Feature Analysis for Emotion Recognition from Mandarin Speech Considering the Special Characteristics of Chinese Language

Yi-hao Kao and Lin-shan Lee

Department of Electrical Engineering
National Taiwan University, Taipei, Taiwan
b91901146@ntu.edu.tw and lslee@gate.sinica.edu.tw

Abstract

Emotion recognition from speech signals is regarded as a critical step toward intelligent human-machine interface. However, feature parameters useful for this purpose may have to do with the special structures of the language. In this paper we present a detailed analysis of the feature parameters for emotion recognition considering the characteristics of the Chinese language, primarily the monosyllable structure and the tone behavior. The analysis is based on the feature parameters on three levels: frame-level, syllable-level, and word-level. The results show that the frame-level and syllable-level ones are good indicators, while taking the ensemble features on all three levels can yield a recognition accuracy of 90.0%. We also found that the pitch and power related features are the most important, and the fourth tone in Mandarin serves as the strongest indicator to emotions. All these findings are consistent with the characteristics of Mandarin Chinese.

Index Terms: emotion recognition, Chinese, feature analysis

1. Introduction

Emotion recognition has been believed to be a critical component in intelligent human-machine interface [1]. The natural applications in spoken dialogue system and automatic customer services are of particular interest [2]. Various research results have been repeated in this area. A typical approach is usually based on a set of linguistic, acoustic, and prosodic feature parameters [3], and a suitable classifier aimed at distinguishing several types of emotion, for example, anger, happiness, fear, sadness and neutral.

Apparently, emotion recognition has to do with language-specific characteristics. For the Chinese language, the most distinguishing characteristics are the monosyllabic structure and tone behavior. Every Chinese character has its own meaning, and is pronounced as a monosyllable. A Chinese word is composed of one to several characters. The number of Chinese syllables is very limited, and many homonym characters with different meanings share the same syllable. This monosyllabic structure is a special characteristic of the Chinese language. Also, every Chinese syllable is assigned a tone carrying lexical meaning. There are five different tones in Mandarin, including four lexical tones and one neutral tone. The same syllable pronounced with different tones has very complicated behavior in spontaneous speech. This is another characteristic of Chinese language.

In this paper we performed a delicate analysis on the feature parameters used for emotion recognition. We investigated feature parameters on three levels, including 630 frame-level features, 229 syllable-level ones, and 223 word-level ones. Since the size of the set is quite large, we applied the well-known data-mining tool Weka [4] to find the most discriminative parameters by ranking their information gains.

Recent studies have suggested that support vector machine (SVM) has superior performance in emotion recognition [5]. Thus we adopted the libSVM [6] as the core classifier. In the experiments to be presented below we yielded a consistent result showing that frame-level and syllable-level features are better emotion indicators, while word-level features carry less information. A good recognition accuracy of 90.0% can be attained by carefully adjusting the feature ensemble. In addition to verifying that the pitch and power related features are very important in emotion recognition, we also compared the significance of different tones, and the results showed that tone 4 is the best emotion indicator. All of these are quite consistent with the characteristics of the Chinese language summarized above.

The paper is organized in the following manner. Section 2 describes how we constructed the database for emotional speech. Section 3 presents the three levels of feature sets in detail. Section 4 discusses our feature selection principles and approaches. Section 5 then reports the achievable recognition performance of different sets of features through experiments. Section 6 summarizes our conclusion.

2. Mandarin emotional speech corpus and experimental setup

Our corpus is based on a script including 36 text sentences, which have typical lengths from ten to twenty characters and cover five daily life scenarios from customer services to casual conversation. To achieve a better balance of the distribution of the intonation, we take three general speaker intentions into account when designing the text, including request, inform and inquire. Each text sentence includes a keyword. Two male and two female students in the Department of Drama and Theatre were asked to produce each of the 36 sentences with five types of emotion, including neutral, anger, happiness, sadness, and fear, with two utterances for each type. Every utterance was judged by a group and those unqualified were reproduced. The eventual corpus has a total size equal to 1,440 utterances. All of them were recorded in a typical office environment with 16 kHz sampling rate and 16 bits per sample. The SNR value is roughly 30dB. Because the main purpose of this research is to investigate the feature parameters, we assumed the correct transcriptions of each utterance are known in advance. The tool HTK was applied to find the starting/ending time of each syllable in every utterance. This is particularly important to investigate the syllable-level and word-level features. We randomly chose two thirds of the database for training and the remaining for evaluation, while keeping the distribution of the five types of emotion balanced.

3. Feature extraction

Since every text sentence was produced in five types of emotion, in this research we excluded the linguistic features but focused on the acoustic and prosodic ones, including pitch, power, the first three formant frequencies and their bandwidths, MFCC, durations, and pauses. All of these were evaluated with frames of 25 ms every 10 ms and organized into three levels given below.

Frame-level features: These features were generated using the numeric values of the above feature parameters for all the frames within a particular segment of the sentence. The particular segments can be the keyword, head, tail, or the whole sentence. We extracted the average, variance, and linear-regression slope in these segments, and normalized versions of these values were conditionally computed. A total of 630 feature parameters were obtained.

Syllable-level features: The numeric values of each of the above features parameters were first averaged within each syllable. The average, variance, linear-regression slope, and first-difference for these averaged values were then evaluated across all syllables within each sentence. The duration of syllables and the pauses between them were also included. A total of 229 feature parameters were obtained.

Word-level features: Similar to the syllable-level features except with respect to the words instead of syllables, the numeric values of each of the above feature parameters were first averaged within each word and then derived into a total of 223 feature parameters.

We may conceive that the frame-level features present some kind of continuous behavior while the word-level and syllable-level ones present the discrete behavior. All of the features here are independent of the tones. Another set of tone-dependent features were be analyzed and discussed later on separately.

4. Feature selection

4.1 The significance of the individual features

The total size of the feature parameters is 1082, which introduced serious concerns about computation load. More importantly, some of these features may be noisy, providing little information, even disturbing the classifier. To remove the redundant ones and select the more discriminative ones is therefore a very important step. We thus appealed to data mining techniques by sorting these features according to the information gain calculated by the tool Weka, and then the selection of the features was based on their ranking. Table 1 lists the top 15 features that have the highest information gain.

Quite interesting phenomenon can be observed from Table 1. First, the top three discriminating features are the pitch averages at three levels. This may have to do with the tones in Mandarin. Second, among the top 15 features, 7 are at the frame-level, which carries more complete information. Besides, the syllable-level ones are usually better than word-level ones, due to the monosyllabic structure of Chinese. Also, among the top 15 features, 12 of them are related to the prosody, including pitch, power, and MFCC-13(energy). Prosody certainly plays a very important role here. The other 3 are derived from formant structures, which have to do with not only the phonemic structures of the sentences, but the emotion types as well. However, it still requires more investigations to understand why the third formant is more significant.

Table 1. Top 15 features with the highest information gains.

Rank	I.G.	Feature Name
1	0.4643	Frame-level pitch average
2	0.4583	Syllable-level pitch average
3	0.4512	Word-level pitch average
4	0.4341	Frame-level MFCC-13 average
5	0.4199	Syllable-level F3 average (norm.)
6	0.4028	Frame-level power average
7	0.3886	Word-level MFCC-13 average
8	0.3821	Word-level F3 average (norm.)
9	0.3804	Frame-level power tail average
10	0.3775	Frame-level MFCC-13 tail average
11	0.3729	Syllable-level MFCC-13 average
12	0.3619	Syllable-level power average
13	0.3583	Word-level power average
14	0.3084	Frame-level F3 tail average (norm.)
15	0.3046	Frame-level power slope

We also plotted the distributions of the frame-level pitch and power average in Figure 1. It is clear that anger and happiness are separated with higher pitch and power, while those for sadness and fear are closer to neutral. Considering the twp-dimension emotion space shown in Figure 2, both pitch and power have strong correlation with activation. Sadness and fear, on the other hand, have more to do with other feature parameters. That is why in Figure 1 they are closely overlapped with the neutral type.

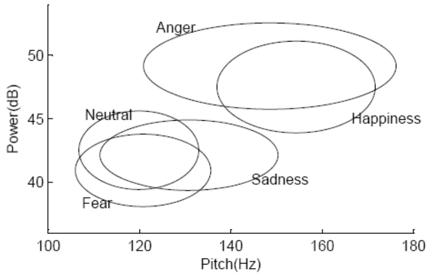

Figure 1: The illustration of the five types of emotion in terms of frame-level pitch and power averages. The centers are the mean values, while the semi-axes are the standard deviations. For simplicity, we include the data of male speakers only.

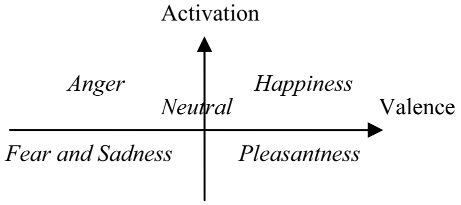

Figure 2: Two-dimension emotion space.

4.2 The significance of feature sets on different levels

To identify the relative importance of the three feature sets on different levels, we introduced a quantitative measure referred to as the significance factor F here. For example, the frame-level feature set has a total of 630 features out of the base pool of 1082 for all three levels, so the share in the base pool is 58%. But out of the top 100 important features 54 are on the frame-level, so its share in the top 100 features is 54%. We define its significance factor F to be 0.93(54/58). That is, for a given feature set S and an integer N,

$$F = \frac{\text{The share of S in the top N features}}{\text{The share of S in the base feature pool}} \qquad (1)$$

The basic idea is that the total numbers of features on the three levels are quite different in the base pool, but the relative importance among them becomes clear with the measure F. Table 2 shows the significance factors of the three levels of feature sets for different N. It can be seen from this table that while frame-level features are still dominant, syllable-level ones are actually vital. This is again consistent with the monosyllabic structure of the language mentioned above.

Table 2. The significance factors F for N=100, 200, and 300.

		Frame	Syllable	Word
Share in Base Pool		58.2%	21.2%	20.6%
N = 100	Share in Top 100	54.0%	28.0%	18.0%
	F	0.93	**1.32**	0.87
N = 200	Share in Top 200	56.5%	28.5%	15.0%
	F	0.97	**1.35**	0.73
N = 300	Share in Top 300	57.7%	30.0%	12.3%
	F	0.99	**1.42**	0.60

5. Emotion recognition results

5.1 Classifier

Recent studies indicate that SVM has superior performance in emotion recognition for speech signals. Consequently, we apply libSVM and its accompanied tool to select suitable parameter for C-SVC method. A grid search algorithms based on cross-validation finds out that, in the case of radial basis kernel function with feature size around 100 to 300, cost=8.0 and γ=0.03125 yielded best recognition accuracy.

5.2 Comparison of feature sets on different levels

To analyze the significance of the feature sets on the three levels, we performed three experiments for emotion recognition. First, we selected different numbers of feature parameters from a single level based on their ranking given by Weka. Figure 3 shows the classification results. It can be found that the frame-level features are the most useful, and syllable-level ones are better than word-level ones. The classification accuracy actually saturated at roughly 86% when 150 features were used. Secondly, we selected the features from the combinations of features on two out of the three levels, namely, frame-level plus syllable-level, syllable-level plus word-level, and word-level plus frame-level. Finally, we also tried to select the features from the ensemble of all three levels. Figure 4 shows the complete results. Apparently the syllable-level features offered better results compared with word-level ones when combined with frame-level ones. The best accuracy achieved here is 89.2% for a total of 250 features.

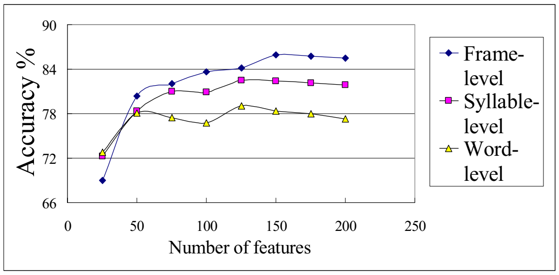

Figure 3: The recognition accuracies yielded by the features on a single level.

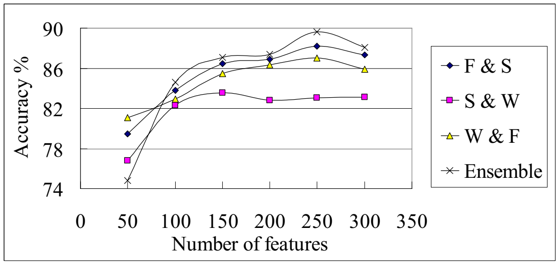

Figure 4: The recognition accuracies of different combinations of feature sets on different levels, where the symbols F, S, and W represent the frame-, syllable-, and word-level respectively.

5.3 The Best Ensemble

All of the above results were based on the ranking given by Weka, but it was not necessarily the best. If we deliberately adjust the share of the three levels of features in the ensemble, better accuracy may be attained. Since there are 146 frame-level features and 71 syllable-level ones in the top 250 features leading to the best result in Figure 4, we performed a grid search on the nearby region of these numbers by varying the share of the two sets while leaving the total feature size at the constant of 250. Table 3 summarizes the results. The peak accuracy of 90.0% was yielded by an ensemble of 140 frame-level features, 65 syllable-level ones, and 45 word-level ones. The confusion matrix of this result is given in Table 5. It can be found from the table that happiness and anger are slightly easier to be confused, so are sadness and fear. This is to some extent in consistency with those shown in Figure 1.

5.4 Tone-dependent features

As mentioned previously, one of the distinctive characteristics of the Chinese language is tones. It is believed that tone-dependent features will be helpful in emotion recognition, but probably more data are needed to verify this concept. Here we try to analyze the tone-dependent features within a limited scope, so the discriminative power of such features can be at least discussed to a reasonable extent. For syllables of each tone we derived a total of 49 features. Again we sorted these features according to their information gains and selected 13 from them, including the average, variance, and slope of the pitch and power, plus the values and normalized versions of formant frequencies and bandwidths. Thus we have 5 sets of tone-dependent features each containing 13 features. To analyze the performance of them, we appealed to the summation of Fisher criteria, which indicates the degree a feature set is able to spread the samples linearly. The summation of Fisher criteria of a feature set T is

$$S.F.C(\mathrm{T}) = \sum_{\mathrm{R}\in\mathrm{T}} \sum_{\mathrm{X}\neq\mathrm{Y}} \left| \frac{\mu_R^X - \mu_R^Y}{\sigma_R^X + \sigma_R^Y} \right| \qquad (2)$$

where X, Y are emotion types, R is a feature in T, μ_R^X and σ_R^X are the mean and standard deviation of R in class X, and so on. The summation is taken over all pairs of five types of emotion and over all features in the set. To verify the discriminative power of different tone-dependent feature sets, we chose the top 87 features from the previous 1082 ones, merged them with the five sets of 13 selected tone-dependent features to result in five sets of 100 features, and then conducted classification based on them. The results are summarized in Table 5. It can be seen from these data that tone 4 has the highest summation of Fisher criteria and the best recognition accuracy, so it is relatively vital in emotion recognition for Chinese.

Table 3: The accuracies under different compositions.

Number of frame-level features	Number of syllable-level features				
	60	65	70	75	80
120	87.1	88.1	87.0	86.5	87.3
130	88.4	89.2	88.5	87.9	86.7
140	88.8	**90.0**	88.0	87.8	87.7
150	88.5	88.9	88.5	87.1	87.4
160	86.1	87.6	87.2	86.7	86.1

Table 4: Confusion matrix of the best result.

Recognized Emotion	Emotion for the speaker					Acc(%)
	Neu.	Ang.	Fear	Hap.	Sad.	
Neutral	90	2	0	2	4	91.8
Anger	2	90	2	8	0	88.2
Fear	1	0	83	2	7	89.2
Happiness	1	4	2	84	0	92.3
Sadness	2	0	9	0	85	88.5
Acc(%)	93.8	93.8	86.5	87.5	88.5	90.0

Table 5: The summation of Fisher criteria and recognition accuracy for different sets of tone-dependent features.

Tone	1	2	3	4	5
S.F.C.	21.4	19.9	21.0	**23.0**	20.9
Acc(%)	84.5	84.1	84.5	**85.7**	84.6

6. Conclusion

When only the feature set on a single level is taken into account, the frame-level features outperform the others, and the syllable-level ones perform better than the word-level ones. This is consistent with our intuition that the frame-level features provide more complete information. The combination of frame-level plus syllable-level feature sets has very close classification accuracy to the ensemble of all three levels, indicating the importance of frame-level and syllable-level features while depreciating the role of word-level ones. This is consistent with the monosyllabic structure of Chinese language mentioned previously.

The best classification accuracy was achieved by the ensemble of 140 frame-level features, 65 syllable-level ones, and 45 word-level ones. The vital roles of frame-level and syllable-level features are verified again. Besides, the preliminary results showed that tone 4 may play significant role in the emotion, although more investigation is still needed to achieve a stronger conclusion.

The features used in this research do not always work well in distinguishing between happiness and anger, both expressing strong feeling, and between fear and sadness, both regarded as weak emotion. In addition, since prosodic features are most suitable for measuring the activation level of emotion, the performance of them in two-dimension emotion space classification is our further interest.

7. Acknowledgements

The authors would like to thank Prof. Chu of National Taiwan University for her help in finding the students of the Department of Drama and Theatre to produce the utterances in the emotional speech corpus. They also thank Mr. I. H. Wang for the help in building the speech corpus, and Mr. C. K. Lin for the help in pre-processing the data. This paper is supported by NSC of Taiwan, R.O.C. under grant 94-2815-C002-089-E.

8. References

[1] Elizabeth Shriberg, "Spontaneous speech: how people really talk and why engineers should care", *Interspeech 2005, Lisbon, Portugal,* pp. 1781-1784, 2005

[2] Valery A. Petrushin, "Emotion recognition in speech signals: Experimental study, development, and application", *Proc. ICSLP 2000*, 2000

[3] Schuller, B., Muller, R., Lang, M., and Rigoll, G., "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles". *Interspeech 2005, Lisbon, Portugal,* pp. 805-808, 2005

[4] Ian H. Witten, Eibe Frank,"Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[5] Luengo, I., Navas, E., Hernaez, I., and Sanchez J., "Automatic recognition using prosodic parameters," *Interspeech 2005, Lisbon, Portugal,* pp. 493-496, 2005

[6] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines, 2005, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[7] Kienast, M., Sendlmeier, W.F., "Acoustical analysis of spectral and temporal changes in emotional speech", *Proceedings of the ISCA ITRW on Speech and Emotion, Newcastle*, 5-7 September 2000, Belfast, Textflow, pp. 92-97

[8] Vogt, T., André, E., "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition", *IEEE International Conference on Multimedia & Expo (ICME 2005)*, 2005

[9] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for SVMs", *Advances in Neural Information Processing Systems, volume 13.* The MIT Press, 2001.

[10] D Ververidis, C Kotropoulos, "A State of the Art Review on Emotional Speech Databases", *I Pitas - Proc. 2004 IEEE Int. Conf. Acoustics, Speech and Signal,* 2004