Effect of Dimensional Emotion in Discrete Speech Emotion Classification

Jian Huang[1,3], Ya Li[1], Jianhua Tao[1,2,3]

[1]National Laboratory of Pattern Recognition,
[2]CAS Center for Excellence in Brain Science and Intelligence Technology,
Institute of Automation, Chinese Academy of Sciences, China
[3]School of Computer and Control Engineering, University of Chinese Academy of Science, China
{jian.huang, yli, jhtao}@nlpr.ia.ac.cn

Abstract

Speech emotion recognition is crucial for future human-machine interfaces to be more natural and harmonious. On the basis of close relationship between dimensional emotion and discrete emotion, this paper investigates how dimensional emotion information contributes to discrete speech emotion classification. The experimental results based on discrete emotion labels and manual dimensional emotion ratings of IEMOCAP database, show that the supplementary of dimensional ratings-based features can improve discrete classification performance significantly. In addition, dimensional information provides a substantial boost to "angry" and "happy" than "neutral" and "sad" specially. To integrate dimensional information into a fully automatic emotion recognition system, binary dimensional emotion predication experiments are constructed, and the results show their feasibility. Automatic dimensional emotion prediction is added to discrete speech emotion classification system, which could obtain 8% accuracy improvement over acoustic-only baseline results, also 3% better than other methods.

Index Terms: discrete speech emotion classification, binarization, automatic dimensional emotion prediction

1. Introduction

Speech emotion recognition (SER) is becoming more and more essential for many applications related to human-machine interactions, such as speech recognition systems and spoken dialogue systems [1]. The increasing application of SER makes it a core component in the next generation of computer system, in which natural human machine interface requires a good appreciation of the emotional state of a user [2]. However, robust and accurate SER is still a challenging problem due to complex factors such as the variations of speakers and contents, and environment distortion [3][4].

SER system consists of feature extraction step followed by classification step. Various spectral and prosodic features have been applied to SER in the literature [5][6]. Speech emotional features, which are different types of low-level descriptors (LLD) [7], are extracted based on frame-level. Then, various functionals, such as mean, maximum, minimum, variance etc., are applied to these LLDs across utterance-level to get final fixed speech emotional feature vector.

In terms of the emotion model, most researches concentrate on two major emotion models [8], discrete emotion model, dimensional emotion model. Discrete emotion model describes an emotion state as discrete labels such as "sad", "happy" etc. It is intuitive and simple but difficult to express complex affective states. Dimensional emotion model considers an emotion state as a point in a continuous dimensional space. Hence, dimensional emotion model can model subtle, complicated, and continuous affective behavior, but it is hard to understand. Typically, an emotion state is described by three dimensions: valence-activation-dominance [8]. Some researchers also use two dimensional arousal-valence space to denote dimensional emotion state [9].

Discrete model and dimensional model are two different descriptions of emotion, which have their strengths and weaknesses. They have a close connection with each other, as shown in Figure 1 [8]. For example, "happy" lies in first quartile of arousal-valence space that arousal and valence are great positive. Conversely, the region which owns great positive arousal and valence corresponds to "happy". Valence indicates a measure of pleasure and has a stronger correlation with the semantic context of what is spoken than how it is prosodically spoken [10][11], which relates "happy", "excited" et al. Activation is described as a conscious affective experience based on a varied degree of subjective mental activation. Increased activation is represented by stronger acoustic formant intensity and fundamental frequency vocal tension, resulting in a perceivable rise in pitch [12], which relates "sad", "depressed" et al. Dominance is an individual's perceived assertiveness authority, and aggressive vocal characteristics. A high degree of dominance is useful in parenting, emergency, or threatening situations, which relates "surprise", "fear" et al.

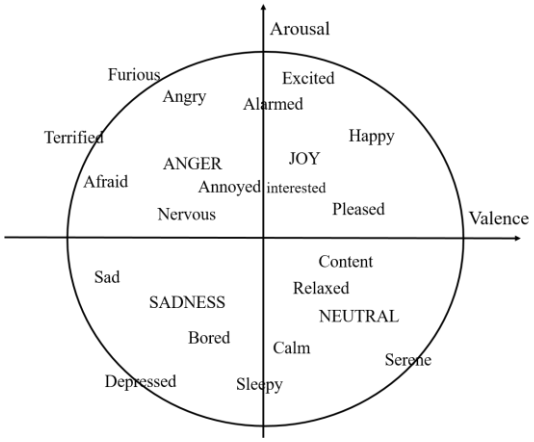

Figure 1: *distribution of the discrete emotions in dimensional arousal-valence space*

Christie and Friedman [13] use discriminant function to describe the location of discrete emotions within dimensional affective space for autonomic nervous system and find valence more accurately portray the structure of self-reported emotion. Barrett [10] points that dimensional emotion is related to the co-occurrences of discrete emotional state. He also suggests that one of emotion theories may not apply to all people due to difference of individuals. Stasak et al. [14] explore that features derived from dimensional affect ratings carry complementary information to conventional acoustic features and achieve the performance gains when classifying depression via speech. Inspired by Stasak's work, we explore the benefits features derived from dimensional ratings could introduce to discrete emotion classification system.

As mentioned above, there is close relationship between dimensional emotion and discrete emotion. In this study, we investigate the effect of dimensional emotion information in discrete speech emotion classification. Practically, continuous dimensional emotion ratings are converted to binary value by setting suitable threshold, and then the binarization value combined with acoustic features is input to classifier. Dimensional ratings-based features are utilized to improve the performance of discrete speech emotion classification. Lastly, a fully automatic discrete speech emotion classification system based on binary dimensional emotion prediction is carried out.

In the following, Section 2 presents database and acoustic features. Section 3 briefly introduces the proposed method. Section 4 describes experimental results and discussion. Section 5 concludes this paper.

2. Database and Feature Set

2.1. Database

In this study, we use Interactive Emotional Dyadic Motion Capture (IEMOCAP) [15] to evaluate our proposed method. This corpus has approximately 12 hours of audiovisual data, including video, speech, motion capture of face, and text transcriptions [15]. It has 10 professional actors (5 males and 5 females) acting in two different scenarios: scripted play and spontaneous dialog, in their dyadic interactions. Each interaction is around 5 minutes in length, and is segmented into sentence levels. Every sentence corresponds a discrete emotion label and a three-dimensional point rating. We use four emotion categories in this study: "angry", "happy", "sad", and "neutral", similar to most prior studies using this corpus. Note that we merge "happy" and "excited" in the original annotation into the "happy" class. Meanwhile, we also use their corresponding three-dimensional ratings. Only the utterances with majority agreement are used in the experiments. In total we use 5,531 utterances. The class distribution is: 20.0% "angry", 19.6% "sad", 29.6% "happy", and 30.8% "neutral". The experiment protocol is leave-one-speaker-out which means there is no speaker overlap between training and testing set.

2.2. Acoustic Features

We add the first dimension of the Mel Filterbank Cepstral Coefficients (MFCC 0), the first order derivatives of all the LLDs, as well as the second order derivatives of MFCC 0-14 to the 65 LLDs of the INTERSPEECH 2014 Computational Paralinguistics Challenge [16]. The resulting features 147 LLDs extracted by openSMILE [17] are listed in Table 1, where △ denotes the first order [18].

Table 1: *Low-Level Descriptors (LLDs) features*

8 energy related LLD
Sum of auditory spectrum (loudness) +Δ Sum of RASTA-filtered auditory spectrum +Δ RMS Energy +Δ, Zero-Crossing Rate +Δ
127 spectral LLD
RASTA-filtered auditory spectrum, bands 1-26 (0-8 kHz) +Δ MFCC 0-14 +Δ + ΔΔ, Spectral Centriod +Δ Spectral energy 259-650 Hz, 1k-4kHz +Δ Spectral Roll Off Point 0.25,0.50,0.75,0.90 +Δ Spectral Flux, Entropy, Variance, Skewness, Kurtosis, Slope, Psychoacoustic Sharpness, Harmonicity +Δ
12 voicing related LLD
F0 (SHS + Viterbi smoothing), Probability of voicing +Δ Logarithmic HNR, Jitter (local, delta), Shimmer (local) +Δ

3. Proposed method

3.1. Discrete Emotion Classification Using Binary Dimensional Emotion Rating

Dimensional ratings information combined with acoustic features are utilized to improve discrete speech emotion classification. We choose three functionals, namely maximum, minimum and mean to explore the distinction of different functionals. These functionals are applied to LLDs across utterance-level in acoustic feature extraction. In IEMOCAP, an audio sample corresponds to a discrete emotional label and a continuous three-dimensional point rating. To simplify the complexity of continuous value, dimensional ratings are converted to binary value, which is easily applied for automatic binary dimensional emotion prediction. Binarization dimensional ratings are complemented to baseline acoustic features to investigate discrete emotion classification. The influence of single dimension from valence, activation or dominance is also explored by adding its ratings to acoustic features individually. Three conventional classifiers, logistic regression, random forest and SVM are utilized in this study.

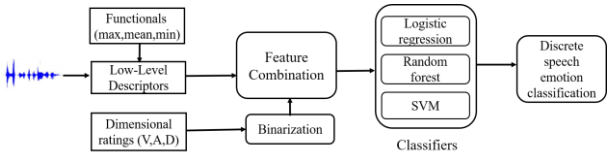

Figure 2: *System framework of proposed method. All experiments evaluate LLD, binary dimensional emotion ratings and classifiers with various combinations.*

3.2. Discrete Emotion Classification Based on Automatic Dimensional Emotion Prediction

Unluckily, the system of Figure 2 can't be applied in automatic discrete emotion classification because of no accessibility of manual emotion ratings in actual applications. To develop the function of dimensional information, binary dimensional emotion predication experiments are constructed, and the results show their feasibility. Therefore, automatic prediction of binary dimensional ratings module is introduced to replace manual ratings, as shown in Figure 3. In this modified system, whose input is only speech, dimensional emotion features predicted by baseline acoustic features are

combined with acoustic features for discrete speech emotion classification.

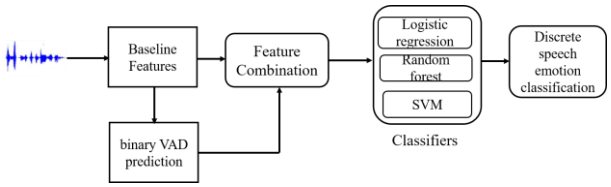

Figure 3: *Automatic discrete emotion classification using predicted binary dimensional rating-based features and baseline features*

4. Results and Discussion

4.1. Discrete Emotion Classification

After extracting 147 dimensional frame-level acoustic features shown as Table 1, three functionals mean, maximum and minimum are respectively applied across utterance-level to get speech emotional feature vector. The classification accuracies are shown in Table 2, which indicates that mean functional can achieve better performance than other functionals. In the following, we combine mean functional with others, for example the combination of mean functional and maximum functional, which are represented by "Mean_max" in Table 2. And "LR", "RF" and "SVM" represent logistic regression, random forest and SVM respectively. In random forest, the number of trees vary from range [50, 500] and the maximum depth of the tree vary from range [5, 10]. SVM adopts RBF kernel and searches the parameter C and gamma from range [0.0001, 1000]. Experimental results indicate that single mean functional has achieved good performance and the supplement of maximum and minimum functional only bring a little improvement. Besides, random forest achieves best performance among three classifiers.

Table 2: *Classification accuracies of different functionals using only acoustic features*

Accuracy	LR	RF	SVM
Mean	0.502	0.562	0.518
Max	0.492	0.515	0.517
Min	0.491	0.524	0.513
Mean_max	0.519	0.565	0.526
Mean_min	0.506	0.564	0.533
Mean_max_min	0.537	**0.566**	0.546

4.2. Binary Dimensional Emotion Prediction

The dimensional ratings are continuous value ranging from 1 to 5.5 in IEMOCAP. The median of all ratings is set as the threshold, then we compare each dimensional rating value with the median to give a binary label:'0' or '1'. '0' denotes below the median value and '1' means above the median value. Therefore, origin continuous dimensional ratings are refined as binary ones, which reduces a regression problem into a classification one. Same as experimental configurations of section 4.1, we get binary three dimensional prediction results as shown in Table 3, which achieve satisfied results. In particular, the accuracy of activation dimension is 0.784, higher than other dimensions. We also observe that optimal functional are various in different dimension and random forest can achieve good performance compared with other classifiers.

4.3. Effect of Binary Dimensional Emotion Ratings

As mentioned earlier, prior studies indicate a close connection between dimensional emotion and discrete emotion [10][13]. We conjecture that the supplementary of dimensional information can guide discrete emotion classification. To verify whether binary dimensional ratings are beneficial to discrete speech emotion classification, we combine binary three-dimensional ratings with acoustic features, according to Figure 2. The experimental results, are shown in Table 4, indicate that the accuracy of the proposed method is 0.741, which uses SVM classifier based on mean functional. The results verify the advantage of using binary dimensional ratings-based features. In a result, it provides us the directory to improve classification performance considering supplementary of binary dimensional information.

Then, we explore the influence of every dimension on discrete speech emotion classification. Binary dimensional emotion ratings from valence, activation and dominance are added to acoustic features individually. The experimental results, shown in Figure 4(b)(c)(d), indicate that the supplementary of valence dimension can achieve best performance among three dimensions, which is comparable to Figure 4(a). Among these classifiers, we observe interesting phenomenon from Table 2 and Figure 4 that SVM can achieve best performance if valence dimensional ratings exist, otherwise random forest does.

Table 4: *Classification accuracies of different functionals using acoustic and three-dimensional ratings-based features*

Accuracy	LR	RF	SVM
Mean	0.702	0.703	**0.741**
Max	0.696	0.680	0.724
Min	0.712	0.687	0.728
Mean_max	0.721	0.688	0.733
Mean_min	0.710	0.692	0.723
Mean_max_min	0.721	0.672	0.732

Table 3: *Binary dimensional emotion prediction accuracies of three dimensions with different functionals using acoustic features*

Accuracy	Valence			Activation			Dominance		
	LR	RF	SVM	LR	RF	SVM	LR	RF	SVM
Mean	0.674	0.677	0.645	0.737	0.757	0.725	0.654	0.670	0.677
Max	0.675	0.669	0.679	0.766	0.765	0.757	0.668	**0.677**	0.664
Min	0.679	0.675	0.674	0.765	0.773	0.765	0.664	0.655	0.656
Mean_max	0.662	**0.686**	0.674	0.772	0.778	0.760	0.669	0.669	0.667
Mean_min	0.671	0.683	0.678	0.774	**0.784**	0.765	0.661	0.658	0.662
Mean_max_min	0.682	0.680	0.683	0.778	0.771	0.776	0.670	0.658	0.671

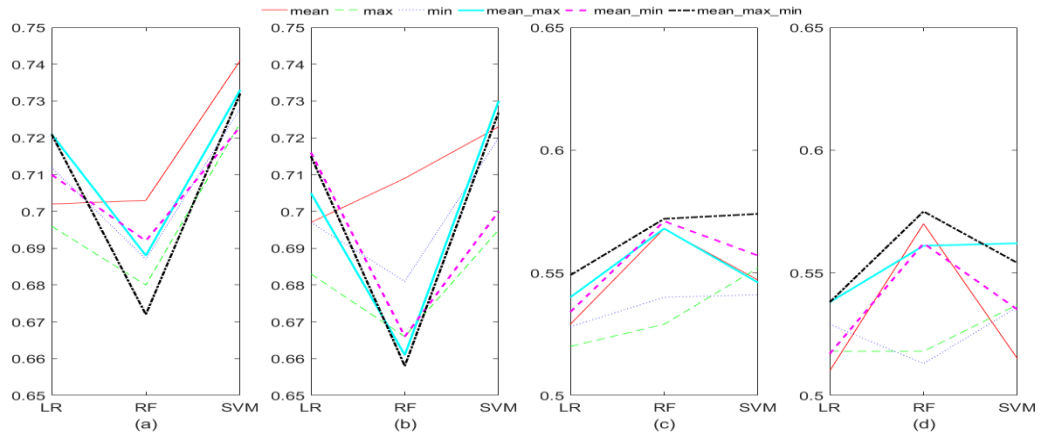

Figure 4: *Classification accuracies of different functionals using acoustic features and dimensional ratings-based features. (a) acoustic features and binary three-dimensional emotion ratings (b) acoustic features and binary valence dimensional ratings (c) acoustic features and binary activation dimensional ratings (d) acoustic features and binary dominance dimensional ratings*

Having realized the contribution of dimensional ratings-based features, we explore their influence on specific speech emotion category. Table 5 shows each category accuracy of only acoustic features from Table 2 and acoustic features with dimensional ratings from Table 4. When only using acoustic features, "neutral" and "sad" have better performance than "angry" and "happy". With the supplementary of dimensional ratings-based features, the performance of "angry" and "happy" is improved significantly, but the performance of "neutral" and "sad" is decreased slightly. Therefore, dimensional emotion information provides a substantial boost to "angry" and "happy" particularly in IEMOCAP.

Table 5: *Each emotion category accuracy of only acoustic features from Table 2 and acoustic features with dimensional ratings from Table 4*

Model	Angry	Happy	Neutral	Sad
Acoustic features	0.518	0.391	0.695	0.710
Acoustic features with dimensional ratings	0.788	0.894	0.602	0.649

4.4 Effect of automatic Dimensional Emotion Prediction

The insights in Section 4.3 point that the supplementary of binary dimensional ratings can improve classification performance largely. However, this method can't be applied in automatic discrete emotion classification due to no accessibility of manual ratings in actual applications. In section 4.2, the experimental results have verified the feasibility of binary dimensional emotion prediction. To explore whether the effectiveness of binary manual dimensional ratings could be replicated to an automatic process, automatic binary dimensional prediction module is added to the system, as shown in Figure 3. Feature combination includes baseline acoustic features and dimensional ratings-based features generated from baseline acoustic features as well.

Section 4.2 has determined optimal functional and classifier for binary dimensional emotion prediction. Similarity, Section 4.3 chooses optimal functional and classifier of discrete emotion classification using binary dimensional emotion ratings. We follow these configurations of corresponding module in the modified system. The result raises the accuracy to 0.644, yielding almost 8% improvement over acoustic-only baseline result. Although the performance is declined compared with direct use of manual dimensional ratings, it still achieves better performance than acoustic-only baseline results, which verifies the effectiveness of automatic binary dimensional prediction.

We also compare the proposed method with two other methods of the literature. Xia and Liu [19] propose to combine deep belief network and i-vector space for speech emotion recognition with IEMOCAP, which achieves 0.596 accuracy, yielding 2% improvement compared with standard i-vector. Denoising autoencoder is also utilized to generate robust feature representations for SER [20], whose accuracy is 0.615, yielding 3% improvement compared with 1584 static acoustic features. The above two papers use elaborate design and complex neural networks, which can achieve better performance than baseline acoustic features. However, the proposed method can boost classification performance of SER, achieving 3% better than these methods.

Table 5: *Classification accuracies of proposed method and other methods*

Model	Accuracy
DBN-ivector Framework [19]	0.596
Denoising Autoencoder [20]	0.615
Acoustic features	0.566
Proposed method	0.644

5. Conclusion

Given dimensional ratings information, this study suggests that dimensional ratings-based features can boost the performance of discrete speech emotion classification. Considering the complexity of continuous dimensional ratings and no accessibility of manual ratings in actual applications, we simplify continuous dimensional ratings by converting them to binary ones. The experimental results demonstrate that the supplementary of binary dimensional ratings can achieve large performance improvement, especially for valence dimension. Further, automatic binary dimensional prediction

combined with baseline acoustic features raises the accuracy to 0.644, yielding almost 8% accuracy improvement over acoustic features baseline, also 3% greater than other methods significantly. The results verify the effectiveness of the proposed method. The optimization selection of different functionals and classifiers are also explored to provide improvements for discrete emotion classification. We observe that optimal functional are various in different situations and SVM can achieve best performance if valence dimensional rating exists, otherwise random forest does. In addition, the supplementary of dimensional information contributes to performance of "angry" and "happy" than "neutral" and "sad" particularly. In the future, we will apply this method to other databases with transfer learning to improve the performance of discrete speech emotion classification.

6. Acknowledgements

This work is supported by the National High-Tech Research and Development Program of China (863 Program) (No. 2015AA016305), the National Natural Science Foundation of China (NSFC) (No.61425017，No.61403386), the Strategic Priority Research Program of the CAS (Grant XDB02080006) and partly supported by the Major Program for the National Social Science Fund of China (13&ZD189).

7. References

[1] J. Tao, T. Tan, "Affective computing: A review," International Conference on Affective Computing and Intelligent Interaction, pp. 981-995, 2005.

[2] D. A. Sauter, F. Eisner, A. J. Calder, et al, "Perceptual cues in nonverbal vocal expressions of emotion," The Quarterly Journal of Experimental Psychology, vol. 66, no. 11, pp.2251-22722,2010.

[3] Z. Zeng, M. Pantic, G. I. Roisman, et al, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 1, pp. 39–58, Jan. 2009.

[4] E. A. Moataz, M. S. Kamel, and K. Fakhri, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recog., vol. 44, no. 3, pp. 572–587, 2011.

[5] K. S. Rao, S. G. Koolagudi, "Robust emotion recognition using spectral and prosodic features," Springer Science & Business Media, 2013.

[6] I. Luengo, E. Navas, and I. Hernandez, "Feature analysis and evaluation for automatice motion identification in speech," IEEETrans.Multimedia, vol. 12, no. 6, pp. 490–501, Oct. 2010.

[7] F. Eyben, M. Wöllmer, B. Schuller, "OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit," Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on. IEEE, 2009: pp. 1-6.

[8] H. Gunes, "Automatic, dimensional and continuous emotion recognition," 2010.

[9] H. Gunes, B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," Image and Vision Computing, vol. 31, no. 2, pp. 120–136, 2013.

[10] L. F. Barrett, "Discrete emotions or dimensions? The role of valence focus and arousal focus," Cognition & Emotion, vol. 12, no. 4, pp. 579-599, 1998.

[11] S. G. Karadoğan, J. Larsen, "Combining semantic and acoustic features for valence and arousal recognition in speech," Cognitive Information Processing (CIP), 2012 3rd International Workshop on. IEEE, pp.1-6, 2012.

[12] J. A. Hall, J. A. Harrigan, R. Rosenthal, "Nonverbal behavior in clinician—patient interaction," Applied and Preventive Psychology, vol. 4, no. 1, pp. 21-37, 1995.

[13] I. C. Christie, B. H. Friedman, "Autonomic specificity of discrete emotion and dimensions of affective space: A multivariate approach," International journal of psychophysiology, vol. 51, no. 2, pp. 143-153,2004.

[14] B. Stasak, J. Epps, N. C. R. Goecke, "An Investigation of Emotional Speech in Depression Classification," Interspeech 2016, pp. 485-489, 2016.

[15] C. Busso, M. Bulut, C. C. Lee, et al, "IEMOCAP: Interactive emotional dyadic motion capture database," Language resources and evaluation, vol. 42, no. 4, pp. 335–359, 2008.

[16] B. Schuller, S. Steidl et al, "The INTERSPEECH 2014 computational paralinguistics challenge: cognitive & physical load", Interspeech, pp. 427-431, 2014.

[17] F. Eyben, F. Weninger, F. Gross et al, "Recent developments in opensmile, the munich open-source multimedia feature extractor", Proceedings of the 21st ACM international conference on Multimedia, pp.835-838, 2013.

[18] X. Xia, L. Guo et al, "Audio Visual Recognition of Spontaneous Emotions In-the-Wild", Chinese Conference on Pattern Recognition, pp. 692-706, 2016.

[19] R. Xia, Y. Liu, "DBN-ivector Framework for Acoustic Emotion Recognition," Interspeech 2016, pp. 480-484, 2016.

[20] R. Xia, Y. Liu, "Using denoising autoencoder for emotion recognition," Interspeech 2013.