Improving Speech Emotion Recognition Based on ToBI Phonological Representations

Lingjie Shen ,Wei Wang
Machine Learning & Cognition lab, School of Education Science, Nanjing Normal University
Nanjing, China
Email: 602249910@qq.com, 769370106@qq.com

***Abstract*—The improvement of Speech Emotion Recognition (SER) relies on the classifiers and features. In terms of feature selection, so far, most of the research only uses a large set of acoustic features which cannot shed lights on the relationship between emotion and phonology. In our study, we improve SER by combining acoustic features and phonological representations together. We improve the SER on the public IEMOCAP database by combing acoustic and phonological features together under leave-one-speaker-out cross validation framework. Support vector machine, logistic regression, multi-layer perceptron and deep learning method of convolutional neural network (CNN) are used in our experiment. With phonological representations, CNN provides 60.22% of unweighted average recall (UAR) on categorical emotion recognition on utterance level which is now the state-of-the-art. When compared to the conventional baseline system based only on acoustic features, the proposed system with combing features gets 7.15% improvement of UAR in four basic emotion classification.**

Keywords-speech emotion recognition; acoustic features; phonology; deep learning.

I. INTRODUCTION

Automatic emotion recognition from speech has been an active research area in past years, which is of great interest for human computer interactions. It has wide applications ranging from computer tutoring applications to mental health diagnostic application [1]. Since the speech recognition has already changed people's life, detecting emotion from the speech is another challenge to improve the user-friendly human machine interaction.

Automatic Speech Emotion Recognition (SER) has been an active research area in past decades and is of great interest for human computer interactions. An efficient human emotion recognition system will help to make the interaction between human and computer more natural and friendly. It has wide applications ranging from computer tutoring applications to mental health diagnostic applications [1].

Accuracy of speech emotion recognition mainly relies on two factors, i.e., classifiers and features. In terms of features used in SER, acoustic features have been used as the dominant features in the literature. These acoustic features include frame-level features called Low Level Descriptors (LLDs) and their corresponding functionals which are used to map LLDs in the segment level to the utterance level. Most research of automatic emotion recognition usually relies on a large set of features and the reasons are as follows. First, so far there is no "standard" feature set for generic speech SER. Second, it is not clear which speech features are the most powerful in distinguishing emotions. Third, the acoustic variability introduced by the existence of different sentences, speakers, speaking styles, and speaking rates adds another obstacle to feature selection because these properties directly affect most of the common extracted speech features such as pitch and energy contours [2]. Therefore, most studies apply a large "brute-force" feature selection method which captures the dynamic temporal character of the contours of acoustic features over segments corresponding to different tasks [3], and this has been shown to outperform modeling the temporal dynamics on the classifier level [4]. During the last ten years, different acoustic feature sets used for various speech tasks have been proposed and have become widely-used feature sets that are beneficial for researchers in comparing their results on the same task [4][5][6][7][8].

Although there is a clearly perceived connection between emotions and phonology [9], researchers still have not formed a satisfactory model linking the emotion and prosody though these large feature sets are correlative with phonology[10]. There is not an accurate mapping between emotions and phonology. Hence, our goal of this study is to find out the emotionally salient phonological features with ToBI label systems and figure out the relationship between phonology and emotions. Then, as indicated by Liscombe [9], paralinguistic information can be conveyed via both segmental information and suprasegmental information that describes phonological information, such as pitch, intonation stress, rhythm and duration. We therefore combined acoustic features obtained from segmental information and phonological representations obtained from suprasegmental information to further improve the speech emotion recognition.

In this paper, we present experiments on the IEMOCAP dataset conveying four basic emotions. The extracted feature vectors are used to develop a support vector machine, logistic regression, multi-layer perceptron and a convolutional neural network as classifiers to recognize the emotional state in the offline system. Two different classes of feature vectors were evaluated: (1) acoustic features and (2) fusion features of acoustic and ToBI [11] features.

The paper is organized as follows. Section 2 provides us with the related work and literature on the IEMOCAP

 ISBN: 978-1-61208-612-5

database. Section 3 gives the methodology including emotional models, features and classifiers used in our study. Experiments and results are presented in Section 4. Experiments include two parts. Finally, discussion and conclusion are presented in Section 5.

II. RELATED WORK

Recent studies on the IEMOCAP database, including the classifiers, features, labels and results of classification are presented in Table Ⅰ. From the Table, we can observe that on the utterance level, the best unweighted average recall (UAR) on the IEMOCAP database is 58.46% using hierarchical binary Bayesian logistic regression [12]. On the frame level, the best UAR is 60.89% using a convolutional neural network [13]. The algorithm research on emotion recognition changes from a traditional machine learning method to a deep learning method (i.e., convolutional neural network, recurrent neural network, etc.). Some studies extract emotionally salient parts of speech by the attention mechanism method [14], which is successfully applied in image and speech recognition fields. Most of the research is starting to focus on solving some problems which might be encountered in the wild rather than exploring new features to improve accuracy. Further, more modals like face, gesture and linguistic information are being added in emotion recognition. Another observation from the literature review is that the Geneva Minimalistic Acoustic Parameter Set (eGeMAPS feature set) [5] performed better than the low complexity Logmel filter-banks. This is contrary to results from the field of computer vision, where in recent years features extracted from raw data by convolutional layers have outperformed hand-crafted features and achieved state-of-the-art results in various tasks [15].

In terms of phonological features about emotions, Busso et al. [16] explore what aspects of the pitch contour are the most emotionally salient. This study presents an analysis of the statistics derived from the pitch contour. The results indicate that gross pitch contour statistics such as mean, maximum, minimum and range are more emotionally prominent than features describing the pitch shape. The study explores the devotion of pitch features to speech emotion recognition and forms an emotional profile from acoustic features. However, we still have no explainable results from this research to interpret the relationship between emotion and phonology.

To find the interpretable relation between emotion and phonology, there has been some research using the ToBI system to find the salient cues of emotions. Iliev et al. [17] use ToBI features to recognize angry, happy and sad. The authors also combine the acoustic features together with ToBI features to improve speech emotion recognition. However, they only use ToBI features relating to tonal information while omitting the break indices which also carry information about emotion. They also neglect the sequential information of ToBI features encoded in an utterance. Cao et al. [10] explore the phonological cues from the ToBI system to study the relationship between acted perceptually unambiguous emotion and phonology. They aim to analyze the predictive power of discrete characterizations of intonations in the ToBI framework to discriminate specific emotions. The study indicates that the discrete features from the ToBI system are comparable to the acoustic features but are not robust for sentence-independent emotion classification tasks. Another limitation of this study is that the database is not public, therefore the outcome is not objective. However, this study provides us with some hints about the relation between phonological cues and specific emotions. Our study is inspired by this work and we further attempt to improve speech emotion recognition based on the IEMOCAP database with deep learning method by incorporating phonological representations.

TABLE Ⅰ. THE LITERATURE OF EMOTION RECOGNITION ON IEMOCAP DATABASE

Classifiers	UAR (%)
Hierarchical binary Bayesian logistic regression [12]	58.46
Support vector machine (SVM) [18]	50.64
Convolutional neural network (CNN) [13]	58.28
Bidirectional Long-short term Memory Recurrent Neural Network (BLSTM) [14]	58.8

III. METHODOLOGY

We introduce the dataset and the features used in the experiment in this section. The features include acoustic features and phonological representations, respectively. We combine these two kinds of features together to improve speech emotion recognition and compare the classification performance with baseline system using the acoustic features only.

A. Data description

The database used in this work is the interactive emotional dyadic motion capture (IEMOCAP) database which contains approximately 12 hours of audio-visual data from five mixed gender pairs of actors [19]. Each recorded session lasts approximately 5 minutes and consists of two actors interacting with each other in scenarios that encourage emotional expression. In this study, we only focus on the audio channel to perform speech emotion recognition. We use the categorical tags of this database. Specifically the categorical tags that we are considering in the IEMOCAP corpus are: neutral, angry, happy, sad (we merge happy and excitement together as happy). In total, the data used in our experiments comprises 5531 utterances with an average duration of 4.5 s.

B. Acoustic features

The openSMILE toolkit [20] is chosen to extract the acoustic features and the baseline feature set of Interspeech 2010 paralinguistic challenge [7] is used for our tests. This extension intends to better reflect a broader coverage of paralinguistic information assessment. As shown in Table 3, it consists of 38 basis LLDs. 21 functionals are applied to the above 34 LLDs and their corresponding delta coefficients, while 19 functional are applied to 4 F0 related LLDs and their corresponding delta coefficients. In addition, the durations and F0 onsets are also considered and included into the feature set. Thus, the final acoustic features vector has a dimension of 1582 as shown in Eq. (1):

$$f_a = (a_1, a_2, a_3, \dots, a_{1582}) \quad (1)$$

where $a_1, a_2, a_3, \dots, a_{1582}$ are the values of 1582 acoustic features.

C. Phonological representations

We use ToBI [21] to generate Phonological representation. TOBI labels encode the underlying phonological representation of an utterance primarily in terms of perceived pitch targets (H) high and (L) low and disjunctures between words (break 0-4 from minimal to strong). Perceptually prominent syllables, primarily due to pitch excursions but also lengthening and intensity, are associated with pitch accents that could consist of single tonal targets (H,!H*,L*), or bi-tonal combinations, most commonly L+H*,L*+H,H+!H*;!H represent a target down stepped from a preceding H target, and "*" corresponds to the tone aligned with the stressed syllable. For prosodic chunking, breaks 0 and 1 correspond to regular fluent word transitions, 2 to a perceived disjuncture with no salient tonal marking, 3 marks an intermediate phrase associated with H-,L-, or !H- targets. Regarding the break indices, diacritics describing uncertainty and disfluency were not used since the nature of the data elicitation minimized these phenomena and we wanted to mitigate data sparsity. Additionally, break 0 was not used. Every sentence's prosodic features consists of the times of every phonological representation in this sentence generated from the AuToBI [11] and become the one-hot-vector with fixed length. We have 141 phonological representations in total from AuToBI and Table Ⅱ presents the list of these features. The phonological feature is formulated as Eq. (2):

$$f_p = (p_1, p_2, p_3, \dots, p_{141}) \quad (2)$$

where $p_1, p_2, p_3, \dots, p_{141}$ are 141 prosodic events.

IV. EXPERIMENT AND RESULTS

We compare four classifiers' performance on speech emotion recognition based on acoustic features and fused features. The classifiers consist of traditional machine learning algorithms and deep learning method. The outcome shows that CNN outperforms other classifiers with fused features.

A. Experiment: emotion recognition using both two kinds of features

The motivation of this experiment is twofold. First, this experiment will compare the results with acoustic features only and with fused features (i.e., acoustic and phonological features) respectively. Second, we will implement several kinds of machine learning methods to improve the performance of speech emotion recognition on classifier level.

A ten-fold leave-one-speaker-out cross-validation scheme was employed in experiments using the nine speakers as training data and the one speaker as test data. Normalization is a critical step in emotion recognition.

The normalization method has an effect to experiment results. The goal of normalization is to eliminate speaker and recording variability while keeping the emotional discrimination. For this analysis, z-score normalization is implemented on all data, meaning that our speech emotion recognition is speaker-independent.

TABLE Ⅱ. LIST OF PHONOLOGICAL REPRESENTATIONS BASED ON TOBI LABELS.

Phonological Representations	Examples	Numbers
Break indices	Break indices 1 Break indices 3 Break indices 4	3
Phrasal tones	L- H- !H-	5
Pitch accent	H* !H* L+H*	6
Bigrams – pitch accent	H*,H* H*,!H* !H*,!H*	27
Bigrams – pitch accent with phrasal tones	H*,L- L*+H,INTONATIONAL_BOUNDARY !H*,INTONATIONAL_BOUNDARY	30
Bigrams – phrasal tones with pitch accents	L-,H* L-,!H* H-,!H*	48
Bigrams – phrasal tones	L-,L- L-,H- H-,H-	22

The classifiers used in our experiment are SVM with complexity 1, Logistic Regression (LR), Multi-Layer Perceptron (MLP) and CNN. Multi-layer perception has two hidden layer and hidden size is 50 and 20 respectively with Rectified Linear Unit (ReLU) activation function. The architecture of CNN is shown in Figure 1 and configuration is as follows. We use one-dimension convolution because the feature is one dimension. We have two convolutional layers each followed with one max pooling layer, one dense layer with 200 neuros. To provide a probabilistic interpretation of the model's output, the output layer utilizes a softmax nonlinearity instead of the nonlinear function used in previous layers. The activation function used in CNN is the ReLU due to its advantage over other activation functions, such as computational simplicity and faster learning convergence [22]. The base learning rate is set to 10-4 and optimizer is Adam [23]. The epoch is 10 and the training batch size is 32.

For support vector machine, logistic regression and multi-layer perception classifiers which are not deep learning methods, to avoid the curse of dimensionality, feature reduction is a necessary preprocessing. We use principle component analysis (PCA) to reduce the acoustic feature and fused feature (i.e., the concatenated acoustic and phonological features with dimension 1723) and we choose 100 and 120 as the number of components of acoustic feature and concatenated fused feature respectively. The number of components of PCA is chosen based on the best performance under SVM, LR and MLP in our pre-experiment. For CNN, we do not have to reduce the feature dimension due to CNN's advantages of sparse connectivity and shared weights.

The baseline system in our experiments is the classifier with acoustic features only. Under the baseline system, it is

 ISBN: 978-1-61208-612-5

objective to validate the predictive power of phonological features in speech emotion recognition and see the improvement of UAR after phonological features are added.

As it is standard practice in the field of automatic speech emotion recognition, results are reported using Unweighted Average Recall (UAR) as Eq. (3) to reflect imbalanced classes [7].

$$\mathrm{UAR} = \frac{1}{N}\sum_{i=1}^{N}\frac{c_i}{n_i} \quad (3)$$

where c_i is the number of correct examples of class i predicted by the classifier, n_i is the total number of examples of class i and N is the number of classes.

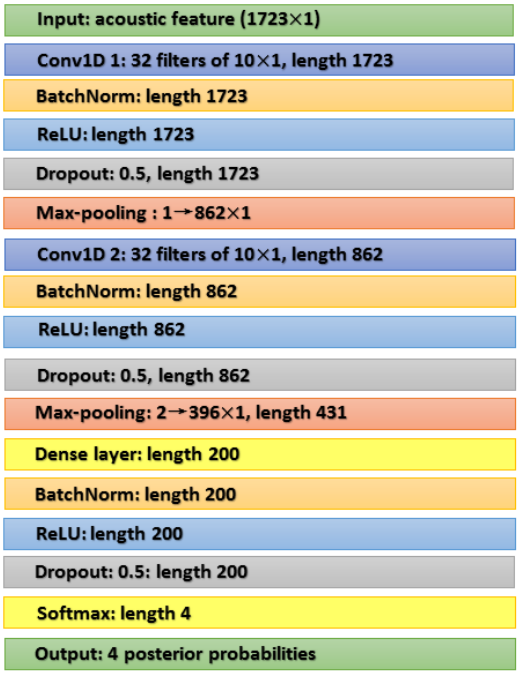

Figure Ⅰ. Topology of CNN

B. *Results*

The experiment results are shown in Table Ⅲ. In our four basic emotion recognition, our proposed method to improve speech emotion recognition with acoustic and phonological features using deep learning method (i.e., convolutional neural network) provides 60.22% of UAR which becomes the state-of-the-art on the utterance level speech emotion recognition and achieves 3.1% of improvement compared with the same classifier with acoustic feature only. The best UAR of the same database is 60.89% on frame level using CNN [12] and 58.46% on utterance level with hierarchical binary decision tree with speaker-dependent normalization [11]. The greatest improvement on four emotion recognition is 7.15% of UAR with multi-layer perception.

In general, the performance of the SVM is the worst. The performance of the deep learning method (i.e., convolutional neural network) is the best and outperformed significantly the other classifiers on four basic emotions recognition.

The improvement of speech emotion recognition by adding phonological features means that expertise knowledge can help machines improve their recognition rate because it is close to human perception and this complementary information is from thousands of years of humans' summarization which is more abstract but more discriminative and useful.

TABLE Ⅲ. THE RESULT OF EXPERIMENT.

Classifiers	**Features**	
	Acoustic	***Acoustic and Phonological Representations***
	UAR (%)	UAR (%)
SVM	49.91	51.35
LR	55.18	57.33
MLP	51.40	58.55
CNN	57.12	**60.22**

V. DISCUSSION AND CONCLUSION

From the result we can see that our proposed method to emotion recognition reaches the state-of-the-art using deep learning method of CNN by adding phonological representations. Phonological features represent people's knowledge summarization about prosodic representations and proves to be correlative to emotion. Therefore, adding expert knowledge to speech emotion recognition could further improve SER and shed light on the perceptual relationship between emotions and phonology. Our work presents evidence that discrete phonological representations have the potential to inform future feature development for emotion recognition and can lead to overall improved performance.

However, the limitation of this study is that the number of efficient phonological features is not rich. The reasons might be that the open source code to automatically recognize phonological representations is not so complete and sometimes it cannot recognize some very salient and evident phonological representations.

In the future, we will try to analyze which kinds of phonological features have discriminative power to specific emotions. We will also explore the cross-language, cross-culture and cross-humans speech emotion recognition to improve the SER' generalization.

ACKNOWLEDGMENT

This study is supported by the National Social Science Foundation of China (BCA150054)

REFERENCES

[1] Jin, Q., Li, C., Chen, S., & Wu, H., Speech emotion recognition with acoustic and lexical features. 2015: p. 4749-4753.

[2] Banse, Rainer, and Klaus R. Scherer. "Acoustic profiles in vocal emotion expression." Journal of personality and social psychology 70.3 (1996): 614.

[3] Schuller, B., Batliner, A., Steidl, S., & Seppi, D., Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication, 2011. **53**(9): p. 1062-1087.

[4] Schuller, B., S. Steidl, and A. Batliner. The Interspeech 2009 Emotion Challenge. in INTERSPEECH 2009, Conference of the International Speech Communication Association. 2009.

 ISBN: 978-1-61208-612-5

[5] Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & Truong, K. P., The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Transactions on Affective Computing, 2016. 7(2): p. 190-202.

[6] Schuller, B., Steidl, S., Batliner, A., Schiel, F., & Krajewski, J. The INTERSPEECH 2011 Speaker State Challenge. in INTERSPEECH 2011, Conference of the International Speech Communication Association. 2011.

[7] Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., & Narayanan, S. S. The INTERSPEECH 2010 paralinguistic challenge. in INTERSPEECH 2010, Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September. 2010.

[8] Schuller, B. W., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., ... & Mohammadi, G. The INTERSPEECH 2012 speaker trait challenge. in INTERSPEECH 2012, Conference of the International Speech Communication Association. 2012.

[9] Liscombe, J.J. Prosody and Speaker State: Paralinguistics, Pragmatics, and Prociency. 2007.

[10] Cao, H., Benus, S., Gur, R. C., Verma, R., & Nenkova, A. Prosodic cues for emotion: analysis with discrete characterization of intonation. in Speech prosody. 2014.

[11] Rosenberg, A. AuToBI - A tool for automatic ToBI annotation. in INTERSPEECH 2010, Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September. 2010.

[12] Lee, C.C., et al., Emotion recognition using a hierarchical binary decision tree approach. Speech Communication, 2011. 53(9–10): p. 1162-1171.

[13] Fayek, H.M., M. Lech, and L. Cavedon, Evaluating deep learning architectures for Speech Emotion Recognition. Neural Networks, 2017.

[14] Mirsamadi, S., Barsoum, E., & Zhang, C. (2017, March). Automatic speech emotion recognition using recurrent neural networks with local attention. In Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on (pp. 2227-2231). IEEE.

[15] Keren, G., & Schuller, B. (2016, July). Convolutional RNN: an enhanced model for extracting features from sequential data. In Neural Networks (IJCNN), 2016 International Joint Conference on (pp. 3412-3419). IEEE.

[16] Busso, C., S. Lee, and S. Narayanan, Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection. IEEE Transactions on Audio Speech & Language Processing, 2009. 17(4): p. 582-596.

[17] Iliev, A. I., Zhang, Y., & Scordilis, M. S. (2007, June). Spoken emotion classification using ToBI features and GMM. In Systems, Signals and Image Processing, 2007 and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services. 14th International Workshop on (pp. 495-498). IEEE.

[18] Mariooryad, S. and C. Busso, Exploring Cross-Modality Affective Reactions for Audiovisual Emotion Recognition. IEEE Transactions on Affective Computing, 2013. 4(2): p. 183-196.

[19] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S., IEMOCAP: interactive emotional dyadic motion capture database. Language Resources and Evaluation, 2008. 42(4): p. 335.

[20] Eyben, F., Wöllmer, M., & Schuller, B. (2010, October). Opensmile: the munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM international conference on Multimedia (pp. 1459-1462). ACM.

[21] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., ... & Hirschberg, J. ToBI: A standard for labeling English prosody. in International Conference on Spoken Language Processing, Icslp 1992, Banff, Alberta, Canada, October. 1992.

[22] Glorot, X., A. Bordes, and Y. Bengio, Deep Sparse Rectifier Neural Networks, in Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, G. Geoffrey, D. David, and D. Miroslav, Editors. 2011, PMLR: Proceedings of Machine Learning Research. p. 315--323.

[23] Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

 ISBN: 978-1-61208-612-5