RISP
NCSP'16
International Workshop on
Nonlinear Circuits, Communications and Signal Processing
Honolulu, Hawaii, USA, March 6-9, 2016

Automatic Speech Emotion Recognition in Chinese Using a Three-layered Model in Dimensional Approach

Xingfeng Li and Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology
1–1, Asahidai, Nomi, Ishikawa, 923–1292 Japan
E-mail: lixingfeng@jaist.ac.jp, akagi@jaist.ac.jp

Abstract

In this paper, we improve the speaker independent emotion classification of the CASIA Mandarin emotional speech corpus, which is provided by Chinese-LDC covering four basic emotions, angry, happy, neutral, and sad. We achieve this by restoring the human processing on emotion perception with a three layered model. The three layered model is constructed with acoustic features in the bottom layer, semantic primitives in the middle layer, and emotion dimensions in the top layer. To implement the proposed system, we first investigate the optimal acoustic feature set that is related to each emotion dimension, then mapping these acoustic features to emotion dimensions through the estimated semantic primitives by using Fuzzy Inference System (FIS). In addition, with the highly predicted emotion dimensions, emotional classification procedure is addressed using the knowledge of commonalities and differences of humans emotion perception. The experimental results show that improved estimation performance compared to previous study is furnished.

1. Introduction

Speech is the fastest and the most natural means of communications among human subjects in our daily life. It can be seen as two-channel mechanisms, involving actual meaning of the communication and several prosodic distinctions [1]. The linguistic channel is the main focus for research in the past which refers to the process of converting human speech into a sequence of texts. However, despite the great process made in speech recognition, having a natural interaction between man and machine is still a challenging. One of reasons is that machine cannot understand emotions of speakers. Hence, researchers now pay more attention to the study of the second implicit channel: speech emotion recognition, which aim at extracting emotional states of a speaker from his or her speech.

To present emotion, linguists have defined categories of the emotional states, most encountered in our lives. Many researchers followed the palette theory, which sates that any emotion can be decomposed into primary emotions. These are the most obvious and distinct emotions, namely anger, disgust, fear, joy, sadness, and surprised [2]. Whereas, besides emotional categories, speech is always colored with changeable intensity of a certain emotion. Obviously, a single label or any small number of inventories could not fully reflect the varied emotions conveyed in the daily communication [3]. For this reason, it is widely advocated the use of dimensional approach to characterize human emotion, where emotional states are not assorted into one of the primary emotional categories, but estimated on a numerical scale in a multiple dimensional space [4].

Up to now, a hybrid speaker independent emotion recognition system has been achieved by using spectral and prosodic feature set in combination with GMM-based subsystem and/or SVM-based subsystem from the point of view of categorical model [5]. In this paper, we improve the emotional classification result for the same Chinese database from the point of view of dimensional approach by restoring the human processing on emotion perception using a three layered model, which consists of acoustic features in the bottom layer, semantic primitives in the model layer, and emotion dimensions in the top layer.

This paper is structured as follows: firstly, database and elements of the three layers are introduced in section 2, afterwards selection method of optimal feature set to each emotion dimension and system implementation are presented in section 3. Finally, the contribution ends with result comparison of proposed system and related research, discussion, and conclusion in sections 4 and 5.

2. Data Source

The CASIA emotional speech dataset is a Mandarin emotional speech dataset developed by the Institute of Automation affiliated with Chinese Academy of Sciences [6]. It was recorded by 2 male and 2 female speakers with neutral and 5 categories of acting emotions, including angry, happy, sad, fear and surprise. The contents consist of two parts, naming dominant and spontaneous. The utterances of the domi-

nant contents have at least one dominant word, e.g. ”anger” or ”annoyed” for angry, ”pleased” or ”joyful” for happy, and ”sad” for sad, etc. There are 100 utterances for each emotion. The utterances of the spontaneous contents are picked from news articles, conversations and essays without emotional-rich words. There are 300 utterances in this part. Each speaker utters (100 + 300)×6 = 2400 sentences in total. Among them, the spontaneous speech can be taken as content aligned data and the dominant speech are content discriminative. Between these two categories, the neutral speech can be used as a complete dataset of 400 utterances because there is no dominant word for neutral. The speech waveforms were recorded utterance by utterance, sampled at 16 kHz, digitized using 16-bit data and stored in single channel files.

In this paper, 200 utterances from 4 speakers covering 4 emotions (neutral, happy, sad, and anger) are selected as initial step, 50 utterances for each emotion. But due to almost all the existing emotional speech databases do not well enough simulate emotions in a natural and clear way, and additionally, our system is proposed by imitating human emotion perception processing on spoken utterances. Therefore, all 200 selected utterances were annotated again using the categorical method by 11 Chinese native speakers (5 females and 6 males). Experimental results finally show that 68 utterances were recognized as neutral speech, 30, 50, and 50 utterances were grouped as happy, anger, and sad respectively. 2 spoken utterances can not be classified into any one of the above four emotional categories. Hence, 198 human-annotated utterances are used in our work ultimately.

2.1 Acoustic Features

In this research, as inputting parameters, acoustic features are a very crucial part to be studied. Therefore, the most relevant acoustic features which have been successfully used in related studies were selected. These used acoustic features can be grouped into five subgroups: 4 F0-related features, 4 Power envelop-related features, 5 power spectrum-related features, 3 duration-related features, and 5 voice quality-related features. Eventually, a set of 21 extracted acoustic features were collected using STRAIGHHT following [7].

2.2 Semantic Primitives and Emotion Dimensions Evaluation

The human perception model as described by Scherer [8] is a multiple perceptual processing, it was adopted by Elbarougy and Akagi as a three layered model [7], in which they assumed that human emotional perception not directly come from a change of acoustic features, but from a smaller perception, namely adjectives describing emotional voice. These smaller percepts or adjectives can be used to recognize emotions of the speaker. These adjectives are: Bright, Dark, High, Low, Strong, Weak, Calm, Unstable, Well-modulated, Monotonous, Heavy, Clear, Noisy, Quiet, Sharp, Fast, and Slow. They are originally from the work [9]. For the evaluation, human listening test is carried out by 10 Chinese native speakers (5 females and 5 males). Each emotional speech was evaluated 17 times by subjects, once for one semantic primitive. Subjects were asked to rate each of the 17 semantic primitives on a five-point scale: ”1-Does not feel at all”, ”2-Seldom feels”, ”3-Feels a little”, ”4-feels”, ”5-Feels very much”.

Moreover, in this study affective content is characterized with two emotional dimensions, valence and activation. So the CASIA emotional database should be annotated using the dimensional approach by doing human listening test by the same subjects in semantic primitives experiment. For emotion dimensions evaluation, a five-point scale -2, -1, 0, 1, 2 was used: valence (from -2 very negative to +2 very positive) and activation (from -2 very calm to +2 very excited). There were two sessions, one for each emotion dimension. Subjects were asked to evaluate one emotion dimension for the whole database in one session.

Before starting the two experiments, the basic theory of semantic primitive and emotion dimension are explained to subjects. Then they took a training session to listen to an example set composed of 20 utterances, which covered the five-point scale. The purpose of this training set is to let subjects understand these adjectives. All stimulus were presented randomly through binaural headphones at a comfortable sound pressure level in a soundproof room. Subjects were asked to evaluate their perceived impression from the way of speaking, not from the content itself, and then choose a score on the five point scale for each adjective. Moreover, the inter-rater agreement was measured by means of pairwise Person’s correlation between two subjects’s rating, for each semantic primitives respectively. Subject’s rating with correlation value greater than 0.8 will be used. The average of the subjects’s rating for each adjective and emotion dimension was calculated per utterance.

3. Features Selection and System Implementation

Reduction of features also often leads to a higher classification performance. To find features highly relevant to each emotion dimension, a correlation-based features selection procedure is performed. Firstly, correlation coefficients between the elements of the top layer (emotion dimension) and the middle layer (semantic primitive) were calculated by correlation function after [7]. Subsequently, the correlation coefficients between elements of the middle layer (semantic primitive) and the bottom layer (acoustic feature) are calculated in a similar way. Thereafter, different highly relevant features have been respectively selected for each semantic primitive and emotion dimension after experimentation. We found that, all 17 semantic primitives are of critical impor-

Table 1: Related acoustic features for each adjective($\surd$: used; $-$: not used)

Feature	MH_A	MH_I	MH_O	MH_U	F0_RS	F0_HP	PW_RTH	SP_F2	SP_F3	SP_TL	SP_SB	DU_TL	DU_CL
Bright	$\surd$	$-$	$\surd$	$-$	$\surd$	$\surd$	$\surd$	$\surd$	$-$	$-$	$\surd$	$\surd$	$\surd$
Dark	$\surd$	$-$	$\surd$	$-$	$\surd$	$\surd$	$\surd$	$\surd$	$-$	$-$	$\surd$	$\surd$	$\surd$
High	$\surd$	$-$	$\surd$	$-$	$\surd$	$\surd$	$\surd$	$\surd$	$-$	$\surd$	$\surd$	$\surd$	$\surd$
Low	$\surd$	$-$	$\surd$	$-$	$\surd$	$\surd$	$\surd$	$\surd$	$-$	$-$	$\surd$	$\surd$	$\surd$
Strong	$\surd$	$-$	$\surd$	$-$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
Weak	$\surd$	$-$	$\surd$	$-$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
Calm	$\surd$	$-$	$\surd$	$-$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
Unstable	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
Wellmodulated	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$-$	$\surd$	$\surd$
Monotonous	$-$	$\surd$	$-$	$\surd$	$\surd$	$\surd$	$\surd$	$-$	$\surd$	$\surd$	$-$	$\surd$	$\surd$
Heavy	$\surd$	$-$	$\surd$	$-$	$\surd$	$\surd$	$\surd$	$-$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
Clear	$\surd$	$-$	$\surd$	$-$	$-$	$\surd$	$\surd$	$-$	$\surd$	$\surd$	$-$	$\surd$	$\surd$
Noisy	$\surd$	$-$	$\surd$	$-$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
Quiet	$\surd$	$-$	$\surd$	$-$	$\surd$	$\surd$	$\surd$	$\surd$	$-$	$-$	$\surd$	$\surd$	$\surd$
Sharp	$\surd$	$-$	$\surd$	$-$	$\surd$	$\surd$	$\surd$	$-$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
Fast	$\surd$	$-$	$\surd$	$-$	$\surd$	$\surd$	$\surd$	$-$	$-$	$\surd$	$\surd$	$\surd$	$\surd$
Slow	$\surd$	$-$	$\surd$	$-$	$\surd$	$\surd$	$\surd$	$-$	$-$	$-$	$\surd$	$\surd$	$\surd$

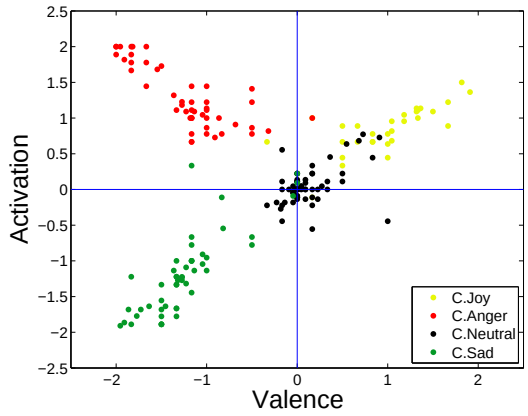
tance to valence and activation. Whereas, for each semantic primitive, the best optimal feature set is adjective-dependent. Table 1 has elucidated 13 acoustic features, which are regarded as highly correlated features to semantic primitives. Those with checking symbols represent selected and used acoustic features, differently, acoustic features with horizontal lines are helpless ones.

For constructing proposed speech emotion recognition (SER) system, FIS is utilized to establish the mapping from acoustic features to emotion dimensions through semantic primitives. To obtain estimated emotion dimensions, first of all, each of the 17 semantic primitives in the middle of three layered model should be predicted separately from specified checking acoustic features using FIS. Beyond that, the estimation of emotion dimension can be done from 17 estimated adjectives in the previous part with another FIS.

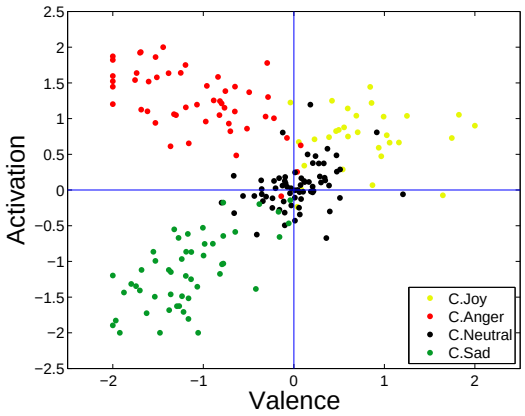
4. Classification

This section addresses the comparison of classification results of one previous study and our proposed approach. Firstly, we present the predicted emotion dimensions for all utterances. Then, scatter plots in dimensional space from human responses and system's estimations are discussed. Besides, an emotional classification method in [9] is applied in valence-activation space based on the accurate estimation of emotion dimension. Ultimately, improved recognition rates are attested compared with previous strategy in [5].

4.1 Emotion Dimensions Estimation



(a) Human responses



(b) Estimated by the proposed system

Figure 1: Distribution of the speech utterances in the valence-activation space

Figure 1(a) and Figure 1(b) have greatly clarified the distribution of emotional speech utterances in the valence-activation space. For the reason that proposed system in this paper is aiming at restoring the human emotion perceptual processing on affective speech. To evaluate how closed the predicted values to the human responses. The mean absolute error of human responses and system's estimations is calculated following Eq. (1).

$$MAE^{(j)} = \frac{\sum_{i=1}^{n} \left| \widehat{x}_n^{(i)} - x_n^{(i)} \right|}{n} \tag{1}$$

where $j \in \{Valence, Activation\}$, $\widehat{x}_n^{(i)}$ is output of the proposed system, $x_n^{(i)}$ is the human responses by using human subjects, and n is the number of utterances in our database. Simultaneously, to obtain a baseline for evaluating the estimating precision of proposed system, two native Chinese speakers (1 male and 1 female) are asked to assess valence and activation for all utterances eight times. Afterwards, the mean standard deviation (MSTDEV) among human responses from subjects in the listening experiments is seemed as baseline,

Table 2: Comparison of human responses and system estimation

Emotion Dimension	Valence	Activation
MSTDEV male	0.37	0.28
MSTDEV female	0.36	0.32
MAE	0.40	0.28

which calculated following Eq. (2).

$$MSTDEV^{(j)} = \frac{\sum_{i=1}^{n} \sqrt{\frac{\sum_{j=m}^{n1} (x_m - \overline{\mu})^2}{n1}}}{n} \quad (2)$$

where j and n have the same definition as Eq. (1), $n1$ is the number of times of listening tests ($n1$=8), and $\overline{\mu}$ is the average value of one utterance from per subject for 8 times.

As seen from Table 2, MAE of valence from the system is greater than MSTDEV of human subjects from experiments, in which the biggest difference is 0.04. But, in view that the difference of activation dimension of MSTDEV between male and female subjects has also reach to 0.04. This can powerfully proved that this small differences in valence of MAE and MSTDEV is acceptable and rational. Furthermore, MAE of activation from proposed system is less than MSTDEV of female and male from listening tests, which reflects that all estimated values of activation are within the scope of human response. Hence with verified results, we can conclude that, the proposed three-layered model has well restored the processing of human emotional perception. The system can easily predict emotion dimensions as human percept.

4.2 Emotion Classification

With the precise estimation of emotion dimensions, our former investigation in [9] has shown that the direction and degree from neutral position to other emotional sate's position in valence-activation space can be used as distinct features to recognize basic emotions. So that using obtained accurate emotion dimensions in previous section, in this part, according to Eq. (3) and Eq. (4) we will extract these two features to detect emotion.

$$angle = \arctan(\frac{y_E - y_N}{x_E - x_N}) \quad (3)$$

$$d(E, N) = \sqrt{(x_E - x_N)^2 + (y_E - y_N)^2} \quad (4)$$

where (x_E, y_E) is the center position of the emotional state E, and (x_N, y_N) is that of the neutral state N.

The Support Vector Machine classifier with 10-fold cross-validation method is used to map the extracted direction and degree into emotional categories. We finally achieve an improvement of 23% reduction of classification error in comparison to a previous study on Chinese speech emotion recognition [5].

Table 3: Recognition Rates for emotional classification of proposed system and previous approach after [5]

R.R. [%]	Neut.	Hapi.	Anger	Sad	Ave.
Proposed	96	97	88	92	93.25
Y. Zhou, J. Li(IEICE Trans.,2010)	98	82.25	90.25	94.5	91.25

5. Conclusion

In this paper, we proposed a three layered model based SER system, which has the ability to precisely estimate emotion dimensions just as well as humans do. From accurate estimated positions of different emotions in dimensional space, motivated by the knowledge of commonalities and differences on human emotional perception, two features in valence-activation space, i.e. direction and degree are mapped into basic emotion categories. We attained 23% noisy reduction rate of emotional classification of the former related study on Chinese SER [5] using the proposed system.

References

[1] K. Scherer, "Vocal communication of emotion: A review of research paradigms," Speech Communication, 40(1–2), pp. 227-256, April, 2003.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, S. Kollias, W. Fellenz, and J. Taylor, Emotion recognition in human–computer interaction, IEEE Signal Process. Mag. 18(2001) 32–80.

[3] I. Albrecht, M. Schroder, J. Haber, and H. P. Seidel, "Mixed feelings: Expression of non-basic emotions in a muscle-based talking head," Virtual Reality, Vol. 8(4), pp .201–212, 2005.

[4] D. Wu, T. D. Parsons, and S. Narayanan, "Acoustic features analysis in speech emotion primitives estimation," Proc. InterSpeech 2010, pp. 785–788, 2010.

[5] Y. Zhou, J. Li, Y. Sun, J. Zhang, Y. Yan, and M. Akagi "A hybrid speech emotion recognition system based on spectral and prosodic features." IEICE Transactions on Information and Systems 93.10 (2010): 2813–2821.

[6] "Mandarin emotional speech corpus," http://www.chineseldc.org/doc/CLDC-SPC-2005-010/intro.htm, 2005. Institute of Automation, Chinese Academy of Sciences.

[7] R. Elbarougy, and M. Akagi. "Improving speech emotion dimensions estimation using a three-layer model of human perception," Acoustical Science and Technology 35.2 (2014): 86-98.

[8] K. Scherer,"Personality inference from voice quality: The loud voice of extroversion." European Journal of Social Psychology 8.4 (1978): 467-487.

[9] X. LI, and M. Akagi, "Toward Improving Estimated Accuracy of Emotion Dimensions in Bilingual Scenario based on Three-layered Model," "Oriental COCOSDA/CASLRE," Paper85, Shanghai, China, 2015.