# Classifying Emotion in Chinese Speech by Decomposing Prosodic Features

*Dan-Ning Jiang, Lian-Hong Cai*

Department of Computer Science and Technology
Tsinghua University, China
jdn00@mails.tsinghua.edu.cn, clh-dcs@tsinghua.edu.cn

## Abstract

Prosodic features have been proven important to discriminate between different speech emotions, but they also have a fundamental linguistic function. Variations caused by linguistic contexts act as noises in emotion classification and should be eliminated. The paper proposes a novel method to decompose the raw "mixed" prosodic features into features determined by linguistic contexts and those responsible for emotionality, and the latter are further used exclusively in emotion classification. In the method, features determined by linguistic contexts are first predicted based on the analysis of neutral speech through Generalized Regression Neural Network (GRNN), and Linear Discriminant Analysis (LDA) is then applied to accomplish the decomposition. Experiments on Chinese emotional speech have shown that the emotional features estimated through feature decomposition have a better discrimination between different emotions, and could achieve much higher classification accuracy than raw features.

## 1. Introduction

Human speech consists of not only words and meanings, but also the information about emotion that resides in the way words are spoken. It would be helpful if a computer had the ability of recognizing what emotion is implied in a given utterance. To tackle the task, it is necessary to examine speech features and find out which ones convey emotion information and could discriminate between different emotions well. In most relevant literatures, the importance of prosodic features in emotional speech is evident [1][2]. Experiments have also demonstrated their efficiency on recognizing emotion [3].

While convey emotion information, prosodic features also have a fundamental linguistic function, and are partly determined by linguistic contexts. For example, questions are often concerned with rising pitch contours; accents are implemented by linked increase of pitch, duration, and intensity in general. Moreover, in Chinese, there exist four specific syllable tone patterns: high-level, rising, falling-rising, and high-falling. As an example, figure 1 shows pitch contour of a neutral Chinese utterance "ni3 bu4 xi3 huan1", which means, "You dislike it". The first and third syllables have the falling-rising pattern, the second syllable has the high-falling pattern, and tone pattern of the last syllable is high-level. It could be observed that both pitch level and contour shape are highly dependent on the syllable tone pattern. Thus variations caused by linguistic contexts would be confounded with those associated with emotionality, and act as noises in emotion classification.

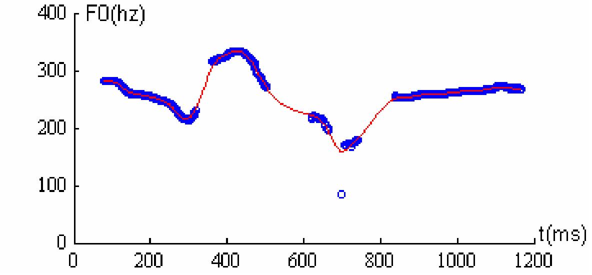

*Figure 1:* Pitch contour of the neutral Chinese utterance "ni3 bu4 xi3 huan1", which means, "You dislike it".

To get more efficient features for emotion classification, the paper proposes a novel method to decompose raw "mixed" prosodic features into features determined by linguistic contexts and those responsible for emotionality, and the latter are further used exclusively in emotion classification. Features determined by linguistic contexts are first predicted based on the analysis of neutral speech through Generalized Regression Neural Network (GRNN), and Linear Discriminant Analysis (LDA) is then applied to accomplish the decomposition. Finally, experiments are performed to demonstrate efficiency of the emotional features estimated through feature decomposition on classifying speech emotion.

The rest of the paper is organized as follows. Section 2 explains the proposed feature decomposition method. Section 3 describes the emotion classification procedure, including feature extraction and classifier selection. Finally, in section 4, experiment results are shown to evaluate the decomposition method.

## 2. Feature Decomposition

Suppose $(f_1, f_2, \cdots, f_D)$ represents the raw prosodic feature vector extracted directly from emotional speech, then each dimension $f_i$ $(0 < i \le D)$ is regarded as a linear combination of the feature determined by linguistic contexts (represented as $f_i^l$ ) and that responsible for emotionality (represented as $f_i^e$ ). That is,

$$f_i = k_i(f_i^l + a_i f_i^e) \tag{1}$$

Where $k_i$ and $a_i$ are both non-zero constants. To estimate $f_i^e$ , formula (1) is transformed as:

$$f_i^e = k_i^*(f_i + a_i^* f_i^l) \tag{2}$$

$$k_i^* = \frac{1}{k_i a_i} \tag{3}$$

$$a_i^* = -k_i \tag{4}$$

So, the decomposition could be divided into two procedures: a. Predict $f_i^l$ based on the analysis of neutral speech; b. Estimate $a_i^*$ in formula (2) to accomplish the decomposition ( $k_i^*$ is just a scale factor and thus not important).

### 2.1. $f_i^l$ Prediction

In Text-to-Speech research field, prosodic characteristics of the synthesized speech have could be predicted successfully by linguistic context parameters through some data-driven models [4], such as Artificial Neural Network (ANN). The problem is similar with the $f_i^l$ prediction, so analogous models could be used for reference.

Generalized Regression Neural Network (GRNN), which is often used in function approximation, is applied to predict $f_i^l$ . Inputs of the network are linguistic context parameters associated with the feature dimension (represented as $CP_i$ in figure 2), and the output is $f_i^l$ . GRNN is a two-layer structured network. The first layer is a radial basis one, whose neurons have the transfer function as below:

$$radbas(p) = e^{-(\|w-p\|b)^2} \tag{5}$$

Where $p$ is the input vector, $w$ is the weight vector, and $b$ is the bias. The function outputs a value based on the distance between the input vector and the weight vector. As the distance decreases, the output of the function increases. The bias $b$ allows the sensitivity of the radial basis neuron to be adjusted. The first layer has as many neurons as there are input/target vector pairs in training set, and the weight vectors are set to be the training input vectors.

The second layer of GRNN is a linear one, whose neuron number is also equal to number of input/target vector pairs in training set. The layer-2 weight vectors are set to be the target vectors. Thus, when a testing input vector is closed to a layer-1 neuron weight, one of the neurons in the first layer produces a layer-1 output closed to 1, and the others are closed to 0. This leads to a final output closed to the associated target vector. In experiments, GRNN is implemented by MATLAB.

### 2.2. $a_i^*$ Estimation

Since $f_i^e$ represents the feature responsible for emotionality, it should have the best discrimination between different emotions. So Linear Discriminant Analysis (LDA), which is a feature extraction and compression method designed to preserve as much discriminant information as possible [5], is used to find the direction with the best discrimination.

Suppose $x$ and $y$ represent the original feature vector and the feature vector after transform respectively, then LDA maps $x$ to $y$ through some linear basis functions $\{f_j,\ j = 1,2,\cdots d\}$ , where $d$ is the dimension number of $y$ :

$$y_j = x^{\mathrm{T}} f_j,\ j = 1,2,\cdots,d \tag{6}$$

The basis functions in LDA are designed to maximize the linear discrimination between classes. Measurement of the discrimination is based on the between-class covariance matrix $S_b$ and the within-class covariance matrix $S_w$ . $S_b$ reflects how much the feature vectors between classes vary, represented as below:

$$S_b = \sum_{k=1}^{N} P_k[(m - m_k)(m - m_k)^{\mathrm{T}}] \tag{7}$$

Where $P_k$ , $m_k$ are prior probability and mean vector of the k-th class respectively, $N$ is the class number, and $m$ is mean of all $m_k$ . $S_w$ reflects how much the feature vectors within one class vary, represented as:

$$S_w = \sum_{k=1}^{N} P_k \Sigma_k \tag{8}$$

$$\Sigma_k = E[(x - m_k)(x - m_k)^{\mathrm{T}} \mid x \in C_k], \tag{9}$$

Where $C_k$ represents the k-th class. Then the discrimination between classes could be measured as:

$$F = trace(S_w^{-1} S_b) \tag{10}$$

To maximize $F$ , basis functions $\{f_j\}$ should be the eigenvectors of $S_w^{-1} S_b$ associated with the first $d$ largest eigenvalues. It should be noted that the between-class covariance matrix $S_b$ has a maximum rank of $N-1$ , so $d$ is no more than $N-1$ .

To estimate $a_i^*$ , the original feature $x$ should be set as $(f_i, f_i^l)^{\mathrm{T}}$ , and then $a_i^*$ is $f_{12}/f_{11}$ .

## 3. Emotion Classification

Figure 2 shows the emotion classification paradigm. First, raw prosodic feature vector $(f_1, f_2, \cdots, f_D)$ is extracted from speech signal $s(t)$ . Then for each feature dimension, the "mixed" raw feature $f_i$ and linguistic context parameters $CP_i$ are input to the feature decomposition module to estimate the emotional feature $f_i^e$ . Finally, the emotional feature vector $(f_1^e, f_2^e, \cdots, f_D^e)$ is used exclusively in classification.

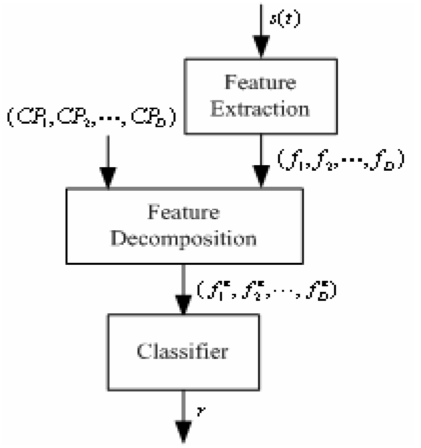
$s(t)$
Feature Extraction
$(CP_1, CP_2, \cdots, CP_D)$
$(f_1, f_2, \cdots, f_D)$
Feature Decomposition
$(f_1^e, f_2^e, \cdots, f_D^e)$
Classifier
$r$

*Figure 2:* Emotion classification paradigm.

### 3.1. Feature Extraction

In the classification, all features are derived from f0 and duration. Basic frame-based parameters are first extracted. F0 parameter is extracted by YIN, a modified auto-correlation algorithm [6]. Syllable boundaries are labeled through the software SPEECH, which could first estimate the boundaries and then permit manual adjustment on them. Thus the syllable duration could be easily extracted.

Features used in classification are statistics computed throughout each utterance. They are listed as below:

Statistics on f0 contour: mean, maximum, range, and standard deviation;

Statistics on derivative of f0 contour: mean and standard deviation;

Statistics on syllable duration: mean and standard deviation.

*Table 1:* Linguistic context parameters used to predict the linguistically determined features.

| Prosodic Features | Linguistic Context Parameters |
|---|---|
| Mean and std of f0 contour; and those of derivative of f0 contour; | Sentence type; total syllable number; number of syllables with the four tone patterns respectively; |
| Maximum of f0 contour; | Sentence type; position, tone pattern and accent type of syllable associated with the maximal f0, its pre-syllable and post-syllable; |
| Range of f0 contour; | Sentence type; position, tone pattern and accent type of syllable associated with the maximal f0, and those of syllable with the minimal f0; |
| Mean and std of syllable duration; | Sentence type; total syllable number; number of accented syllables; |

Table 1 lists detailed linguistic context parameters used to predict the linguistically determined features for each dimension (see details in section 2.1).

### 3.2. Classifier

In the work, two neural networks designed for classification problems are used. One is the familiar Multiple Layer Perceptron (MLP) with one hidden layer. The input is acoustic feature vector, and each output-layer neuron represents one class. In training stage, output of the neuron which is associated with the class that input vector belongs to is set to 1, others are set to 0. BP algorithm is then performed to train the network. In testing stage, the testing input vector is classified into the class associated with the output-layer neuron that has the maximal output value.

The other model is Probabilistic Neural Network (PNN). The network has a structure of two layers. Similar with GRNN, the first layer is a radial basis one, whose action is to compute distances between the testing input vector and the training input vectors (as weights of the first layer), and produces a vector that indicates how close the testing input is to them. The second layer has class number-equaled linear neurons; each weight has a value of 1 only when the layer-2 neuron associated with the particular class of the layer-1 neuron, and 0 otherwise. Thus, it sums these contributions for each class of inputs to produce as its output a vector of probabilities. Finally, a compete transfer function on the output layer picks the maximum of these probabilities, and outputs 1 for that class and 0 for other classes. Different from MLP, PNN classifies testing input vectors based on their similarities with training inputs stored in the network, and do not need a complicated training algorithm. In experiments, these two networks are both implemented by MATLAB.

## 4. Experiments

### 4.1. Database for Experiments

The emotional speech database contains six emotion classes: anger, fear, happiness, sadness, surprise, and neutral emotion. There are about 200 sentence level utterances for each class, which are produced by an amateur actress. Texts for different classes are not all the same, and they all include different sentence types (statements and questions), syllable tone patterns, and accent distributions. All the utterances are recorded in a relative quiet environment, and saved in mono wave files with 16 kHz sample rate and 16 quantitative bits. In classification, 80 percent of the utterances are used as training data, and the other 20 percent are testing data.

There is also another neutral database, which is used as the training data to estimate the features determined by linguistic contexts (see section 2 for details). This neutral database has about 300 utterances, which also include different sentence types, syllable tone patterns, and accent distributions.

### 4.2. Classification Results

In experiments, the leave-one-out cross-validation technique is used. The whole dataset is equally divided into five subsets. Then the classification is performed five times, at each time using one unique subset for testing, and the other four subsets for training. The classification results shown below are those averaged across all the subsets. Table 2 lists the average classification accuracy by using the emotional features estimated through feature decomposition (abbreviated as emotional features) and the raw "mixed" features respectively. It is shown that the accuracy associated with the emotional features is much higher than that with raw features, 17.0% higher by MLP, and 11.5% higher by PNN. The highest accuracy reaches 93.7%. The result suggests that the proposed decomposition method could reduce the influence of linguistic contexts on emotional features efficiently.

*Table 2:* Average classification accuracy.

| Accuracy (%) | MLP | PNN |
|---|---|---|
| Emotional Features | 82.4 | 93.7 |
| Raw Features | 65.4 | 82.2 |

More detailed results are further shown to analyze the confusion between different emotions. Only the PNN results are shown, for those of MLP are much similar. Table 3 and table 4 list confusion matrix associated with the emotional features and raw features respectively. Each emotion class is represented by its first one or several letters.

*Table 3:* Confusion matrix with the emotional features.

| (%) | A | F | H | Sad | Sur | N |
|---|---|---|---|---|---|---|
| A | 82.3 | 5.6 | 4.2 | 0.0 | 7.0 | 0.9 |
| F | 3.9 | 95.6 | 0.0 | 0.5 | 0.0 | 0.0 |
| H | 3.8 | 0.0 | 95.7 | 0.0 | 0.5 | 0.0 |
| Sad | 0.0 | 0.0 | 0.0 | 98.6 | 0.0 | 1.4 |
| Sur | 8.2 | 0.0 | 0.9 | 0.0 | 90.9 | 0.0 |
| N | 0.0 | 0.0 | 0.0 | 1.1 | 0.0 | 98.9 |

*Table 4:* Confusion matrix with the raw features.

| (%) | A | F | H | Sad | Sur | N |
|---|---|---|---|---|---|---|
| A | 67.0 | 6.5 | 18.1 | 0.0 | 6.1 | 2.3 |
| F | 5.4 | 93.7 | 0.0 | 0.5 | 0.0 | 0.5 |
| H | 14.7 | 0.0 | 65.9 | 0.0 | 17.5 | 1.9 |
| Sad | 0.0 | 0.0 | 0.0 | 98.1 | 0.0 | 1.9 |
| Sur | 6.8 | 0.0 | 21.4 | 0.0 | 71.8 | 0.0 |
| N | 0.0 | 0.4 | 1.1 | 1.8 | 0.0 | 96.7 |

To illustrate the results more clearly, $I_{ij}$ is defined as the confusion degree between the i-th and j-th emotion classes:

$$I_{ij} = \frac{P(r = i \mid x \in C_j) + P(r = j \mid x \in C_i)}{2} \qquad (11)$$

Where $r$ is the classification result of the input vector $x$. An emotion class pair is regarded as "most confused" when the associated $I_{ij}$ is larger than 10%. Then table 3 indicates that none of the emotion pairs is most confused by using the emotional features, while in table 4, emotion pair (anger, happiness) and (happiness, surprise) are most confused by using raw features. After feature decomposition, confusion degrees of the above two emotion pairs decrease from 16.4% to 4.0% and 19.5% to 0.7% respectively.

## 5. Conclusions

The paper proposes a novel method to decompose the raw "mixed" prosodic features into features determined by linguistic contexts and those responsible for emotionality, and the latter are further used exclusively in emotion classification. Features determined by linguistic contexts are first predicted based on the analysis of neutral speech through GRNN, and then LDA is applied to accomplish the decomposition. Classification experiments have been performed to evaluate efficiency of the feature decomposition method. By using the emotional features estimated through feature decomposition, the classification average accuracy is improved at least 11.4 percent when compared with the raw features, and reaches a best performance of 93.7%. More detailed results indicate that the most evident decrease of confusion happens at the emotion pair (anger, happiness) and (happiness, surprise).

## 6. References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatoulis, etc, "Emotion Recognition in Human-Computer Interaction," *Signal Processing Magazine*, Vol. 18, No. 1, IEEE Publisher, New York, pp. 32-80, Jan. 2001.

[2] A. Paeschke, W.F. Sendlmeier, "Prosodic Characteristics of Emotional Speech: Measurements of Fundamental Frequency Movements," *Proceedings of the ISCA Workshop on Speech and Emotion*, Northern Ireland, pp. 75-80, Sep. 2000.

[3] F. Dellaert, T. Polzin and A. Waibel, "Recognizing Emotion in Speech," *Proceedings of the International Conference on Spoken Language Processing*, pp. 1970-1973, 1996.

[4] J.H. Tao, L.H. Cai, S.X. Zhao, etc, "The Study of the Trainable Prosodic Model for Chinese Text to Speech System," *Shengxue Xuebao/Acta Acustica (in Chinese)*, Vol. 26, No. 1, Beijing, pp. 67-72, Jan. 2001.

[5] N. Malayath, H. Hermansky, "Data-Driven Spectral Basis Functions for Automatic Speech Recognition," *Speech Communication*, No. 40, Elsevier Science, pp. 449-466, 2003.

[6] A.D. Cheveign, H. Kawahara, "Yin, a Fundamental Frequency Estimator for Speech and Music," *J. Acoust. Soc. Am.* Vol. 111, No. 4, Apr. 2002, pp. 1917-1930.