Information Processing and Management 45 (2009) 315–328

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

# Acoustic feature selection for automatic emotion recognition from speech

Jia Rong, Gang Li *, Yi-Ping Phoebe Chen *

*School of Engineering and Information Technology, Deakin University, 221 Burwood Highway, Melbourne, VIC 3125, Australia*

ARTICLE INFO

*Article history:*
Received 30 May 2008
Received in revised form 18 September 2008
Accepted 18 September 2008
Available online 31 October 2008

*PACS:*
43.72.Ne
43.71.Bp
43.71.Ft

*Keywords:*
Emotion recognition
Feature selection
Machine learning

ABSTRACT

Emotional expression and understanding are normal instincts of human beings, but automatical emotion recognition from speech without referring any language or linguistic information remains an unclosed problem. The limited size of existing emotional data samples, and the relative higher dimensionality have outstripped many dimensionality reduction and feature selection algorithms. This paper focuses on the data preprocessing techniques which aim to extract the most effective acoustic features to improve the performance of the emotion recognition. A novel algorithm is presented in this paper, which can be applied on a small sized data set with a high number of features. The presented algorithm integrates the advantages from a decision tree method and the random forest ensemble. Experiment results on a series of Chinese emotional speech data sets indicate that the presented algorithm can achieve improved results on emotional recognition, and outperform the commonly used Principle Component Analysis (PCA)/Multi-Dimensional Scaling (MDS) methods, and the more recently developed ISOMap dimensionality reduction method.



## 1. Introduction

Emotional recognition is a common instinct for human beings, which has been studied by researchers from different disciplines for more than 70 years (Fairbanks & Pronovost, 1939, 1941). Fairbanks et al's pioneering work on emotional speech (Fairbanks & Pronovost, 1939, 1941) revealed the importance of vocal cues in the expression of emotion, and the powerful effects of vocal emotion expression on interpersonal interaction. Understanding the emotional state of the speaker during communication can help the listeners to catch more information than is represented by the content of the dialogue sentences, especially to detect the 'real' meaning of the speech hidden between words. The practical value of emotion recognition from speech is suggested by the rapidly growing number of areas to which it is being applied, such as humanoid robots, the car industry, calling centers, etc. (Cowie & Cornelius, 2003; Lee & Narayanan, 2004; Lee et al., 2004; Pantic & Rothkrantz, 2003; Schuller et al., 2005).

Although machine learning and data mining techniques have obtained flourishing applications (Mitchell, 1997), only a few works have utilized these powerful tools and achieved better performance in emotion recognition from speech. Here a serious encumbrance is the lack of available emotional speech data. There is only few public benchmark databases available for research purpose.

A sufficient number of training examples is the premise for most machine learning and data mining algorithms to work well. When there are only a few training examples, it is very possible to have the problem of *overfitting*, which means that a model can be trained with perfect performance on the training set, but can hardly generalize well on new examples. In prac-

* Corresponding authors.
*E-mail addresses:* jrong@deakin.edu.au (J. Rong), gang.li@deakin.edu.au (G. Li), phoebe@deakin.edu.au (Yi-Ping Phoebe Chen).

0306-4573/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2008.09.003

tise, how many training examples will be adequate is task-dependent; for example, the task of learning the XOR function, four different training examples are sufficient, while for more complex tasks such as emotion recognition, thousands of training examples might still be insufficient. In general, if a data set can not fully cover the whole variable space then the data set is referred to as a *small data set*. In this sense, the data sets collected for emotion recognition are small because the typical size of the data set is less than 1000, while the number of features is close to 100. Such kind of data scarcity usually outstrips many machine learning and data mining algorithms (Vapnik, 1995).

There are two obvious ways to overcome the problem of data scarcity: one is to collect more data while the second is to design techniques that can deal with small data sets. Considering the fact that further data collection is manual, cost intensive and hard to achieve, it is more feasible and desirable for the second way. Based on this recognition, this paper presents a novel feature selection algorithm, *ERFTrees*, to extract effective features from small data sets. There are two facets of benefits by using this algorithm for emotion recognition: firstly, the irrelevant data can be removed and the dimensionality of the training data can be reduced; secondly, with a reduced data set, most existing machine learning algorithms which do not work well on small data set, can now produce better recognition accuracy. The empirical results on Chinese (Mandarin) emotional data sets indicate that the presented algorithm outperforms other linear and non-linear dimensionality reduction methods, including *Principle Component Analysis* (PCA), *Multi-Dimensional Scaling* (MDS), and ISOMap.

The rest of the paper is organized as follows: we introduce the background and the related work in Section 2. The algorithm, *ERFTrees*, is presented in Section 3. The experiment design and empirical results are presented in Section 4, and finally in Section 5, we conclude the paper with a perspective analysis of possible future work.

## 2. Background and related work

### 2.1. Theory of human emotions

Constructing an automatic emotion recognizer depends on a sense of what emotion is. Most people have an informal understanding, but there is a formal research tradition which has probed the nature of emotion systematically. It has been shaped by major figures in several disciplines – philosophy, biology, and psychology – but conventionally it is called the 'psychological tradition'.

In psychology, the theories of emotion are grouped into four main traditions, each making different basic assumptions about what is central to the nature of emotion. As summarized by Cornelius (1996), there are four perspectives focusing on different aspects of emotions: *Darwinians* are interested in the evolutionary organization of emotion; *Jamesians* are interested in the bodily organization of emotion; *Cognitive*-emotion theorists are interested in the psychological organization of emotion; and *social constructivists* are interested in the social-psychological and sociological organization of emotion.

In the field of emotion recognition from speech, different research groups usually use different emotion states, as shown in Table 1. Considering the general agreement in the *Darwinian* and the *Jamesian* traditions of emotion research that some full-blown emotions are more foundational than others, it might be more desirable to focus on those foundational emotions. Based on the study of Russell (1980) and Banse and Scherer (1996), the most foundational emotions are defined as follows:

*Anger*: Anger is often a response to the perception of threat due to a physical conflict, injustice, negligence, humiliation, or betrayal. The state of anger includes emotional states such as tense, alarmed, angry, afraid, annoyed, mad and so on. There are two common types of anger: active and passive. In this paper, we focus on the 'active' anger due to its stronger characteristics.

**Table 1**
Emotional states used in emotional speech recognition

| Research group | Emotion states |
|---|---|
| Mozziaconacci (1995) | Joy, anger, fear, sadness, boredom, indignation, neutral |
| Klasmeyer and Sendlneier (1995) | Happiness, anger, fear, sadness, boredom, disgust, neutral |
| McGilloway et al. (1995) | Happiness, anger, fear, sadness |
| Nicholson et al. (1999) | Joy, anger, fear, sadness, teasing, disgust, surprise, neutral |
| Nwe et al. (2001) | Anger, fear, sadness, dislike, disgust |
| Schuller et al. (2003) | Joy, anger, fear, sadness, disgust, surprise, neutral |
| Kwon et al. (2003) | Happiness, anger, sadness, bored, neutral |
| Cai et al. (2003) | Happiness, anger, sadness, surprise |
| Park et al. (2003) | Anger, laugh, surprise, neutral |
| Litman and Forbes-Reley (2003) | Negative, positive, neutral |
| Song et al. (2004) | Joy, anger, fear, sadness, disgust, surprise, neutral |
| Bhatti et al. (2004) | Happiness, anger, fear, sadness, surprise, disgust |
| Hyun et al. (2005) | Joy, anger, sadness, neutrality |
| Lee et al. (2004) | Negative, positive, neutral |

*Happiness*: Happiness is an emotional or affective state that is characterized by feelings of enjoyment, pleasure and satisfaction. The state of happiness includes well-being, delight, excitement, inner peace, health, safety, contentment, love and so on.

*Fear*: Fear is an emotional state that is expressed when people is feeling danger, which is a natural response to a particular negative stimulus. Fear is related to the emotional states including worry, anxiety, terror, fright, paranoia, horror, panic, persecution complex and dread.

*Sadness*: Sadness is a mood that displays feeling of disadvantage and loss. Sadness is considered as an opposite feeling to happiness. The state of sadness includes emotional states such as sad, sorry, miserable, gloomy, depressed, bored, droopy, tired and so on.

*Neutrality*: Neutrality is a nonemotional state, including sleepiness, calmness, relaxation, satisfaction, content, and so on.

In this work, we are going to use the above five emotional states for emotion recognition from speech.

### 2.2. A machine learning framework for emotion recognition

Depending on the understanding of various researchers, different emotion recognition projects usually adopt different methods. However, most existing work on emotion recognition can be summarized into the following general procedure, as shown in Fig. 1:

*Feature extraction stage* to extract the whole acoustic feature set from the original speech corpus and transform these features into an appropriate format for further processing.

*Data preprocessing stage* to select the most relevant subset of the whole candidate feature set or reduce the size of the speech data set into fewer dimensions.

*Emotion recognition stage* to apply machine learning methods on the processed speech data set from the previous stage to recognize the emotional states in speech.

#### 2.2.1. Feature extraction

In the feature extraction stage, the raw speech wave data examples are preprocessed to extract the basic acoustic or linguistic features, such as *pitch*-related, *intensity*-related, *duration*-related, *spectral*-related or *contour*-related, *tone*-based and/or *vowel*-related features. In addition, some transform functions are often employed to convert the speech features between different data domains, such time and frequency domains. The result of this stage is a speech data set represented by a set of high-dimensional speech features.

Defining emotional states in psychology is complex, it is also not easy to give a universal and effective criteria to guide the selection of features (see Table 2). From Table 2, some of the research groups used only the raw acoustic features extracted directly from the pure speech signals, such as *pitch*, *intensity* and *duration* parameters without further processing applied. Other researchers, like Kwon, Chan, Hao, and Lee (2003) tried to do more. They transformed an original speech wave from the time-domain scale to the frequency-domain scale, in order to highlight the potential change trend (Lee et al., 2004; Song, Bu, Chen, & Li, 2004).

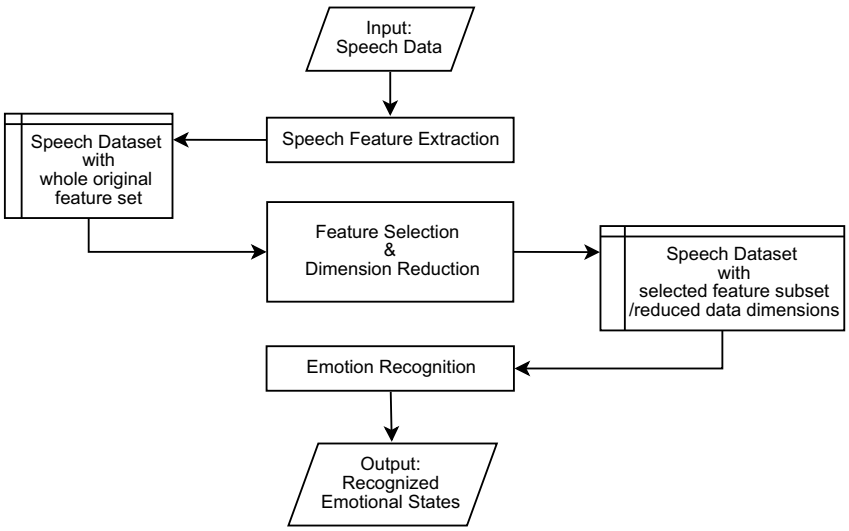


**Fig. 1.** A machine learning framework for emotion recognition from speech.

**Table 2**
Feature vectors used in emotional speech recognition

| Study group | Feature vector |
|---|---|
| Dellaert et al. (1996) | Pitch, rhythm, voiced parts, slope |
| Petrushin (2004) | Pitch, bandwidth, energy, duration, formant |
| Amir (2001) | Pitch, intensity, duration |
| Kwon et al. (2003) | Pitch, log energy, formant, MFCC |
| Schuller et al. (2003) | Pitch, energy |
| Cai et al. (2003) | Duration, pitch, formant, power |
| Litman and Forbes-Reley (2003) | Pitch, energy, temporal, turn |
| Lee et al. (2001), Lee and Narayanan (2003) | Pitch, energy, duration, formant |

Table 3 gives a summary of the previous study of acoustic features which have been used to encode emotional states by psychologists and human behavior biologists since 1930's. The correlation between the speech signals and the archetypal states of the four emotions can also be found in the table.

Although there are many unsolved problems in psychology, Table 3 is still widely used as the most important theoretical foundation by a large number of computing-background research groups who have strenuously engaged with the emotional speech recognition in the past decade. The basic acoustic features were the primary choices in the early days. Most of the feature vectors were composed with the simple extracted *pitch*-related, *intensity*-related, and *duration*-related attributes, such as the maximum, minimum, median, range, and variability values (Amir, 2001; Dellaert, Polzin, & Waibel, 1996; Lee, Narayanan, & Pieraccini, 2001; Petrushin, 2004). Other groups preferred to use the pre-processed attributes from different mathematic transforms, such as *Linear Prediction Cepstral Coefficients (LPCC)* (Song et al., 2004), *Log Frequency Power Coefficients (LFPC)* (Song et al., 2004) and *Mel-frequency Cepstral Coefficients (MFCC)* (Lee et al., 2004; Song et al., 2004). However, it is evident that there are some contradictory results in the existing work, for example, the increase change of the pitch mean value (see '*mean* ↑' in Table 3) may cause a negative emotion in some cases (Coleman & Williams, 1979; Williams & Stevens, 1969), while it can also cause a positive emotion in others (Bezooijen, 1984).

### 2.2.2. Data preprocessing

Unlike the previous stage which extracts the related speech features with potential value for emotion recognition, the second stage aims to reduce the size of the speech data set by selecting the most relevant subset of features and removing the irrelevant ones, or by generating few new features that contains most of the valuable speech information. The problem of selecting a subset of relevant features in a potentially overwhelming quantity of data is classic and found in many branches of science (Wolf & Shashua, 2005). Feature selection, as a preprocessing step to machine learning, has been proven very effective in reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving comprehensibility

**Table 3**
Emotion states and acoustic features

| Features | Anger | Happiness | Fear | Sad |
|---|---|---|---|---|
| Pitch | Mean ↑ Davitz (1964), Fónagy (1978), Williams and Stevens (1969), | Mean ↑ Davitz, 1964, Fónagy (1978), Bezooijen (1984), Öster and Risberg (1996), | Mean ↑ Fónagy (1978), Coleman and Williams (1979), | Below normal mean |
| | Range ↑ Fairbanks and Pronovost (1939), Williams and Stevens (1969), Höffe (1960), | Range ↑ Havrdova and Moravek (1979), Fónagy and Magdics (1963), Sedlacek and Syhra (1963), | Range ↑ Williams and Stevens (1969), | Williams and Stevens (1969), Sedlacek and Syhra (1963), Coleman and Williams (1979), |
| | Variability ↑Fairbanks and Pronovost (1939), Havrdova and Moravek (1979), | Variability ↑ Havrdova and Moravek (1979), Sedlacek and Syhra (1963) | Variability ↑ Williams and Stevens (1969), | Below normal range |
| | Median ↑ Fairbanks and Pronovost (1939) | | Perturbation ↑ Williams and Stevens (1969), Sulc (1977) | Fónagy (1978), Williams and Stevens (1969) |
| Intensity | ↑ Davitz (1964), Williams and Stevens (1969), Bezooijen (1984), Kotlyar and Mozorov (1976) | ↑ Bezooijen (1984), Huttar (1967) | Normal | ↓ Davitz (1964), Muller (1960) |
| Duration | High rate Davitz (1964), Muller (1960), Fónagy (1978) | Rate ↑ Davitz (1964), Coleman and Williams (1979), | ↑ Kotlyar and Mozorov (1976) Coleman and Williams (1979) | Slightly slow Fónagy (1978), Johnson et al. (1986), |
| | Rate ↓ Williams and Stevens (1972), Sulc (1977) | Slow tempo Bezooijen (1984) | Rate ↓ Sulc (1977) | Long pitch falls Davitz (1964) |
| Spectral | High midpoint for av spectrum for nonfric portions McGilloway et al. (1995) | High-frequency energy ↑ Bezooijen (1984), Kaiser (1962) | High-frequency energy ↑ | High-frequency energy ↓ Kaiser (1962), Hargreaves et al. (1965) |

(Liu, Motoda, & Yu, 2002; Talavera, 1999). Dellaert et al. (1996) used *Promising First Selection (PFS)* and *Forward Selection (FS)* method in 1996 and combined them with *k-Nearest Neighbor* (*k-NN*) classifiers in their experiments. Another method called *Sequential Forward Selection (SFS)* was employed by Bhatti, Wang, and Guan (2004), whom achieved improved results than prior work without a feature selection process.

Another category of data preprocessing methods is dimensionality reduction, this contains linear methods such as *Principle Component Analysis (PCA)* and *Multidimensional Scaling (MDS)* (Lattin, Carroll, & Green, 2003), and non-linear methods such as *Non-linear ICA* (Hyvärinen, 1999), *ISOMap* (Tenenbaum, de Silva, & Langford, 2000) and *Self-Organizing Map (SOM)* (Kohavi & John, 1997).

#### 2.2.3. Emotion recognition as a classification problem

The last stage in this procedure is to train and build a classification model using machine learning algorithms to predict the emotional states in the speech data instances. The key task of this stage is to choose an appropriate method which can provide an accurate predicted result for emotion recognition efficiently.

In 1990s, most of the emotion recognition models are based on the simple *Maximum Likelihood Bayes algorithm (MLB)* (Dellaert et al., 1996; Han & Kamber, 2000) and *Linear Discriminant Classification (LDC)* (Dellaert et al., 1996). *Neural Network (NN)* classification methods, which became popular around 2000, is also a good choice for emotion recognition applications (Nicholson, Takahashi, & Nakatsu, 1999; Petrushin, 2004; Park, Wook Lee, & Sim, 2002; Park et al., 2003). After 2002, more attention was turned to *Support Vector Machine (SVM)* (Chuang & Wu, 2004; Hoch, Althoff, McGlaun, & Rigoll, 2005; Lee, Narayanan, & Pieraccini, 2002; Lee et al., 2004; Schuller, Rigoll, & Lang, 2004) and *Hidden Markov Model (HMM)* (Inanoglu & Caneel, 2005; Shafran, Riley, & Mohri, 2003; Schuller, Rigoll, & Lang, 2003; Song, Chen, & You, 2004; Song et al., 2004). Each classification method has its advantages and shortfalls, and the researchers are still working for the better ones. No matter which classification model to be used in this stage, the main purpose is to analyze the prepared data set produced in the previous two stages, and to find out valuable patterns that can predict the speech data instance to a certain emotional state accurately.

### 2.3. Summary

Machine learning is an important method to achieve automatic emotion recognition from speech. The performance is highly related to the training data set and data preprocessing.

However in the area of emotion recognition from speech, as the size of training data is relatively small considering the number of features, existing work is not as efficient as expected.

(1) Most attentions were focused on how to improve the accuracy of emotional speech recognition by building a better classification model. Little work has been done to provide a clear summary of which feature subset will be the most effective for a classification model.
(2) Many feature selection algorithms in machine learning, such as *PFS*, *FS* (Dellaert et al., 1996) and *SFS* (Bhatti et al., 2004), need to be trained from a huge data set that tried to cover adequate samples in real-life. However, due to the difficulties in collecting emotional speech samples, the available training data set does not satisfy this requirement. Therefore, a new method is expected to handle those small data sets.
(3) *PCA* and *MDS* are widely used dimensionality reduction methods in the emotion recognition research, both of which are linear methods. Whether some more recent non-linear development like ISOmap could be used to further improve performance remains an unclosed problem.

In this paper, we aim to develop solutions for application with small data sets to produce better data pre-processing results for the future classification process.

## 3. The feature selection algorithm: *ERFTrees*

From the last section, we know that a data preparation step is important for the performance of emotion recognition algorithms. However, the small size of the emotion data samples, usually with tens of dimensions, has outstripped the capability of many existing feature selection algorithms which requires adequate samples. To address this challenge, a novel method, called *Ensemble Random Forest to Trees (ERFTrees)*, is introduced to do the feature selection task by integrating the random forest ensemble and simple decision tree algorithm together. The structure of *ERFTrees* model is shown in Fig. 2.

The *Ensemble Random Forest to Trees (ERFTrees)* model consists of two major components: *feature selection* and *voting* strategy. The feature selection strategy uses two algorithms (*C4.5* Decision Tree and *Random Forest* ensemble) to select two subsets of the candidate features from the original feature set, while the voting strategy uses *voting-by-majority* method to combine these two subsets of candidate features to work out the final output of feature selection model.

### 3.1. Feature selection strategy

In feature selection process, the training data set is supplied into two learners separately. The first one is based on a single decision tree algorithm, whose output is a single decision tree with flow-chart-like structure. Each node in this tree denotes a

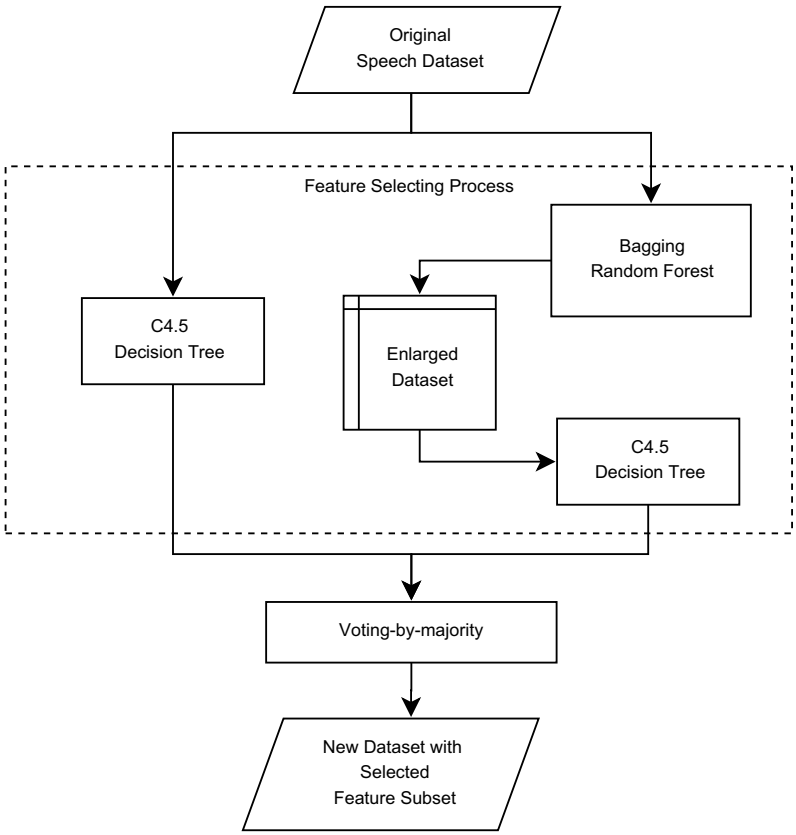


**Fig. 2.** Ensemble random forest to trees model *(ERFTrees)*.

test on a variable, each branch represents an outcome of the test, and leaf nodes represent classes. The decision tree will select the 'best' attributes that contain most of the available information.

Suppose the training data set $X$ contains $n$ instances, and each instance has $m$ variables (features). Input $X$ into a *C4.5* decision tree classifier to grow a decision tree, which returns a set of selected features $F_{DT}$. However, this selected feature subset $F_{DT}$ can not be considered as the output of our feature selection model, because the decision tree algorithm has some weakness. For example, if the training data set is too small, then overfitting problem may happen when single decision tree algorithm is applied. Unfortunately, the training data set used in this study is a small data set with high-dimensional speech examples, and due to the limitation of the resource available, it is not possible to collect more instances for the experiment. Therefore, other methods need to be involved to solve this problem. Our focus is centralized on the random forest ensemble as described in *RF2TREE* method (Zhou, 2004).

In random forest ensemble, a set of base learners generated by the *RandomForests* (Breiman, 2001) algorithm will be trained using the *'Bagging'* ensemble strategy to avoid the overfitting problem caused by the small training data set. This ensemble is used to enlarge the training data set through randomly generating new virtual instances, classifying them by the ensemble and putting these new instances into the original training data set. The generated random forest consists of many decision trees, and each tree is grown in the same way: suppose the number of training examples is $n$, each example has $m$ variables; and then randomly sample $n$ examples with a replacement from the original training data set; specify a number $l \ll m$ so that at each tree node, $l$ variables are randomly selected out of the $m$ variables and the best one of these $l$ variables is used to split the node; therefore, each tree is grown to the largest extent possible, and the new instances are generated in a style that all the possible values of different variables that have not appeared in the original training data set are tried. Finally, a decision tree is grown over the enlarged training data set, which returns the feature subset selected by the random forest ensemble. Then, each of the selected candidate features will be ranked by a *voting-by-majority* strategy.

### 3.2. Voting strategy

The voting strategy is used to combine the multiple results obtained from the feature selection strategy and determine which subset of the features will be selected as the final output of the feature selection model. Here, we use the *voting-by-majority* method to complete the voting task.

**Algorithm: Ensemble Random Forest to Trees.**
**Input:**
- Training dataset, $X$, which contains $n$ training instances with $m$ variables;
- $Generate\_decison\_tree$, Decision Tree algorithm
- $Generate\_random\_forest$, RandomForests algorithm
- $Voting\_by\_majority$, a voting procedure

**Output:** A set of selected variables.
**Method:**
(1) generate a decision tree, $DT$, from the training dataset $X$ by $Generate\_decision\_tree(X)$;
(2) let $F$ be the set of the selected variables in the nodes of $DT$;
(3) create a set of base learners, $L = \{l_1, l_2, \ldots, l_s\}$;
(4) **for each** learner $l_i$
(5) generate a random forest, $RF$, from the training dataset $X$ by $Generate\_random\_forest(X)$;
(6) **for each** training instance $x_i \in X$
(7) select $k$ variables from $m$ variables randomly to generate a set of new virtual instances, $x'_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,j}\}$ as more as possible;
(8) label each instance $x_{i,j}$ with the majority class that is classified by most number of trees in the forest;
(9) let $X'$ be the enlarged training dataset with new generated virtual instances;
(10) add the labelled instance $x_{i,j}$ into $X'$;
**end for**
**end for**
(11) apply $Generate\_decision\_tree(X')$ to grow a decision tree, $DT'$, from the enlarged training dataset;
(12) let $F'$ be the set of the selected variables in the nodes of $DT'$;
(13) combine $F$ and $F'$ using $Voting\_by\_majority(F, F')$ to get the final output set of the selected data variables $F*$
(14) return $F*$

**Fig. 3.** Ensemble random forest to trees (ERFTrees) algorithm.

In the *voting-by-majority*, each learner has the same priority or importance, and it will contribute one voice to the candidate features. The voting result is determined by the number of votes received and the features with the majority vote. For example, in this case, we have two learners: a single *C4.5* decision tree and a random forest ensemble. Let $F_{\mathrm{DT}}$ be the resulting subset of features, which is selected by the single decision tree algorithm, and $F_{\mathrm{ELRF}}$ be the resulting subset that is selected by the random forest ensemble. If one candidate feature $f_i$ appears in both $F_{\mathrm{DT}}$ and $F_{\mathrm{ELRF}}$, that is, $f_i$ gets two voices ($v_i = 2$), then it has a higher chance to be chosen as one of the final selected features over other candidate features that have one voice ($v_j = 1$) or even no voice ($v_k = 0$). The pseudo-code of *ERFTrees* algorithm is shown in Fig. 3. The final output of *ERFTrees* is a set of selected variables, these are considered as the most efficient features for emotion recognition in this study.

### 3.3. Justification

Suppose we denote $X$ as the input space and the set of the class labels is $Y$ specially, then our target is to work out a function: $F{:}X \to Y$. Let $F_T$ denote the function implemented by a decision tree trained on a given training data set, and its probability to approach $F$ can be expressed as

$$P_{F_T} = P_{F=F_T} = 1 - P_{F\neq F_T} = 1 - \mathrm{err}_T, \tag{1}$$

where $\mathrm{err}_T$ denotes the error rate of the decision tree. In the same way described in (Zhou, 2004), $\mathrm{err}_T$ can be broken into three parts: $\mathrm{err}_T^c$, $\mathrm{err}_T^n$ and $\mathrm{err}_T^s$.

$$\mathrm{err}_T = \mathrm{err}_T^c + \mathrm{err}_T^n + \mathrm{err}_T^s, \tag{2}$$

$\mathrm{err}_T^c$ is an error term caused by the limited learning ability of the decision tree; $\mathrm{err}_T^n$ is the an error term caused by the noise contained in the training data set; while $\mathrm{err}_T^s$ is an error term caused by the limitation of the finite samples.

Since $\mathrm{err}_T^c$ can be extremely small, and the noise can be removed by data pre-processing, it is obvious that the performance of a decision tree is usually restricted by the training data set that may not contain a sufficient amount of the data samples to capture the target distribution. That is, $\mathrm{err}_T$ can be dominated by $\mathrm{err}_T^s$ principally. On the other hand, we can also obtain the probability to approach the function $F_E$ implemented by a random forest ensemble trained on the same training data set with the following equations:

$$P_{F_E} = P_{F=F_E} = 1 - P_{F\neq F_E} = 1 - \mathrm{err}_E, \tag{3}$$

$$\mathrm{err}_E = \mathrm{err}_E^c + \mathrm{err}_E^n + \mathrm{err}_E^s. \tag{4}$$

Unlike the simple decision tree, the error term caused by limitation of the finite samples, $\mathrm{err}_E^s$ is much smaller than $\mathrm{err}_T^s$. Due to the fact that the original training data set is enlarged by using a random forest ensemble, $\mathrm{err}_E^s$ is decreased at any rate.

However, the error rate caused by the limited learning ability may be enhanced in the generated training data set, which does not contain all possible feature vectors. That is, assuming the noise error rate can be ignored, $\mathrm{err}_E$ is not only dominated by $\mathrm{err}_E^s$, but also by $\mathrm{err}_E^c$.

If we use $F_{TE}$ as the function implemented by the combined model for both *RF2TREE* and decision tree, then the probability to approach $F_{TE}$ can be expressed as

$$P_{F_{TE}} = P_{F=F_{TE}} = 1 - P_{F\neq F_{TE}} = 1 - \mathrm{err}_{TE}, \tag{5}$$

where

$$\mathrm{err}_{TE}^{*} = \mathrm{err}_{TE}^{c*} + \mathrm{err}_{TE}^{s*}. \tag{6}$$

According to the above justification, $P_{F_{TE}}$ can be greater than either $P_{F_T}$ or $P_{F_E}$ so far as the following equations are satisfied:

$$\mathrm{err}_{T}^{s*} < \mathrm{err}_{T}^{s}, \tag{7}$$

$$\mathrm{err}_{TE}^{c*} < \mathrm{err}_{E}^{c}. \tag{8}$$

Therefore, it shows that the performance can be improved if we combine a decision tree with the random forest ensemble method used in *RF2TREE* together. The experiments described in the next section also verify this.

## 4. Experiment design and result analysis

In this section, we evaluate the performance of the presented feature selection algorithm against other common methods. According to the common sources of collecting emotional speech data, the data used in this work contains two speech corpora from different sources: (1) acted speech corpora and (2) natural speech corpora. The language spoken in both speech corpus is Chinese (Mandarin).

### 4.1. Data sets and original acoustic features

The first data corpus contains speech examples with acted emotions which are expressed by a group of well-selected actors; while the second speech corpus contains speech examples with natural emotions recorded from the daily dialogues between humans in the real life. Table 4 summarizes the key facts about the data sets used in the experiment. In this table, these 9 data sets are divided into three groups: (1) acted data sets (data sets 1–3); (2) natural data sets (data sets 4–6); and (3) overall data sets contains both acted and natural data sets (data sets 7–9). Each group has three individual data sets (female, male and both) which are separated based on the gender of the speakers, and the purpose is to test the importance of gender-dependency in the emotion recognition task.

No matter what features were used in the previous work by any research group, they are all derived from a set of basic acoustic features. The basic acoustic features are those features extracted directly from raw speech signals, and this is in accordance with most studies in this field. In our experiment, we use all 32 basic acoustic features, plus 52 transformed features using *Discrete Fourier Transform* (DFT) and *Mel-scale Frequency Cepstral Coefficients* (MFCC). Table 5 gives the list of all 84 features, which will be used to represent the speech data samples in all 9 data sets as shown in Table 4.

### 4.2. Evaluation of feature selection results

Following the three-stage framework as shown in Fig. 1, once the original speech data examples are represented by the 84 acoustic features from the first stage, then the *ERFTrees* algorithm is then used in the *Data Preprocessing Stage*. For convenience of discussion, in the last *Emotion Recognition Stage*, two typical classification methods: *C4.5* Decision Tree algorithm and *Random Forest* algorithm, are applied to evaluate the quality of the feature subset that are selected by *ERFTrees* model, and the classification results on the selected subset will be compared with that on the original 84 feature set. Due to the

**Table 4**
Data sets used in the experiments

| Data set | Source | Gender | ♯Sent. | Emotion states | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | Angry | Happy | Fear | Sad | Neutral |
| 1 | Acted | Female | 567 | 115 | 117 | 112 | 113 | 110 |
| 2 | Acted | Male | 318 | 67 | 67 | 62 | 63 | 59 |
| 3 | Acted | Both | 885 | 182 | 184 | 174 | 176 | 169 |
| 4 | Natural | Female | 312 | 200 | 12 | 22 | 78 | – |
| 5 | Natural | Male | 178 | 91 | 66 | 3 | 18 | – |
| 6 | Natural | Both | 490 | 291 | 78 | 25 | 96 | – |
| 7 | Both | Female | 879 | 315 | 129 | 134 | 191 | 110 |
| 8 | Both | Male | 496 | 158 | 65 | 133 | 81 | 59 |
| 9 | Both | Both | 1375 | 473 | 194 | 267 | 272 | 169 |

**Table 5**
The raw features

| Feature | Description | Attributes in experiment | |
|---|---|---|---|
| | | Normal | DFT |
| Pitch | is a very sensitive factor responds to the auditory sense, also called fundamental frequency, refers to the periodic time of a wave pulse generated by air compressed through the glottis from the lungs. In our work, we employed Autocorrelation Function (ACF) method, which is one of the popular method in pitch tracking process to extract *pitch*-related features. It helps us to find out the similarity between the signal and a shifted version of itself to have peaks in multiples of the fundamental frequency | pitch_max<br>pitch_mean<br>pitch_min<br>pitch_median<br>pitch_range<br>pitch_std<br>pitch_var<br>pitch_changerate | fft_pitch_max<br>fft_pitch_mean<br>fft_pitch_min<br>fft_pitch_median<br>fft_pitch_range<br>fft_pitch_std<br>fft_pitch_var<br>fft_pitch_changerate |
| Intensity | Energy – also known as power or energy, the intensity of a voice can be physically detected through the pressure of sounds or a subjective level of noisiness. Normally, the simple intensity is the sum of the absolute values for each data frame | energy_max<br>energy_mean<br>energy_min<br>energy_median<br>energy_range<br>energy_std<br>energy_var<br>energy_changerate | fft_energy_max<br>fft_energy_mean<br>fft_energy_min<br>fft_energy_median<br>fft_energy_range<br>fft_energy_std<br>fft_energy_var<br>fft_energy_changerate |
| | PowerDb – the relative intensity value, usually used in speech signal processing for its familiarity with the normal sense of hearing. The range of sound that human's ear can hear is from the ratio of the sound pressure that causes permanent damage from short exposure to the limit that (undamaged) ears can hear is more than a million. In order to simplify the representation of a large range, logarithmic units are used | powerDb_max<br>powerDb_mean<br>powerDb_min<br>powerDb_median<br>powerDb_range<br>powerDb_std<br>powerDb_var<br>powerDb_changerate | fft_powerDb_max<br>fft_powerDb_mean<br>fft_powerDb_min<br>fft_powerDb_median<br>fft_powerDb_range<br>fft_powerDb_std<br>fft_powerDb_var<br>fft_powerDb_changerate |
| Zero cross rate | is considered as one of the *duration*-related feature to be used in our experiment, which represents the number of times that the speech signals crossing the zero point. It can be easily calculated by counting the times that the wave touches the level zero reference. Instead of speech rate mentioned in the psychology, ZCR is more appropriate for language-independent speech recognition | zcr_max<br>zcr_mean<br>zcr_min<br>zcr_median<br>zcr_range<br>zcr_std<br>zcr_var<br>zcr_changerate | fft_zcr_max<br>fft_zcr_mean<br>fft_zcr_min<br>fft_zcr_median<br>fft_zcr_range<br>fft_zcr_std<br>fft_zcr_var<br>fft_zcr_changerate |
| Spectral feature | Mel-scale Frequency Cepstral Coefficients (MFCC), a very common groups of features used in automatic speech recognition (ASR), which convert the basic features (pitch, powerDb, and phase features) into a 12 MFCC features. They are synthesis features may have no physical meaning | mfcc1<br>mfcc2<br>mfcc3<br>mfcc4<br>mfcc5<br>mfcc6 | mfcc7<br>mfcc8<br>mfcc9<br>mfcc10<br>mfcc11<br>mfcc12 |

limited number of the available experimental data samples, a 10-fold cross validation is applied and the average accuracy is reported for comparison.

The intuition of the experimental design is that: if the *ERFTrees* algorithm performs well, it will produce a reduced data set, on which all the classification methods should achieve improved accuracy, in comparison to the original 84-feature set.

#### *4.2.1. Selected features*

From the original 84 features, the *ERFTrees* algorithm extracts 16 features as shown in the Table 6. From this table, we can see that:

- A majority (68 out of 84) of original features are identified as irrelevant and may be ignored in practice. This will not only save storage memory and processing time, but also make it possible for many machine learning algorithms to work efficient.

**Table 6**
The selected subset of the specifying acoustic features

| Feature | Selected attributes | | |
|---|---|---|---|
| Pitch-related | pitch_mean<br>pitch_std | pitch_min<br>pitch_changerate | pitch_median |
| Intensity-related | energy_mean<br>powerDb_std | energy_min<br>powerDb_changerate | energy_median<br>fft_energy_median |
| Duration-related | zcr_mean | zcr_min | zcr_median |
| MFCC | mfcc1 | mfcc12 | |
| Phase-related | – | | |

- *Pitch*-related, *intensity*-related (*Energy*-related features and *PowerDb*-related features) are identified as the most important among all the features, which is the same conclusion as mentioned in psychology studies. 11 out of 16 selected features are from these three groups;
- None of the *Phase*-related features has been selected by the algorithm;
- Only three transformed features are identified as relevant by the algorithm.

#### 4.2.2. Quality of the selected features

In order to evaluate the quality of the selected features, we compare it with the original 84 features. Two classification algorithms, *C4.5* Decision Tree algorithm and *RandomForests* algorithm, are applied on both 84-feature data sets, and the selected 16-feature data sets. Both classification algorithms run on each data set for 100 runs, and the average accuracy is shown in Table 7.

In Table 7, the numbers in bold show which feature set has better performance by comparing the emotion recognition accuracy of the data sets represented by the original 84-feature set with the data sets represented by the selected 16-feature subset. About 2/3 data sets (12 out of 18) achieved higher results (increased about 2.32% on the overall data sets in average) on the 16-feature subset than on the 84-feature set. Therefore, the results indicate that the selected feature subset can provide better performance than the whole feature set for emotional speech recognition.

#### 4.2.3. Gender dependency

By comparing the result values of the separate data sets recorded by female (data sets 1, 4, 7) and male speakers (data sets 2, 5, 8 in Table 7) with the results of the data sets containing both gender types (data sets 3, 6, 9), it is found that the emotion recognition performance on gender-dependent data sets is better than the performance on gender-independent data sets.

The classification accuracy on most of the data sets when considering gender dependency were improved by about 3.99% for 84-feature set and 4.69% for 16-feature set in average. Especially, from the comparison of the male and overall natural data sets (data sets 5, 6), which gave remarkable improvements of 8.28% (84-feature set) and 9.25% (16-feature set) by the *RandomForests* algorithm. Therefore, it has advantages to involve gender information which provides better performance on emotion recognition.

In addition, the results on the 84-feature and 16-feature data sets both imply that male and female speakers express their emotions differently, and females are more 'emotional'. For example, for the data sets containing all data instances (data sets 7, 8), the correct recognition rates of female speakers (from 62.97% to 73.19%) are obviously higher than those of male speakers (from 57.61% to 71.22%).

#### 4.2.4. Acted vs. natural emotions

In general, the correct recognition accuracy was higher for natural data sets (data sets 4–6) than for acted data sets (data sets 1–3). The results of recognizing natural emotions on data sets with both female and male speakers was approximately 3.18% better than the results of acted emotions. For this reason, if the available speech data instances are limited, it is possible to apply the model trained based on the acted emotional speech data to recognize natural emotions as well.

Furthermore, unlike the recognition performance on the data sets of acted emotions, the natural data sets gave a lower accuracy for the 16-feature subset than for the 84-feature set (see Table 7). The recognition results decreased about 3% on the overall natural emotions (data set 6) by replacing the whole 84-feature set with the selected 16-feature subset. This may be due to the complexity of natural emotions. In some real-world cases, even if a major emotional state is expressed in speech, it is also possible to have other emotions hidden inside. For example, people may feel both sad and fear when they do something wrong that might hurt the others. Therefore, it needs more information to recognize the certain natural emotional states, contained in the speech examples than to recognize the states contained in the acted emotions.

**Table 7**
Experiment results of feature selection models (%)

| Data set | Source | Gender | C4.5 | | RandomForests | |
|---|---|---|---|---|---|---|
| | | | (84-feature) | (16-feature) | (84-feature) | (16-feature) |
| 1 | Acted | Female | 60.21 | **65.01** | 67.74 | **74.52** |
| 2 | Acted | Male | 60.34 | 60.67 | 66.70 | **69.50** |
| 3 | Acted | Both | 57.61 | **61.19** | 65.34 | **71.25** |
| 4 | Natural | Female | 65.83 | **66.92** | **74.85** | 73.34 |
| 5 | Natural | Male | 64.58 | **69.60** | **79.12** | 78.82 |
| 6 | Natural | Both | **65.18** | 62.53 | **70.84** | 69.57 |
| 7 | Both | Female | 62.97 | **66.49** | 68.95 | **73.19** |
| 8 | Both | Male | **58.08** | 57.61 | 67.71 | **71.22** |
| 9 | Both | Both | 58.50 | **61.40** | 65.45 | **68.90** |
| Average | – | – | 61.48 ± 3.22 | **63.49** ± **3.77** | 69.63 ± 4.62 | **72.25** ± **3.13** |

#### 4.2.5. Analysis of confusion matrix

Generally, the ability to recognize an individual emotional state is improved by using the subset of the selected 16 features rather than the whole acoustic feature set. The 5 confusion matrices displaying correct recognition accuracy of each individual emotion are shown in Table 8, respectively: the whole data set with all data instances (data set 9), the overall female data set (data set 7), the overall male data set (data set 8), the overall acted data set (data set 3) and the overall natural data set (data set 6). The columns show the emotional states as the class labels in the experimental data sets; while the rows show the certain emotional states that the model classified the data instances as. Therefore, the values (in Bold) on the diagonal are the correct recognition rate of each emotional states on the special data set.

In Table 8, the *'Angry'* emotion got the best recognition accuracy on all data sets, especially, for the female data set with 16 features, it achieved the highest correct rate of 82.54%; while the *'Happy'* emotion performed worst, which only has 16% on 84-feature natural data set and 20% on 16-feature natural data set because at least 20% of *'Happy'* speech instances were misclassified as *'Angry'* ones. This may be caused by the nature of the speech data instances used in the experiment, with most only lasting for 2 or 3 s duration. Even for humans, it is hard to distinguish the similar emotions like anger and happiness in such a short time.

#### 4.2.6. Summary

In summary, the feature selection process with the presented *ERFTrees* algorithm, is well suited for the task of emotion recognition. A well-selected acoustic feature subset can represent the original speech data examples using fewer features but containing most of the useful available information for an emotion recognition task. The experiment results also show the advantages of the applying feature selection process on the original speech data set before undertaking additional classification tasks.

### 4.3. Comparison with other methods

*PCA/MDS* and *ISOMap* are widely used dimensionality reduction methods in emotional speech recognition and related areas. In this part, we compare the performance of our presented algorithm with *PCA/MDS* and *ISOMap*. All compared algorithms are applied to the original data sets with 84 features to extract or select best features respectively. Since the *ERFTrees* algorithm found 16 features, we set the lower-dimension for both PCA/MDS and the ISOMap algorithms as 16. The *k-NN* algorithm is then applied on both the original data sets and the selected/extracted features data sets, and its performance

**Table 8**
Feature selection results shown in confusion matrices (%)

| | Angry | Happy | Sad | Fear | Neutral | Angry | Happy | Sad | Fear | Neutral |
|---|---|---|---|---|---|---|---|---|---|---|
| | Overall 84-feature data set | | | | | Overall 16-feature data set | | | | |
| Angry | **70.19** | 10.36 | 9.3 | 9.3 | 0.85 | **78.44** | 7.82 | 7.82 | 5.29 | 0.63 |
| Happy | 21.11 | **39.20** | 16.58 | 18.09 | 5.03 | 29.15 | **49.25** | 9.05 | 8.04 | 4.52 |
| Sad | 18.75 | 7.72 | **45.96** | 19.85 | 7.72 | 15.07 | 10.29 | **47.43** | 15.81 | 11.40 |
| Fear | 14.12 | 10.31 | 20.99 | **48.85** | 5.73 | 16.41 | 6.87 | 21.76 | **51.53** | 3.44 |
| Neutral | 1.18 | 4.14 | 13.02 | 8.88 | **72.78** | 1.78 | 2.96 | 13.61 | 9.47 | **72.19** |
| | Female 84-feature data set | | | | | Female 16-feature data set | | | | |
| Angry | **73.65** | 9.21 | 12.70 | 4.13 | 0.32 | **82.54** | 6.98 | 8.57 | 1.59 | 0.32 |
| Happy | 25.37 | **52.99** | 11.19 | 6.72 | 3.73 | 26.12 | **52.24** | 11.19 | 7.46 | 2.99 |
| Sad | 22.51 | 5.76 | **51.83** | 12.04 | 7.85 | 19.90 | 9.95 | **48.17** | 11.52 | 10.47 |
| Fear | 11.63 | 5.43 | 23.26 | **55.04** | 4.65 | 6.20 | 10.08 | 20.16 | **62.79** | 0.78 |
| Neutral | 3.64 | 5.45 | 19.09 | 6.36 | **65.45** | 5.45 | 6.36 | 14.55 | 2.73 | **70.91** |
| | Male 84-feature data set | | | | | Male 16-feature data set | | | | |
| Angry | **68.35** | 9.49 | 5.70 | 15.19 | 1.27 | **82.28** | 5.70 | 2.53 | 8.86 | 0.63 |
| Happy | 23.08 | **29.23** | 23.08 | 15.38 | 9.23 | 29.23 | **40.00** | 12.31 | 12.31 | 6.15 |
| Sad | 9.88 | 11.11 | **46.91** | 28.40 | 3.70 | 3.70 | 16.05 | **41.98** | 28.40 | 9.88 |
| Fear | 17.29 | 7.52 | 12.78 | **60.90** | 1.50 | 17.29 | 3.76 | 13.53 | **62.41** | 3.01 |
| Neutral | 1.69 | 13.56 | 16.95 | 3.39 | **64.41** | 3.39 | 8.47 | 8.47 | 8.47 | **71.19** |
| | Acted 84-feature data set | | | | | Acted 16-feature data set | | | | |
| Angry | **71.98** | 18.68 | 1.65 | 6.04 | 1.65 | **73.08** | 18.13 | 3.30 | 2.75 | 2.75 |
| Happy | 20.69 | **45.98** | 8.05 | 12.64 | 12.64 | 24.14 | **48.85** | 11.49 | 8.62 | 6.90 |
| Sad | 1.70 | 9.66 | **50.57** | 25.00 | 13.07 | 1.70 | 10.80 | **58.52** | 17.61 | 11.36 |
| Fear | 7.07 | 14.13 | 20.11 | **51.09** | 7.61 | 7.07 | 10.87 | 17.39 | **57.07** | 7.61 |
| Neutral | 1.18 | 10.65 | 11.83 | 10.06 | **66.27** | 3.55 | 5.33 | 11.83 | 6.51 | **72.78** |
| | Natural 84-feature data set | | | | | Natural 16-feature data set | | | | |
| Angry | **78.69** | 3.09 | 13.06 | 5.15 | – | **79.38** | 4.81 | 10.65 | 5.15 | – |
| Happy | 52.00 | **16.00** | 24.00 | 8.00 | – | 48.00 | **20.00** | 8.00 | 24.00 | – |
| Sad | 32.29 | 7.29 | **41.67** | 18.75 | – | 31.25 | 4.17 | **39.58** | 25.00 | – |
| Fear | 23.08 | 1.28 | 20.51 | **55.13** | – | 26.92 | 2.56 | 29.49 | **41.03** | – |
| Neutral | – | – | – | – | – | – | – | – | – | – |

**Table 9**
Experiment results for dimension reduction models (%)

| Data set | Source | Gender | ♯Sent. | None (84-dimensional) | MDS (16-dimensional) | ISOMAP (16-dimensional) | ERFTrees (16-dimensional) |
|---|---|---|---|---|---|---|---|
| 1 | Acted | Female | 567 | **64.74** ($k = 7$) | 58.40 ($k = 17$) | 53.95 ($k = 9$) | 66.18 ($k = 5$) |
| 2 | Acted | Male | 318 | 63.25 ($k = 9$) | 55.76 ($k = 19$) | 53.78 ($k = 15$) | 71.41 ($k = 3$) |
| 3 | Acted | Both | 885 | 60.51 ($k = 7$) | 50.49 ($k = 14$) | 50.93 ($k = 5$) | 64.74 ($k = 5$) |
| 4 | Natural | Female | 312 | 70.99 ($k = 7$) | 69.40 ($k = 9$) | 70.52 ($k = 5$) | 72.73 ($k = 5$) |
| 5 | Natural | Male | 178 | 80.24 ($k = 7$) | 76.87 ($k = 5$) | 76.85 ($k = 5$) | 79.42 ($k = 9$) |
| 6 | Natural | Both | 490 | 68.43 ($k = 9$) | 65.14 ($k = 11$) | 66.92 ($k = 5$) | 65.88 ($k = 19$) |
| 7 | Both | Female | 879 | 62.94 ($k = 5$) | 58.86 ($k = 13$) | 58.06 ($k = 5$) | 64.62 ($k = 3$) |
| 8 | Both | Male | 496 | 65.24 ($k = 12$) | 60.67 ($k = 7$) | 59.12 ($k = 5$) | 73.85 ($k = 3$) |
| 9 | Both | Both | 1375 | 59.83 ($k = 7$) | 55.02 ($k = 18$) | 53.51 ($k = 18$) | 64.09 ($k = 3$) |
| Average | – | – | – | 66.24 ± 6.33 | 61.18 ± 8.10 | 60.40 ± 8.98 | 69.21 ± 5.37 |

will reflect the quality of the prepared data sets. Furthermore, the *k-NN* algorithm is also applied on *PCA/MDS* and *ISOMap*, and the results are used for performance comparison to the presented *ERFTrees* algorithm. In our experiment, we tune the parameter *k* for the *k-NN* algorithm by trying different values from the range (Fairbanks & Pronovost, 1939, 2003). Table 9 gives the classification results on data sets preprocessed by different algorithms.

When comparing the results from *PCA/MDS* and *ISOMap* algorithms with the original data sets, we can see that these two algorithms fail to produce better reduced feature sets, as indicated in the fifth (*None*) to seventh columns (*ISOMap*). Comparing the performance of the linear dimension reduction method, *PCA/MDS*, and the non-linear dimension reduction method *ISOMap*, it can not observe the obvious difference between them. Although for the natural data sets (data sets 4–6), *ISOMap* outperforms by 2%, this may come from the fact that natural speech data is more complex than acted speech data, accordingly the non-linear method is more suitable for these cases.

The last column (*ERFTrees*) of the Table 9 gives the accuracy of *k-NN* algorithm on the data sets preprocessed by the *ERFTrees* algorithm. It is evident that the *ERFTrees* algorithm outperforms the other dimensionality reduction algorithms on most data sets. On average, the performance has been improved 8% over those data sets preprocessed by the dimensionality reduction methods. Moreover, the comparison between the 5th column (*None*) and the 8th column (*ERFTrees*) further confirms that the *ERFTrees* algorithm can extract a small but more relevant feature set which can lead to improved emotion recognition accuracy.

## 5. Conclusion

It is no doubt that introducing advanced machine learning techniques into emotion recognition is beneficial. However, since collecting a large set of emotion speech samples is time consuming and labor intensive, machine learning algorithms that can work with only a small number of training examples are desirable. In this paper, the *Ensemble Random Forest to Trees (ERFTrees)* is presented, which can be used to extract effective features from small data sets.

The small size of the training data sets has been identified as an important factor, impacting the learning performance (Vapnik, 1995), and the presented *ERFTrees* algorithm exhibits a better way to the preprocessing of small data sets. The algorithm works in a ensemble-learning style, and integrates the advantages from the simple decision tree and the random forest ensemble. It is justified that this method can reduce error rates caused by both the limitation of the finite data samples of the decision tree algorithm and the restricted learning ability of the random forest ensemble.

A series of experiments were done on 9 emotional speech data sets where the language is Chinese (Mandarin), and 16 acoustic features were selected out from the original 84 feature set. Most of them are the basic *pitch*-related and *intensity*-related acoustic features. It is proven by the experiment results that the data sets with the selected 16-feature subset can provide a higher recognition accuracy than those with the original 84-feature set. The recognition performance was improved by 2.01% for *C4.5* Decision Tree classifiers, while the best recognition accuracy 72.25% can be achieved by the *Random Forest* classifier.

## Acknowledgements

The related work is partially supported by Deakin CRGS grant 2008. The authors would like to thank Sam Schmidt for proof reading the English of the manuscript.

## References


Amir, N. (2001). Classifying emotions in speech: A comparison of methods. In *Proceedings of European conference on speech communication and technology (EUROSPEECH'01)*. Escandinavia.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70*(3), 614–636.

Bezooijen, R. V. (1984). *Characteristics and recognizability of vocal expressions of emotions*. Dordrecht, The Netherlands, Foris: Walter de Gruyter Inc.

Bhatti, M. W., Wang, Y., & Guan, L. (2004). A neural network approach for human emotion recognition in speech. In *Proceedings of the 2004 international symposium on circuits and systems (ISCAS'04)* (Vol. 2). Canada: Vancouver.

Breiman, L. (2001). Random forest. *Machine Learning, 45*(1), 5–32.

Cai, L., Jiang, C., Wang, Z., Zhao, L., & Zou, C. (2003). A method combining the global and time series structure features for emotion recognition in speech. In *Proceedings of international conference on neural networks and signal processing (ICNNSP'03)* (Vol. 2, pp. 904–907).

Chuang, Z.-J., & Wu, C.-H. (2004). Emotion recognition using acoustic features and textual content. In *Proceedings of IEEE international conference on multimedia and expo (ICME'04)* (Vol. 1, pp. 53–56). IEEE Computer Society.

Coleman, R., & Williams, R. (1979). Identification of emotional states using perceptual and acoustic analyses. *Transcript of the eighth symposium: Care of the professional voice* (Vol. 1). New York: The Voice Foundation.

Cornelius, R. R. (1996). *The science of emotion: Research and tradition in the psychology of emotion*. Upper Saddle River, NJ: Prentice Hall.

Cowie, R., & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech communication special issue on speech and emotion* (Vol. 40, pp. 5–32). Amsterdam, The Netherlands, The Netherlands: Elsevier Science Publishers B.V.

Davitz, J. R. (1964). Personality, perceptual, and cognitive correlates of emotional sensitivity. *The Communication of Emotional Meaning*, 57–68.

Dellaert, F., Polzin, T., & Waibel, A. (1996). Recognizing emotion in speech. In *Proceedings of fourth international conference on spoken language processing (ICSLP'96)* (Vol. 3, pp. 1970–1973).

Fairbanks, G., & Pronovost, W. (1939). An experimental study of the pitch characteristics of the voice during the expression of emotion. *Speech Monograph, 6*, 87–104.

Fairbanks, G., & Pronovost, W. (1941). An experimental study of the durational characteristics of the voice during the expression of emotion. *Speech Monograph, 8*, 85–91.

Fónagy, I. (1978). A new method of investigating the perception of prosodic features. *Language and Speech, 21*, 34–49.

Fónagy, I. (1978). Emotions, voice and music. *Language and Speech, 21*, 34–49.

Fónagy, I., & Magdics, K. (1963). Emotional patterns in intonation and music. *Zeitschrift für Phonetik SprachWissenschaft und Kommunicationsforschung, 16*, 293–326.

Han, J., & Kamber, M. (2000). *Data mining concepts and techniques* (1st ed.). Publishers, USA: Morgan Kaufman.

Hargreaves, W., Starkweather, J., & Blacker, K. (1965). Voice quality in depression. *Journal of Abnormal Psychology, 7*, 218–220.

Havrdova, Z., & Moravek, M. (1979). Changes of the voice expression during suggestively influenced states of experiencing. *Activitas Nervosa Superior, 21*, 33–35.

Hoch, S., Althoff, F., McGlaun, G., & Rigoll, G. (2005). Bimodal fusion of emotional data in an automotive environment. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing (ICASSP'05)* (Vol. 2, pp. 1085–1088). IEEE Computer Society.

Höffe, W. L. (1960). On the relation between speech melody and intensity. *Phonetica, 5*, 129–159.

Huttar, G. L. (1967). Relations between prosodic variables and emotions in normal american english utterances. *Journal of the Acoustical Society of America, 11*, 481–487.

Hyun, K. H., Kim, E. H., & Kwak, Y. K. (2005). Improvement of emotion recognition by Bayesian classifier using non-zero-pitch concept. In *Proceedings of IEEE international workshop on robots and human interactive communication, no. 0-7803-9275-2/05* (pp. 312–316). IEEE Computer Society.

Hyvärinen, A. (1999). Survey on independent component analysis. *Neural Computing Surveys, 2*, 94–128.

Inanoglu, Z., & Caneel, R. (2005). Emotive alert: Hmm-based motion detection in voicemail messages. In *Proceedings of the 10th international conference on intelligent user interfaces (IUI'05), no. 585*. San Diego, California, USA: ACM Press.

Johnson, W., Emde, R., Scherer, K., & Klinnert, M. (1986). Recognition of emotion from vocal cues. *Arch Gen Psychiatry, 43*, 280–283.

Kaiser, L. (1962). Communication of affects by single vowels. *Synthese, 14*(4), 300–319.

Klasmeyer, G., & Sendlneier, W. F. (1995). Objective voice parameteres to characterize the emotional content in speech. In *Proceedings of the 13th international congress of phonetic sciences (ICPhS'95)* (Vol. 3, pp. 181–185). Stockholm, Sweden.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*(1-2), 273–324.

Kotlyar, G., & Mozorov, V. (1976). Acoustic correlates of the emotional content of vocalized speech. *Journal of Acoustical Academy of Sciences of the USSR, 22*, 208–211.

Kwon, O.-W., Chan, K., Hao, J., & Lee, T.-W. (2003). Emotion recognition by speech signal. In *Proceedings of the eighth European conference on speech communication and technology (EUROSPEECH'03)*, Geneva, Switzerland.

Lattin, J., Carroll, J. D., & Green, P. E. (2003). *Analyzing multivariate data*. USA: Books/Cole.

Lee, C. M., & Narayanan, S. (2003). Emotion recognition using a data-driven fuzzy inference system. In *Proceedings of the eighth European conference on speech communication and technology (EUROSPEECH'03)*. Genena, Swtzerland.

Lee, C. M., Narayanan, S., & Pieraccini, R. (2002). Combining acoustic and language information for emotion recognition. In *Proceedings of the seventh international conference on spoken language processing (ICSLP'02)*. Denver, CO, USA.

Lee, C. M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., et al. (2004). Emotion recognition based on phoneme classes. In *Proceedings of international conference on speech language processing (ICSLP'04)*. Jeju, Korea.

Lee, C. M., & Narayanan, S. (2004). Towards detecting emotions in spoken dialogs. *IEEE transactions on speech and audio processing* (Vol. 13, pp. 293–302). IEEE Computer Society.

Lee, C. M., Narayanan, S., & Pieraccini, R. (2001). Recognition of negative emotions from the speech signal. In *Proceedings of IEEE workshop on automatic speech recognition and understanding* (pp. 240–243). Trento, Italy: IEEE Computer Society.

Litman, D., & Forbes-Reley, K. (2003). Recognizing emotion from student speech in tutoring dialogues. In *Proceedings of IEEE workshop on automatic speech recognition and understanding (ASRU'03)* (pp. 25–30). IEEE Computer Society.

Liu, H., Motoda, H., & Yu, L. (2002). Feature selection with selective sampling. In *Proceedings of the 19th international conferenceon machine learning (ICML'02)* (pp. 395–402). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

McGilloway, S., Cowie, R., & Douglas-Cowie, E. (1995). Prosodic signs of emotion in speech: Preliminary results from a new technique for automated statistical ananlysis. In *Proceedings of the 13th international congress of phonetic sciences (ICPhS'95)* (Vol. 3, pp. 250–253). Stockholm, Sweden.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill Education Co.

Mozziaconacci, S. (1995). Pitch variations and emotions in speech. In *Proceedings of the 13th international congress of phonetic sciences (ICPhS'95)* (Vol. 3, pp. 178–181). Stockholm, Sweden.

Muller, A. (1960). *Experimental studies on vocal portrayal of emotion*. Ph.D. thesis, University of Gottingen, Germany.

Nicholson, J., Takahashi, K., & Nakatsu, R. (1999). Emotion recognition in speech using neural networks. In *Proceedings of sixth international conference on neural information processing (ICONIP'99)* (Vol. 2, pp. 495–501).

Nwe, T. L., Wei, F. S., & Silva, L. D. (2001). Speech based emotion classification. In *Proceedings of IEEE region 10 international conference on electrical and electronic technology* (Vol. 1, pp. 297–301).

Öster, A.-M., & Risberg, A. (1986). The identification of the mood of a speaker by hearing impaired listeners. In *Quarterly progress status report 4*. Speech Transmission Lab, Stockholm.

Pantic, M., & Rothkrantz, L. (2003). Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE, Special Issue on Human-Computer Multimodal Interface, 91*, 1370–1390.

Park, C.-H., Wook Lee, D., & Sim, K.-B. (2002). Emotion recognition of speech based on rnn. In *Proceedings of international conference on machine learning and cybernetics (ICMLC02)* (Vol. 4, pp. 2210–2213).

Park, C.-H., & Sim, K.-B. (2003). Emotion recognition and acoustic analysis from speech signal. In *Proceedings of the international joint conference on neural networks (IJCNN'03)* (Vol. 4, pp. 2594–2598).

Petrushin, V. A. (2000). Emotion recognition in speech signal: Experimental study, development, and application. In *Proceedings of sixth international conference on spoken language processing (ICSLP'00)*.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*, 1161–1178.

Schuller, B., Reiter, S., Müller, R., Al-Hames, M., Lang, M., & Rigoll, G. (2005). Speaker independent speech emotion recognition by ensemble classification. In *Proceedings of IEEE international conference on multimedia and expo (ICME'05)*. IEEE Computer Society.

Schuller, B., Rigoll, G., & Lang, M. (2003). Hidden markov model-based speech emotion recognition. In *Proceedings of the 28th IEEE international conference on acoustic, speech and signal processing (ICASSP'03)* (Vol. 2, pp. 1–4). IEEE Computer Society.

Schuller, B., Rigoll, G. & Lang, M. (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *Proceedings of the 28th IEEE international conference on acoustic, speech and signal processing (ICASSP'04)* (Vol. 1, pp. 577–580). IEEE Computer Society.

Sedlacek, K., & Syhra, A. (1963). Speech melody as a means of emotional expression. *Folia Phoniatrica, 15*, 89–98.

Shafran, L., Riley, M., & Mohri, M. (2003). Voice signatures. In *Proceedings of The eighth IEEE automatic speech recognition and understanding workshop (ASRU 2003)*.

Song, M., Bu, J., Chen, C., & Li, N. (2004). Audio-visual based emotion recognition-a new approach. In *Proceedings of IEEE computer society conference on computer vision and pattern recognition (CVPR04)* (Vol. 2, pp. 1020–1025). IEEE Computer Society.

Song, M., Chen, C., & You, M. (2004). Audio-visual based emotion recognition using tripled hidden markov model. In *Proceedings of IEEE international conference on acoustic, speech and signal processing (ICASSP'04)* (Vol. 5, pp. 877–880). IEEE Computer Society.

Sulc, J. (1977). Emotional changes in human voice. *Activitas Nervosa Superior, 19*, 215–216.

Talavera, L. (1999). Feature selection as a preprocessing step for hierarchical clustering. In *Proceedings of the 16th international conference on machine learning (ICML 1999)* (pp. 389–397). Morgan Kaufmann.

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 2319–2323.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Berlin: Springer.

Williams, C. E., & Stevens, K. N. (1969). On determining the emotional state of pilots during flight: An exploratory study. *Aerospace Medicine, 40*, 1369–1372.

Williams, C. E., & Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. *Journal of the Acoustical Society of America, 52*, 1238–1250.

Wolf, L., & Shashua, A. (2005). Feature selection for unsupervised and supervised inference: The emergence of sparisity in a weight-based approach. *Journal of Machine Learning Research, 6*, 1855–1887.

Zhou, Z.-H. (2004). *Mining extremely small data set on software reuse*. Technique report, Nanjing University.