Significance of Phonological Features in Speech Emotion Recognition

Wei Wang, Paul A. Watters, Xinyi Cao, Lingjie Shen & Bo Li

International Journal of Speech Technology

ISSN 1381-2416

Int J Speech Technol
DOI 10.1007/s10772-020-09734-7

Springer

International Journal of Speech Technology
https://doi.org/10.1007/s10772-020-09734-7

Significance of Phonological Features in Speech Emotion Recognition

Wei Wang[1] · **Paul A. Watters**[2] · **Xinyi Cao**[1] · **Lingjie Shen**[1] · **Bo Li**[3]

Received: 27 December 2019 / Accepted: 8 July 2020

Abstract
A novel Speech Emotion Recognition (SER) method based on phonological features is proposed in this paper. Intuitively, as expert knowledge derived from linguistics, phonological features are correlated with emotions. However, it has been found that they are seldomly used as features to improve SER. Motivated by this, we set our goal to utilize phonological features to further advance SER's accuracy since they can provide complementary information for the task. Furthermore, we will also explore the relationship between phonological features and emotions. Firstly, instead of only based on acoustic features, we devise a new SER approach by fusing phonological representations and acoustic features together. A significant improvement in SER performance has been demonstrated on a publicly available SER database named Interactive Emotional Dyadic Motion Capture (IEMOCAP). Secondly, the experimental results show that the top-performing method for the task of categorical emotion recognition is a deep learning-based classifier which generates an unweighted average recall (UAR) accuracy of 60.02%. Finally, we investigate the most discriminative features and find some patterns of emotional rhyme based on the phonological representations.

Keywords Speech emotion recognition · Phonological features · Feature analysis · Acoustic features

1 Introduction

Automatic Speech Emotion Recognition (SER) has been an active research area during the past several decades, and is of great interest for the human computer interaction community. An accurate and efficient human emotion recognition system will help make the interaction between humans and computers more natural and friendlier. Automatic SER has wide applications ranging from computer tutoring to mental health diagnosis (Jin et al. 2015).

The accuracy of speech emotion recognition mainly relies on two factors—features and classifiers. In terms of features used in SER, different acoustic features have dominated the literature, primarily the large set of acoustic features characterizing prosodic, voice quality and spectral related features. These acoustic features consist of frame-level features that are often referred as low-level descriptors (LLDs), and their corresponding functions are used to map LLDs at the segment level to a space at the utterance level. Most research uses a "brute-force" feature selection method (Jin et al. 2015; Shen and Wang 2018) for different classification tasks. However, these feature sets are often different in diverse tasks, and there are no other features as salient as Mel-frequency cepstral coefficient (MFCC) in SER (Han (2014)).

Currently, there are many limitations in automatic SER. Firstly, since most related research works only use acoustic features in SER, it is very difficult for them to find out whether there are existing some specific speech phonological patterns for different emotions. Secondly, although some research works have used deep learning methods to classify speech emotions with raw acoustic data, it is still quite challenging to interpret the relationship between emotions and

✉ Bo Li
bo.li@usm.edu; li.bo.ntu0@gmail.com

Wei Wang
njnuwangwei@qq.com

Paul A. Watters
dr.paul.watters@gmail.com

[1] School of Education Science, Nanjing Normal University, Nanjing, JS 210097, China

[2] Department of Computer Science and Information Technology, La Trobe University, Melbourne, VIC 3350, Australia

[3] School of Computer Sciences and Computer Engineering, University of Southern Mississippi, 730 East Beach Blvd, Long Beach, MS 39560, USA

Published online: 15 July 2020

phonology based on the abstract features extracted from raw acoustic data.

Motivated by the challenges mentioned above, we propose a novel method which combines acoustic features and phonological representations in speech emotion recognition. Our aim is two-fold. Firstly, we investigate whether adding the phonological features that represent the linguistic expert knowledge can achieve information gain to improve SER's performance based on a deep learning method. Secondly, we explore the relationship between different emotions and phonological representations. We are able to obtain a set of specific phonological patterns from a series of discrete phonological representations that are actually derived from the original acoustic features.

Evaluative and comparative experiments have been conducted on the IEMOCAP dataset (Busso et al. 2008) which convey four basic emotions in three-dimensional emotion space. The experiments include two parts. The first part is to improve SER by combining acoustic features and phonological representations together. The second part is to analyze the discriminative power of acoustic features and phonological representations and to find out related feature patterns.

This project is based on but significantly extended beyond a previous work (Shen and Wang 2018) on SER. In (Shen and Wang (2018)), they have demonstrated that the ToBI phonological representation can improve the speech emotion recognition performance, while we make a further study on an optimal classification method to balance phonological and continuous features in this project. Furthermore, this research also explores how the phonological representation can be utilized to improve the performance and their different impact in emotion dimensions.

The paper is organized as follows. Section 2 reviews the related work and literature on the IEMOCAP database. Section 3 presents the methodology. Experimental results are presented in Sect. 4. Finally, we conclude in Sect. 5.

2 Related work

2.1 Emotion models

There are two popular emotion models that dominate the research (Han 2014). The first is the discrete emotion model. It claims that only a few discrete emotions exist. Categorial labels are the most popular ones to differentiate distinct emotions. However, the size of an emotion lexicon is huge. To facilitate emotion recognition research, a set of six basic emotions (happiness, sadness, anger, fear, surprise and disgust) proposed by Ekman (Ekman (1992)), are found to be more universal. Other emotions can be regarded as a combination or variation of these six. The second one is the dimensional emotion model which is an alternative to the discrete emotion model (Han 2014). It states that emotions can be distinguished based on a set of certain characteristics (dimensions). Based on this model, emotions can be labeled by specifying a value for each dimension. It is well-known that emotion can be characterized using two dimensions: activation and valence (Fernandez 2004). Activation refers to the amount of energy required to express a certain emotion while valence refers to the subjective feeling of pleasantness or unpleasantness.

2.2 Phonological representations

There are few reports about how to improve SER by utilizing the features from the Tones and Break Indices (ToBI) framework in publicly available studies. Existing related research mainly focuses on the impact of some discrete features extracted based on ToBI for SER. For instance, Iliev et al. (2007) used the ToBI features to recognize angry, happy and sad emotions. The authors also combined acoustic features together with ToBI features to improve SER. However, they only used ToBI features related to tonal information, while ignoring the break indices which also carry information about emotions. They did not consider the sequential information of ToBI features encoded in an utterance, as well. Cao (2014) explored the phonological cues from the ToBI system to study the relation between phonological cues and specific emotions. The study indicates that the discrete features from the ToBI system are comparable to the acoustic features but not robust for sentence-independent emotion classification tasks. However, the dataset utilized is not publicly available, and only a few discrete phonological features related to breaks, boundary tones and types of pitch accent were extracted manually and involved in research. This study provides some tantalizing hints about the relationship between phonological and specific emotions. Our research is inspired by this work, while focusing on improving the speech emotion recognition further based on the IEMOCAP database. Our main research tasks include: (1) To prove the phonological features can provide supplementary emotion information for acoustic features in public database. (2) To propose a hybrid feature model to improve the UAR performance of SER. Compared to Cao (2014), we plan to use AutoToBI to extract more phonological features in the ToBI framework. (3) To analyze the Top 3 acoustic features and Top 3 phonological features in four basic emotions model and three dimensions emotions model, respectively. We also want to analyze the different effect of these features in different emotion dimensions.

2.3 Related work on the IEMOCAP database

We summarize the results of recent related studies on the IEMOCAP database in Table 1. We observe that the best

Table 1 Comparison of several selected discrete emotion recognition related work based on the IEMOCAP database, in terms of classifiers, features and UAR

Classifiers	Features	UAR (%)
Hierarchical binary Bayesian logistic regression (Lee et al. 2011)	Acoustic	58.46
Support vector machine (SVM) (Mariooryad and Busso 2013)	Acoustic	50.64
Replicated softmax model (RSM) + SVM (Shah et al. 2014)	Acoustic	57.39
CNN (Fayek et al. 2017)	Acoustic	58.28
Attention based bidirectional long short term memory recurrent neural network (Badshah et al. 2017)	Acoustic	58.8
Attentive CNN (Huang et al. 2017)	Acoustic	56.1

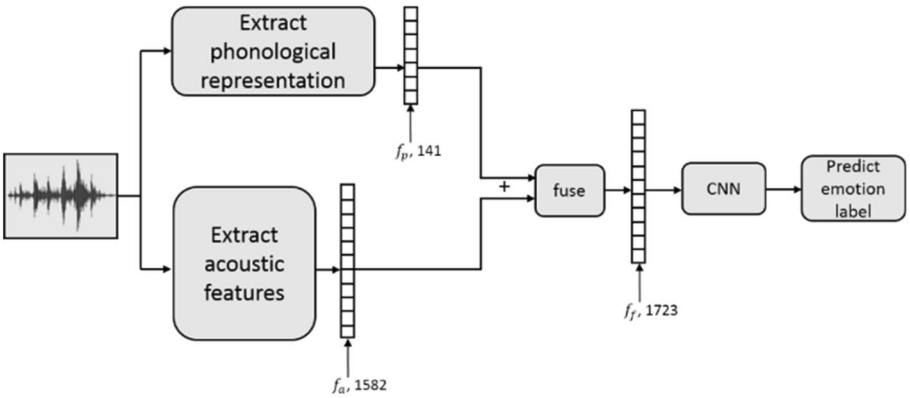

Fig. 1 Our proposed method to improve speech emotion recognition

unweighted average recall (UAR) on the IEMOCAP database is 58.46% using hierarchical binary Bayesian logistic regression (Lee et al. 2011). Support vector machines are the most popular algorithm to classify emotions due to good performance on small datasets and its ability to deal with high dimensions. However, it performs worse than other classifiers in this situation, including logistic regression and deep learning. Currently, the focus on algorithmic research in emotion recognition has changed from traditional machine learning methods to deep learning methods (Fayek et al. 2017; Badshah et al. 2017; Huang et al. 2017). Some studies extract emotionally salient parts of speech based on an attention mechanism (Mirsamadi et al. 2017), which has been successfully applied in image and speech recognition fields. However, to the best of our knowledge, all previous research based on the IEMOCAP database has only used acoustic features in emotion recognition.

3 Methodology

As demonstrated in Fig. 1, we propose to incorporate phonological features to further improve the performance of SER. Our aim is to determine whether phonological representations can improve SER performance which will validate the significance of using phonological representations.

Table 2 List of the acoustic features

LLDs	Number
(Δ)Loudness	1
(Δ)MFCC(Mel frequency cepstral coefficients)[0–14]	15
(Δ)Log Mel frequency band[0–7]	8
(Δ)LSP[0–7]	8
(Δ)F0	1
(Δ)F0 envelope	1
(Δ)Voicing probability	1
(Δ)Jitter local	1
(Δ)Jitter consecutive frame pairs	1
(Δ)Shimmer local	1

3.1 Acoustic features

We extract acoustic features of speech in terms of prosodic features, voice quality and spectral related features as well as their global statistics with openSMILE toolkit (Eyben 2010). The baseline feature set of the Interspeech 2010 paralinguistic challenge (Müller 2010) is used for our experiments. As shown in Table 2 (Beckman et al. 2005), the feature set used in our experiment consists of 38 basic Low-Level Descriptors (LLDs). 21 functions are applied to 34 of the above LLDs and their corresponding delta coefficients, while 19 functions are applied to the remaining four F0-related LLDs and their corresponding delta coefficients. In addition, the durations and F0 onsets are

also considered and included in the feature set. Thus, the dimension of the emotional feature vector is: 21*34 (LLDs) + 21*34 (delta coefficients of LLDs) + 19*4 (LLDs) + 19*4 (delta coefficients of LLDs) + 1 (durations) + 1 (F0 onsets) = 1582. The final acoustic feature vector has a dimension of 1582,

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

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

3.2 Phonological features

ToBI is a system for transcribing phonological patterns and other aspects of the prosody of English utterances (Beckman et al. 2005), though it has been adapted for other languages as well. We use ToBI to generate phonological representations. Tone tier and break tier elements are used to depict the chosen distinct emotional states. Phonological features are defined as sequences of abstract prosodic events aligned to syllables under the ToBI framework. The phonological features consist of the times of every prosodic event and is a sparse one-hot vector with fixed length. AuToBI system is a freely distributed tool for non-commercial use. It is used to analyze automatically Standard American English prosody. After feeding a segmentation of a speech sample, AuToBI can extract the requisite features from the speech signal and generate predictions for each element of the ToBI standard using the stored models. In our experiments, 141 phonological representations based on ToBI labels are extracted, and the detailed list of those 141 features are shown in Table 3. We use AuToBI (Rosenberg 2010) to automatically extract these prosodic events to avoid time-consuming manual labeling. Instead of only counting these discrete prosodic events, we also incorporate the bigram of neighboring prosodic events which takes the sequence of phonological representations into account. Finally, we obtain 141 phonological representations in total, as listed in Table 3 (Schröder 1992). The phonological feature is represented as,

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

where p_1, …, p_{141} are the prosodic events. Finally, we combined both the acoustic and phonological features together to form a hybrid feature vector,

$$f_f = (f_1, f_2, f_3, ..., f_{1723}) \quad (3)$$

Table 3 List of the 141 phonological representations based on ToBI labels

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

4 Experiments and discussions

4.1 Data description

The database used in this work is the interactive emotional dyadic motion capture database (IEMOCAP) database. It is collected by the Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (USC). This database records around 12 h of the audio-visual data of five pairs of mixed-gender actors during a set of spoken communication scenarios with scripts (Busso et al. 2008). In this study, we only focus on the audio channel to perform speech emotion recognition.

We use two tags of the database: the categorical and dimensional tags. Specifically, the categorical tags that are under consideration for the IEMOCAP corpus are: neutral, angry, happy and sad (for simplicity, we combine happy and excitement into one: happy). In total, the data used in our experiments comprise 5531 utterances with an average duration of 4.5 s. We split the dimensional tags (valence, activation and dominance) into three levels: Level 1 contains ratings in the range [1,2), Level 2 for the range [2,4), and Level 3 for the range (Ekman 1992; Fernandez 2004). Correspondingly, these levels intuitively correspond to low, medium and high activation; or negative, neutral and positive valence; or weak, medium and strong dominance. In Table 4, we show their distributions on the database. As can be seen, it is an imbalanced dataset, either in terms of the number of samples in each class (1084–1708) or dimensional tags (647–3764).

Table 4 Distribution of emotion categories in valence, activation and dominance

	Valence			Activation			Dominance			Total
	Negative	Neutral	Positive	Low	Medium	High	Weak	Medium	Strong	
Neutral	99	1465	144	304	1358	46	244	1343	121	1708
Angry	856	237	10	14	608	481	7	407	689	1103
Happy	19	439	1178	50	1109	477	138	1033	465	1636
Sad	591	468	25	358	689	37	258	686	140	1084
Total	1565	2609	1357	726	3764	1041	647	3469	1415	5531

4.2 Data preprocessing

The normalization method has a nontrivial effect on the experiment results. The goal of normalization is to eliminate speaker and recording variabilities while keeping their emotional discriminations. For our experiments, Z-score normalization is applied to all the data to produce a distribution with a mean value of 0 and standard deviation of 1, making our speech emotion recognition speaker-independent.

Z-score normalization is a general normalization method to avoid the influence of range of different features values. In SER, there are a lot of various features with different dimensions, it is necessary to make a Z-score normalization. To calculate the z value for a feature observation value x_i of feature x, we use the following equation:

$$Z = \frac{x_i - \bar{x}}{s},$$

where $\bar{x}$ and s are the mean and standard deviation values of feature x, respectively.

4.3 Experiment 1: Improving speech emotion recognition by adding phonological representations

4.3.1 Experiment setup

The classifiers used in our experiments include support vector machine (SVM) with complexity 1, logistic regression (LR) and convolutional neural networks (CNNs), representing the most successful classifiers that have been used in SER. SVM and LR are the classical methods of machine learning which achieve great performance. CNN is the method which layered architecture matches what we find in the cortex. The architecture and configurations of our CNN model are as follows. We use one-dimensional convolution because the feature vector that we have extracted with OpenSmile is one-dimensional. Firstly, we have two convolutional layers each followed by BatchNorm, ReLU, dropout and a maximum pooling layer. Then, one dense layer is added followed by BatchNorm and dropout. Finally, to provide a probabilistic interpretation of the model's output, the output layer utilizes a softmax nonlinearity instead of the nonlinear function used in the previous layers. The base learning rate is set to 10–4 and the optimizer is Adam. The epoch is 10 and the training batch size is 32. The L2 regularization ($\lambda = 0.01$) is applied on the convolutional and dense layers. All filter sizes in the experiments are set to 10 and the max-pooling size is 2. We have four CNNs with different topologies to explore CNN's performance in depth by changing the number of hidden neuros in the dense layers and the number of filters in the convolution layers according to experience and experiments.

A ten-fold leave-one-speaker-out cross-validation scheme (Schuller et al. 2009a) is employed in our experiments using nine speakers as the training data and one speaker as the testing data. As per standard practice in the field of automatic SER, experimental results are evaluated using the unweighted average recall (UAR) metric to reflect imbalanced classes (Schuller et al. 2009b),

$$UAR = \frac{1}{N}\sum_{i=1}^{N}\frac{c_i}{n_i} \tag{4}$$

where c_i is the number of examples in class i ($i = 1, 2, 3, 4$) that are correctly predicted by the classifier, while n_i is the total number of examples in class i ($n_1 = 1708$, $n_2 = 1103$, $n_3 = 1636$, $n_1 = 1084$) and N is the total number of classes in the dataset ($N = 4$). The main reason to use UAR is to avoid that the accuracies of large classes will dominate the final evaluation result. UAR will first calculate each class' recall value which does not rely on the size of the class. Then, by averaging all the recall values across all the classes it generates the final UAR values. On the other hand, weighted average recall (WAR) will consider all the classes as a whole, and calculate the recall based on the ratio between total number of correctly predicted examples (hits) and total number of examples in the dataset,

$$WAR = \frac{\sum_{i=1}^{N} c_i}{\sum_{i=1}^{N} n_i} \tag{5}$$

This will make large classes' recall values dominate the final result.

4.3.2 Baseline SER methods

The baseline SER methods in our experiments are SVM, logistic regression and convolutional neural network with the acoustic features only. Under these baseline methods, the objective is to validate the predictive power of the phonological features in speech emotion recognition and find out the improvement in UAR after adding those phonological features.

4.3.3 Results and analysis

The experiment results are shown in Table 5. CNN classifier training took the longest time. It takes 186 s to train the CNN model based on 1582 acoustic features, 187 s based on 1723 fused features, and 153 s based on 141 prosodic features.

Figure 2 compares the UAR performance of acoustic only, prosodic only, and fused features. As can be seen from the results, we can find that the UAR performance for prosodic features only based method (Cao 2014) is very low. However, after fusing them with the acoustic ones, we can achieve the best UAR performance, which demonstrates the complementary component contributions from the prosodic features towards SER. For the four basic emotions recognition task, our proposed method achieves a UAR accuracy of 60.22% by using both acoustic and phonological features based on a deep learning method (i.e., convolutional neural network), which is a 3.1% improvement beyond our approach with acoustic features only. Compared with the research on the same database, our performance is significantly better than the state-of-the-art (58.46%) achieved by Lee et al. (Lee et al. 2011), who utilized a hierarchical binary decision tree with speaker-dependent normalization (Fayek et al. 2017). Similarly, for dimensional emotion recognition tasks, the best results

Table 5 The results of Experiment 1

Tasks	Classifiers	Features		
		Acoustic	Phonological	Acoustic and Phonological
		UAR (%)	UAR (%)	UAR (%)
Four basic emotions	SVM	49.91	32.31	51.35
	LR	55.18	30.05	57.33
	#32*64/ 100/4	59.04	30.72	59.59
	#32*64/200/4	**57.12**	**32.14**	**60.22**
	#16*32/100/4	59.05	28.83	59.59
	#16*32/200/4	59.05	29.27	59.59
Activation	SVM	53.21	38.81	54.48
	LR	54.69	40.07	58.77
	#32*64/100/3	59.76	44.26	59.65
	#32*64/200/3	59.43	46.07	59.49
	#16*32/100/3	58.32	39.78	60.59
	#16*32/200/3	**58.58**	**40.35**	**60.78**
Dominance	SVM	44.82	37.43	46.06
	LR	45.99	38.86	49.50
	#32*64/100/3	**48.72**	**41.43**	**49.89**
	#32* 64/200/3	48.37	42.12	49.37
	#16*32/100/3	47.82	39.94	47.91
	#16*32/200/3	48.03	40.09	49.88
Valence	SVM	44.54	37.25	44.59
	LR	49.68	37.34	53.58
	#32*64/100/3	56.07	37.41	56.53
	#32*64/200/3	**55.44**	**37.54**	**56.62**
	#16*32/100/3	55.97	36.10	56.22
	#16*32/200/3	55.64	35.46	55.79

The best runs are indicated in bold font

We show results on three classifiers: SVM, LR and CNN with four different topologies. FC(n_0) denotes a fully connected layer of n_0 units. Conv1D(m×j) denotes a one-dimensional convolutional layer of m filters of size j with a stride of 2. Softmax(n_0) denotes a softmax output layer of n_0 units

#a*b/c/d means the CNN parameters setup as Conv1D (*a***b*)-FC (*c*)-Softmax (*d*)

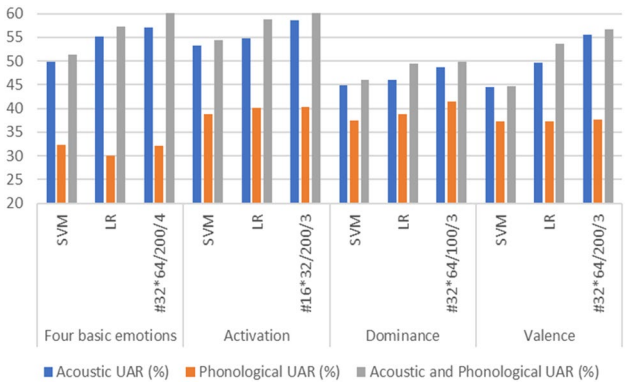

Fig. 2 SER performance comparison for acoustic only, prosodic only and fused features in terms of UAR

for activation, dominance and valence classification are 60.78%, 49.89% and 56.62% in UAR respectively, which are also achieved by our CNN classifier and indicate a UAR improvement of 4.08%, 3.51% and 3.9% respectively, if compared with the logistic regression classifier.

Typically, the deep learning-based approach (CNN) achieves the best performance across all the four recognition tasks. The improvements in SER by adding phonological features have demonstrated that this type of expertise knowledge's information gain is helpful for machines to improve their recognition performance if we consider the fact that it is related to human perception. This complementary information is concluded from thousands of years' humans' summarization, which is more abstract, general, discriminative, and useful.

Finally, we employ the *t*-test (confidence level: 0.05) on independent samples to statistically evaluate the significance of the SER performance improvement of our fused features-based method over the acoustic features only based method in terms of UAR. We select the best run (marked with a bold font in Table 5) for each classification task and use independent examples for the *t*-test. The *p*-values for the four classification tasks listed in Table 5 are: 0.042, 0.033, 0.033, and 0.028, accordingly. This indicates that the *p*-values for either the categorical or the dimensional tag data are always less than 0.05, which suggests that there is a significant improvement in the SER performance after integrating the phonological features.

4.4 Experiment 2: feature analysis

4.4.1 Experiment setup

This experiment analyzes the discriminative power of acoustic and phonological features related to emotion recognition. We design four sub-experiments, namely, category, valence, activation and dominance classification. Acoustic and phonological features are respectively used in these four tasks. Each time only one feature is used in a classification task. All the experiments use the leave-one-speaker-out cross validation framework which uses nine speakers as the training data and one speaker as the testing data. The classifier we used in the experiments is logistic regression because it is a well-known technique to model binary or dichotomous variables while CNN is better to deal with high-dimensional features. Then we rank the features according to the results of the classification. Three top features are picked up in the list as three most discriminative features in the tasks. Then we compare these features by Pearson correlation coefficient to check if these features are significantly different among emotional classes in descriptive and dimension tags classification tasks. A p-value of smaller than 0. 001 means that the correlation is very strong. Otherwise, if the p-value is less than 0.01, the correlation is moderately significant. A p-value of between 0.01 and 0.05 means that the correlation is significant. A p-value of larger than 0.05 indicates that there is no statistic correlation.

We use the Pearson correlation coefficient (Pearson 1895) to explore the mutual effect of acoustic and phonological features on emotion prediction in terms of dimensional labels. The Pearson correlation coefficient (PCC) is to measure the linear relationship between two variables X and Y. It has a value between -1 and $+1$, where '1' means total positive linear correlation, '0′ indicates no linear correlation, and '-1' is total negative linear correlation. The formula for the Pearson correlation coefficient $r_{X,Y}$ is defined as follows,

$$r_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X, \sigma_Y} \tag{6}$$

where cov is the covariance of two variables X and Y, σ_X and σ_Y are the standard deviation of X and Y, respectively.

4.4.2 Results and analysis

Table 6 compares the results of Experiment 2. We can find that all single acoustic and phonological feature is not significantly correlated with each individual emotion task. In dimension classification tasks, the Δloudness (standard deviation), Δloudness (99%-percentile) and Δloudness (1%-percentile) are strongly correlated with activation dimension. The phonological features such as Pitch accent (! H*), Break indices (break indices 1) and Pitch accent (H*) also show strong correlation with the valence dimension classification task. All of Break indices (break indices 3), Break indices (break indices 4) and Pitch accent (!H*) are strongly correlate to the dominance dimension. Pitch accent (L* + H), (L* + H, intonational boundary) and Pitch accent (!H*) are

Table 6 Top 3 acoustic features (with their functions) and top 3 phonological features according to UAR based on logistic regression, as well as their Pearson correlation coefficients

Class	Acoustic Feature (with functions)	PCC	Phonological representation	PCC
Four basic emotions	ΔlogMelFreqBand (quartile1)	–	Bigram—pitch accent with phrasal tones (! H*, intonational boundary)	–
	Δloudness (standard deviation)	–	Pitch accent (! H*)	–
	Loudness (standard deviation)	–	Pitch accent (H + ! H*)	–
activation	Δloudness (standard deviation)	0.52***	Pitch accent (! H*)	0.30***
	Δloudness (99%-percentile)	0.51***	Break indices (break indices 1)	0.25***
	Δloudness (1%-percentile)	−0.51***	Pitch accent (H*)	0.25***
dominance	MFCC (percentile range)	0.49***	Break indices (break indices 3)	0.27***
	logMelFreqBand (percentile range)	0.47***	Break indices (break indices 4)	0.273***
	logMelFreqBand (quartile3)	0.44***	Pitch accent (! H*)	0.28***
valence	logMelFreqBand (percentile range)	−0.02*	Pitch accent (L* + H)	−0.04***
	Δloudness (percentile range)	−0.07***	Bigram-pitch accent with phrasal tones (L* + H, intonational boundary)	−0.04***
	logMelFreqBand (99%-percentile)	0.01**	Pitch accent (! H*)	−0.07***

*p-value smaller than 0.05

**p-value smaller than 0.01

***p-value smaller than 0. 001. An absent symbol indicates a p-value larger than 0.05. The value of PCC is the mean value of the feature

strongly correlated with the valence dimension. All of these demonstrate some phonological feature-based emotional patterns. Similarly, we can also find that Table 6 also shows that the acoustic features are related to the emotional tasks.

We summarize the findings as follows. (1) some acoustic features and certain phonological representations characterize the same information of some emotions during their classifications. However, these phonological representations help us further explain the phonological pattern of different speech emotions. The phonological representations are the combination of different acoustic features which derive from the rule-based linguistic system. Loudness and pitch accent are salient features to classify discrete emotions, but the phonological representations imply different pitch accent types in classifying emotions. In activation classification, loudness is the most discriminative acoustic feature, while pitch accents H* and !H* are salient phonological representations. However, pitch accent, which is related to loudness and pitch, not only indicates that the stress of some words is salient, but also implies the patterns of pitch rhythm for different emotions. (2) Phonological representations provide complementary information which implies the intonation and rhythm. For instance, in terms of dominance, logMelFreqBand is the most discriminative acoustic features while break indices are the most discriminative phonological representations. These features imply that the more disfluent the speech is, the more likely that the dominance has a strong level, which makes the speech sound firmer. In terms of activation, break index '1′ has a positive correlation with activation, indicating that the more fluent the speech is, the more likely the activation is at a high level. However, we cannot draw this type of general conclusions based on the acoustic features alone.

In summary, phonological representations indeed explain the intuitive relations between phonology and emotions. The reason is that although phonological representations are derived from acoustic features, they encode related expert knowledge which will be helpful for related application, such as Bhowmik and Mandal derived Bengali phoneme knowledge from acoustic features (Bhowmik and Mandal 2018). That is, phonological representations are more directly related to emotions than the original acoustic features. Although there is an information loss in the phonological representations due to their intermediate level, the phonological features provide complementary information for us to understand emotions and their patterns.

This method could improve the UAR of SER, but there is a little more time cost associated with the added features. Fortunately, this time is related to off-line cost, since it has happened during the training stage. It has little impact on the recognition speed. Meanwhile, the deep learning method can effectively eliminate redundant information existing in the fed features. If there are no new information gain by contributing supplemental information by the newly added features, the performance of SER will seldomly be improved.

5 Conclusions and future work

In this paper, we have proposed a SER approach which incorporates both phonological and acoustic features. From the results, we can see that our proposed method has

achieved the best performance by using a CNN deep learning model. The proposed method can help us explain the relationship between phonology and emotions. The results of our feature analysis experiment indicate that the phonological representations, for example, pitch accent and break indices, are predictive for emotions. Phonological features represent human's expert knowledge about prosody and are correlated with emotions. Therefore, adding this type of expert knowledge to speech emotion recognition could further improve SER performance and sheds light on finding the perceptual relationship between emotions and phonology. With the help of phonological features, we can better interpret the results, and easily find the patterns for different speech emotions which we cannot observe based on the original acoustic features only. Our work has also demonstrated that discrete phonological representations are beneficial to improve emotion recognition performance.

In the future, after further analyzing the features that are discriminative for different emotions, we plan to develop an end-to-end deep learning-based method for emotion recognition since we believe that if a large amount of data is collected and used, it is promising to significantly improve the accuracy and generalize speech emotion recognition with by utilizing a deep learning method. In addition, the correlation between different features and emotions in this study is based on univariate linear regression. Therefore, we plan to have an in-depth analysis of information gain using different feature sets and their complementary power in emotion recognition. We also tend to explore the applications of our method in cross-language, cross-culture, or cross-human speech emotion recognition to find out possible phonological features that can generally represent certain emotional states regardless of the speaker's cultural background, language, and accent, etc.

Funding The project was funded by the Innovative Research Group Project of the National Natural Science Foundation of China (CN) (Grant No. BCA150054) and the Faculty Startup Funds Award of the University of Southern Mississippi (US).

References

Appiah, A. Y., Zhang, X., Ayawli, B. B. K., & Kyeremeh, F. (2019). Long short-term memory networks based automatic feature extraction for photovoltaic array fault diagnosis. *IEEE Access, 7*, 30089–30101.

Badshah, A. M., Rahim, N., Ullah, N., Ahmad, J., Muhammad, K., Lee, M. Y., S. KwonSung, Baik, W. (2017). Deep features-based speech emotion recognition for smart affective services. Multimedia Tools and Applications, pp 1–19.

Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The Original ToBI System and the Evolution of the ToBi Framework. Prosodic Typololgy: The Phonology of Intonation and Phrasing, pp. 9–54.

Bhowmik, T., & Mandal, S. K. D. (2018). Manner of articulation based Bengali phoneme classification. *International Journal of Speech Technology, 21*(2), 233–250.

Busso, C., Bulut, M., & Lee, C. C. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation, 42*(4), 335.

Cao, H. (2014). Prosodic cues for emotion: analysis with discrete characterization of intonation. Speech prosody. pp. 1147–1152.

Ekman, P. (1992). Are there basic emotions? *Psychological Review, 99*(3), 550–553.

Eyben, F. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *ACM International Conference on Multimedia*, pp. 1459–1462.

Fayek, H. M., Lech, M., & Cavedon, L. (2017). Evaluating deep learning architectures for Speech Emotion Recognition. Neural Networks, S089360801730059X.

Fernandez, R. (2004). A computational model for the automatic recognition of affect in speech. Massachusetts Institute of Technology

Han, W. J. (2014). Review on speech emotion recognition. Journal of Software.

Huang, Z., Xue, W., Mao, Q., & Zhan, Y. (2017). Unsupervised domain adaptation for speech emotion recognition using PCANet. *Multimedia Tools & Applications, 76*(5), 6785–6799.

Iliev, A. I., Zhang, Y., & Scordilis, M. S. (2007). Spoken emotion classification using ToBI features and GMM. In *International Workshop on Systems, Signals and Image Processing, 2007 and Eurasip Conference Focused on Speech and Image Processing, Multimedia Communications and Services*, pp. 495–498.

Jin, Q. Li, C., Chen, S. (2015). Speech emotion recognition with acoustic and lexical features. In *IEEE International Conference on Acoustics*, pp. 4749–4753

Lee, C. C., Mower, E., & Busso, C. (2011). Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication, 53*(9–10), 1162–1171.

Mariooryad, S., & Busso, C. (2013). Exploring cross-modality affective reactions for audiovisual emotion recognition. *IEEE Transactions on Affective Computing, 4*(2), 183–196.

Mirsamadi, S., Barsoum, E., Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2227–2231.

Müller, C. (2010). The INTERSPEECH 2010 paralinguistic challenge. INTERSPEECH 2010, Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September, pp. 2794–2797.

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London, 58*, 240–242.

Rosenberg, A. (2010). AuToBI—A tool for automatic ToBI annotation. INTERSPEECH 2010, Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, pp. 146–149.

Schröder, M. (1992). ToBI: A standard for labeling English prosody. In *International Conference on Spoken Language Processing*, ICSLP 1992, Banff, Alberta, Canada.

Schuller, B., Steidl, S., & Batliner, A. (2009). The Interspeech 2009 Emotion Challenge. In *INTERSPEECH 2009, Conference of the International Speech Communication Association*, pp. 312–315.

Schuller, B., Vlasenko, B., & Eyben, F. (2009) Acoustic emotion recognition: A benchmark comparison of performances. In *IEEE Workshop on Automatic Speech Recognition & Understanding*. ASRU 2009, pp. 552–557, 2009.

Shah, M., Chakrabarti, C., Spanias, A. (2014). A multi-modal approach to emotion recognition using undirected topic models. IEEE International Symposium on Circuits and Systems, pp. 754–757,.

Shen, L., & Wang, W. (2018). Improving speech emotion recognition based on tobi phonological representation. In PATTERNS 2018, The Tenth International Conference on Pervasive Patterns and Applications, pp.1–5.