±3.57	70.04±4.04	59.15±5.98	68.46±4.32

Average confusion matrix for the best parameter (%)

	Predicted / Actual	E1	E2	E3	E4	E5
Single	E1	83.53	5.29	6.47	2.35	2.35
	E2	12.55	75.29	3.73	0.78	7.65
	E3	3.73	3.33	83.53	5.49	3.92
	E4	0.19	1.35	10.77	77.5	10.19
	E5	0.56	10.37	4.81	6.30	77.96
Multiple	Predicted / Actual	E1	E2	E3	E4	E5
	E1	83.14	4.31	5.10	5.29	2.16
	E2	15.49	70.20	3.33	3.14	7.84
	E3	5.69	3.33	79.41	7.06	4.51
	E4	0.58	3.27	9.62	75.96	10.58
	E5	0.37	6.85	2.59	7.59	82.59

Table. 6-9 DES database, hold-out cross-validation with 10 iterations, male samples in ACS (%). S means the results in classification with single judgment, and M means with multiple judgments.

Operator		Lukasiewicz	Hamacher			
Parameter		-	γ=0.5	γ=3	γ=5	γ=10
Mean rate	S	73.38±2.16	79.13±2.50	80.91±1.91	80.84±1.51	80.33±1.54
	M	73.42±2.21	82.18±2.02	81.93±2.08	81.38±1.58	80.76±1.81
Operator		Yager				Weber-Sugemo

Parameter		p=1	p=3	p=5	p=10	λ_T=-0.5
Mean rate	S	74.33±2.30	79.35±1.08	75.24±2.19	71.82±3.90	68.58±2.08
	M	74.36±2.29	81.35±1.36	79.13±2.06	76.55±3.72	68.87±2.32
Operator		Weber-Sugemo				
Parameter		λ_T=0	λ_T=0.5	λ_T=2	λ_T=4	λ_T=5
Mean rate	S	71.89±1.60	73.78±1.78	78.55±1.47	79.38±2.33	80.40±2.07
	M	71.90±1.59	73.90±1.75	78.95±1.88	80.07±2.77	80.73±2.24
Operator		Schweizer & Sklar				
Parameter		q=0.2	q=0.4	q=0.6	q=0.8	q=1
Mean rate	S	78.62±2.79	81.05±1.60	81.02±1.96	**81.42±1.98**	81.09±1.14
	M	79.17±2.56	82.07±1.71	82.36±2.00	**82.91±1.66**	82.73±1.33
Operator		Schweizer & Sklar		Frank		
Parameter		q=3	q=5	s=2	s=5	s=8
Mean rate	S	77.27±2.48	66.54±7.90	50.95±0.68	66.25±5.15	66.80±3.86
	M	80.69±2.96	68.09±7.64	53.48±0.73	67.78±4.85	67.37±4.08
Operator		Frank			Average	Geometric Average
Parameter		s=10	s=12	s=15	-	-
Mean rate	S	68.11±3.67	69.96±1.88	71.67±1.89	59.53±6.25	68.84±4.64
	M	69.02±3.60	70.36±1.85	72.07±1.93	69.09±4.00	74.69±4.56

Average confusion matrix for the best parameter (%)

	Predicted / Actual	E1	E2	E3	E4	E5
Single	E1	66.86	22.35	0.39	2.35	8.04
	E2	1.48	93.89	1.85	2.04	0.74
	E3	1.64	10.73	76.00	5.64	6.00
	E4	0.37	1.30	12.78	84.26	1.30
	E5	9.51	1.97	1.31	2.30	84.92
Multiple	Predicted / Actual	E1	E2	E3	E4	E5
	E1	67.12	23.92	2.16	4.18	9.54
	E2	3.02	94.20	3.95	3.46	1.79
	E3	3.39	15.15	77.03	11.27	9.21
	E4	1.54	4.63	14.51	84.69	5.19
	E5	11.69	4.64	2.73	4.97	85.36

Table. 6-10 DES database, hold-out cross-validation with 10 iterations, all samples in ACS: single judgment
Rate 1: with gender classification; Rate 2: without gender classification

Operator	Lukasiewicz	Hamacher				Yager
Parameter	-	γ=0.5	γ=3	γ=5	γ=10	p=1
Rate 1 (%)	64.18±1.95	74.36±1.81	73.99±0.98	75.34±0.77	75.24±1.08	64.98±1.56
Rate 2 (%)	48.45±1.52	49.96±1.81	52.64±2.06	52.58±1.87	52.47±2.00	49.61±2.31
Operator	Yager			Weber-Sugemo		
Parameter	p=3	p=5	p=10	λ_T=2	λ_T=4	λ_T=5
Rate 1 (%)	70.92±1.74	69.81±1.64	61.48±3.35	69.64±1.72	74.63±1.00	74.31±2.08
Rate 2 (%)	51.29±2.26	45.67±2.49	42.60±3.08	50.79±1.65	53.75±2.67	52.85±1.54
Operator	Schweizer & Sklar					
Parameter	q=0.4	q=0.6	q=0.8	q=1	q=3	
Rate 1 (%)	73.13±1.43	76.91±1.80	76.37±1.38	75.71±1.17	71.03±1.62	
Rate 2 (%)	52.70±1.77	52.96±2.10	51.67±3.23	51.40±2.68	47.27±2.81	
Operator	Frank			Average	Geometric Average	
Parameter	s=10	s=12	s=15	-	-	
Rate 1 (%)	62.77±2.59	63.90±1.86	62.40±2.42	49.64±3.76	58.37±3.83	
Rate 2 (%)	47.25±2.42	49.34±3.04	48.46±3.35	34.55±3.40	41.25±4.93	

Best confusion matrix with gender classification		E1	E2	E3	E4	E5
	E1	67.45	15.69	3.82	3.14	9.90
	E2	9.05	81.14	3.14	1.71	4.95
	E3	4.53	7.36	77.17	5.75	5.19
	E4	0.66	2.36	10.94	80.85	5.19
	E5	5.91	8.09	3.13	5.30	77.57

Best confusion matrix without gender classification		E1	E2	E3	E4	E5
	E1	55.29	16.08	5.20	2.65	20.78
	E2	19.33	58.95	4.86	1.62	15.24
	E3	10.00	16.42	51.13	3.96	18.49
	E4	5.75	5.38	25.57	46.32	16.98
	E5	6.61	21.30	8.00	7.22	56.87

Table. 6-11 DES database, hold-out cross-validation with 10 iterations, all samples in ACS: multiple judgments
Rate 1: with gender classification; Rate 2: without gender classification

Operator	Lukasiewicz	Hamacher				Yager
Parameter	-	γ=0.5	γ=3	γ=5	γ=10	p=1

Rate 1 (%)	64.36±1.92	78.33±1.24	75.36±1.23	76.22±0.72	75.92±0.54	65.09±1.66
Rate 2 (%)	48.75±1.49	58.63±1.71	54.83±2.17	54.64±1.79	53.63±1.64	50.04±2.61
Operator	Yager			Weber-Sugemo		
Parameter	p=3	p=5	p=10	λ_T=2	λ_T=4	λ_T=5
Rate 1 (%)	75.54±1.96	74.85±1.60	69.12±2.88	70.41±1.43	75.62±1.31	75.30±2.29
Rate 2 (%)	**59.14±1.81**	55.92±2.25	54.21±3.45	52.04±1.01	55.04±2.45	54.48±1.47
Operator	Schweizer & Sklar					
Parameter	q=0.4	q=0.6	q=0.8	q=1	q=3	
Rate 1 (%)	74.83±1.40	78.48±1.76	**78.50±1.43**	78.28±0.89	76.55±1.62	
Rate 2 (%)	55.77±1.60	56.65±2.47	56.40±3.15	57.58±2.35	57.36±2.34	
Operator	Frank			Average	Geometric Average	
Parameter	s=10	s=12	s=15	-	-	
Rate 1 (%)	64.48±2.84	64.74±1.92	63.07±2.40	62.79±3.00	66.89±3.22	
Rate 2 (%)	50.26±3.32	51.01±3.23	49.85±3.07	50.90±4.07	53.95±4.53	

	Predicted / Actual	E1	E2	E3	E4	E5
Best confusion matrix with gender classification	E1	71.24	16.67	6.80	4.25	9.61
	E2	13.46	78.51	7.08	3.14	7.94
	E3	5.31	12.48	78.08	9.31	8.08
	E4	2.08	5.47	16.07	78.36	9.31
	E5	9.10	8.93	4.78	8.81	78.84
Best confusion matrix without gender classification	Predicted / Actual	E1	E2	E3	E4	E5
	E1	58.60	13.70	10.29	5.65	28.92
	E2	32.40	47.10	8.70	3.75	30.73
	E3	20.10	13.70	51.04	6.98	26.35
	E4	12.30	8.52	27.33	47.58	21.92
	E5	16.90	17.60	8.99	11.19	64.46

Fig. 6-8 shows the best classification rates of the tested operators with single judgment, and Fig. 6-9 shows the best classification rates with multiple judgments. The error bars in the figures show the root mean square errors of the classification rates.

In the case of single judgment, the best result is 79.54%±2.51% for female samples with Weber-Sugemo operator when λ_T=5, 81.42%±1.98% for male samples with Schweizer & Sklar operator when q=0.8, 76.91%±1.80% for mixed genders with

gender classification with Schweizer & Sklar operator when q=0.6 and 53.75%±2.67% for mixed genders without gender classification with Weber-Sugemo operator when λ_T=4.

In the case of multiple judgment, the best result is 80.85%±1.46% with Schweizer & Sklar operator when q=1 for female samples, 82.91%±1.66% with Schweizer & Sklar operator when q=0.8 for male samples, 78.50%±1.43% with Schweizer & Sklar operator when q=0.8 for mixed genders with gender classification and 59.14%±1.81% for mixed genders without gender classification with Yager operator when p=3.

Same as the case with the approach of hierarchical classification and with the Berlin dataset, the operators with convex curve surfaces of properties give better results than the others.

The improvement caused by the introduction of the multiple judgments on the DES dataset is not so significant, which may indicate that the speech samples in the DES dataset are rather typical with the emotions. This can also be suggested by correct classification rate on the DES dataset that is much higher than the result on the Berlin dataset with the same feature set and the same classification algorithms, even though there are fewer emotion classes with the DES dataset. Actually, since these public available datasets are recorded with the aim to obtain typical emotional states, new datasets need to be recorded specially considering the subjective judgments of emotions.

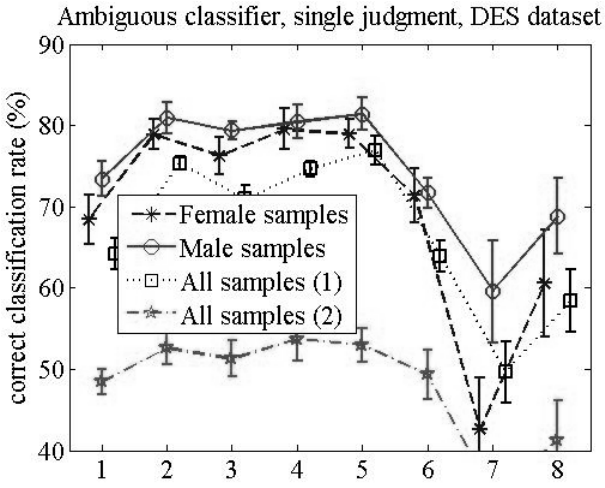

Fig. 6-8 Classification rate with ACS with single judgment for DES dataset

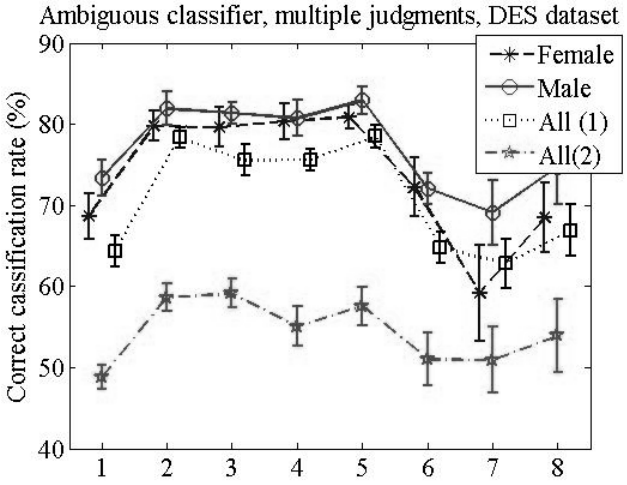

Fig. 6-9 Classification rate with ACS with multiple judgments for DES dataset

The best features are analyzed for the two genders respectively. The importance of the features is evaluated by the numbers of times of the features being selected in these sub-classifiers. The indexes of the most frequently selected features are listed according to the feature groups in Table. 6-12. Up to five features and the percentages of the features in the selection are listed for each group, and the features are ordered according to the frequency of selection. The Zipf features are calculated together with the harmonic features because there are only 2 Zipf features in our feature set. The description of the most important features is shown in Annex A.

Table. 6-12 Most frequently selected features for the ACS for DES dataset

Feature Groups	Harmonic	Zipf	Frequency	Energy & rhythm	MFCC
Female	11,3,20	225,226	31,29,32,41,27	44,46,43,60,73	87,124,209,120,123
Male	3,7,15	225,226	27,28,23,31,29	44,46,74,70,61	100,183,89,91,107

Percentages of number of features in each group		
group	Female	Male
Harmonic & Zipf	2	3
Frequency	34	34
Energy & rhythm	20	26
MFCC	44	37

Similar features are selected as in the hierarchical classifier and with the two approaches on Berlin dataset.

The best result on the DES dataset for the two genders and the mixed cases with the two approaches are illustrated in Fig. 6-10. The performances of the hierarchical classifiers and the single judgment of the ambiguous classifiers are quite close to each other. The results of both genders with gender classification are obviously better than the results without gender classification.

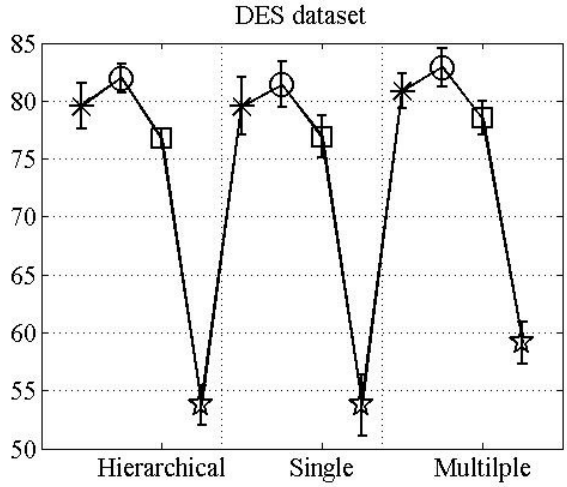

Fig. 6-10 Best classification rates for DES dataset, the star markers stand for female samples, the circle markers stand for male samples, the square markers stand for mixed samples with gender classification, and the pentagram markers stand for mixed samples without gender classification

The results from work of Ververidis [Ver05a] [Ver05b] are also listed in Table. 6-13. Their results are also obtained on DES dataset with 90% training data and 10% testing data with 10 groups of cross-validation. The results for single genders and the result for both genders with gender classification are higher than the results in [Ver05b], and the result for both genders without gender classification is close to the result of both genders in [Ver05b].

Table. 6-13 Best Classification rate for DES dataset (%)

	HCS	ACS Single	ACS Multiple	[Ver05b]
Female	79.54	79.54	80.85	60.00
Male	81.96	81.42	82.91	66.00
All (1)	76.74	76.91	78.50	-
All (2)	53.75	53.75	59.14	54.20

The summary of the results shown in Table. 6-13 shows that significant improvements can be obtained by dividing the complicated classification problem into a series of simple sub-problems, and that gender information may influence greatly in the emotion classification because of the different feature distributions for the two genders.

6.3.3 Synthesis on the results from the two datasets

The automatic ACS is experimented in this section on Berlin dataset and DES dataset. The results on Berlin dataset are also compared with the results from the human testing. Between the two ways of evaluation in the ACS, the results from the single judgment are actually also absolute classification in traditional way, and have no much difference from other methods, such as the hierarchical classification described in the previous chapter. The results from multiple judgments are preliminary attempt in the subjective emotion recognition as close as possible to human judgments.

We make major summaries in this subsection about the ACS on the two datasets from the multi-label patterns in recognizing, and also make comparison with the hierarchical classifiers tested in the previous chapter on the most efficient features and the performance of the operators.

As a first study on the ambiguous classification emotions, we tried in this approach with multiple emotion labels to simulate the human manner in subjective recognizing of emotions. In the experiments on Berlin dataset, the machine judged multi-label emotions are compared with the results in human testing. The multiple emotion labels present common patterns in both human testing and machine recognizing. The utterances with original emotion labels as anger and happiness are frequently judged as anger or happiness at the same time, especially for the happy utterances in machine recognition in the ambiguous approach. Several emotions, such as happiness, fear and boredom, have relatively high chance to be recognized as neutral, especially for boredom. The speech samples in passive emotional states, sadness and boredom, are often judged as with both emotions in the ambiguous machine recognizing, while in human testing, only boredom is frequently assigned with both emotions for male utterances, and sadness is seldom judged as also boredom.

Two rough summaries can be drawn from these results. First, these patterns in the ACS with multiple judgments fit the distribution of the emotional states in the dimension space: the emotions close to each other have more chance to be assigned simultaneously to a same utterance, and all the emotions might be assigned as with no particularly emotional states as neutral that is located in the center of the dimensional emotion space. Thus, it proves that the mapping of the discrete emotions into dimensional space is reasonable in the emotion classification. Second, the similarity between the multi-label patterns in human testing and machine recognition shows the potential to simulate the human manner in subjective judgment of vocal emotions, even if only the absolute emotion labels are applied in the learning process of building the ambiguous classifiers.

Compared to the synthesis of the results in the automatic hierarchical approach in section 5.4, the HCS and ACS show consistency in the most efficient acoustic features and performance of the operators.

Several typical features are selected in the two approaches on the two dataset of emotional speech with different number and types of emotional states. These features include features 3, 7, 15, 20 (harmonic features), 23, 27, 31, 44, 46, 78, 80 (frequency or energy based features), 81, 89, 92, 100, and 107 (MFCC features) (see detailed feature descriptions in Annex A). Among those features, there are some harmonic features which concern the timber structure and the energy ration on different spectral band that reflect the tonality, and some MFCC features that simulate characteristics of human perception. This consistency on the two datasets proves again that these features can well reflect the inherent clues of emotions in speech signals.

The performances of the operators are also consistent in the ambiguous classifications as compared to the hierarchical ones. Similar shapes of curves appear in Fig. 6-5, Fig. 6-6, (Berlin dataset) and Fig. 6-8, Fig. 6-9 (DES dataset) as they appear in Fig. 5-13 and Fig. 5-15 with better performance with Yager, Weber-Sugemo, and Schweizer & Sklar operators that have convex curve surfaces. The convex properties of the operators enhance the discriminatory power of the combined features.

6.4 Conclusion

An automatic approach as ambiguous classification scheme (ACS) of emotional speech that allows to label the emotional utterances with multiple emotions is proposed and experimented on Berlin dataset and DES dataset in this chapter.

As only a preliminary attempt on automatic ambiguous classification of emotions, the ambiguous classifier proposed in this chapter still needs to be greatly improved. The aim of proposing the ambiguous classifier of the emotional speech is to produce machine judgments to emotions as close as possible to the human ones. Two main aspects of improvement are needed in our future work.

First, in the learning process of the ACS, the original emotional labels from the datasets with single emotion are still used as ground truth. Only the classification results are allowed to be labeled with multiple emotions, and the results on Berlin dataset are evaluated according to the multiple emotional labels obtained from human testing. In our future work, the ground truth of emotions for both the learning process and the evaluation will be modified to multiple emotional labels. Since the current public emotional speech datasets such as Berlin dataset and DES dataset are both recorded for the analysis for typical emotions by professional actors, new datasets taking into account the subjective vocal emotions will be needed.

Second, as we mentioned in the beginning of this chapter, an emotion state is not only subjective, which may lead to multiple possibilities in judgment, but also quite fuzzy and should be continuous, which may lead to the different degrees within the same emotional family, such as explosive happiness vs. calm happiness, or hot fury vs. cold anger. This continuity in the emotions is not yet considered in our work, and should be a topic of interest in the future investigations. Further to the two dimensional model with an arousal and an appraisal dimension used in our work, a more precise model of the position of the emotions in the dimensional space is needed for the research of continuous emotions, and other dimensions might be necessary to be introduced to make more reliable model of continuous emotion.

Chapter 7
Application to Music Mood Analysis

We have so far developed audio features and several classification techniques for vocal emotion analysis. In this chapter, we propose to extend our previous work to another type of audio signal, namely music signal, which also conveys emotional meanings. In the case of music signal, instead of emotional analysis the term "mood analysis" is rather preferred and its automatic recognition also has many potential applications such as intelligent music browsing based on music mood or music recommendation.

This chapter is organized as follows. We first study in section 7.1 properties of music signal, especially in terms of music mood and acoustic correlates. We then present in section 7.2 a short overview of related work and introduce in section 7.3 our approach. In section 7.4, we define our feature set for music mood analysis. Section 7.5 describes the experimental results on our hierarchical and ambiguous classification schemes on music mood recognition. Finally, concluding remarks and future work are sketched in section 7.5.6.

7.1 Music Signal and Music mood

As compared to speech signal, music signal is likely more stationary but characterized by some very specific properties such as multiple pitches, melody, tonality, intervals, etc. Automatic music mood analysis needs correct music mood taxonomy and acoustic correlates. However, before a description of music mood taxonomy and acoustic correlates, we need first to define music signal and its properties.

7.1.1 About music signal

Music is an form of art consisting of sound and silence expressed through time. Elements of sound as used in music are pitch (including melody and harmony), rhythm (including tempo and meter), structure, and sonic qualities of timbre, articulation, dynamics, and texture [Wiki].

Melodies are usually sequences of pitches, which are created in western music in respect to scales and modes and having a certain rhythm. Scales and modes are notions of music theory that describe a set of notes involved in the play. In western music, there are 12 notes. An interval between neighbor notes is called *semitone.*

For better understanding of particularity of music signal, let's consider the following Fig. 7-1 containing a musical excerpt expressed in a form of pattern.

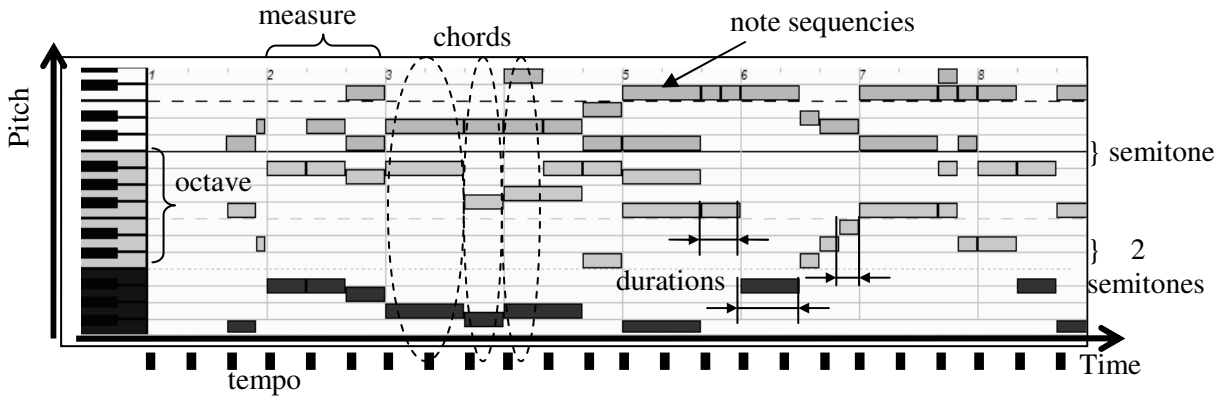

Fig. 7-1. Typical music pattern.

The term *pitch* is directly related to frequency, higher is pitch – higher is frequency. The relation between pitch and frequency is logarithmic.

$$p = 69 + 12 \cdot \log_2\left(\frac{f}{440Hz}\right) \qquad (7.1)$$

In music, pitch is the perceived *fundamental frequency.* Sounds of real instruments do not have single frequency in their spectrum, but a packet of frequencies – fundamental and its harmonics. The Fig. 7-1 would represent a spectrogram of a musical signal if instruments in it had a single frequency in their spectrum.

As we have noticed, musical signal have two main properties – temporal and frequency granulation. In reality, we do not speak about frequency granulation of musical signal since there could be signing voice or percussion instruments in it. Speeding up or slowing down the signal also leads to shifts of note frequencies from their theoretical values corresponding to integer pitches.

In comparison to speech signals, music signals can be assumed to be more stationary (during one note or one chord the spectrum of the signal does not change much). Duration of notes can be considered to be around 120-250 ms for 1/8 - 1/16 (quaver - semiquaver) at the most popular tempo of 120 BPM (the minimal duration of a note which could be found in music at 120 BPM is then 65 ms for 1/32). Nevertheless, music signals may contain percussion instruments rapidly changing in time or of a very short duration. Signing voice of one or many persons may be also present. The second grand specificity of a musical signal is *multipitch.* Music is usually represented by multiple simultaneous note events, such as chords. A chord may contain from two to many (more than ten) pitches at the same time. Some instruments may already contain multiple pitches in one note. These facts make musical spectrum much more complicated in the meaning of frequency contents in comparison to speech signals. Such complexity can be referred to *timbre* of the musical sound. Timbre characterizes each voicing instrument defined by its spectrum (presence of harmonic and inharmonic components) and envelop. Timbre allows us to distinguish different instruments.

Unlike voicing instruments, percussion instruments in music may not have its fundamental frequency. Thus, they are undetermined-frequency instruments (no pitch can be perceived). As an example, we can mention snare drums, hi-hats and cymbals. These instruments have mainly noisy components in their sound that are then superposed to voiced instruments and form a more complex spectrum.

7.1.2 Music mood taxonomy

As compared to emotion carried out by vocal signal, emotional clues delivered by music works convey expressions over relatively longer time duration and the term music mood is preferred instead of music emotion. Music mood taxonomy can also rely on a discrete description model or a dimensional one.

According to a discrete description model, music mood response is characterized by a set of adjectives, such as pathetic, hopeful and gloomy. However, these adjectives vary quite freely in different researches and there is currently no standard mood taxonomy system accepted by all. A checklist of adjective circle by Hevner [Hev36] presented in 1930s has served as the basis for some subsequent research on mood response to music, see Fig. 7-2. This checklist is composed of 67 adjectives from eight clusters, which include Sober, Gloomy, Longing, Lyrical, Sprightly, Joyous, Restless and Robust [Lu06].

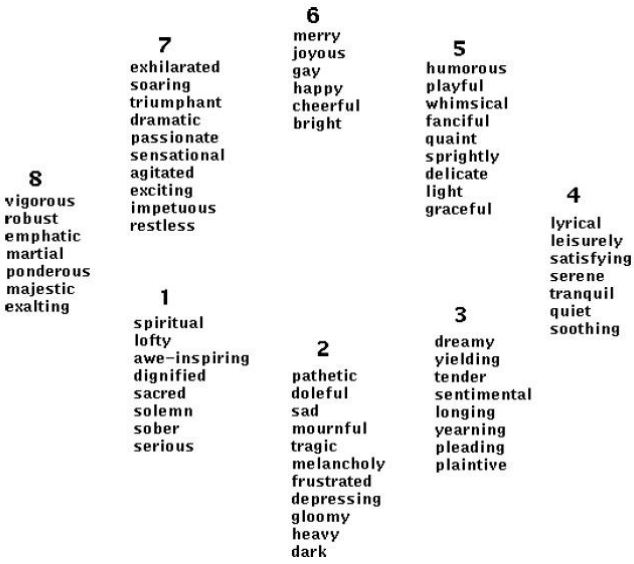

Fig. 7-2 Adjective Circle according to K. Hevner [Wie05b] [Hev36]

Another adjective checklist was proposed by Farnsworth [Far58] with ten adjective groups. Three additional groups are added by latter work [Li03a]. The adjective groups are listed in Fig. 7-3, including the additional groups K, L, M.

A	cheerful,gay,happy	H	dramatic, emphatic
B	fanciful, light	I	agitated, exciting
C	delicate,graceful	J	frustrated
D	dreamy,leisurely	K	mysterious, spooky
E	longing, pathetic	L	passionate
F	dark, depressing	M	bluesy
G	sacred, spiritual		

Fig. 7-3 Adjective groups by Farnsworth [Far58] [Li03a]

However, for both models of Hevner and Farnsworth, it is very difficult to discriminate these adjectives one from others because the adjectives in the same

cluster have approximately the same meaning. The underlying stimulus that influences the mood responses cannot be indicated by these adjectives either.

In the late 1990s, Thayer [Tha89] proposed a two-dimensional mood model. It was applied in the work of Microsoft Research Asia [Lu06]. This dimensional approach describes the mood with two factors: Stress dimension (happy/anxious) and Energy dimension (calm/ energetic), and divides music mood into four clusters according to the four quadrants in the 2-D space: "Contentment", "Depression", "Exuberance" and "Anxious/Frantic", as shown in Fig. 7-4.

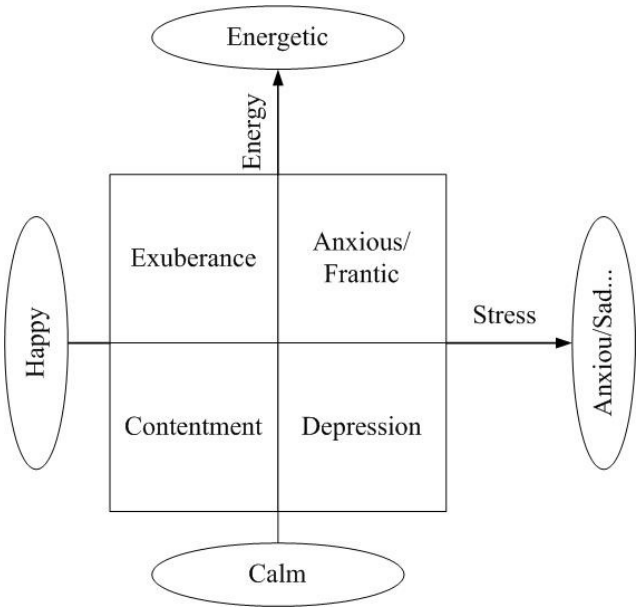

Fig. 7-4 Thayer's model of mood [Tha89] [Lu06]

In this model, "Contentment" refers to happy and calm music; "Depression" refers to calm and anxious music; "Exuberance" refers to happy and energetic music; and "Anxious/Frantic" refers to anxious and energetic music. Such definitions of the four clusters are clear and have high discriminatory power. Thus, the corresponding dimensional mood model allows to divide the whole music emotion space into four meaningful quadrants and facilitates rough music mood categorization. For these reasons, this model has been adopted in our work.

7.1.3 Acoustic correlates of music mood

There exist significant acoustic correlates of music mood as evidenced by the following very short overview of the literature.

The correlates between the music and mood effects can be prove by a nonverbal musical-mood induction procedure proposed by Hevner [Hev35b] [Hev36] [Hev37] and Pignatiello, Camp, and Rasar [Pig86]. Nonlyrical music selections from classical and popular music with characteristics of pitch, rhythm, mode, loudness, *etc.* were successful in inducing emotional effects of depressed and related.

Early synthetic analysis on the correlates of music mood is explored by Hevner [Hev37] between six musical elements and eight emotional categories. The six musical features explored in their work include mode, tempo, pitch (register), rhythm, harmony, and melody [Mey07]. The eight categories of emotions are shown in an adjective circle as in Fig. 7-2. The correlates are summarized in Table. 7-1, evaluated by related degrees of the musical elements in numbers.

Table. 7-1 Hevner's weighting of musical characteristics in 8 affective states [Hev37]

Musical elements	dignified/ solemn	sad/ heavy	dreamy/ sentimental	serene/ gentle	graceful/ sparkling	happy/ bright	exciting/ elated	vigorous/ majestic
Mode	major 4	minor 20	minor 20	major 3	major 21	major 24	-	-
Tempo	slow 14	slow 12	slow 16	slow 20	fast 6	fast 20	fast 21	fast 6
Pitch	low 10	low 19	high 6	high 8	high 16	high 6	low 9	low 13
Rhythm	firm 18	firm 3	flowing 9	flowing 2	flowing 8	flowing 10	firm 2	firm 10
Harmony	simple 3	complex 7	simple 4	simple 10	simple 12	simple 16	complex 14	complex 8
Melody	ascend 4	-	-	ascend 3	descend 3	-	descend 7	descend 8

The effect of the major and minor mode in music as a mood induction procedure was studied in [Hin96] stating that "On a structural level, music mode has been examined as a possible explanation for the effectiveness of musical mood induction." Hevner suggested that among the structural aspects of music, mode has a greater influence on mood than rhythm or tempo. Mode is represented by the placement of an octave's eight diatonic tones [Hev35b] [Hev36]. The primary distinction between the major and minor modes is the placement of the mediant, or third. The third in the major mode is composed of four semitones, while there are only three in the minor third [Rad88]. Hevner's study on mood associations with major and minor modes showed that the minor mode is associated with feelings of grief and melancholy whereas the major mode is associated with feelings of joy and happiness [Hev35b].

Meyer's theory of deviations from expectations supports Hevner's finding [Mey56]. It is suggested in this theory that the major mode presents the human

affective states of joy and happiness because it contains expectations of more regular and normative melodic and harmonic progressions;, while music in the minor mode contains complex and forceful departures that deviate from the expectations in the major mode, and results in corresponding to feelings of sadness [Hin96].

The information of the major/minor mode thus corresponds strongly to the stress dimension in the Thayer's model of music mood, while it is also indicated by Hinn [Hin96] that the mode is very difficult to obtain from acoustic data. Thus, the tonality of the music can only be used as a subsidiary reference in automatic mood detection.

The correlates of rhythm, intensity to the music mood are investigated from the aspect of *signal processing* recently in the literature [Kru02] [Liu03] [Yan04].

Krumhansl observed that the rhythm and the intensity influence obviously the music mood, that the music excerpts have slow tempos, minor harmonies and constant ranges of pitch and intensity, fear music excerpts have rapid tempos and large variations of intensity and pitch, while happy music excerpts have rapid tempos, dancelike rhythms [Kru02]. These statements are proved by the experimental results given by Liu [Liu03] and Yang [Yan04].

7.2 Related works

There exist several works on automatic detection of music emotion and music mood. However, the mood definitions used in these works are often very different. Thus, these related works concern classification problems on music mood with simple discrete definitions, multi-label emotions, or joint description with discrete and dimensional moods.

Yang *et al.* focused on the Negative Affect according to the Waston model [Tel99] (Fig. 2-2) applying data fusion with acoustic features and text features using Support Vector Machine (SVM) regression over the WEKA [Wit05] platform [Yan04]. The emotion types include hostility, sadness, guilt, love, excitement, pride, attentive, reflective, calm, *etc*. The acoustic features include Beats per Minute (BPM), low-level standard descriptors from the MPEG-7 audio standard, spectral centroid, spectral rolloff, spectral flux, and spectral kurtosis, *etc*. The Spectral Kurtosis (SK) of a signal is defined as the kurtosis of its frequency components [Vra03]. The kurtosis is a measure of the "peakedness" of the probability distribution of a real-valued random

variable, higher kurtosis means that most of the variance is due to infrequent extreme deviations [Wiki]. The text features include typical words derived from the lyrics of songs, such as expletives, not, get, got, want, never for hostility, and life, time, say, slowly, hold, feel for love, *etc*. The best acoustic features found in their work were BPM (Beats per Minute), Sum of Absolute Values of Normed Fast Fourier Transform (FFT), and Spectral Kurtosis. The text features were extracted from the lyrics of the songs. The result drawn from their work was accuracy of successful classification of 82.8% with mean error of 0.0252.

Li and Ogihara applied a SVM-based multi-label classification for two problems of classification into thirteen adjective groups and classification into six super groups on their 499 music samples with 50% as training data and 50% as testing data [Li03a]. The goal of their work was to take the emotion detection problem as a multiclass classification problem. Three categories of features were used in their work with timbral texture features, rhythmic content features, and pitch content features. Timbral features include MFCC features and other features such as spectral centroid and spectral rolloff features [Li04]. Daubechies wavelet co-efficient histograms (DWCH) [Li03b] features were proposed on several frequency subbands for a synthesized presenting for all the aspects as timbral, rhythmic and pitch. Two terms for the accuracy as micro-averaging and macro-averaging were defined in their work according to the process of the averaging over the classifiers with or without weighting in calculating precisions and recalls in the cross-validation. The half-way point between the precision and the recall was 46% in micro-averaging and 43% in macro-averaging for the thirteen groups, and 50% in micro-averaging and 49% in macro-averaging for the six super groups.

Similar experiments were driven by Wieczorkowska in [Wie05a] with the same six super groups as applied by Li. They performed similar experiments on their own dataset with 303 pieces of music and also on the data obtained from the research group of Li and Ogihara. Each piece of music in Wieczorkowska's work was labeled with only one emotion label. Acoustic features included dominating fundamental frequency, maximal level of sound, contents of even and odd harmonics in the spectrum, brightness, irregularity of the spectrum, and the 10 most prominent peaks in the spectrum. The innovation in their feature set was that they analyzed the harmonics with even and odd harmonics separately, while the drawback in their representing of the feature without indicating which kind of music properties such as timbre, rhythm,

etc. were contained in the related features. Cross validation with 80% of data as training set and 20% of data of testing set using K-NN was tested and the best recognition rate for each of the six super groups varies from 62.67% to 92.33% with different K, while the super groups were not balanced represented. Their accuracy with 6 classes tested in parallel yielded only 37% of correctness.

Another kind of music mood taxonomy with fewer classes while joint with the dimensional description based on the Thayer's model was adopted in the work of Microsoft Research Asia [Lu06]. They used three types of features related to the intensity, timbre and rhythm of the music signals. The timbre features included both spectral shape features and spectral contrast features; the intensity was approximated by the signal's root mean-square (RMS) level in decibels; and the rhythm features included the strength and regularity of rhythm, and the tempo of the music. Their work suggested the correlation between the acoustic features and the music mood dimensions so that intensity corresponds to "energy", while both timbre and rhythm are related to "stress". An empirical hierarchical framework was proposed for mood detection in their work, as shown in Fig. 7-5. The feature sets were modeled by using GMM (Gaussian Mixture Model). A cross validation evaluation where the 25% of the dataset were used for testing and 75% were used for training was done with an overall classification accuracy for the hierarchical framework up to 86.3%, which is about 5.7% better than the non-hierarchical framework.

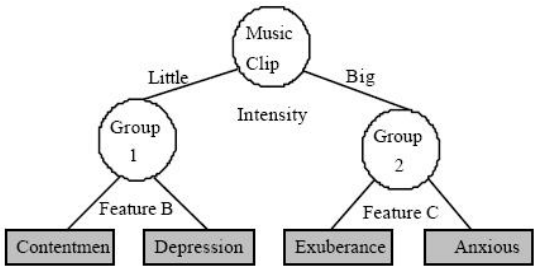

Fig. 7-5 The hierarchical mood detection framework by Microsoft Research Asia [Liu03] [Lu06]

We summarize the related works mentioned above in two aspects: the features related to the music moods and the classification schemes.

The content-based acoustic features are categorized as timbral texture features, rhythmic content features, and pitch content features by Tzanetakis [Tza02]. Typical timbral features include spectral shape features such as spectral centroid, spectral

rolloff, and spectral flux [Yan04] [Li03a] [Li04] [Liu03] [Lu06], spectral contrast features [Liu03] [Lu06], and cepstral features such as Mel-Frequency Cepstral Coefficients (MFCCs) [Li03a] [Li04]. Rhythmic features contain the regularity of the rhythm, the beat and tempo information, such as Beat per Minute (BPM) [Yan04]. Pitch features deals with the frequency information of the music signals. There are also some features used in several works but not so commonly in the classification of music mood, such as Spectral Kurtosis [Yan04], DWCH features [Li04], even and odd harmonics [Wie05a], and text features derived from lyrics of songs [Yan04]. Summarizing these features, we propose a feature set including three categories as music features, timbre features, and octave-based perceptual features. The feature set will be introduced in section 7.4.

Classification schemes such as SVM [Yan04] [Li03a] and K-NN [Wie05a] are used in the early work in music mood classification. A hierarchical framework is built by Liu and Lu [Liu03] [Lu06] according to the two dimensions of the moods in each layers. This hierarchical framework is similar to the idea of our automatic hierarchical approach with emotional speech while it is built empirically. As the research on the music mood faces the same problem as the one encountered in the case of emotional speech concerning the fact that there is no universe agreement on the mood taxonomy, we apply the automatic approaches proposed for classification of emotional speech (section 5.3 and section 6.2) on the problem of music mood, and make experiment on the moods states according to the Thayer's model.

7.3 Our approach

As stated in the previous section on music mood taxonomy, we have adopted in this work the Thayer's dimensional model that divides the music mood space into four rough mood quadrants. As compared to the dimensional model, which showed its relevance for emotional speech analysis (Fig. 4-7), the energy dimension corresponds to the arousal dimension, and the stress dimension corresponds to the appraisal dimension (with inversed direction).

As music signal is quite different from speech signal for its specific properties such as multiple pitches, melody, harmony, rhythm, etc., the previous set of audio features specifically designed for vocal emotion analysis, in particular F0 detection and formant features, is not relevant anymore. This irrelevance was further confirmed

by a straightforward experiment that we carried out on a small corpus with several segments of songs and instrumental music pieces, using the same audio feature set and the same algorithms as we did for vocal emotion analysis. As expected, this experimental setting resulted in a poor accuracy rate of around 50% for the four mood states.

Our basic assumption is that efficient music mood analysis needs to rely on music features, such as rhythm, tonality, interval, timbre, etc., as suggested our short overview on acoustic correlates of music mood. While intensity and rhythmic features are quite straightforward as strongly related to music mood, timbre feature helps to characterize instruments involved in a music excerpt, music interval in octave simulates the human perceptual behavior and music mode as characterized by music tonality is clearly related to music mood as evidenced in our overview on acoustic correlates. As compared to the dimensional Taylor's music mood model, intensity is clearly related to the energy dimension whereas the other music features should be related to the stress dimension.

Concerning the machine learning engine, we have chosen to make use our automatic hierarchical classification scheme as described in Chapter 5.

7.4 Extracting Music and Perceptual Feature set

According to the related works concerning features extraction, we have identified two points of possible improvement. Thus, we have developed in this work the features in two aspects considering the characteristics of music and the perceptual to the music by human respectively. Features from these two aspects are categorized into two groups, namely music features characterizing rhythm and tonality, and perceptual features including timbre features and octave-based perceptual features. While similar timbre features and octave-based perceptual features have already been used in Microsoft's work [Lu06], the group of music features is directly derived from the teamwork of Paradzinets who proposed in his thesis a Variable Resolution Transform (VRT) for music feature extraction and their applications in music retrieval by similarity [Par06] [Par07].

7.4.1 Music features

The music features carry the music characteristics. In this work, we consider two kinds of music features, namely rhythm and tonality. A beat is a pulse on the metric level and is heard as the basic unit. Thus, it works as the basic time unit of a piece defining the music rhythm. Tonality is a system of music in which certain hierarchical pitch relationships are based on a key "center" or tonic. Nowadays, the term tonality is most often used to refer to Major-Minor tonality (also called diatonic tonality or functional tonality) [Wiki].

7.4.1.1 Rhythmic features

Our rhythmic features are based on a 2D beat histogram from the spectrogram of a Variable Resolution Transform (VRT) applied on a music signal as described in [Par07]. The Variable Resolution Transform (VRT) has a variable time-frequency resolution grid with a high frequency resolution and a low time resolution in the low frequency band and an opposite resolution in the high frequency band. This time-frequency resolution characteristic is similar to the human ear according to [Tza01].

A “wavelet-like” function with logarithmic frequency scale was used to follow the musical note system in [Par07]:

$$\psi(x,a^*) = H_{x,m(a^*)} e^{jw(a^*)x} \qquad (7.1)$$

where a^* presents the relative scale of wavelet, and $H(x,m)$ is the function of Hamming window of length m.

$$m(a^*) = L_{\max} k_1 \cdot e^{-k_2 a^*} \qquad (7.2)$$

$$w(a^*) = L_{\max}{}^{a^*} \Big/ L_{\min}{}^{a^*+1} \qquad (7.3)$$

where k_1 and k_2 are time resolution factors, L_{max} and L_{min} are range of wavelet absolute scales.

The time/frequency scale of the transform is shown in Fig. 7-6.

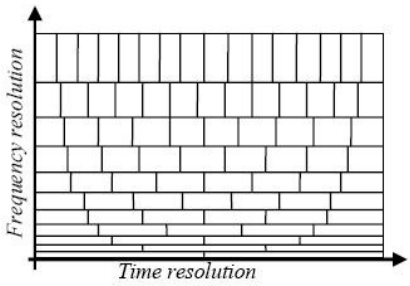

Fig. 7-6 Time-frequency resolution of the VRT [Par06] [Par07]

A 2D form beat histogram can then be built based on this VRT spectrogram with a beat period on the X axis and with amplitude (strength) of a beat on the Y axis, as shown in Fig. 7-7.

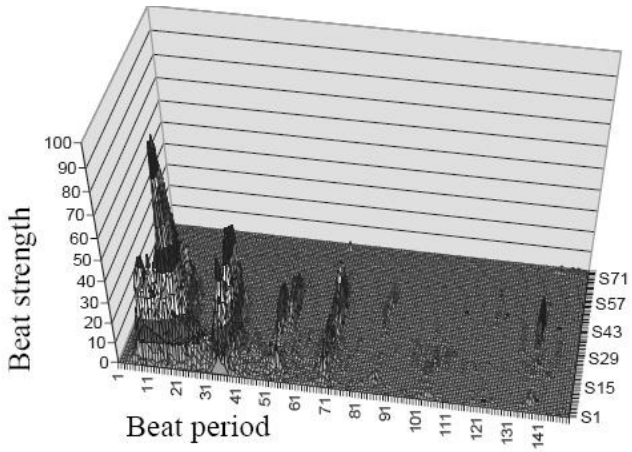

Fig. 7-7 A 2-D beat histogram [Par07]

Several features can be derived from this kind of beat histogram. The features as tempo (presented by Beats per minute (BPM)), the position of the peaks, the distance between the peaks, and the width of the peaks are considered as musical features corresponding to beat information.

Since the axis of "Beat period" in the beat histogram is labeled with time, a FFT is applied to the beat histogram into frequency domain in order to summarize inherent characteristics of the beat histograms. The averages of the FFT to the beat histogram of the four moods are plotted in Fig. 7-8.

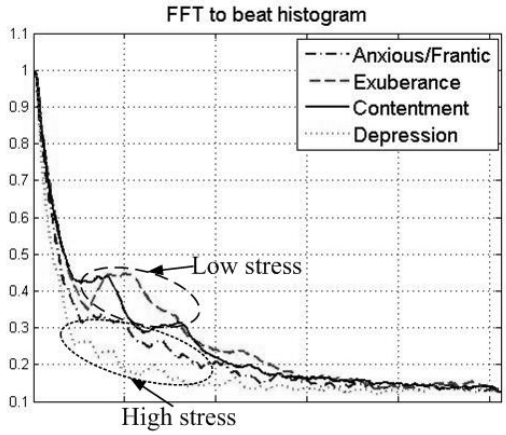

Fig. 7-8 Average FFT to the beat histograms for the four moods

As shown in Fig. 7-8, in the middle part of the curves of FFT to the beat histograms (marked with dashed ellipses), the moods with low stress (Exuberance and Contentment) have higher values, and the moods with high stress (Anxious/Frantic and Depression) have lower values. The statistics of this part are also taken as features according to the beat information.

7.4.1.2 Tonality features

As we discovered in the subsection on acoustic correlates, music major and minor mode has great significance in arousing music mood. Nowadays, the term tonality is most often used to refer to Major-Minor tonality (also called diatonic tonality or functional tonality) [Wiki]. Tonality is a system in which certain hierarchical pitch relationships are based on a key "center" or tonic. The tonality can be estimated from a multiple F0 estimation algorithm with a note profile. In this work, we made use of the multiple F0 estimation algorithm and note profile (histogram), both from the thesis work of Paradzinets [Par06] [Par07].

The multiple F0 estimation algorithm proposed in [Par06] [Par07] is a technique inspired by harmonic pitch models. The analysis procedure includes two parts (Fig. 7-9). The first part consists in the model generation. The model is built using a serie of peaks at the frequency of *F0* and its harmonics 2**F0*, 3**F0*, with the forms of the peaks obtained by using the VRT applied on sine waveforms with appropriate frequencies. The second part analyzes the input wave signals for transcription by moving the harmonic structure across the frequency scale of the VRT spectrogram slice and computing the correlation between the model and the

spectrogram. The place where the correlation has a maximum value on the spectrogram is assumed to be an F0 candidate. The procedure is repeated until one of the following conditions is satisfied: 1) no more harmonic-like structures are found in the spectrum (above the certain threshold), 2) the maximum number of harmonic structures to be searched within the algorithm is reached.

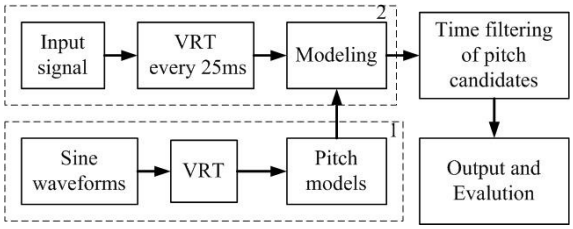

Fig. 7-9 Transcription process [Par06]

Note profiles (histogram) that represent the frequency of each of the 12 notes appearing in the music melodies can then be extracted once multiple F0s have been estimated. The tonality of the music can be determined from this profile.

Indeed, each tonality has its own distribution of notes and can be obtained from the note histogram [Chu05]. Examples of note profiles of major mode and minor mode are shown in Fig. 7-10.

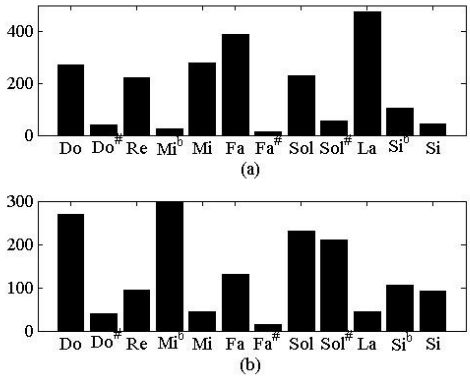

Fig. 7-10 Approximate note profiles for the two tonalities
(a) major (C-dur), (b) minor (C-mol) [Par07]

As the 12 notes make a circle and as it is difficult to determine the beginning point of these 12 notes, the note profiles may be not detected properly and thus the tonality features could be not precise enough in representing music mood. For example, it is rather hard to decide the mode of music as major or minor automatically [Hin96], we can only estimate from the note profile that a music piece is more likely to be major or minor. Thus we apply the probability to be major of a piece of music as

a music feature in our work, when the probability is larger than 0.5, the piece of music is more likely to be major, and otherwise more likely to be minor. This probability is derived in the work of Paradzinets [Par07] according to the distribution of the note profile. Since the typical probability values derived from this work are normally between 0.4 and 0.6, which are not strong enough to represent the tonality of the music, the distribution of the note profile is also applied as music features to represent the tonality. For example, the note Re tends to appear more frequently in majors than in minors, while the note Mi^{b} ($Re^{\#}$) tends to appear less frequently in majors than in minors.

We can summarize the list of music features as follows:

- Beats per minute (BPM)
- The position of the peaks, the distance between the peaks, and the width of the peaks in the beat histogram
- Mean value in the middle part of the FFT result to the beat histogram
- Probability to be major
- The distribution of the note profile

According to the analysis in section 7.1.3 on acoustic correlates of music mood, the rhythmic and tonality are strongly related to the mood states, among which the most intuitive ones are BPM (influence in both dimensions) and music mode (major or minor, influence more in stress dimension), which are not so accurate with current algorithms. Although the beat histogram and note profiles are efficient in the analyzing of music similarities by calculating the distance between two musical compositions instead of the rhythmic and tonality values directly, the application of these features may not be efficient enough in the recognition of music mood. Thus, the performance of such features could be improved with the development of a method allowing to extract the tempo and the tonality.

7.4.2 Perceptual features

The mood contained in music pieces can be affected by various elements, such as the use of different instruments or emphasizing different parts in play. The same music piece may convey completely different mood in different plays even though the melody does not change. Thus, the previous rhythmic and tonality features may not be

enough to reflect all the clues related to the mood in the music. Other features, in particular timbre features characterizing music instruments involved within a music excerpt, are needed to complete the previous rhythmic and tonality features.

The perceptual features are linked to the way humans perceive properties of music signals. Two kind of perceptual features, namely timbre features and octave-based features are presented in this subsection.

7.4.2.1 Timbre features

The timbre represents the quality of a musical note or sound that distinguishes different types of sound production even if with the same pitch and amplitude [Wiki] (See also description of timbre in section 4.2.1). Several features according to the spectral shape, thus related to global timbre properties, including the spectral centroid, the spectral roll-off, and the spectral flux [Lu06], are included in this group of timbre features.

The spectral centroid is strongly correlated with the brightness of a sound as the "balancing point" of the spectrum [Mar98]. It measures the "center of gravity" of the sound spectrum using both the frequency and the amplitude information. The spectral centroid can be obtained as

$$spec\_cent = \frac{\sum_{k=1}^{N-1} kX[k]}{\sum_{k=1}^{N-1} X[k]} \qquad (7.4)$$

where $X[k]$ is the magnitude corresponding to bin k and N is the length of the FFT. The increase of the amount of the high frequency that causes a higher value of the spectral centroid can make the sound brighter in the sense of timbre.

The spectral roll-off measures the point where frequency that is below certain percentage of the power spectrum resides [Lu06] [Nor05]. This measure reflects the degree of concentration of energy in the spectrum thus also defines the brightness of the music signal using another form. Two levels of spectral roll-off are considered at 50% and 95% respectively in this work.

The spectral flux is a measure of how quickly the power spectrum of a signal is changing [Lu06], calculated by comparing the power spectrum for one frame

against the power spectrum from the previous frame. The music frames are defined by 20ms analysis windows with 10ms overlapping. The spectral flux is calculated as the 2-norm distance between the two spectra:

$$spec\_flux = \left\| |X_k| - |X_{k-1}| \right\| \qquad (7.5)$$

The examples of the spectral shapes of eight seconds segments with the mood types "Exuberance" and "Depression" are shown in Fig. 7-11.For the three spectral shape aspects, the mood with high energy and low stress (e.g. "Exuberance") tends to have higher values.

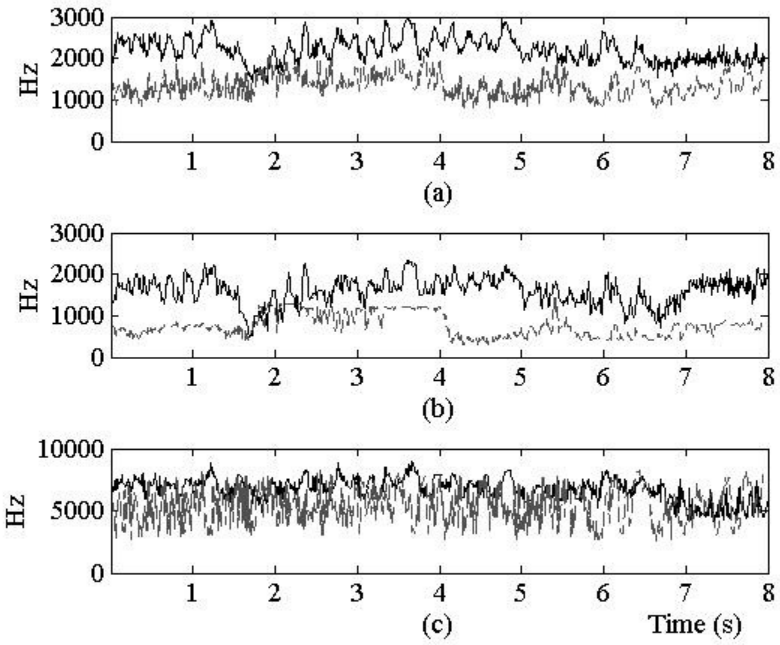

Fig. 7-11 Examples of spectral shape features (a) Spectral centroid (b) Spectral roll-off at 50% (c) Spectral roll-off at 95%. Black curves stand for a sample of "Exuberance", grey curves stand for a sample of "Depression"

We can thus summarize the list of timbre features:

- Mean values, minima and maxima values, ranges and variances of spectral centroid
- Mean values, minima and maxima values, ranges and variances of spectral roll-off

- Mean values, minima and maxima values, ranges and variances of spectral flux

All the timbre features are computed as statistics over the length of selected music samples.

7.4.2.2 Octave-based features

In music, an octave is the interval between one musical note and another with half or double its frequency. The human ear tends to hear both notes as being essentially "the same". For this reason, notes on different octaves are given the same note names in the western system of music notation. For example, if one note has a frequency of 400 Hz, the note on an octave above has a frequency of 800 Hz, and the note an octave below is at 200 Hz. For modern western music, the most common tuning system is twelve-tone equal temperament (12-TET), which is tuned relative to a standard pitch of 440 Hz (the A above middle C, ISO 16-1975, [ISO75]), and the frequency of each successive pitch is derived by multiplying the previous by the twelfth root of two. A music keyboard in 12-TET with three octaves is shown in Fig. 7-12. In the subbands processing of audio signals, though MFCC with Mel-scale filter bank is more commonly used for general auditory model, the octave-scale filter bank is more suitable for music processing [Jia02].

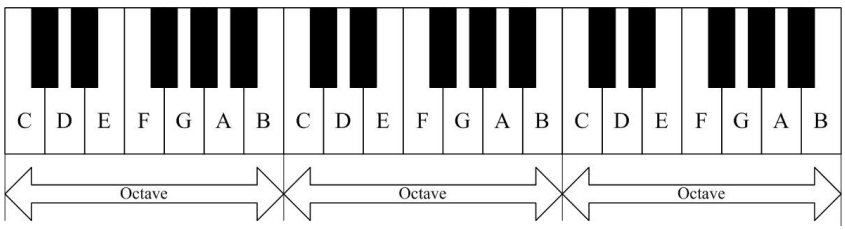

Fig. 7-12 Musical keyboard with 3 octaves

In order to make consistence with the literature [Jia02], the frequency domain is divided in our work into six octave-based subbands as shown in Fig. 7-13. The frequency ranges of the subbands are listed in Table. 7-2. The subbands approximately correspond to the music octaves starting from the notes G. The music samples in our experiments are with sampling frequency of 22050 Hz, which contain information from 0 to 11025 Hz. As the band between 6400 Hz and 11025 Hz is not an entire octave subband (the upper limit of this subband should be at 12800 Hz), this part of signal is ignored in this group of features.

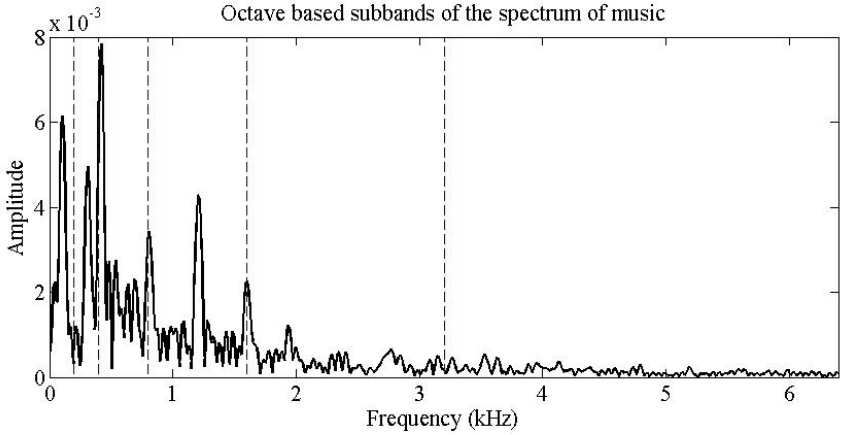

Fig. 7-13 Octave based subbands of music spectrum

Table. 7-2 The frequency ranges of the octave-based subbands

Subband No.	1	2	3	4	5	6
Freq range (Hz)	0~200	200~400	400~800	800~1600	1600~3200	3200~6400
Octave scale	~- $F^{\#}3$	G3–$F^{\#}4$	G4–$F^{\#}5$	G5–$F^{\#}6$	G6–$F^{\#}7$	G7–$F^{\#}8$

The octave-based subbands energy and spectral contrast are considered as features. The energy itself contributes to the recognition of the mood in the arousal dimension (energy dimension in Thayer's model), and the distribution of the energy according to the spectrum contributes to the recognition in the appraisal dimension (stress dimension in Thayer's model). Due to the perception property of music that the human ear tends adapt according to the octave-scales, the octave-based subbands are used for analyzing the music energy. As evidenced in the classification of emotional speech, the features related to the harmonic structure are rather efficient for the emotion classification in the appraisal dimension, while the method of extracting the harmonic features based on the estimation of F0 cannot be applied directly on music signal because there exist generally multiple F0s [Par06] [Par07] in music signals. The spectral contrast could roughly reflect the distribution of harmonic and non-harmonic components by representing the relative spectral characteristics in each octave-based subband separately [Jia02].

The subbands energy features include the total energy and energy ratio in each of the subbands, and the minima, maxima, mean, median values, and variances of energy in each of the subbands.

The strength of spectral peaks and valleys and their differences in the octave-based subbands are estimated for the spectral contrast. The feature extraction

is also based on the 20ms frames with 10ms overlapping. For each frame, FFT is performed to get the spectral components and it is then divided into six octave-based subbands. The average around a small neighborhood of the peaks and the valleys of the subbands are taken into consideration. A neighborhood factor α as the ratio of the bandwidth of the subbands is set to define the small neighborhood [Jia02]. Suppose there are N elements in the FFT vector of the k^{th} subband (k=1, 2, 3, 4, 5, 6), and it is noted as $\{x_{k\_1}, x_{k\_2}, \ldots, x_{k\_N}\}$, then the vector is sorted in descending order as $\{x'_{k\_1}, x'_{k\_2},\ldots, x'_{k\_N}\}$, which satisfies $x'_{k\_1} \geq x'_{k\_2} \geq \ldots \geq x'_{k\_N}$. The subband peaks and subband valleys can be estimated as

$$Peak_k = \log\{\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} x_{k\_i}\} \quad (7.1)$$

$$Valley_k = \log\{\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} x_{k\_N-i+1}\} \quad (7.2)$$

And the difference between the subband peak and valley is

$$D_{pvk} = Peak_k - Valley_k \quad (7.3)$$

According to Jiang [Jia02], varying the neighborhood factor α in the range from 0.02 to 0.2 does not influence the performance significantly. The strength of spectral peaks and valleys and their differences are estimated over a certain duration of music signals, and the mean values, median values and variances are taken as spectral contrast features.

We summarize the list of octave-based perceptual features as follows:

- Energy levels in each of the subbands
- Energy ratios in each of the subbands
- Mean, maxima and minima values of energy
- Mean values, minima and maxima values, ranges and variances of the subbands peaks, subbands valleys, their difference and subbands averages

7.4.3 Synthesis

For summary, two groups of features, namely music features including rhythmic features and tonality features, and perceptual features including timbre features and octave-based features are extracted in this work with the aim to characterize the music mood. Most of these features have already been used in the previous work on classification of music mood except the tonality features. In our work, we recombine the features according to two categories as the characteristics of music itself and human perception to the music respectively.

7.5 Experiments and results

In this section, we propose to benchmark the effectiveness of our approach, using the two approaches of automatic emotion classification scheme previously elaborated for vocal emotion analysis, namely hierarchical classification and ambiguous classification, with the feature set proposed in the previous section. As music mood within an excerpt may vary with time, the influence of the duration of music clips as respect to classification accuracy is also studied and a simple method on music mood tracking is also proposed. But, first of all, we introduce in subsection 7.5.1 our music mood dataset used as ground truth for the assessment of our approach.

The experiments and the results on classification of music mood are displayed in this section from subsection 7.5.2 to subsection 7.5.5. Then, we synthesize the experimental results on music mood in subsection 7.5.6.

7.5.1 Music mood dataset

As a branch of the multidisciplinary research domain “Music Information Retrieval” (MIR) [ISMIR], automatic analysis of music mood faces the common difficulty of the MIR: there exists unfortunately no standard universally agreed music mood dataset on which the research community can scientifically compare and contrast their work [Dow03] [Lem]. For this reason, the current work in the literature all builds its own dataset as ground truth. Resources of music mood analysis may include pure music with only instruments or songs with human voice.

In the work of Li [Li03a], 499 sound files of 30 seconds duration in mp3 format covering four major music types (Ambient, Classical, Fusion, and Jazz) were labeled with ground truth of multi-label applying the Farnsworth adjective groups. A

dataset which consists of 500 randomly-chosen rock song segments of 20 seconds for disambiguating music emotions was built by Yang in [Yan04]. 303 songs were collected by Wieczorkowska *et al.* [Wie05b] considering the specific features of music with six classes described by adjective groups. Microsoft Research Asia collected 200 representative music clips of 20 seconds long each labeled for each of the four mood clusters, namely Contentment, Depression, Exuberance and Anxious, resulting in a dataset of 800 music clips. These 800 representative music clips were selected from about 250 pieces of music composed mainly in the classical period and romantic period [Lu06]. Unfortunately, none of these datasets can be available to the public.

Thus, we have also built a dataset for music mood recognition. In order to avoid the influence by the lyrics on mood judgment in labeling the ground truth, we have selected classical musics pieces that contain no human voice.. Music pieces are labeled according to the four clusters of the Thayer's model enabling conjoined application of discrete and dimensional description of mood states. Our music dataset contains about 60 pieces of classical music with sampling rate of 22050 Hz in mp3 format. The pieces were labeled into the four mood states by two persons. As the aim of music mood analysis in our work is mainly multimedia search in daily life, the persons labeling the mood states are people with no expert knowledge in music. Since the mood states usually changes within a whole piece of classic music, we developed four versions of the dataset with durations of 4s, 8s, 16s and 32s respectively in order to study the influence of the duration of music segments on automatic recognition of music mood. According to Kamien [Kam92], a musical paragraph is usually composed of 16 bars and a very fast tempo is about 1 bar/second in classical music, thus the duration of 4s of a music segment is already only about ¼ musical paragraphs and it is considered as the shortest segment duration in our dataset. There are totally 1205 music samples in the version of 4s, 603 samples in the version of 8s, 416 samples in the version of 16s, and 323 samples in the version of 32s. The samples of each duration length have been labeled with their corresponding mood state. Only consistent segments considered as having stable mood states are selected into our datasets.

In order to study the effectiveness of our mood tracking algorithm, four pieces of classical music annotated in the work of Microsoft Research Asia [Liu03] are also selected in our dataset.

7.5.2 Experimental results by HCS

The HCS classifier generated for the four classes of music mood according to Thayer's model is shown in Fig. 7-14. The same structure is generated with all the operators and parameters tested in our experiment. It appears that this automatically generated hierarchical classification framework has the same structure as the one proposed by Microsoft Research Asia [Lu06]. The energy dimension is divided into two parts at the first stage with classifier 1, and then the high-energy mood states and the low-energy mood states are separated according to the stress dimension respectively in the second stage. Note that in our process of generation of the hierarchical classifier, the pair of mood states subsets with the highest correct classification rate is selected as the sub-classifier in the first stage. This proves that the two dimensions in the mood space truly exist, and the energy dimension is easier to classify than the stress dimension.

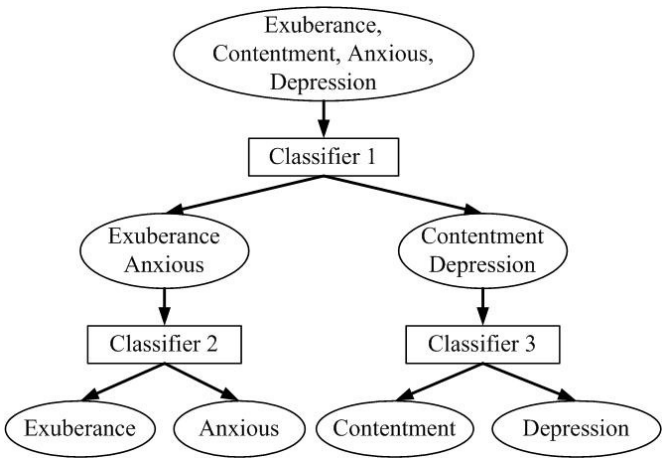

Fig. 7-14 HCS classifier for music dataset

The average classification rates and the root mean square errors of the classification rates are computed for the use of several combination operators with different parameters. The detailed results are listed in Table. 7-3. The mood states are indexed in the confusion matrix for the moods as: M1=Exuberance, M2=Anxious, M3=Contentment, and M4=Depression. The experiments are performed with hold-out cross-validation with 10 iterations. In each iteration, 50% of data are used as training set, and the other 50% are used as test set.

Table. 7-3 Music mood, hold-out cross-validation with 10 iterations, HCS (%)

Operator	Lukasiewicz	Hamacher			
Parameter	-	γ=0.5	γ=3	γ=5	γ=10

Mean rate	4s	78.86±2.70	80.59±1.91	81.35±2.53	81.41±2.58	81.15±2.2
	8s	81.79±2.74	82.72±1.99	83.55±2.58	83.80±2.49	83.42±2.49
	16s	78.29±4.09	82.41±1.93	81.95±2.16	82.21±3.01	81.47±3.37
Operator		Yager				Weber-Sugemo
Parameter		p=1	p=3	p=5	p=10	λ_T=-0.5
Mean rate	4s	78.97±2.74	81.10±2.53	80.20±1.87	79.58±1.62	79.19±1.58
	8s	83.42±2.49	83.47±2.08	81.96±1.74	81.41±3.07	81.43±2.74
	16s	78.32±4.33	82.36±5.89	82.00±2.65	81.49±2.53	80.46±4.45
Operator		Weber-Sugemo				
Parameter		λ_T=0	λ_T=0.5	λ_T=2	λ_T=4	λ_T=5
Mean rate	4s	78.86±2.70	79.81±2.03	81.09±2.12	81.48±2.08	81.81±2.20
	8s	81.76±2.74	82.14±2.74	82.95±2.32	83.55±3.24	83.55±2.33
	16s	78.29±4.09	78.94±4.33	80.36±4.33	81.18±4.21	81.54±3.37
Operator		Schweizer & Sklar				
Parameter		q=0.2	q=0.4	q=0.6	q=0.8	q=1
Mean rate	4s	81.36±2.04	81.34±2.04	81.78±2.08	81.19±1.95	81.33±2.45
	8s	83.67±2.91	83.57±2.15	83.8±2.66	83.48±1.82	83.82±2.15
	16s	82.02±2.53	81.85±3.01	83.03±2.40	82.79±2.17	82.93±2.29
Operator		Schweizer & Sklar		Frank		
Parameter		q=3	q=5	s=2	s=5	s=8
Mean rate	4s	80.16±1.99	80.23±2.37	71.72±1.00	71.72±1.00	75.28±1.21
	8s	81.89±2.57	81.68±2.57	68.91±5.47	74.56±5.39	77.98±3.32
	16s	82.14±3.13	81.76±3.01	66.73±5.05	76.73±6.37	78.99±4.57
Operator		Frank			Average	Geometric Average
Parameter		s=10	s=12	s=15	-	-
Mean rate	4s	77.34±2.32	77.84±2.37	78.22±1.33	77.78±2.04	78.96±2.33
	8s	80.25±3.48	80.66±2.74	80.71±3.07	80.81±2.57	80.71±3.49
	16s	79.95±2.65	79.90±3.73	80.34±1.68	80.65±4.69	81.47±1.68

Average confusion matrix for the best parameter (%)

	Predicted / Actual	M1	M2	M3	M4
4s	M1	81.66	0.66	17.43	0.25
	M2	1.39	87.57	1.42	9.63
	M3	29.25	1.59	68.49	0.67
	M4	0.21	13.58	0.74	85.47

	Predicted / Actual	M1	M2	M3	M4
8s	M1	83.14	0.50	16.03	0.33
	M2	1.02	85.72	0.80	12.46
	M3	22.94	1.35	75.56	0.16
	M4	0.06	11.6	0.00	88.34
16s	Predicted / Actual	M1	M2	M3	M4
	M1	86.26	2.20	11.43	0.11
	M2	1.37	78.31	1.85	18.47
	M3	18.77	2.10	79.14	0.00
	M4	0.33	10.25	1.33	88.08

Fig. 7-15 shows the best classification rates of the tested operators. The error bars in the figure show the root mean square errors of the classification rates.

The best result for the dataset of 4s is 81.81±2.20% with Weber-Sugemo operator when λ_T=5, 83.82±2.15% with Schweizer & Sklar operator when q=1 for the data version of 8s and 83.03±2.40% with Schweizer & Sklar operator when q=0.6 for the data version of 16s. The result for the version of 32s is almost the same with the ambiguous classifier with single judgment and the result will be listed later in section 7.5.3.

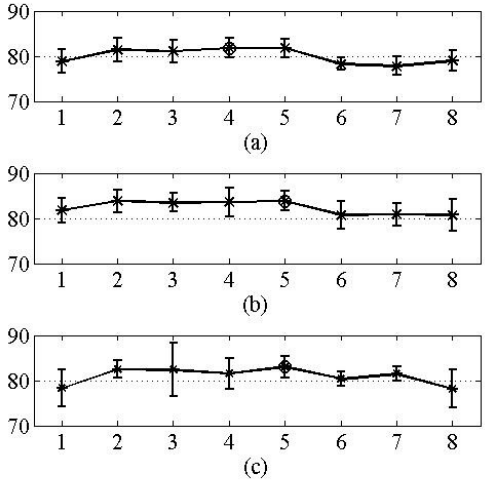

Fig. 7-15 Classification rate with hierarchical classifier for music dataset (a) 4s (b) 8s (c) 16s. The indexes on the X axis stand for the operators: 1 – Lukasiewicz, 2 – Hamacher, 3 – Yager, 4 - Weber-Sugemo, 5 - Schweizer & Sklar, 6 – Frank, 7 - Average, 8 - Geometric Average. Y axis represents the correct classification rate, expressed in percentage

Similar hierarchical classifier based on the taxonomy with Thayer's model is also implemented by Microsoft [Liu03] [Lu06], with classification accuracy of 86.3% with a hierarchical classifier based on GMMs and 80.6% with a non-hierarchical classifier. As their dataset is not publicly available, we cannot compare our classification results with theirs. However, their work suggests that the hierarchical structure of classifier may achieve better performance than the non-hierarchical one.

The features that best represent the mood states in the hierarchical classifier for the music are selected. The most frequently selected 10 features for the 3 sub-classifiers are listed in Table. 7-4. As we mentioned in section 7.4.1 describing the music features, they may not be accurate enough, which explains that this group of features does not appear as very important in these results.

Table. 7-4 Most frequently selected features for the hierarchical classifier for the music mood. The last column presents the groups of the feature: "M" for music features (rhythmic and tonality features), "T" for timbre features, and "O" for octave-based perceptual features

	Feature index	Feature description	
Classifier 1	57	mean value of spectral flux	T
	183	mean value of subband average of the 6th subband	O
	141	normalize variance of subband peak of the 4th subband	O
	132	range of subband average of the 3rd subband	O
	187	variance of subband average of the 6th subband	O
	168	range of subband average of the 5th subband	O
	156	range of subband peak of the 5th subband	O
	126	range of subband valley of the 3rd subband	O
	63	normalized mean value of spectral flux	T
	184	maximum value of subband average of the 6th subband	O
Classifier 2	120	range of subband peak of the 3rd subband	O
	132	range of subband average of the 3rd subband	O
	35	mean value of the middle part of FFT to beat histogram	M
	28	maximum value of beat histogram	M
	149	minimum value of subband average of the 4th subband	O
	133	variance of subband average of the 3rd subband	O
	8	power ratio of the 2nd subband	O
	167	minimum value of subband average of the 5th subband	O

	46	value of Mi^b in the note histogram	M
	173	minimum value of subband peak of the 6th subband	O
Classifier 3	7	power ratio of the 1st subband	O
	28	maximum value of beat histogram	M
	122	normalized variance of subband peak of the 3rd subband	O
	163	variance of subband valley of the 5th subband	O
	159	mean value of subband valley of the 5th subband	O
	98	normalized variance of subband average of the 1st subband	O
	59	maximum of spectral flux	T
	179	minimum of subband valley of the 6th subband	O
	116	normalized variance of subband average of the 2nd subband	O
	35	mean value of the middle part of FFT to beat histogram	M

Table. 7-5 Distribution of the best 20 features for hierarchical classifier for music mood

		Classifier 1	Classifier 2	Classifier 3	All
Music features		0	3	3	4
Timbre features		2	1	2	4
Octave subbands	Energy	0	1	2	3
	1	1	0	1	2
	2	2	1	3	6
	3	3	6	2	8
	4	4	3	1	7
	5	5	3	4	10
	6	3	2	2	7
	Peak	2	5	3	10
	Valley	4	4	3	9
	Average	12	6	7	21

The distribution of the features in each group of features among the first 20 most frequently selected features in the 3 sub-classifiers is listed in Table. 7-5. This list involves totally 51 features.

Among those 51 features, there are 4 music features, 4 timbre features, 3 subband energy features, and 40 subbands spectral contrast features. For classifier 1 (classifier in energy dimension, “Exuberance” & “Anxious/ Frantic” vs. “Depression”

& "Contentment"), no music features are frequently selected, the timbre features appear more often. For classifiers 2 and 3 (classifiers in stress dimension), the music features are more frequently selected; the features 28 (maximum value of beat histogram) and 35 (mean value of the middle part of FFT to beat histogram) appear in both classifiers for the stress dimension.

The distribution of the selected features shows the contribution of the 3 groups of features to the 2 dimensions of the mood states. The music features do not have evident influence on the energy dimension in the mood space; they represent mainly the characteristics in the stress dimension. The effect of timbre features on the two dimensions shows no significant difference. The octave based perceptual features, including subband energy features and spectral contrast features, are the most important group of features for both of the dimensions. The subband energy features, presented in the introduction of the octave-based subbands features (section 7.4.2.2), work better in the energy/arousal dimension, and the subband spectral contrast features are more suitable in the classification in the stress/arousal dimension because they represent the distribution of harmonic and non-harmonic components of music signal.

7.5.3 Experimental results by ambiguous classification

The ambiguous classifier generated for the four classes of music mood according to Thayer's model is shown in Fig. 7-16. The same structure is generated with all the operators and parameters tested in our experiment. Three sub-classifiers are selected to form 2 steps of fusions to get the belief masses of the 4 mood states. A particularly interesting phenomenon in Fig. 7-16 is that the two mood states on the diagonal of the mood space – "Exuberance" and "Depression" – appear together in the classifier 3. This shows that there might be other relationships among the mood states besides the dimensions.

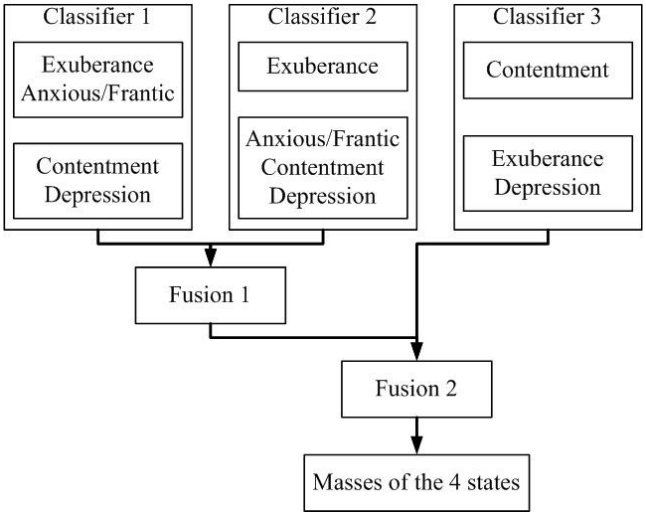

Fig. 7-16 Ambiguous classifier for music dataset

The average classification rates and the root mean square errors of the classification rates are computed for several combination operators with different parameters in two ways of evaluations: single judgment and judgment of multiple possibilities. The detailed results are listed in Table. 7-6.

Table. 7-6 Music mood, hold-out cross-validation with 10 iterations, ACS (%)

Operator		Lukasiewicz	Hamacher			
Parameter		-	γ=0.5	γ=3	γ=5	γ=10
Single	4s	78.86±2.70	80.5±1.75	81.33±2.53	81.40±2.58	81.15±2.20
	8s	81.76±2.74	82.72±1.99	83.55±2.58	83.80±2.49	83.42±2.49
	16s	78.17±4.21	82.31±2.05	81.97±2.16	82.19±3.01	81.47±3.37
	32s	62.54±9.91	64.74±7.12	68.39±2.94	67.34±5.11	66.84±7.12
Multi	4s	78.92±2.67	83.74±1.79	81.76±2.67	81.71±2.49	81.35±2.12
	8s	81.79±2.71	85.44±1.60	83.93±2.66	84.08±2.44	83.65±2.33
	16s	78.26±4.21	84.32±2.29	82.44±2.17	82.49±2.89	81.7±3.29
	32s	62.58±9.91	70.48±5.68	68.88±2.74	67.57±5.11	67.04±7.17
Operator		Yager				Weber-Sugemo
Parameter		p=1	p=3	p=5	p=10	λ_T=-0.5
Single	4s	78.97±2.74	81.06±2.53	80.08±1.87	79.29±1.74	78.52±1.83
	8s	83.42±2.49	83.47±2.08	81.96±1.74	81.41±3.07	81.43±2.74
	16s	78.20±4.45	82.28±5.89	81.61±2.64	81.23±2.77	80.05±4.68
	32s	62.57±9.45	66.47±5.42	60.71±6.66	58.08±6.35	62.26±3.87
Multi	4s	79.02±2.73	84.01±2.06	84.27±1.59	83.80±1.47	78.90±1.90

	8s	83.65±2.33	85.24±2.24	85.29±1.13	85.38±2.18	81.68±2.76
	16s	78.29±4.49	84.03±5.13	84.61±2.61	84.77±2.65	80.31±4.68
	32s	62.60±9.45	71.15±2.84	68.59±5.37	66.75±4.90	62.91±4.18
Operator		Weber-Sugemo				
Parameter		λ_T=0	λ_T=0.5	λ_T=2	λ_T=4	λ_T=5
Single	4s	78.86±2.70	79.81±2.03	81.09±2.12	81.48±2.08	81.81±2.20
	8s	81.76±2.74	82.14±2.74	82.95±2.32	83.55±3.24	83.55±2.33
	16s	78.29±4.09	78.94±4.33	80.36±4.33	81.15±4.09	81.51±3.25
	32s	62.54±9.91	63.25±7.90	65.98±7.74	66.32±6.35	67.37±5.42
Multi	4s	78.92±2.67	79.88±1.99	81.24±2.12	81.70±2.02	82.08±2.13
	8s	81.79±2.71	82.31±2.77	83.16±2.43	83.80±3.24	83.88±2.52
	16s	78.26±4.21	79.09±4.33	80.69±4.41	81.59±4.09	82.04±3.17
	32s	62.58±9.91	63.31±7.90	66.08±7.74	66.60±6.30	67.81±5.32
Operator		Schweizer & Sklar				
Parameter		q=0.2	q=0.4	q=0.6	q=0.8	q=1
Single	4s	81.34±2.04	81.38±2.15	81.78±2.08	81.13±1.91	81.25±2.41
	8s	83.67±2.91	83.57±2.15	83.80±2.66	83.48±1.82	83.82±2.15
	16s	82.02±2.65	81.80±2.88	83.08±2.40	82.74±2.17	82.91±2.29
	32s	68.42±2.64	68.42±4.96	68.98±4.34	68.52±4.34	68.39±4.03
Multi	4s	81.80±2.01	82.26±2.27	82.90±2.16	82.78±1.48	83.41±2.46
	8s	84.14±2.71	84.20±2.49	84.57±2.85	84.68±1.77	85.25±2.38
	16s	82.42±2.81	82.31±2.77	83.64±2.40	83.7±2.57	84.01±2.37
	32s	69.10±2.37	69.41±4.29	70.33±4.18	70.54±3.72	70.94±2.84
Operator		Schweizer & Sklar		Frank		
Parameter		q=3	q=5	s=2	s=5	s=8
Single	4s	80.07±1.87	80.07±2.24	67.93±3.20	75.09±1.16	76.17±1.41
	8s	81.89±2.57	81.68±2.57	68.91±5.47	74.56±5.39	77.98±3.32
	16s	82.00±3.01	81.42±3.01	64.71±5.65	72.91±4.57	76.40±3.25
	32s	63.28±5.42	60.19±8.05	48.82±4.95	53.9±8.05	58.82±6.04
Multi	4s	83.87±1.63	84.19±1.55	74.08±2.98	80.75±0.94	79.45±2.38
	8s	85.22±1.83	85.39±1.85	75.27±4.01	79.78±3.57	78.88±3.04
	16s	84.90±2.61	84.74±2.77	70.97±3.69	75.71±4.13	77.19±3.33
	32s	70.40±3.92	68.62±5.16	57.71±4.60	63.73±6.35	64.25±3.15
Operator		Frank			Average	Geometric

						Average
Parameter		s=10	s=12	s=15	-	-
Single	4s	76.63±2.08	77.73±1.74	78.60±2.08	77.42±1.87	78.61±2.32
	8s	80.25±3.48	80.66±2.74	80.71±3.07	80.83±2.49	80.53±3.32
	16s	78.66±2.89	79.26±3.37	80.00±0.84	80.17±4.33	80.27±2.89
	32s	59.10±6.04	60.80±2.64	62.11±3.56	57.93±7.28	55.45±7.74
Multi	4s	78.25±2.93	78.29±1.85	78.93±2.12	83.71±1.51	83.53±1.84
	8s	80.64±3.71	80.93±2.71	80.81±3.10	86.70±1.85	85.23±2.57
	16s	79.04±3.09	79.58±3.29	80.21±0.96	85.65±4.13	84.78±2.17
	32s	61.57±2.79	61.88±2.27	62.62±3.41	67.32±5.78	65.39±5.27
Average confusion matrix for the best parameter (%)						
Single 4s	Predicted / Actual	M1	M2	M3	M4	
	M1	81.66	0.66	17.43	0.25	
	M2	1.39	87.57	1.42	9.63	
	M3	29.25	1.59	68.49	0.67	
	M4	0.21	13.58	0.74	85.47	
Single 8s	Predicted / Actual	M1	M2	M3	M4	
	M1	83.22	0.41	16.03	0.33	
	M2	1.02	85.72	0.80	12.46	
	M3	22.94	1.35	75.56	0.16	
	M4	0.00	11.66	0.06	88.28	
Single 16s	Predicted / Actual	M1	M2	M3	M4	
	M1	86.26	2.20	11.43	0.11	
	M2	1.29	78.39	1.85	18.47	
	M3	18.77	2.10	79.14	0.00	
	M4	0.25	10.17	1.42	88.17	
Single 32s	Predicted / Actual	M1	M2	M3	M4	
	M1	75.95	12.78	6.33	4.94	
	M2	27.42	67.58	3.71	1.29	
	M3	7.65	4.31	53.63	34.41	
	M4	4.38	1.38	11.50	82.75	
Multi	Predicted / Actual	M1	M2	M3	M4	

4s	M1	81.75	1.02	30.07	0.43
	M2	2.34	86.81	1.37	18.90
	M3	39.98	2.01	72.42	1.13
	M4	0.41	21.25	0.75	86.09
Multi 8s	Predicted / Actual	M1	M2	M3	M4
	M1	82.02	0.69	36.18	0.36
	M2	1.50	85.74	2.89	26.51
	M3	38.94	1.00	77.84	0.65
	M4	0.11	29.59	0.44	89.34
Multi 16s	Predicted / Actual	M1	M2	M3	M4
	M1	86.75	3.89	25.52	0.48
	M2	4.66	77.32	3.27	30.79
	M3	35.58	2.30	77.11	1.98
	M4	0.73	24.73	1.83	91.58
Multi 32s	Predicted / Actual	M1	M2	M3	M4
	M1	73.19	21.27	8.10	7.09
	M2	33.74	69.87	4.52	1.58
	M3	9.10	7.80	54.00	39.57
	M4	4.85	2.78	17.90	84.43

Fig. 7-17 shows the best classification rates of the tested operators with single judgment, and Fig. 7-18 shows the best classification rates with multiple judgments. The error bars in the figures show the root mean square errors of the classification rates.

In the case of single judgment, the best result for the dataset of 4s is 81.81±2.20% with Weber-Sugemo operator when λ_T=5, 83.82±2.15% with Schweizer & Sklar operator when q=1 for the data version of 8s, 83.08±2.40% with Schweizer & Sklar operator when q=0.6 for the data version of 16s and 68.98±4.34% with Schweizer & Sklar operator when q=0.6 for the data version of 32s.

In the case of multiple judgment, the best result for the dataset of 4s is 84.27±1.59% with Yager operator when p=5, 86.70±1.85% with average operator for the data version of 8s, 85.65±4.13% with average operator for the data version of 16s and 71.15±2.84% with Yager operator when p=3 for the data version of 32s.

Similar to the results in the classification of emotional speech with Berlin and DES datasets, the operators with convex curve surfaces of properties such as Weber-Sugemo, Schweizer & Sklar give better results than the others.

For both of the cases, the performances of the versions of 8s and 16s are quite closed to each other. The performance of the version of 4s is a little lower than those results and the performance of the version of 32s decreases obviously. Lu assumed in [Lu06] that the minimum segment duration without mood state changing as 16 seconds which is proposed according to the musical theory of [Kam92] indicating that one musical paragraph is usually composed of 16 bars and a very fast tempo is about 1 bar/second in classical music. For the short music clips of 4s, the music within 4 bars may be not long enough to sufficiently present the mood states, which leads to a little decrease in the classification accuracy, while the mood state in the long music clips of 32s may already change or in the transition between different mood states, thus the recognition accuracy of the longer clips decrease obviously. The stability of music mood within 16s and the possibility to vary in longer durations will also be proved in the next section.

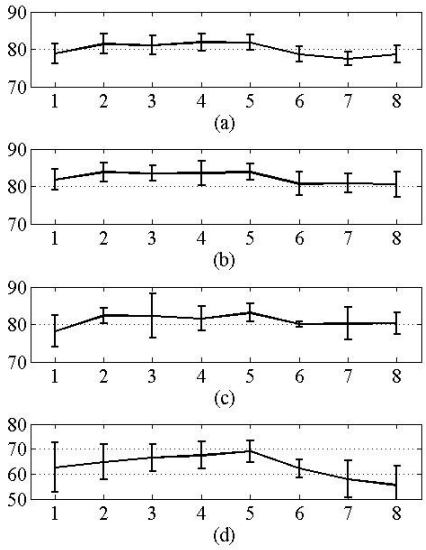

Fig. 7-17 Classification rate with ACS with single judgment for music dataset (a) 4s (b) 8s (c) 16s (d) 32s

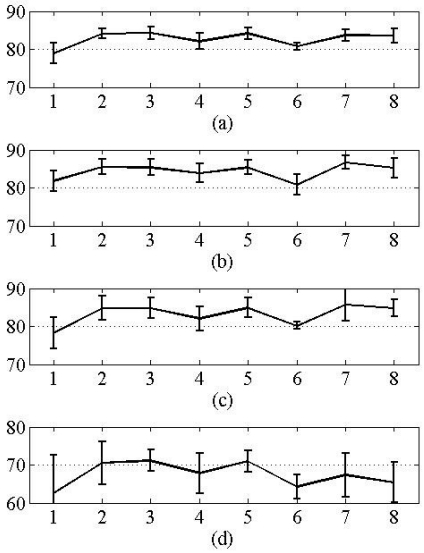

Fig. 7-18 Classification rate with ACS with multiple judgment for music dataset (a) 4s (b) 8s (c) 16s (d) 32s

The features which best represent the mood states in the ambiguous classifier for the music are selected. The 20 most frequently selected features are listed in Table. 7-7. The letters in the last column of the table indicate the feature groups as mentioned in Table. 7-4.

Table. 7-7 Most frequently selected features for the ACS for the music mood

Feature index	Feature description	
55	mean value of spectral centroid	T
187	variance of subband average of the 6th subband	O
178	maximum value of subband valley of the 6th subband	O
156	range of subband peak of the 5th subband	O
162	range of subband valley of the 5th subband	O
174	range of subband peak of the 6th subband	O
28	maximum value of beat histogram	M
35	mean value of the middle part of FFT to beat histogram	M
183	mean value of subband average of the 6th subband	O
77	maximum value of spectral roll off at 50%	T
184	maximum value of subband average of the 6th subband	O
150	range of subband average of the 4th subband	O
108	range of subband valley of the 2nd subband	O
47	value of mi in the note histogram	M

179	minimum value of subband valley of the 6th subband	O
7	power ratio of the 1st subband	O
173	minimum value of subband peak of the 6th subband	O
27	possibility to be major	M
144	range of subband valley of the 4th subband	O
78	range of spectral roll off at 50%	T

Number of features selected in each group

Music features	4	Octave based subbands	1	0	Peak	3
Timbre features	3		2	1	valley	5
Octave subbands energy	1		3	0	average	4
			4	2		
			5	2		
			6	7		

Among these 20 features, there are 4 music features, 3 timbre features, and 13 octave-based perceptual features. Most of these selected features also appear in the features selected in the hierarchical classifier. For the octave-based perceptual spectral contrast features, the features in the high frequency band (especially in the 6th subband) occupy higher proportion than in the low frequency band.

Global classifiers with the same scheme of feature selection are also tested for the dataset of 8s and 32s. The global classifiers are built with the same feature selection and classification scheme as those introduced in section 5.2, allowing here to perform in one step the classification into the four classes. The detailed results are listed in Annex C. The best results for each approach are listed in Table. 7-8. Both the hierarchical approach and the ambiguous approach have significant improvement as compared to the global classifier. The result from the hierarchical classifier implemented with GMMs and non-hierarchical classifier with the work of Microsoft [Liu03] is also displayed in this table.

Table. 7-8 Comparison of best classification rates on DES dataset (%)

	Hierarchical	Ambiguous/ single judgment	Ambiguous/ multiple judgments	Global classifier
8s	83.82	83.82	86.70	72.80
32s	68.90	68.98	71.15	52.01
Work of Microsoft with	86.3	-	-	80.6

segments of 20s [Liu03]				

7.5.4 Influence of the duration of music clips on mood recognition

In order to analyze the influence of the duration of music clips on the recognition of mood, we have performed cross tests. Considering there are only slightly differences between the classification rates of the hierarchical classifiers and the ambiguous classifiers with single judgment, only the ambiguous classification method is applied in this part of work.

The cross tests are implemented by taking one of the versions of the dataset as the training data to get the classification model, and all the 4 versions with different segment durations as testing set to evaluate the performances. Some samples with longer durations may contain one or several samples with shorter durations. The detailed results are listed in Table. 7-9 (single judgment) and Table. 7-10 (multiple judgments), and the best classification rates for the cross tests with different operators are illustrated in Fig. 7-19.

Table. 7-9 Results on cross test, ambiguous classification with single judgment. Operators/parameters are marked above the results. (%)

Test set	4s	8s	16s	32s	4s	8s	16s	32s	4s	8s	16s	32s
	Lukasiewicz				Hamacher							
Model	-				γ=0				γ=0.5			
4s	80.08	75.46	66.83	27.24	81.33	75.95	65.63	25.08	81.83	78.11	66.83	27.86
8s	77.51	82.92	80.05	30.03	78.09	84.58	76.68	32.20	79.00	85.24	79.09	34.67
16s	76.10	76.95	84.13	32.82	70.95	73.96	86.78	33.75	72.78	78.77	87.98	32.51
32s	33.86	34.83	38.22	64.71	33.86	34.00	32.69	60.06	33.53	32.67	35.10	70.28
	Hamacher											
	γ=3				γ=5				γ=10			
4s	84.4	78.44	69.95	28.17	84.73	77.94	70.19	30.96	83.9	77.61	68.99	31.89
8s	79.00	87.23	79.81	30.03	80.08	85.57	80.29	32.51	77.59	86.57	79.57	28.17
16s	74.02	76.62	88.46	28.79	75.85	77.45	88.46	32.51	74.77	76.62	87.74	32.51
32s	30.37	28.86	33.17	71.21	31.37	30.02	33.65	69.97	27.97	27.20	32.69	69.04
	Yager											
	p=1				p=3				p=5			
4s	80.25	75.62	67.07	26.32	81.83	75.95	70.43	30.34	80.58	76.12	68.51	25.39
8s	77.34	83.08	80.29	29.41	81.49	86.24	79.57	34.37	77.59	84.25	76.92	32.20

16s	76.02	76.29	84.13	33.13	76.51	78.94	88.46	32.51	73.03	77.61	87.02	31.27
32s	33.94	34.99	38.46	64.71	34.11	32.84	32.93	69.97	33.61	32.34	32.69	63.47
	Yager				Weber-Sugemo							
	p=10				λ_T=-0.5				λ_T=0			
4s	80.50	75.95	66.11	24.77	78.76	75.79	69.95	29.1	80.08	75.46	66.83	27.24
8s	77.34	84.08	77.88	33.44	76.35	85.07	80.05	28.79	77.51	82.92	80.05	30.03
16s	73.78	76.29	85.82	30.34	77.76	80.10	84.62	29.41	76.10	76.95	84.13	32.82
32s	32.95	31.84	31.97	59.75	31.95	36.65	34.38	60.06	33.86	34.83	38.22	64.71
	Weber-Sugemo											
	λ_T=0.5				λ_T=2				λ_T=4			
4s	81.41	76.95	68.51	27.55	82.99	77.78	70.19	30.34	83.90	79.10	68.99	30.03
8s	78.67	84.41	82.69	29.72	79.25	86.07	78.37	29.72	77.84	85.57	78.37	30.96
16s	71.95	78.77	85.10	32.51	72.45	76.62	86.54	31.58	75.60	77.11	88.22	32.82
32s	33.36	34.99	36.06	67.18	30.71	29.85	33.65	67.80	29.13	27.53	31.73	69.66
	Weber-Sugemo				Schweizer & Sklar							
	λ_T=5				q=1				q=3			
4s	83.90	78.94	68.99	30.03	82.90	78.11	68.99	28.17	80.83	75.12	65.87	26.01
8s	79.92	86.07	81.25	30.65	80.08	87.06	76.68	33.75	80.17	84.91	77.40	34.98
16s	74.02	76.45	88.46	32.51	74.77	78.61	87.74	31.27	70.62	75.95	86.78	32.20
32s	31.62	30.35	35.10	69.04	34.52	31.34	35.10	71.83	34.52	35.16	39.18	66.56
	Schweizer & Sklar				Frank							
	q=5				s=2				s=5			
4s	81.24	75.95	65.38	25.08	66.22	66.83	60.10	33.75	74.27	74.30	73.08	29.72
8s	78.09	84.25	78.85	34.06	71.62	71.14	70.19	30.34	73.61	74.30	72.84	29.41
16s	70.46	73.63	86.54	33.13	57.68	59.87	66.35	32.20	68.55	67.50	76.92	30.34
32s	31.54	29.02	28.13	58.20	13.86	13.27	25.72	51.08	26.56	30.35	29.09	54.8
	Frank											
	s=8				s=10				s=12			
4s	75.52	73.30	69.71	34.06	76.35	74.79	70.67	31.58	76.60	74.46	71.15	34.06
8s	77.68	81.92	76.92	29.10	80.00	82.92	78.61	30.96	76.60	83.42	77.40	28.48
16s	70.04	73.80	78.61	26.93	75.52	76.45	81.01	31.58	71.45	75.95	80.77	32.82
32s	33.61	27.36	33.65	61.61	34.61	33.5	35.82	63.78	32.86	36.48	35.34	63.47
	Frank				Average				Geometric Average			
	s=15				-				-			

4s	77.84	76.12	72.12	26.63	79.00	78.28	67.31	29.41	80.33	76.12	67.07	23.84
8s	78.01	83.08	79.33	28.48	64.98	82.42	70.19	28.79	65.81	82.92	78.13	29.10
16s	71.04	75.46	81.97	31.27	76.10	76.78	83.65	33.44	77.43	76.78	83.89	30.34
32s	35.77	38.81	38.46	64.40	25.23	29.85	31.73	57.89	31.87	32.67	31.25	58.82

Table. 7-10 Results on cross test, ambiguous classification with multiple judgment (%)

Test set	4s	8s	16s	32s	4s	8s	16s	32s	4s	8s	16s	32s
	Lukasiewicz				Hamacher							
Model	-				γ=0				γ=0.5			
4s	80.08	75.46	66.91	27.24	85.17	80.38	70.27	30.13	84.9	81.21	70.35	31.48
8s	77.70	82.92	80.05	30.34	82.38	87.23	80.69	34.88	83.10	88.06	82.61	36.12
16s	76.13	77.11	84.13	32.92	77.98	80.15	89.98	35.81	77.68	82.59	89.34	34.06
32s	33.97	35.10	38.62	64.71	37.87	36.59	36.62	69.14	36.18	34.44	36.54	72.76
	Hamacher											
	γ=3				γ=5				γ=10			
4s	84.67	79.22	70.91	28.69	84.87	78.44	70.99	31.27	83.96	77.83	69.47	31.99
8s	79.75	87.45	80.13	30.65	80.72	85.79	80.85	33.02	78.04	86.73	79.89	28.69
16s	74.47	76.78	88.70	29.00	76.87	78.33	88.62	32.82	75.46	77.22	87.98	32.92
32s	30.68	29.13	33.57	71.41	31.62	30.29	33.81	70.07	27.97	27.31	32.77	69.04
	Yager											
	p=1				p=3				p=5			
4s	80.28	75.62	67.15	26.32	84.9	79.88	73.32	33.33	84.56	80.15	73.00	29.62
8s	77.48	83.08	80.29	29.72	84.07	88.00	81.65	35.5	82.79	87.84	81.49	35.29
16s	76.13	76.51	84.13	33.13	79.92	82.31	89.66	33.33	79.00	82.70	89.74	33.75
32s	34.00	34.99	38.46	64.71	35.96	34.44	34.38	73.79	37.93	35.66	36.86	71.10
	Yager				Weber-Sugemo							
	p=10				λ_T=-0.5				λ_T=0			
4s	84.84	80.65	72.04	29.51	79.00	76.56	71.31	30.65	80.08	75.46	66.91	27.24
8s	82.79	87.84	82.53	36.53	76.65	85.13	80.29	29.00	77.70	82.92	80.05	30.34
16s	80.53	82.03	89.34	33.02	78.45	80.38	84.70	29.51	76.13	77.11	84.13	32.92
32s	37.15	34.72	35.34	67.39	32.86	37.31	34.78	61.20	33.97	35.10	38.62	64.71
	Weber-Sugemo											
	λ_T=0.5				λ_T=2				λ_T=4			
4s	81.41	76.95	68.51	27.66	83.02	77.89	70.51	30.65	83.98	79.27	69.15	30.34

8s	78.76	84.41	82.69	30.13	79.67	86.18	78.85	30.03	78.42	85.79	79.01	31.06
16s	71.98	78.77	85.10	32.71	72.56	76.89	86.70	32.30	76.02	77.28	88.22	33.23
32s	33.50	35.05	36.30	67.18	30.84	29.91	33.81	67.80	29.35	27.69	31.89	69.76
	Weber-Sugemo				Schweizer & Sklar							
	λ_T=5				q=1				q=3			
4s	83.98	79.27	69.07	30.55	85.06	80.54	70.67	30.55	84.62	79.38	69.79	30.13
8s	80.39	86.24	81.57	31.17	82.13	88.06	78.37	34.67	83.68	87.62	80.93	36.74
16s	74.14	76.45	88.46	32.61	78.17	81.81	88.94	32.61	76.63	80.87	88.94	33.95
32s	31.87	30.51	35.26	69.25	36.18	32.67	36.38	74.10	37.70	37.65	41.51	72.34
	Schweizer & Sklar				Frank							
	q=5				s=2				s=5			
4s	85.09	80.38	70.11	30.13	72.61	72.31	67.15	41.07	79.92	78.77	77.48	34.88
8s	82.24	87.01	82.21	36.53	78.31	77.72	75.64	33.95	79.75	80.65	78.29	33.02
16s	77.43	79.82	89.74	36.02	63.85	66.00	71.88	35.81	71.26	69.98	78.69	31.17
32s	36.10	32.67	32.29	67.29	24.90	24.10	35.18	60.58	34.47	36.71	36.46	66.15
	Frank											
	s=8				s=10				s=12			
4s	78.15	75.40	71.23	36.43	77.21	75.18	70.91	32.82	76.96	74.85	71.23	34.67
8s	80.66	83.80	79.41	30.86	80.91	83.03	78.85	31.06	77.10	83.58	77.88	29.41
16s	70.79	74.63	79.33	27.04	76.21	76.73	81.09	31.68	71.87	76.23	80.77	32.82
32s	34.55	28.41	34.54	64.29	35.16	34.27	36.78	64.19	33.44	36.71	36.54	64.50
	Frank				Average				Geometric Average			
	s=15				-				-			
4s	78.15	76.34	72.52	27.04	85.01	84.91	76.52	35.40	84.92	80.93	71.79	28.69
8s	78.45	83.14	79.41	28.48	74.44	87.84	77.96	32.09	70.84	87.67	82.37	31.79
16s	71.26	75.62	81.97	31.27	83.02	83.75	88.46	36.33	83.21	82.64	87.90	33.44
32s	36.15	38.86	38.94	64.81	31.48	33.06	35.18	66.25	38.98	38.97	37.58	69.04

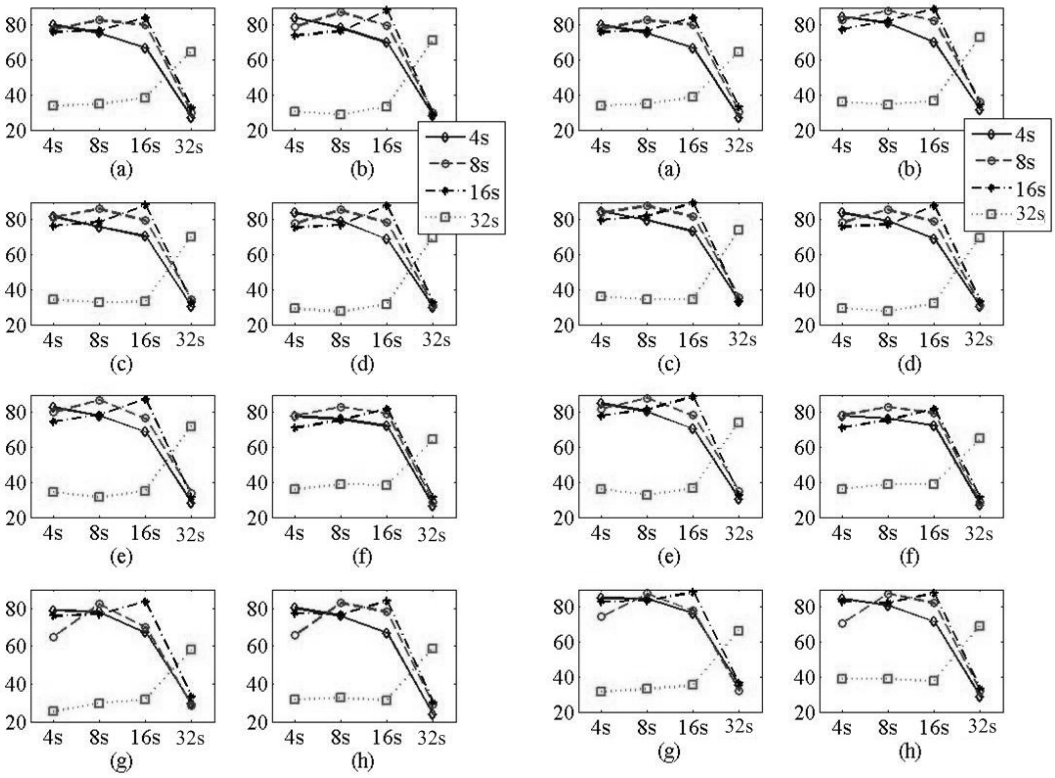

Fig. 7-19 Cross tests for the different duration of music clips with ambiguous classifier. The figure in the left shows the result with single judgment; the figure in the right shows the result with multiple judgments. The legend shows the models correspond to the curves, the X axis shows the duration of the test sets.
(a) Lukasiewicz (b) Hamacher (c) Yager (d) Weber-Sugemo (e) Schweizer & Sklar (f) Frank (g) Average (h) Geometric Average

All the sub figures in Fig. 7-19 have identical tendency with the curves. The performances between the three shorter durations are relatively close to each other (higher than 60%; higher than 70% in most cases), while the cross tests between the shorter versions and the version of 32s normally have correct classification rate between 20% and 40%. The best results in the cross tests without cross validation and the corresponding parameters are listed in Table. 7-11. The best classification rate between the three shorter versions is 82.69% with single judgment and 84.91% with multiple judgments except the rates with the same set in training and testing. Taken the three versions of 4s, 8s, and 16s as an integrated ensemble, the best performance in the cross test between the version of duration of 32s and the shorter versions is 39.18% with single judgment and 41.51% with multiple judgments. The

Weber-Sugemo operator works better for the single judgment and the average operator works better for the cross tests with multiple judgments.

Table. 7-11 Best classification rates and the corresponding operators/parameters in the cross tests with ambiguous classifier

Single judgment									
Classification rate (%)					Operator/parameter				
test set / model	4s	8s	16s	32s	test set / model	4s	8s	16s	32s
4s	84.73	79.10	73.08	34.06	4s	2/5	4/4	6/5	6/8
8s	81.49	87.23	82.69	34.98	8s	3/3	2/3	4/0.5	5/3
16s	77.76	80.10	88.46	33.75	16s	4/-0.5	4/-0.5	3/3	2/0
32s	35.77	38.81	39.18	71.83	32s	6/15	6/15	5/3	5/3
Multiple judgments									
Classification rate (%)					Operator/parameter				
test set / model	4s	8s	16s	32s	test set / model	4s	8s	16s	32s
4s	85.17	84.91	77.48	41.07	4s	2/0	7/-	6/5	6/2
8s	84.07	88.06	82.69	36.74	8s	3/3	2/0.5	4/0.5	5/5
16s	83.21	83.75	89.98	36.33	16s	8/-	7/-	2/0	7/-
32s	38.98	38.97	41.51	74.10	32s	8/-	8/-	5/3	5/1

With the comparison of the values in Table. 7-11, the differences between the shorter segments (4s, 8s, and 16s) and the longer segments (32s) can be observed. First, the performances of cross validation within the dataset of certain duration of segments between the versions of 4s, 8s and 16s are rather close to each other, and are obviously higher than that of the version of 32s. Second, in the cross tests between the four versions of dataset, the models from datasets of 4s, 8s and 16s match well to each other, and conflict with the model from dataset of 32s. Two points can be drawn from these differences. First, the mood states may already change during the duration of 32s; second, the shorter segments have coincident characteristics in presenting the mood cues, while the characteristics change in the longer segments. These results validate the assumption in [Lu06] of the minimum segment duration without mood state changing as 16 seconds.

7.5.5 Music mood tracking

The mood is usually not consistent in a whole piece of classical music [Kam92]. Based on the results of section 7.5.4, the mood can be considered as stable within 16 seconds, while it may change significantly during the range of 32 seconds. Although the main object of this work is the automatic classification/detection of emotional/mood states of audio signals, a simple music mood tracking is also implemented based on the algorithms of mood detection.

The music is first divided into several frames each of them containing a constant mood state, and then the mood state in each frame is detected respectively. The same models with the hierarchical classifier or the ambiguous classifier are applied for the mood tracking. The frame length corresponds to the duration of segments of the models. According to the performances of the the mood classification, the frame length is set to 8 seconds or 16 seconds. The consequent frames have overlaps of half a frame in order to smooth the tracking result as shown in Fig. 7-20.

The mood state for any frame between two frames with the same mood state is forced to be identical with its neighbors, because with the overlap between the frames, all the contents in a frame appear in its neighbor frames, and the identical judgments on its neighbor frames might be highly reliable.

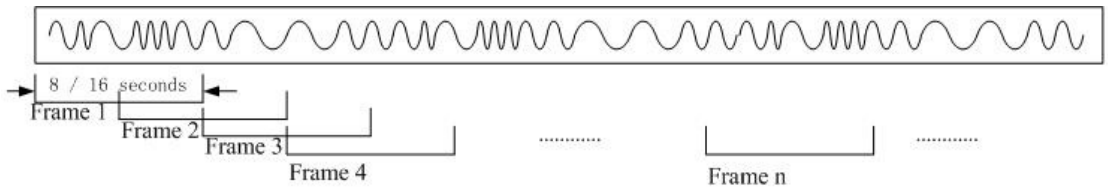

Fig. 7-20 Frames in mood tracking

Several music pieces are tracked in our experiments. As it is a preliminary work, we do not have the exact ground truth of the mood states over the whole pieces. Thus, our goal here is to see the probable trend of the mood tracks.

The tracking results with hierarchical model of 8 seconds to three pieces music with relatively stable mood states and one example with quickly changing mood states are illustrated below. The mood reference to these pieces comes from the comments in the literature [Liu03] [Lu06]. The tracked pieces are: 1) A part taken from Offenbach's "Orpheus in the Underworld", which is famous as "Can Can". The most part of this piece is with the mood state "Exuberance"; 2) A segment from Bach's

"Jesus, Joy of Man's Desiring", with the mood state "Contentment" according to [Lu06]; 3) An anxious example, opening of Stravinsky's "Firebird"; 4) The beginning part of Beethoven's "Fate" with different mood states appearing alternately.

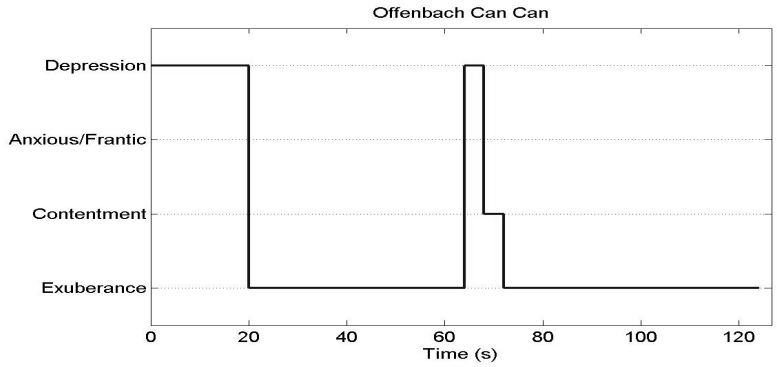

Fig. 7-21 Mood tracking, Offenbach, "Can Can"

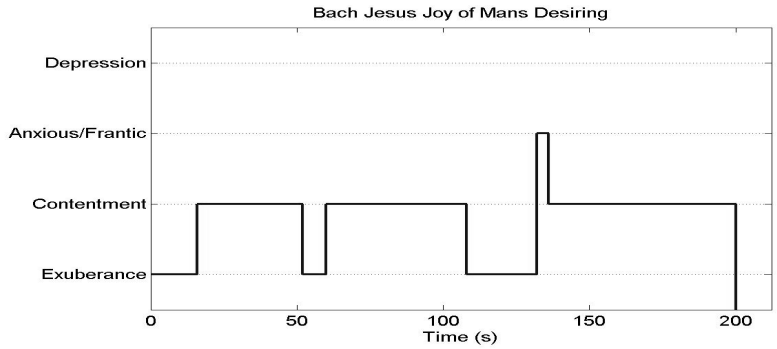

Fig. 7-22 Mood tracking, Bach, "Jesus joy of man's desiring"

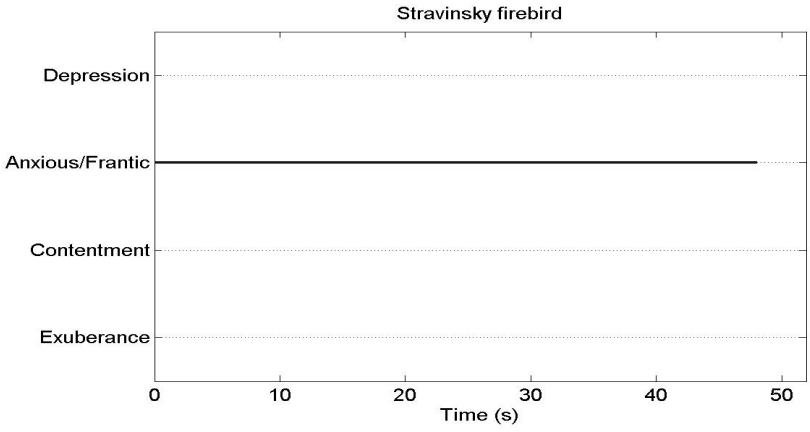

Fig. 7-23 Mood tracking, Stravinsky, "Firebird"

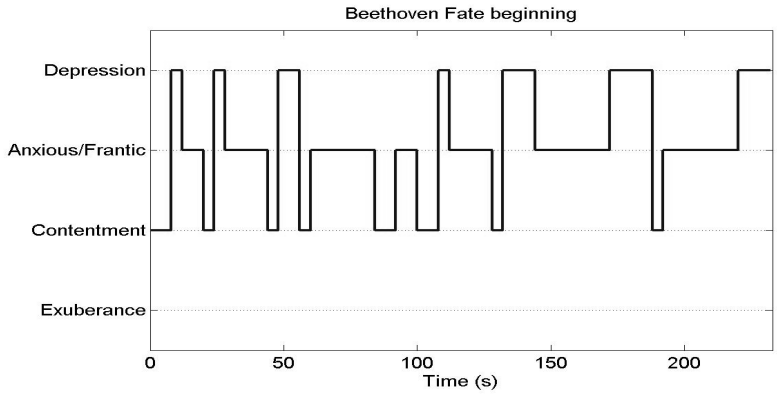

Fig. 7-24 Mood tracking, Beethoven, Beginning of "Fate"

7.5.6 Synthesis

In this subsection, we summarize the experimental results on music mood in the following aspects: the applicability of the two algorithms of classification of audio emotions, the most efficient features in classification of music mood, the influence of the segment durations to the recognition of music mood, and the performances of the combination operators. The results on classification of music mood are also compared with those on vocal emotions in each aspect.

The automatic algorithms proposed for the classification of vocal emotions are experimented on the music mood in this section. Firstly, the same structure of hierarchical classifier is generated automatically in our experiments as the empirical hierarchical framework proposed in the work of Microsoft Asia [Liu03] [Lu06]. Secondly, same as in the case of classification on vocal emotions, the automatically generated hierarchical frameworks for music mood proceeds also first with a sub-classifier in the arousal dimension. Thus, the arousal/energy dimension is presented by more superficial factors in the emotions and is easier to distinguish. The acoustic clues according to the appraisal/stress dimension are more subtle and more difficult to recognize. Thirdly, the performance with the hierarchical classifier are rather close to those with the ambiguous classifier when only a single judgment is taken, while there is no final fusion step in the hierarchical classifier and thus it is simpler in classification. Thus, when only a definite assignment of mood state is needed, the hierarchal classifier may be used, whereas in the situation when vague judgments with multiple possible moods, the ambiguous classifier is more adapted for the classification.

Several features are selected as efficient in the two approaches. These features include rhythmic features such as the maximum value of the beat histogram and the properties of the FFT of the beat histogram, tonality features such as probability to be major, timbre features such as spectral centroid and spectral rolloff, and octave-based perceptual features especially the spectral contrast features in the 5^{th} and 6^{th} octave subband. Moreover, the music features and perceptual features have different importance for the two dimensions in the mood space with Thayer's model. The music features mainly correspond to the stress dimension, the perceptual features correspond to both stress and energy dimensions, while the octave-based features in higher frequency bands are more important in the energy dimension. The perceptual features show more importance in our experiments, especially the octave-based features. The music features do not perform so satisfying, partly because although the music features especially tonality features are important in the acoustic correlates of music moods, they could not be extracted with high accuracy at the current stage (for example, the tonality as major/minor).

The influence of segment durations is presented in the results of the experiments with the two classification schemes in subsection 7.5.2 and 7.5.3, and are particularly analysed in subsection 7.5.4. First, the segments with durations of 8 seconds or 16 seconds are better classified in both approaches. The performance decrease a little for the segments with a duration of 4 seconds, and decrease obviously for the segments with duration of 32 seconds. Second, in the cross test with the segments with different durations, the datasets with segments durations from 4 seconds to 16 seconds show high similarities between each other, while the dataset with duration of 32 seconds is very different from other three versions of durations due to the mood that changes in long music clips. According to Kamien [Kam92], a musical paragraph is usually composed of 16 bars and a very fast tempo is about 1 bar/second, which indicates that an average length of normal musical paragraph should be 16 seconds. For shorter segments such as 4 seconds, the mood state might be considered as consistent while not yet sufficiently expressed in such a length of duration. The reason for the obvious decrease of classification and inconsistent in the classification model with the shorter segment durations for longer segments such as 32 seconds is that there exist more than one music paragraph and the mood state might already change. Thus, the analyses of music mood should be based on shorter segments, in our case, no longer than 16 seconds.

The best operator in the feature combination and selection for the hierarchical classifier and for the single judgment of the ambiguous classifier is Schweizer & Sklar operator, whereas average operator suits better the case of multiple judgments of the ambiguous classifier. Similar to the results obtained with the emotional speech, it is proved again with the music mood that the operators that have properties of convex curve surfaces perform better than the other operators.

Chapter 8
Conclusion and Future Work

8.1 Contributions

This research work addresses the problem of the elaboration of a system allowing the automatic emotion/mood classification of audio signals for both emotional speech and music mood. The system comprises a new feature combination and selection scheme, as well as two approaches of classification algorithms. Our work is evaluated with two datasets of emotional speech and one dataset of music mood.

The contributions in this thesis are discussed as follows.

8.1.1 Feature sets and feature selection scheme

Appropriate features that can efficiently carry the correlative characteristics of signals are of great importance in the problems of pattern classification. Thus, classification algorithms should be preceded by an efficient feature set extraction and feature selection processes.

Independent feature sets are built upon the emotional speech and music clips respectively. Besides the traditional acoustic features concerning the pitch, frequency and energy, the feature sets contain several new features that express the emotional/mood clues of audio signals. New harmonic features based on a sub-band amplitude modulation of the signal and features derived from an analysis according to Zipf laws are proposed for the emotional speech. For the description of the music mood, music features according to the rhythm and the tonality, and perceptual features including timbre features according to the spectral shape and octave-based features have been developed.

An embedded feature selection scheme (ESFS), based on both SFS and the evidence theory is proposed in section 5.2. Its basic idea is to generate new features with the original features by combining the belief masses of features using t-norm operators. The process of feature selection itself serves also as a classifier, which is used as an elementary classifier for the scheme proposed in our automatic classification approaches HCS and ACS..

8.1.2 Algorithms adapting different problems

Due to the lack of universal agreement on the emotion definitions, the number of types of discrete emotional/mood states and their distribution in the dimensional emotion space may change in different problems facing different objectives. The problem of the definition of audio emotion/mood is an interdisciplinary subject that concerns philosophy, psychology, biology and so on. Before it reaches a final uniform definition with universal agreement, the classification and recognition of the audio emotion/mood has to deal with various number and type of classes in the definition of emotions in the level of signal processing. Therefore, we proposed two approaches of classification algorithms that can adapt classification problems with different number and types of classes of emotional/mood states: a hierarchical framework HCS and an ambiguous method ACS.

8.1.3 Automatic hierarchical classification

The combinations of several binary sub-classifiers into a hierarchical framework can improve the performance of the classifications. Hierarchical frameworks have been applied in a few previous works in the audio emotion/mood classification. However, they were performed with empirical structures according to the specific number and types of emotional states in particular problems and the change of the emotion definition may cause a great deal of repeated work for building the new hierarchical structure. Therefore, we propose an approach to automatically build the hierarchical classifier. The performances of sub-classifiers on binary pairs of emotional states can be evaluated by our system and thus the optimized structure of hierarchical classifier can be generated automatically for any given definition of audio emotion/mood according to Chapter 5. The automatic HCS can adapt the different problems, but it does not necessarily give a better classification performance than empirical hierarchical framework. With the experimental results in Chapter 5 and

Chapter 7 on different audio datasets, we proved that the automatically generated hierarchical frameworks can validate perfectly the distribution of the emotional/mood states in the dimensional space, with the first stage of classification in arousal/energy dimension.

8.1.4 Ambiguous classification

As subjective judgments of human beings, the emotion/mood contained in a certain segment of audio signal may be between some states or be a combination of several states. Thus, we proposed an ambiguous classification method ACS as a *first step* considering the multiple emotional labels using the belief masses of each emotional state as output instead of the judgment with yes/no in Chapter 6 to make emotion judgments as close as possible to human judgments. The belief masses are obtained by analyzing the classifiers concerning all the emotional/mood classes or some of the classes and making fusions among the best classifiers. The process of selecting the sub-classifiers can be shared with the method of hierarchical classification. The performance of this approach is evaluated in Chapter 6 and Chapter 7 on different datasets.

8.2 Perspectives for future work

Extensions of this work that we envisage are presented in the following paragraphs.

8.2.1 Further investigation with ambiguous classification of emotions

The ambiguous approach proposed in our work is only a first step in the ambiguous recognition of emotions, which try to simulate the human process for emotion recognition, and it needs to be greatly improved. The two main aspects of improvements are the followings.

First, the original emotional labels from the datasets with single emotion are still used as ground truth in the learning process. In our future work, the ground truth of emotions for the learning process needs to be modified to multiple emotional labels in order to give a more reliable model of ambiguous judgment. A new dataset aiming at the subjective vocal emotions will be needed.

Second, emotional states are quite fuzzy and should be continuous which may lead to the different degree within the same emotional family. This continuity in the emotions should be a topic of interest in the future investigates. Further to the two dimensional model with an arousal and an appraisal dimension used in our work, more precise model of the position of the emotions in the dimensional space is needed for the research of continuous emotions.

8.2.2 For voice-instrumental mixed music

The resources of audio signals concerned in this work are limited to speech and instrumental music, while the songs which contain both human voice and instruments compose a large proportion of music in daily life. The emotion/mood aspects contained in voice-instrumental mixed music will include some of the characteristics of both speech and instrumental music, and some specific features other than the features in either of the two types of audio signals. More complex models should be built for the songs with both human voice and instruments in our future work, for both the feature sets and the description of emotional/mood states. The techniques involved with voice and music separation might be used for reference in this topic [Kil02] [Li07].

8.2.3 For the classification problems with a large number of classes

Although the two approaches of classification in this work are proposed to be applied to adapt different pattern classification problems with audio emotion/mood, the number of classes in the problem affects the computational complexity in our approaches. Although the classification process is rather fast (about 3~5 seconds for a dataset with 500 samples with 4 to 6 classes), the training time for generating the classifiers is influenced seriously by the number of classes. The number of sub-classifiers to be evaluated in the generation of the classifiers increases exponentially with the number of classes. For a dataset with about 500 samples, the training time for evaluating the sub-classifier for problem with 4 classes is about 400 to 500 seconds according to different operators, and the training time increases to 900 to 1000 seconds for problem with 5 classes and 1500 to 2000 seconds for problem with 6 classes. If the number of classes in the classification problems is larger, the training time needed in our approaches will be much longer. The evaluation process should be optimized and simplified in our future work. For example, some obvious

empirical knowledge might be introduced to make a semi-automatic approach by forcing the classes very far to each other to be located in different class subsets when building the initial stages of the hierarchical structure to reduce the number of possible sub-classifiers to be evaluated.

8.2.4 Evaluation of the approaches on other classification problems

The approaches we have developed and presented in this report can also be applied as general classifiers and can be implanted for other classification problems on audio or images. As an example, we envisage to apply them for the classification of music genres. Moreover, these methods could also be applied in security systems when, for instance, alarms have to be started automatically starting the alarm when sounds with dangerous emotions are detected.

Annex A

Feature list for emotional speech

All the features used in our work for the classification of emotional speech including the traditional features (frequency based features, energy based features, and MFCC features) and the new features proposed in our work (harmonic features and Zipf features) are listed in Table A – 1 as follows:

Table A - 1 Feature list for emotional speech

Group	Index	Description
Harmonic features	1	Mean value over entire harmonic space
	2	Variance over entire harmonic space
	3	Normalized variance over entire harmonic space
	4	Mean value in area 1 of harmonic space
	5	Maximum value in area 1 of harmonic space
	6	Variance in area 1 of harmonic space
	7	Normalized variance in area 1 of harmonic space
	8	Mean value in area 2 of harmonic space
	9	Maximum value in area 2 of harmonic space
	10	Variance in area 2 of harmonic space
	11	Normalized variance in area 2 of harmonic space
	12	Mean value in area 3 of harmonic space
	13	Maximum value in area 3 of harmonic space
	14	Variance in area 3 of harmonic space
	15	Normalized variance in area 3 of harmonic space
	16	Mean value in area 4 of harmonic space
	17	Maximum value in area 4 of harmonic space
	18	Variance in area 4 of harmonic space
	19	Normalized variance in area 4 of harmonic space
	20	Ratio of mean value between area 2 and area 1 of harmonic space
	21	Ratio of mean value between area 3 and area 1 of harmonic space
	22	Ratio of mean value between area 4 and area 1 of harmonic space

Frequency features	23	Mean F0
	24	Maximum of F0
	25	Minimum of F0
	26	Median of F0
	27	Variance of F0
	28	Mean of the first formant F1
	29	Maximum of the first formant F1
	30	Minimum of the first formant F1
	31	Median of the first formant F1
	32	Variance of the first formant F1
	33	Mean of the second formant F2
	34	Maximum of the second formant F2
	35	Minimum of the second formant F2
	36	Median of the second formant F2
	37	Variance of the second formant F2
	38	Mean of the third formant F3
	39	Maximum of the third formant F3
	40	Minimum of the third formant F3
	41	Median of the third formant F3
	42	Variance of the third formant F3
Energy features	43	Mean energy
	44	Maximum of energy contour
	45	Minimum of energy contour
	46	Ratio of energy below 250 Hz
	47	Mean duration of silence period
	48	Maximum duration of silence period
	49	Ratio of silence period
	50	Mean duration of energy plateaus
	51	Maximum duration of energy plateaus
	52	Median duration of energy plateaus
	53	Variance of duration of energy plateaus
	54	Mean duration of energy valleys
	55	Maximum value of energy valleys
	56	Median of energy valleys

	57	Variance of duration of energy valleys
	58	Mean of slope of rising edges of energy contour
	59	Maximum of slope of rising edges of energy contour
	60	Median of slope of rising edges of energy contour
	61	Variance of slope of rising edges of energy contour
	62	Mean of duration of rising edges of energy contour
	63	Maximum of duration of rising edges of energy contour
	64	Median of duration of rising edges of energy contour
	65	Variance of duration of rising edges of energy contour
	66	Number of rising edges of energy contour per second
	67	Mean of slope of falling edges of energy contour
	68	Maximum of slope of falling edges of energy contour
	69	Median of slope of falling edges of energy contour
	70	Variance of slope of falling edges of energy contour
	71	Mean duration of falling edges of energy contour
	72	Maximum of duration of falling edges of energy contour
	73	Median of duration of falling edges of energy contour
	74	Variance of duration of falling edges of energy contour
	75	Number of falling edges of energy contour per second
	76	Mean value of cross zero rate
	77	Maximum value of cross zero rate
	78	Median value of cross zero rate
	79	Variance of cross zero rate
	80	Normalized variance of cross zero rate
MFCC features	81-224	Mean value of the MFCC coefficients
		Maximum value of the MFCC coefficients
		Minimum value of the MFCC coefficients
		Mean of the absolute values of the MFCC coefficients
		Variance of the MFCC coefficients
		Normalized variance of the MFCC coefficients
Zipf features	225	Entropy feature of Inverse Zipf of frequency coding
	226	Resampled polynomial estimation Zipf feature of UFD coding

Annex B

Feature list for music mood

All the features used in our work for the classification of music mood including the musical features (rhythmic features, and tonality features) and the perceptual features (timbre features and octave-based features) are listed in Table B – 1 as follows:

Table B – 1 Feature list for music mood

Group	Index	Description
Rhythmic features	1	Tempo
	2	Possibility to be major
	3	Maximum of beat histogram
	4	Position of maximum value in the beat histogram
	5	Position of the first peak in the beat histogram
	6	Number of peaks in the beat histogram
	7	Minimum of distance of peaks in the beat histogram
	8	Maximum of distance of peaks in the beat histogram
	9	Width of the most wide peak in the beat histogram
	10	Mean of middle part of FFT to beat histogram
Tonality features	11	Maximum of note histogram
	12	Position of Maximum in the note histogram
	13	Number of note above middle value note histogram
	14	Variance of note histogram
	15	Value of the first peak in the FFT of the note histogram
	16	Position of the first peak in the FFT of the note histogram
	17	Value of the last peak in the FFT of the note histogram
	18	Position of the last peak in the FFT of the note histogram
	19	Ratio between the last peak and the first peak in the note histogram
	20	Number of peaks in the note histogram
	21	Number of re#
	22	Number of mi

	23	Ratio between re# and mi
Timbre features	24	mean spectral centroid
	25	variance of spectral centroid
	26	Normalized variance of spectral centroid
	27	variance of filtered spectral centroid
	28	variance of difference between the filtered spectral centroid and the spectral centroid
	29	Mean value of the difference of the spectral centroid
	30	Mean of the absolute value of the difference of the spectral centroid
	31	Position of peak of FFT to the difference of the spectral centroid
	32	Mean value of the spectral flux
	33	Minimum value of the spectral flux
	34	Maximum value of the spectral flux
	35	Range of the spectral flux
	36	Variance of the spectral flux
	37	Normalized variance of the spectral flux
	38	Mean value of the spectral flux normalized according to energy
	39	Minimum value of the spectral flux normalized according to energy
	40	Maximum value of the spectral flux normalized according to energy
	41	Range of the spectral flux normalized according to energy
	42	Variance of the spectral flux normalized according to energy
	43	Normalized variance of the spectral flux normalized according to energy
	44	Mean value of spectral roll off at 50%
	45	Minimum value of spectral roll off at 50%
	46	Maximum value of spectral roll off at 50%
	47	Range of spectral roll off at 50%
	48	Variance of spectral roll off at 50%
	49	Normalized variance of spectral roll off at 50%
	50	Mean value of spectral roll off at 95%
	51	Minimum value of spectral roll off at 95%
	52	Maximum value of spectral roll off at 95%
	53	Range of spectral roll off at 95%
	54	Variance of spectral roll off at 95%
	55	Normalized variance of spectral roll off at 95%

Octave-based features	56 - 61	Subband energy of the 6 octave-based subbands
	62 -67	Subband energy ratio of the 6 octave-based subbands
	68	Maximum of energy
	69	Minimum of energy
	70	Mean of energy
	71	Median of energy
	72	Minimum of 20% of highest energy
	73	Mean of 20% of highest energy
	74	Median of 20% of highest energy
	75	Minimum of 40% of highest energy
	76	Mean of 40% of highest energy
	77	Median of 40% of highest energy
	78	Minimum of 50% of highest energy
	79	Mean of 50% of highest energy
	80	Median of 50% of highest energy
	81 - 188	Mean value of subband peak of the 6 octave-based subbands
		Minimum value of subband peak of the 6 octave-based subbands
		Maximum value of subband peak of the 6 octave-based subbands
		Range of subband peak of the 6 octave-based subbands
		Variance of subband peak of the 6 octave-based subbands
		Normalized variance of subband peak of the 6 octave-based subbands
		Mean value of subband valley of the 6 octave-based subbands
		Minimum value of subband valley of the 6 octave-based subbands
		Maximum value of subband valley of the 6 octave-based subbands
		Range of subband valley of the 6 octave-based subbands
		Variance of subband valley of the 6 octave-based subbands
		Normalized variance of subband valley of the 6 octave-based subbands
		Mean value of subband average of the 6 octave-based subbands
		Minimum value of subband average of the 6 octave-based subbands
		Maximum value of subband average of the 6 octave-based subbands
		Range of subband average of the 6 octave-based subbands
		Variance of subband average of the 6 octave-based subbands

		Normalized variance of subband average of the 6 octave-based subbands

Annex C

Experimental results on music mood - Result lists on global classifiers

M1=Exuberance, M2=Anxious, M3=Contentment, and M4=Depression.

Table C - 1 Results on global classifier for 8 seconds and 32 seconds

8 seconds							
Operator	Lukasiewicz	Hamacher					Yager
parameter	-	γ=0	γ=0.5	γ=3	γ=5	γ=10	p=1
Rate (%)	20.07	67.99	69.49	63.52	63.35	63.35	20.07
Operator	Yager			Weber-Sugemo			
parameter	p=3	p=5	p=10	λ_T=-0.5	λ_T=0	λ_T=0.5	λ_T=2
Rate (%)	69.32	72.14	63.52	32.84	20.07	20.07	58.04
Operator	Weber-Sugemo		Schweizer & Sklar			Frank	
parameter	λ_T=4	λ_T=5	q=1	q=3	q=5	s=2	s=3
Rate (%)	63.68	69.49	**72.80**	68.99	71.31	32.84	32.84
Operator	Frank					Average	Geometric Average
parameter	s=5	s=8	s=10	s=12	s=15	-	-
Rate (%)	32.84	32.84	32.84	32.84	32.84	72.47	71.14

Best confusion matrix		M1	M2	M3	M4
	M1	76.03	1.65	22.31	0.00
	M2	0.00	87.17	0.00	12.83
	M3	35.71	0.79	62.70	0.79
	M4	0.00	37.87	0.00	62.13

32 seconds			
Operator	Lukasiewicz	Hamacher	Yager

parameter	-	γ=0	γ=0.5	γ=3	γ=5	γ=10	p=1
Rate (%)	29.10	29.10	29.10	46.13	45.51	46.13	29.10
Operator	Yager			Weber-Sugemo			
parameter	p=3	p=5	p=10	λ_T=-0.5	λ_T=0	λ_T=0.5	λ_T=2
Rate (%)	50.46	51.39	29.10	29.10	29.10	41.18	46.44
Operator	Weber-Sugemo		Schweizer & Sklar			Frank	
parameter	λ_T=4	λ_T=5	q=1	q=3	q=5	s=2	s=3
Rate (%)	29.10	29.10	**52.01**	29.10	29.10	29.1	29.10
Operator	Frank					Average	Geometric Average
parameter	s=5	s=8	s=10	s=12	s=15	-	-
Rate (%)	29.10	29.10	29.10	29.10	29.10	29.10	29.10

Best confusion matrix		M1	M2	M3	M4
	M1	32.91	29.11	25.32	12.66
	M2	25.81	59.68	9.68	4.84
	M3	16.67	4.90	48.04	30.39
	M4	6.25	0.00	23.75	70.00

References

[Abe00] Abelin, A., Allwood, J., Cross-linguistic interpretation of emotional prosody. In: Proceedings of the ISCA Workshop on Speech and Emotion, 2000.

[Alm91] Almuallim, H., Dietterich, T. G., Learning with many irrelevant features, Proceedings of the Ninth National Conference on Artificial Intelligence, p 547 – 552, San Jose, CA: AAAI Press, 1991.

[Alt00] Alter, K., Rank, E., Kotz, S. A., Toepel, U., Besson, M., Schirmer, A., Angela D. Friederici, A. D., Accentuation and emotions - two different systems? Proceedings of the ISCA Workshop on Speech and Emotion, 2000.

[Ami00] Amir1, N., Ron, S., Laor, N., Analysis of an emotional speech corpus in Hebrew based on objective criteria, In: Proceedings of the ISCA Workshop on Speech and Emotion, 2000.

[Aro04] Arauzo-Azofra, A., Benitez, J. M., Castro, J. L., A feature set measure based on Relief, Proceedings of the 5th International Conference on Recent Advances in Soft Computing, p 104 – 109, 2004

[Aud05] Audibert, N., Aubergé, V., Rilliard, A., The prosodic dimensions of emotion in speech: the relative weights of parameters, In INTERSPEECH-2005, 525-528, 2005

[Ban96] Banse, R., Sherer, K.R., Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology 70 (3), 1996, 614–636.

[Bar06] Barra, R., Montero, J.M., Macías-Guarasa, J., D'Haro, L.F., San-Segundo, R., Córdoba, R., Prosodic and segmental rubrics in emotion identification, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2006.

[Bat00] Batliner, A., Fischer, K., Huber, R., Spilker, J., Noth, E., Desperately seeking emotions or: actors, wizards, and human beings, Proceedings of the ISCA Workshop on Speech and Emotion, 2000.

[Bel61] Bellman, R.E., Adaptive Control Processes, Princeton University Press, Princeton, NJ, 1961.

[Blu97] Blum, A. and Langley, P. (1997) Selection of relevant features and examples in machine learning. Artificial Intelligence Journal (97). 245-271.

[Bon85] Bonissone, P.P, Selecting uncertainty calculi and granularity: An experiment in trading off precision and complexity, Proceedings of the first Workshop on Uncertainty in Artificial Intelligence, Los Angeles, 57 – 66, 1985.

[Bos00] Bosh, L.T., 2000. Emotions: what is possible in the ASR framework? In: Proceedings of the ISCA Workshop on Speech and Emotion.

[Bre01] Breazeal, C., 2001. Designing Social Robots, MIT Press, Cambridge, MA.

[Bru56] Brunswik, E., 1956. Perception and the Representative Design of Psychological Experiments, University of California Press, Berkeley.

[Bur00] Burkhardt, F., Sendlmeier, W., Verification of acoustical correlates of emotional speech using formant-synthesis, In: Proceedings of the ISCA Workshop on Speech and Emotion, 2000.

[Bur05] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B., A Database of German Emotional Speech, Proceedings Interspeech 2005, Lissabon, Portugal.

[Chu05] Chuan, C. H., Chew, E., Polyphonic audio key finding using the spiral array CEG algorithm, ICME2005.

[Coh97] Cohen, A., Mantegna,R. N., and Havlin, S., Numerical analysis of word frequencies in artificial and natural language texts, Fractals, vol. 5, no. 1, pp. 95–104, 1997.

[Cor00] Cornelius, R., Theoretical approaches to emotion, In: Proceedings of the ISCA Workshop on Speech and Emotion, 2000.

[Cow00] Cowie, R., Describing the Emotional States Expressed in Speech, ISCA Workshop on Speech & Emotion, Northern Ireland 2000, p. 11-18.

[Dar1871] Darwin, C., The Descent of Man, 1871

[Del04] Dellandrea, E., Makris, P., and Vincent, N., Zipf Analysis of Audio Signals, Fractals, World Scientific Publishing Company, vol. 12(1), p. 73-85, 2004.

[Dem67] Dempster, A.P., Upper and lower probabilities induced by a multivalued mapping, Ann. Math. Statistics, 1967.

[Dem68] Dempster, A.P., – A generalization of Bayesian inference. J. Royal Statistical Soc. Series B, vol. 30, 1968.

[Det00] Detyniecki, M., Mathematical Aggregation Operators and their Application to Video Querying, University of Paris 6, France, Doctoral thesis + LIP6 research report 2001/002, November 2000

[Dev05] Devillers, L., Vidrascu, L., Lamel, L., Emotion detection in real-life spoken dialogs recorded in call center, Journal of Neural Networks, special issue on Emotion and Brain, volume 18, No. 4, 407-422, May 2005.

[Dou00] Douglas-Cowie, E., Cowie, R., Schröder, M., A new emotion database: considerations, sources and scope, Proceedings of the ISCA Workshop on Speech and Emotion, 2000.

[Dou06] Dou, W., Segmentation of multispectral images based on information fusion: application for MRI images, thesis of Ph.D., Université de Caen, 2006.

[Dow03] Downie, J.S., Towards the Scientific Evaluation of Music Information Retrieval Systems, Proceedings of the International Symposium on Music Information Retrieval, Baltimore, MD, USA, 2003.

[Dru00] Druin A., Hendler J., Robots for Kids: exploring new technologies for learning, Morgan Kauffman, Los Altos, CA, 2000.

[Dut96] Duroit, T., Pagel, V., Pierret, N., Bataille, F., Van der Vreken, O ., The MBROLA Project : Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes, Proc ICSLP, 3 : 1393-1396, Philadelphia, 1996

[Ekm82] Ekman P., Emotions in the human face, Cambridge University Press, 1982

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

[Eng96] Engberg, I. S., Hansen, A. V., Documentation of the Danish Emotional Speech Database DES, Aalborg September 1996

[Far58] Farnsworth, P. R., The social psychology of music. The Dryden Press, 1958.

[Fer00] Fernandez, R., Picard, R. W, Modeling drivers' speech under stress, Proceedings of the ISCA Workshop on Speech and Emotion, 2000

[Fio04] Fioretti, G., Evidence Theory: A Mathematical Framework for Unpredictable Hypotheses, Metroeconomica, Volume 55, Number 4, November 2004, pp. 345-366(22)

[Ful96] Fuller, R., OWA Operators in Decision Making, C. Carlsson ed,. Exploring the Limits of Support Systems, TUCS General Publications, No.3, Turku Centre for Computer Science, 1996, p 85 - 104

[Grif] Griffith, S., http://www.case.edu/pubs/cnews/2002/1-17/emotion.htm

[Guy03] Guyon, I., Elisseff, A., An introduction to variable and feature selection, Journal of Machine Learning Research 3, p1157 – 1182, 2003.

[Hal97] Hall, M. A., Smith, L. A., Feature Subset Selection: A Correlation Based Filter Approach, International Conference on Neural Information Processing and Intelligent Information Systems, Springer, p855-858, 1997

[Har03a] Harb, H., Chen, L., Robust Speech Music Discrimination Using Spectrum's First Order Statistics and Neural Networks, Proceedings of the IEEE International Symposium on Signal Processing and its Applications ISSPA2003, July 1-4, Paris - France, 2003

[Har03b] Harb, H., Chen, L., Gender Identification Using A General Audio Classifier, Proc. of the IEEE International Conference on Multimedia & Expo ICME 2003, July 6-9, Baltimore, USA, 2003

[Har04] Harb, H., Chen, L., Mixture of experts for audio classification: an application to male female classification and musical genre recognition, in the proceedings of IEEE International Conference on Multimedia and Expo ICME 2004, 2004

[Har05a] Harb, H., Chen, L., A General Audio Semantic Classifier based on human perception motivated mode, Multimedia Tools and Applications, 2005

[Har05b] Harb, H., Auloge, A.J-Y, Chen, L., Voice-based Gender Identification in multimedia applications, Journal of intelligent information systems, 2005

[Hav95] Havlin, S., The distance between Zipf Plots, Physica A216, pp. 148–150, 1995.

[Her04] Hernandez, E., Recasens, J., Indistinguishability relations in Dempster–Shafer theory of evidence, International Journal of Approximate Reasoning, Volume 37, Issue 3, November 2004, Pages 145-187

[Hev35a] Hevner, K., Expression in music: a discussion of experimental studies and theories. Psychological Review, 42, 186-204, 1935

[Hev35b] Hevner, K. The affective character of the major and minor mode in music. American Journal of Psychology, 47, 1935, p103-118.

[Hev36] Hevner, K.: Experimental studies of the elements of expression in music. American Journal of Psychology 48 (1936) p246–268

[Hev37] Hevner, K., The affective value of pitch and tempo in music, American Journal of Psychology, 49(4):621–630, 1937.

[Hin96] Hinn, D. M., The Effect of the Major and Minor Mode in Music as a Mood Induction Procedure, Thesis (M.A.)--Virginia Polytechnic Institute and State University, 1996.

[Hur00] Huron, D., Perceptual and cognitive applications in music information retrieval, In International Symposium on Music Information Retrieval,2000.

[ISMIR] International Conference on Music Information Retrieval

[ISO75] ISO 16: 1975, Acoustics–standard tuning frequency (standard musical pitch).

[Jia02] Jiang, D.N., Lu, L., Zhang, H.J., Tao, J.H., Cai, L.H., Music type classification by spectral contrast feature, Proceedings of IEEE International Conference on Multimedia and Expo, p113-116, vol.1, 2002.

[Joh04] Johnstone, T., Scherer, K. R., Vocal communication, in M. Lewis & J. M. Haviland (Eds.), Handbook of Emotion, second edition, p 220 - 235, 2004

[Kam92] Kamien, R., Music: an appreciation (5th Edition). McGraw-Hill Inc., 1992

[Kie00] Kienast, M., Sendlmeier, W. F., Acoustical analysis of spectral and temporal changes in emotional speech, Proceedings of the ISCA workshop on Speech and Emotion, 2000

[Kil02] Kilian, J., Hoos, H., Voice Separation – A Local Optimization Approach, in Proceedings of the 3rd International Conference on Music Information Retrieval, p 39-46, 2002.

[Kir92] Kira, K., Rendell, L., A practical approach to feature selection, Proceedings of the ninth International Conference on Machine Learning, p. 249-256, Aberdeen, Scotland: Morgan Kaufmann, 1992.

[Kla00] Klasmeyer, G., Johnstone, T., Bänziger, T., Sappok, C., Scherer, K. R., Emotional voice variability in speaker verification, Proceedings of the ISCA workshop on Speech and Emotion, 2000

[Koh95] Kohavi, R., A study of cross-validation and bootstrap for accuracy estimation and model selection, Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, p 1137 – 1143, San Mateo, CA: Morgan Kaufmann, 1995

[Koh97] Kohavi, R., John, G. H., Wrappers for Feature Subset Selection, Artificial Intelligence, Volume 97, Issue 1-2, Special issue on relevance, p273 – 324, 1997

[Koj00] Kojadinovic, I., Wottka, T., Comparison between a filter and a wrapper approach to variable subset selection in regression problems, ESIT 2000, September 14-15, 2000, Aachen, Germany

[Kop00] Kopeček, I., Emotions and prosody in dialogues: an algebraic approach based on user modeling, Proceedings of the ISCA workshop on Speech and Emotion, 2000

[Kor04] Korycinski, D., Crawford, M. M., Barnes, J. W., Adaptive feature selection for hyperspectral data analysis, Image and Signal Processing for Remote Sensing IX. Edited by Bruzzone, Lorenzo, Proceedings of the SPIE, Volume 5238, pp. 213-225, 2004

[Kru02] Krumhansl, C. L., Music: a link between cognition and emotion. Current Directions in Psychological Science, 11(2), p45-50, 2002

[Kuo05] Kuo, F. F., Chiang, M.F., Shan, M.K., and Lee, S.Y., Emotion-based Music Recommendation based on Association Discovery from Film Music, ACM International Conference on Multimedia, Singapore, 2005

[Kus01] 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

[Lem] Leman, M., GOASEMA – Semantic description of musical audio. Retrieved from http://www.ipem.ugent.be/.

[Les03] Lesage, D., Evidence Theory [episode 2] Implementation Issues and Applications, LRDE seminar, May 28, 2003

[Li03a] Li, T., Ogihara, M., Detecting Emotion in Music, ISMIR2003, p239 – 240, 2003

[Li03b] Li, T., Ogihara, M., Li, Q., A comparative study on content-based music genre classification, in SIGIR'03, pp. 282–289, ACM Press, 2003.

[Li04] Li, T., Ogihara, M., Content-Based Music Similarity Search and Emotion Detection. In Proceedings of the ICASSP 2004, p705-708, 2004

[Li07] Li, Y., Wang, D. L., Separation of singing voice from music accompaniment for monaural recordings, IEEE Transactions on Audio, Speech, and Language Processing, Volume 15, Issue 4, p 1475-1487, May, 2007.

[Liu03] Liu, D., Lu, L., Zhang, H., Automatic mood detection from acoustic music data, ISMIR2003

[Lu06] Lu, L., Liu, D., Zhang, H.J., Automatic Mood Detection and Tracking of Music Audio Signals, Special Issue on Statistical and Perceptual Audio Processing, IEEE Trans. on Audio, Speech and Language Processing, Vol. 14, No. 1, pp. 5-18, 2006

[Luk05] Lukasiak, B. M., Faria, R., Zomer, S., Brereton, R. G., Duncan, J. C., Pattern recognition for the analysis of polymeric materials, The Royal Society of Chemistry, 2005

[Mar98] Martin, K. D., Kim, Y. E., Musical instrument identification: A pattern-recognition approach, the 136th meeting of the Acoustical Society of America, October 13, 1998

[Mas00] Massaro, D. W., Multimodal emotion perception: analogous to speech processes, Proceedings of the ISCA workshop on Speech and Emotion, 2000

[McG00] McGilloway, S., Cowie, R., 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.

[Mei04] Meignier, S., Moraru, D., Fredouille, C., Besacier, L., Bonastre, J.-F., Benefit of prior acoustic segmentation for speaker segmentation systems, International Conference on Acoustics Speech & Signal Processing (ICASSP), Montreal, Canada, May 2004

[Men42] Menger, K., Statistical metrics, Proc. Nat. Acad. Sci, U.S.A. 8, p535-537. 1942.

[Mey56] Meyer, L. B, Emotion and Meaning in Music, Chicago: The University of Chicago Press, 1956

[Mey07] Meyers, O. C., A mood-Based music classification and exploration system, thesis of master, MIT, June 2007.

[Moo97] Moore, B. C.J., An Introduction to the Psychology of hearing, Academic Press, 1997

[Nar77] Narendra, P.M., Fukunaga, K., A branch and bound algorithm for feature selection, IEEE Transactions on Computers, C-26(9): 917-922, September 1977

[Nie00] Niedenthal, P. M., Halberstadt, J. B., Margolin, J., Innes-Ker, A. H., Emotional state and the detection of change in facial expression of emotion, European Journal of Social Psychology, Vol. 30, NO. 2, p211-222, 2000

[Nor05] Norowi, N. M., Doraisamy, S. C., Rahmat, R. W.O.K. - Factors affecting Automatic Genre Classification: An Investigation Incorporating Non-Western Musical Forms, ISMIR 2005, p13 – 20, 11 – 15 September 2005

[Nwe03] Nwe, T. L., Foo, S. W., Silva, L.C, Speech emotion recognition using hidden Markov models, Speech Communication 41, p603 – 623, 2003.

[Oud03] 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.

[Pae00] Paeschke, A., Sendlmeier, W. F., Prosodic Characteristics of Emotional Speech: Measurements of Fundamental Frequency Movements, Proceedings of the ISCA workshop on Speech and Emotion, 2000

[Pao04] Pao, T. L., Chen, Y. T., Lu, J. J., Yeh, J. H., The Construction and Testing of a Mandarin Emotional Speech Database, Proceeding ROCLING XVI, pp. 355-363, Sep. 2004

[Par06] Paradzinets, A., Harb, H., Chen, L., Use of Continuous Wavelet-like Transform in Automated Music Transcription, EUSIPCO06, Florence, Italy, 2006

[Par07] Paradzinets, A., Kotov, O., Harb, H., Chen, L., Continuous Wavelet-like Transform Based Music Similarity Features for Intelligent Music Navigtion, CBMI07, Bordeaux, France, 2007

[Per00] Pereira, C., Dimensions of emotional meaning in speech, Proceedings of the ISCA workshop on Speech and Emotion, p 25 – 28, Newcastle, Northern Ireland, 2000

[Pic97] Picard, R., Affective Computing, MIT Press, 1997

[Pig86] Pignatiello, M. F., Camp, C. J., Rasar, L. A., Musical mood induction: An alternative to the Velten technique, Journal of Abnormal Psychology, 95 (3) p 295-297, 1986.

[Pin06] Pinquier, J,. André-Obrecht, R., Audio Indexing: Primary Components Retrieval - Robust Classification in Audio Documents. In: Multimedia Tools and Applications, Springer-Verlag, Vol. 30 N. 3, p. 313-330, September 2006.

[Pio00] Piot, O., Attitudes and yes-no questions in standard French: testing two hypotheses, Proceedings of the ISCA workshop on Speech and Emotion, 2000

[Pit93] Pittam, J., Scherer, K.R., 1993, Vocal expression and communication of emotion, in M. Lewis & J. M. Haviland (Eds.), Handbook of emotions (pp 185-198), New York: Guilford Press

[Plo70] Plomp, R., Timbre as a multidimensional attribute of complex tones, in Frequency Analysis and Periodicity Detection in Hearing, ed. by R. Plomp and G. F. Smocrenburg, A. W. Sijthoff, Leiden, 1970, p397-414.

[Plu80] Plutchik, R., Emotion: A psycho evolutionary synthesis. New York: Harper and Row, 1980

[Pol00] 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.

[Pra01] PRAAT, a system for doing phonetics by computer. Glot International 5(9/10), 341-345, 2001

[Pud94] Pudil, P., Novovicova, J., Kittler, J., Floating search methods in feature selection, Pattern Recognition Letters 15, pp 1119 – 1125, Nov. 1994

[Qua07] Quang, V-M., Besacier, L., Castelli, E., Automatic question detection: prosodic-lexical features and cross lingual experiments, INTERSPEECH 2007. Antwerp. Belgium, August 2007

[Rad88] Radocy, R. E., & Boyle, J. D. (1988). Psychological foundations of musical behavior, Springfield, IL: Charles C Thomas.

[Rak05] 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

[Roa00] Roach, P., Techniques for the phonetic description of emotional speech, Proceedings of the ISCA workshop on Speech and Emotion, 2000

[Rou05] Rouas, J. L., Farinas, J., Pellegrino, F., André-Obrecht, R., Rhythmic unit extraction and modeling for automatic language identification, In : Speech Communication, Elsevier, Vol. 47 N. 4, p. 436-456, 2005.

[Sae07] Saeys, Y., Inza, I., Larranaga, P., A review of feature selection technique in bioinformations, Bioinformatics advance access, August 24, 2007.

[Sch60] Schweizer. B. and Sklar. A., Statistical metric spaces, Pacific J. Math. 10, p313-334, 1960.

[Sch83] Schweizer B. and Sklar A., Probabilistic metric spaces, North Holland, New York, 1983.

[Sch88] Scherer, K.R., Kappas, A., 1988: Primate vocal expression of affective state, in D. Todt, P. Goedeking, & D. Symmes (Eds.), Primate vocal communication (pp. 171-194). Berlin: Springer

[Sch89] Scherer, K. R., Vocal correlates of emotion, in A. Manstead & H. Wagner (Eds.), Handbook of psychophysiology: emotion and social behaviors (pp.165-197). London: Wiley, 1989

[Sch95] Scherer, K.R. Expression of emotion in voice and music, J. Voice, 9(3), pp. 235-248, 1995.

[Sch00a] Scherer, K R.: "A cross-cultural investigation of emotion inferences from voice and speech: implications for speech technology", In ICSLP-2000, vol.2, 379-382, 2000

[Sch00b] 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.

[Sch00c] Scherer, K.R., Psychological models of emotion, In: Borod, J. (Ed.), The neuropsychology of Emotion. Oxford University Press, Oxford/Newyork, p 137 -162

[Sch01] Scherer, K.R., Schorr, A., Johnstone, T., Appraisal Processes in Emotion: Theory, Methods, Research.Oxford University Press, New York and Oxford, 2001.

[Sch02] Scherer, K, R., Vocal communication of emotion: A review of research paradigms, Speech Communication 40, pp. 227-256, 2002

[Seb04] Sebban, M., Nock, R., A Hybrid Filter/Wrapper Approach of Feature Selection using Information Theory, Proceedings of 2004 International Conference on Machine Learning and Cybernetics, p2537- 2542 vol.4, 2004

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

[Sha76] Shafer, G., – A mathematical theory of evidence, Princeton University Press, 1976.

[Sha90] Shafer, G., Axioms for probability and belief-function propagation (with Prakash Shenoy). Uncertainty in Artificial Intelligence 4, pp. 169-198. R. D. Shachter, T. S. Levitt, L. N. Kanal, and J. F. Lemmer, eds. North-Holland, 1990.

[Sha92] Shafer, G., The Dempster-Shafer theory. Pp. 330-331, Encyclopedia of Artificial Intelligence, Second Edition, Stuart C. Shapiro, editor. Wiley. 1992

[Sla98] 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.

[Spe98] 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.

[Ste37] Stevens, S.S., Volkman, J., Newman, E.B., A scale for the measurement of the psychological magnitude of pitch, Journal of the Acoustical Society of America, 8, 1937, pp. 185-190

[Sti00] Stibbard, R., Automated extraction of tobi annotation data from the reading/leeds emotional speech corpus, Proceedings of the ISCA workshop on Speech and Emotion, 2000

[Tel99] Tellegen, A., Watson, D. & Clark, L.A., On the dimensional and hierarchical structure of affect, Psychological Science, Vol. 10, No. 4, July 1999, p 297-303.

[Tel04] Telmoudi, A., Chakhar, S., Data fusion application from evidential databases as a support for decision making, Information and Software Technology, Vol. 46, No 8, pp. 547-555, 2004

[Tha89] Thayer, R. E., The biopsychology of mood and arousal, Oxford University Press, 1989

[Tic00] Tickle, A., 2000. English and Japanese speaker's emotion vocalizations and recognition: a comparison highlighting vowel quality. Proceedings of the ISCA workshop on Speech and Emotion, 2000.

[Tso90] Tsoi, A. C., Pearson, R. A., Comparison of three classification techniques, CART, C4.5 and multi-layer perceptrons, Proceedings of the 1990 conference on Advances in neural information processing systems, p 963 – 969, 1990

[Tza01] Tzanetakis G., Essl G., Cook P. Audio Analysis using the Discrete Wavelet Transform Proc. WSES Int. Conf. Acoustics and Music: Theory 2001 and Applications (AMTA 2001) Skiathos, Greece

[Tza02] Tzanetakis, G., Cook, P., Musical genre classification of audio signals, IEEE Transactions on Speech and Audio Processing, vol. 10, no. 5, pp. 293–298, July 2002.

[Ver00] Véronique, A., Ludovic, L., The prosody of smile, Proceedings of the ISCA workshop on Speech and Emotion, 2000

[Ver04a] Ververidis D. and Kotropoulos C.; Ioannis Pitas, Automatic emotional speech classification, Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004), pp. 593 – 596, 2004, Montreal, Canada.

[Ver04b] Ververidis D. and 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.

[Ver05a] Ververidis, D. Kotropoulos, C, Emotional speech classification using Gaussian mixture models, IEEE International Symposium on Circuits and Systems, 2005. ISCAS 2005, Vol. 3, p2871- 2874

[Ver05b] 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.

[Vra03] Vrabie, V., Granjon, P., Serviere, C., Spectral kurtosis: from definition to application, 6th IEEE International Workshop on Nonlinear Signal and Image Processing (NSIP 2003), Grado-Trieste, Italy, 2003

[Whi89] Whissell, C. The dictionary of affect in language, In R. Plutchnik & Kellerman, H. (eds), Emotion: Theory, research and experience: vol 4, the measurement of emotions. New York: Acdemic Press, 1989

[Wic00] Wichmann, A., The attitudinal effects of prosody, and how they relate to emotion, Proceedings of the ISCA workshop on Speech and Emotion, 2000

[Wie05a] Wieczorkowska, Alicja., 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

[Wie05b] Wieczorkowska, A., Synak, P., Lewis, R. A., Ras, Z. W., Creating Reliable Database for Experiments on Extracting Emotions from Music, IIS05, Poland, 2005, p395-402

[Wiki] http://en.wikipedia.org

[Wit05] Witten, I.H., Frank, E., "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005

[Xia05] Xiao, Z., Dellandréa, 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.

[Xia06] Xiao, Z., Dellandréa, 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.

[Xia07a] Xiao, Z., Dellandréa, E., Dou, W., Chen, L., Hierarchical Classification of Emotional Speech, research report RR-LIRIS-2007-006, LIRIS UMR 5205 CNRS, 2007

[Xia07b] Xiao, Z., Dellandrea, E., Dou, W., Chen, L., Automatic Hierarchical Classification of Emotional Speech, ismw, pp. 291-296, Ninth IEEE International Symposium on Multimedia Workshops (ISMW 2007), 2007

[Xia07c] Xiao, Z., Dellandréa, E., Dou, W., Chen, L., A Dimensional Emotion Model Driven Multi-stage Classification of Emotional Speech, research report RR-LIRIS-2007-033, LIRIS UMR 5205 CNRS, 2007, submitted to IEEE Transactions on Audio, Speech, and Language Processing

[Yan04] Yang, D., and Lee, W., Disambiguating Music Emotion Using Software Agents, ISMIR, 2004

[Zhu07] Zhu, A., Luo, Q., Study on Speech Emotion Recognition System in E-Learning, HCI International 2007 Part III, China, LNCS 4552, p544 – 552, 22-27 July, Beijing

[Zip49] Zipf, G. K., Human Behavior and the Principle of Least Effort. Addison-Wesley Press, 1949

Publications

During this work, 3 papers have been published in international conferences, 1 paper has been submitted to IEEE Trans. on Audio, Speech, and Language Processing, and two research reports are published in the laboratory LIRIS.

International Conferences

1. Zhongzhe Xiao, Emmanuel Dellandrea, Weibei Dou, Liming Chen, "Automatic Hierarchical Classification of Emotional Speech," ismw, pp. 291-296, Ninth IEEE International Symposium on Multimedia Workshops (ISMW 2007), 2007
2. 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.
3. 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.
4. Zhongzhe Xiao and Zaiwang Dong, Improved GIB Synchronization Method for OFDM Systems, 10th International Conference on Telecommunications ICT'2003, pp 1417-1421, Volume II, February 23-28, 2003, Tahiti, Papeete – French Polynesia.

Submission to International Journal

1. Zhongzhe Xiao, Emmanuel Dellandréa, Weibei Dou, Liming Chen, A Dimensional Emotion Model Driven Multi-stage Classification of Emotional Speech, submitted to IEEE Transactions on Audio, Speech, and Language Processing

Research Report

1. Zhongzhe Xiao, Emmanuel Dellandréa, Weibei Dou, Liming Chen, A Dimensional Emotion Model Driven Multi-stage Classification of Emotional Speech, research report RR-LIRIS-2007-033, LIRIS UMR 5205 CNRS, 2007, submitted to IEEE Transactions on Audio, Speech, and Language Processing

2. Zhongzhe Xiao, Emmanuel Dellandréa, Weibei Dou, Liming Chen, Hierarchical Classification of Emotional Speech, research report RR-LIRIS-2007-006, LIRIS UMR 5205 CNRS, 2007

List of figures

List of tables