Natural Language Engineering (2021), **27**, pp. 747–761
doi:10.1017/S1351324920000339

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

Improving speech emotion recognition based on acoustic words emotion dictionary

Wang Wei[1], Xinyi Cao[1], He Li[1,2], Lingjie Shen[1], Yaqin Feng[1] and Paul A. Watters[2]*

[1]School of Education Science, Nanjing Normal University, Nanjing, 210097, China and [2]Department of Computer Science and Information Technology, La Trobe University, Melbourne, Australia
*Corresponding author. E-mail: P.Watters@latrobe.edu.au

(Received 18 December 2018; revised 11 May 2020; accepted 12 May 2020; first published online 10 June 2020)

Abstract

To improve speech emotion recognition, a U-acoustic words emotion dictionary (AWED) features model is proposed based on an AWED. The method models emotional information from acoustic words level in different emotion classes. The top-list words in each emotion are selected to generate the AWED vector. Then, the U-AWED model is constructed by combining utterance-level acoustic features with the AWED features. Support vector machine and convolutional neural network are employed as the classifiers in our experiment. The results show that our proposed method in four tasks of emotion classification all provides significant improvement in unweighted average recall.

Keywords: Speech emotion recognition; Emotion dictionary; Deep learning; Support vector machine

1. Introduction

Speech is a powerful and efficient way to express emotions. People typically perceive others emotional state mostly from human speech. Hence, due to the importance of human speech, speech emotion recognition (SER) is becoming a research field of growing interest. It has had many real-life applications in the human–computer interaction (HCI) field during the last decade, for example, mental health diagnostics (Yang *et al.* 2012) and computer tutoring applications (Litman and Forbes 2003). Having access to the affective state of a user makes HCI not only more effective but also ensures that any generated behaviors are more human-like.

Since emotion is a suprasegmental phenomenon, many researchers have explored different segments of audio signals, such as frames, words, and utterances (Schuller and Rigoll 2006) in SER. They compute features from different segments with time-continuous acoustic low-level descriptors (LLDs) and various functionals, such as mean and percentile. Mirsamadi, Barsoum, and Zhang (2017) used frame-level LLDs to classify four emotion classes with recurrent neural networks (RNNs) and achieved 58.8% in unweighted average recall (UAR). Fayek, Lech, and Cavedon (2017) classified four emotions with two types of deep learning architectures (feed-forward and recurrent architectures). They used log mel filter banks computed from the frame level as the features to classify discrete emotions, and achieved utterance-level prediction by averaging posterior probabilities across all frames in that utterance. Wollmer, Metallinou, and Katsamanis (2012) proposed a word-level emotion classifier to generate predictions for words within a sentence and then combined the predictions from these words to obtain a sentence-level decision. Their experimental results on two different data sets showed that their proposed method significantly outperformed the standard sentence-based classification approach. Cortes and Vapnik (1995) trained reliable phoneme-level emotional models to improve the SER with very intense emotional content, and the results showed improvement compared with the baseline system at the utterance level. Fernandez and Picard (2011) developed and applied a class of hierarchical directed graphical models on the task of recognizing affective categories, by combining features from words and utterances. One strength of this new approach was the integration and summarization of information from both the word level and the utterance level, which showed better performance than the approach based on the utterance level alone. Cao, Savran, and Verma (2015) divided utterances into regions of interest based on words and proved that this detailed representation, combined with utterance-level features, could provide an improvement in SER.

However, there are some pervasive limitations about SER research on the frame, word, and utterance levels. Although there are results indicating that global features from utterances are highly predictive (Ozkan, Scherer, and Morency 2012; Wollmer *et al.* 2012) in SER, these global features contain some noisy information from parts of speech which is not emotionally salient. Thus, SER on the utterance level is not accurate and not reliable in some cases. In contemporary research about SER on the word level, almost all methods assume that the emotion label for a word is the same as that for the sentence containing the word (Shami and Kamel 2005; Schuller and Rigoll 2006; Cao *et al.* 2015) and consider every word has the same weight of discriminating different emotions. Due to the fact that not all the parts of the utterance are emotionally salient, it is not reasonable to use the same label as the sentence. Since this research does not generate an acoustic words emotion dictionary (AWED), they dont take the effect of different acoustic words on SER into consideration. In research about SER on the frame level, features from the frame do not contain any linguistic information about emotions, and so using frame-level features to classify emotions is a weak method in terms of interpretability. To cope with those problems, we propose a method to quantify the emotionally discriminative power of words by making an emotion dictionary with acoustic words. Then, we select emotional acoustic words from the emotion dictionary to form the utterance representations. Finally, by combining the utterance representations with the emotion dictionary, and the acoustic features extracted from the utterance level, our proposed SER method can provide improvements in accuracy compared to the baseline system with acoustic features from the utterance level alone.

The remainder of this paper is organized as follows. Section 2 introduces related work about features, classifiers, and research on the interactive emotional dyadic motion capture (IEMOCAP) database. Section 3 describes the proposed methodology of SER with the AWED. Section 4 presents the experiments and results of our proposed method and makes comparisons with the baseline system comprising utterance-level acoustic features alone. Finally, Section 5 draws some important conclusions and describes our future work.

2. Related work

In this section, we will introduce the related work of SER in terms of emotion models, classifiers, features, and research on the IEMOCAP database.

2.1 Emotion models

There are two popular emotion models. The first one is a discrete model. It claims that only a few discrete emotions exist. Category labels drawn from every data language are the most familiar for describing emotions. However, the size of an emotion lexicon is remarkable. To make emotion research more feasible, a set of six basic emotions was proposed by Ekman (1992), happiness, sadness, anger, fear, surprise, and disgust, to be universal. The other emotions can be regarded as combinations or variations of these six. The second one is the dimensional model, which is

https://doi.org/10.1017/S1351324920000339 Published online by Cambridge University Press

an alternative to the category model. The dimensional model states that emotions can be distinguished by means of certain characteristics. Based on these dimensional models, emotions can be labeled by specifying a value for each dimension. It is widely thought that emotions can be characterized into only 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 Classifiers in SER

In terms of classifiers in SER, the support vector machine (SVM) (Cortes and Vapnik 1995) is a widely used machine learning method in SER, due to its good performance on small data sets, and its advantage of dealing with linearly nonseparable problems with different kernel functions. Mariooryad and Busso (2013) used SVM to recognize speech emotions on the IEMOCAP data set with 50.64% in accuracy. It provides an observable way to compare the impacts of features extracted and selected on outcomes. However, SVM cannot handle large data sets or form good representations during learning. Deep learning method is another popular and effective way to classify data using acoustic processing. Researchers have proposed a convolutional neural network (CNN) to learn high-level features from the original input data (Mao *et al.* 2014) and to combine RNN with CNN together to learn sequential information from speech data (Keren and Schuller 2016; Trigeorgis *et al.* 2016).

2.3 Features in SER

In terms of features in SER, many research papers and competitions have proposed different standard feature sets for people to compare classification results from the same feature set. Schuller, Steidl, and Batliner (2009) proposed the 384 features from LLDs and functionals for SER. Schuller *et al.* (2010a) proposed 1582 acoustic features obtained by brute-force generation from 38 LLDs and 21 functionals. Eyben *et al.* (2016) proposed 62 parameters that perform similarly with large feature sets on SER. These features are shown to have good performance in SER. Neumann and Vu (2017) observed that the hand-crafted feature set Geneva minimalistic acoustic parameter set (Eyben *et al.* 2016) performs better than the low-complexity log mel filter banks.

2.4 Related work on the IEMOCAP database

In terms of different segment levels, Table 1 lists the results on the IEMOCAP data set with acoustic features only. Most research focuses on features from utterance segments. The best UAR on the IEMOCAP database is 58.46% using hierarchical binary Bayesian logistic regression (Lee *et al.* 2011) on the utterance level. Researchers assume that every frames label is equal to the sentences label (Mirsamadi *et al.* 2017; Neumann and Vu 2017). Then, the posterior class probabilities computed for each frame in an utterance are averaged across all frames in that utterance, and the utterance-based label is selected based on the maximum average class probabilities. However, so far, there has not been any research focus on word-level SER on the IEMOCAP database.

3. Method

In this section, we propose the method of SER from the word level with the AWED. First, M features are selected, and every continuous features value is divided into Q discrete values. Then, the AWED is generated by selecting the top N words using the term-frequency inverse-document-frequency (tf-idf) theory. Finally, utterance-level acoustic features and word-level vectors from the emotion dictionary are concatenated together to improve SER.

https://doi.org/10.1017/S1351324920000339 Published online by Cambridge University Press

Table 1. The literature of SER on the IEMOCAP database with four discrete emotions

Segment	Classifiers	Test UAR(%)
Utterance	Hierarchical binary Bayesian logistic regression (Lee *et al.* 2011)	58.46
Utterance	SVM (Mariooryad and Busso 2013)	50.64
Utterance	Replicated softmax model (RSM)+SVM (Shah, Chakrabarti, and Spanias 2014)	57.39
Utterance	CNN (Fayek *et al.* 2017)	58.28
Frame	Bi-directional long short-term memory RNN with attention mechanism (Mirsamadi *et al.* 2017)	58.8
Frame	Attentive CNN (Neumann and Vu 2017)	56.1

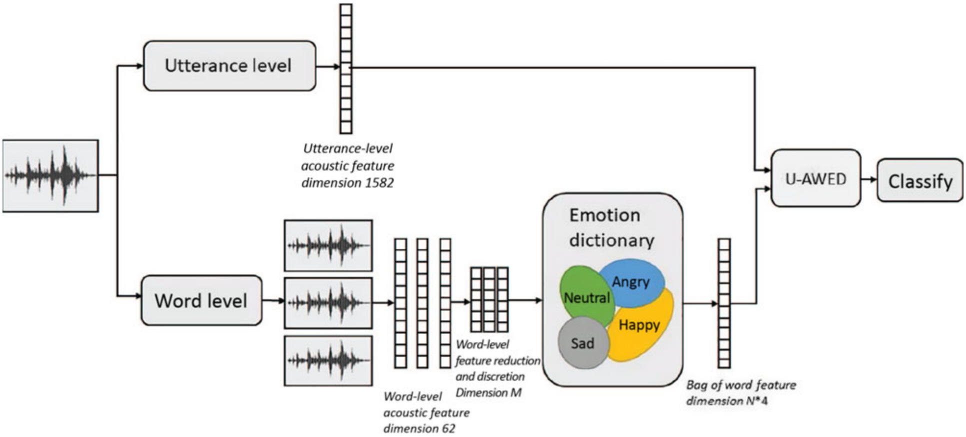
Utterance level
Utterance-level acoustic feature dimension 1582
Word level
Word-level acoustic feature dimension 62
Word-level feature reduction and discretion Dimension M
Emotion dictionary
Angry
Neutral
Happy
Sad
Bag of word feature dimension N*4
U-AWED
Classify

Figure 1. The diagram of the proposed U-AWED method of SER based on the utterance and word levels, respectively.

3.1 Proposed SER architecture

In this section, a method is proposed to combine the utterance-level and word-level SER together as shown in Figure 1. First, we segment the utterance into the word level automatically, and then each words acoustic features are extracted with the Gemaps (Eyben *et al.* 2016) feature set. Due to the fact that these acoustic features are very sensitive and have broad value distribution, even the same acoustic word with the same pronunciation still get different feature values among different persons, even within the same person. So, to describe the significant acoustic characteristics, dimension reduction and feature value discretization are designed to set up a novel acoustic word features subset. The number of features M that we selected will be tested in the experiment. Then, every feature is divided into Q discrete values which will also be tested in the following experiment. After transforming the continuous values to discrete values, the AWED can be generated by counting every words frequency and computing their tf-idf value. We select the top N basic vocabulary acoustic words vector from each emotion class as the emotion dictionary of each class. Next, based on AWED, the final dimension of every utterance is N*4, where N is the number of emotion words. If the basic acoustic word vocabulary vector is in the utterance, then the value in specific position is assigned as 1, otherwise 0. At last, we combine the utterance-level acoustic features and word-level vectors with the emotion dictionary together to classify speech emotions.

https://doi.org/10.1017/S1351324920000339 Published online by Cambridge University Press

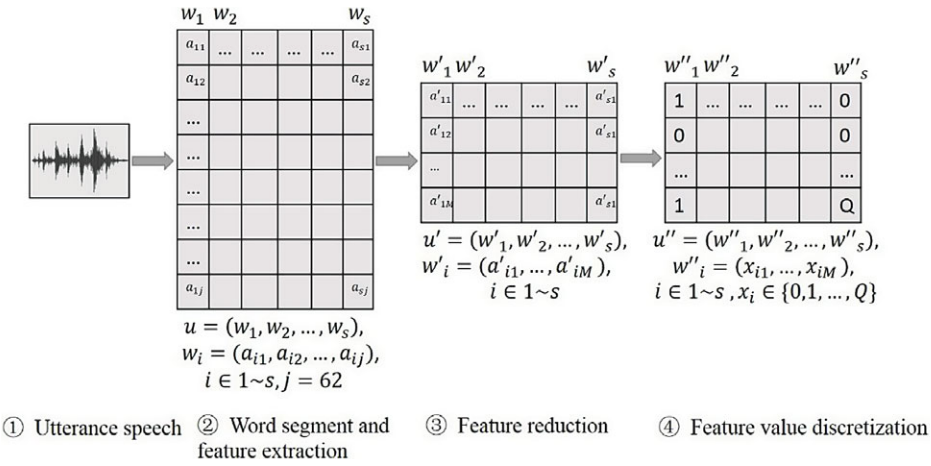
w_1 w_2 w_s
a_{11} … a_{s1}
a_{12} a_{s2}
a_{1j} a_{sj}
$u = (w_1, w_2, \ldots, w_s)$,
$w_i = (a_{i1}, a_{i2}, \ldots, a_{ij})$,
$i \in 1\sim s, j = 62$
w'_1 w'_2 w'_s
a'_{11} a'_{s1}
a'_{12} a'_{s1}
a'_{1M} a'_{s1}
$u' = (w'_1, w'_2, \ldots, w'_s)$,
$w'_i = (a'_{i1}, \ldots, a'_{iM})$,
$i \in 1\sim s$
w''_1 w''_2 w''_s
1 0
0 0
1 Q
$u'' = (w''_1, w''_2, \ldots, w''_s)$,
$w''_i = (x_{i1}, \ldots, x_{iM})$,
$i \in 1\sim s, x_i \in \{0,1,\ldots,Q\}$
① Utterance speech
② Word segment and feature extraction
③ Feature reduction
④ Feature value discretization

Figure 2. The diagram of feature generation. It follows the step from word segment to feature value discretization. First, the utterance u is represented by acoustic word w_i with 62 acoustic features. After feature selection, the utterance u' is represented by acoustic words w'_i with M acoustic features. Finally, after feature value discretization, the utterance u'' is represented by acoustic words w''_i with value from 0 to Q-1.

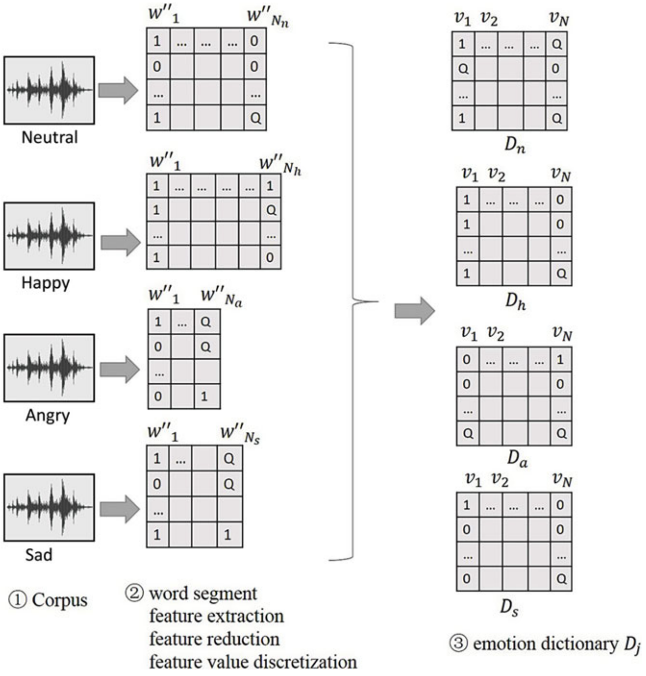

Figure 3. The diagram of AWED generation in each emotion classes. Every speech utterance follows five steps to generate emotion dictionary, namely, word segment, feature extraction, feature selection, feature value discretization, and computing every acoustic words tf-idf. The number of basic acoustic word vocabulary vector representations is N for each emotion.

3.2 Feature dimension reduction and feature value discretization

First, every vocal word is removed from utterance speech based on the forced-alignment transcription provided in the IEMOCAP database. Then, every acoustic words Gemaps acoustic features are extracted with Opensmile (Eyben *et al.* 2010). The acoustic word is represented as

$$w_i = (a_{i1}, a_{i2}, \ldots, a_{ij}) \tag{1}$$

https://doi.org/10.1017/S1351324920000339 Published online by Cambridge University Press

where i is the index of acoustic word in the utterance, s is the total number of words that this utterance contains, and a_{i1} is the first acoustic feature in acoustic word w_i. Next, features dimension reduction is implemented in order to select significant features that can effectively reflect speech emotion characteristics. SVM is used to choose the top M features that are most discriminative. The reduced acoustic word feature is represented as

$$w'_i = (a'_{i1}, a'_{i2}, \ldots, a'_{ij}) \tag{2}$$

where M is the number of features we selected. Then, every continuous feature value is transformed into Q discrete values because it would be extremely difficult to count each words frequency with continuous features. For example, the value is higher than its 50th percentile, then the specific feature value is 1, otherwise 0. The feature then finally becomes

$$w''_i = (x_{i1}, x_{i2}, \ldots, x_{iM}), x_i \in \{0, 1, \ldots, Q\} \tag{3}$$

Finally, the utterance is represented as

$$u'' = (w''_1, w''_2, \ldots, w''_s) \tag{4}$$

The procedure of feature generation is shown as Figure 2.

3.3 Generation of AWED

We follow a similar way of making the emotion dictionary in document classification as we make an AWED. First, utterance speech with the same emotion labels are combined together as the corpus. Then, we follow the previously introduced step to generate every utterance representation from the word level. Next the AWED is generated by computing every words tf-idf value. The term frequency of every word is calculated as

$$tf_{i,j} = \frac{n_{ij}}{\sum_k n_{k,j}} \tag{5}$$

where $n_{i,j}$ is the frequency that acoustic word w''_i appeared in emotion j, n_{kj} is the number of acoustic words in emotion j, that is, N_n in neutral, N_h in happy, N_a in angry, and N_s in sad. Then, we compute the $idf_{i,j}$ of word w''_i, which is formed as

$$idf_{i,j} = log \frac{|c|}{1 + c_{w''_i}} \tag{6}$$

where $|c|$ is the number of emotion classes, and $c_{w''_i}$ is the number of emotion classes that audio word w''_i appeared in.

Next, tf-idf of every acoustic word w''_i in each emotion class j is computed by multiplying $tf_{i,j}$ and $idf_{i,j}$. We select around top N acoustic words in each four emotion classes based on the ranking of acoustic words tf-idf value to generate AWED. Finally, each emotion has its own AWED as follows:

$$D_j = (v_1, v_2, \ldots, v_N) \tag{7}$$

where N is the number of emotion words in each class and v_N is the basic acoustic word vocabulary vector representation.

The procedure of Generation of AWED is shown as Figure 3.

3.4 Emotion classification with U-AWED vectors based on AWED and utterance-level acoustic features

After feature reduction and discretization, every utterance u''' is represented by basic word vocabulary vector representation from each AWED with length 4^*N as follows:

$$u''' = (v'_1, v'_2, \ldots, v'_{4*N}), v'_n \in \{1, 2\} \tag{8}$$

https://doi.org/10.1017/S1351324920000339 Published online by Cambridge University Press

Table 2. The distribution data of each emotions in the IEMOCAP data set.

	Valence			Activation			Dominance			
	Negative	Neutral	Positive	Low	Medium	High	Weak	Medium	Strong	Total
Neutral	568	736	404	924	498	286	808	485	415	1708
Angry	1049	36	18	121	187	795	46	102	955	1103
Happy	61	115	1460	259	397	980	525	310	801	1636
Sad	903	119	62	694	236	154	516	233	335	1084
Total	2581	1006	1944	1998	1318	2215	1895	1130	2506	5531

where 1 stands for the occurrence of a corresponding audio word v_n and 0 stands for non-occurrence. Finally, concatenated with utterance acoustic feature $A = (A_1, A_2, \ldots, A_{1582})$, the fused feature is $F = (a_1, a_2, \ldots, a_{1582}, v'_1, v'_2, \ldots, v'_{4*N})$. We use this fused feature from AWED and utterance-level acoustic feature to classify four emotions with SVM and CNN.

4. Experiments and results

In this section, we will introduce the data, preprocessing, and training steps used in this paper.

4.1 Data and preprocessing

The database used in this work is the IEMOCAP database which contains approximately 12 h of acoustic-visual data from five mixed gender pairs of actors (Busso *et al.* 2008). Each recorded session lasts approximately 5 min and consists of two actors interacting with each other in scenarios that encourage emotional expression. In this study, we only focus on the acoustic channel to perform SER. The time-aligned boundaries of each word are provided in the IEMOCAP corpus. We can identify acoustic words according to the timing markers.

We use two tags from the database: the categorical and dimensional tags. The categorical tags used were neutral, angry, happy, and sad (we merge happy and excitement together as happy) (Mariooryad and Busso 2013; Jin, Li, and Chen 2015; Fayek *et al.* 2017). In total, the data used in our experiments comprised 5531 utterances with an average duration of 4.5 s. We split the valence, activation, and dominance into three levels: level 1 contains ratings in the range [1, 3), level 2 contains ratings equal to 3, and level 3 contains ratings in range (3, 5]. These levels intuitively correspond to low, medium, and high activation, to negative, neutral, and positive valence, and weak, medium, and strong dominance. This was done in order to have a clearer view of the relation between emotions and features and to reduce the influence of imbalanced classes (Tian, Moore, and Lai 2015). In Table 2, we show how the utterances of each categorical class break down into three dimensions.

In previous work on the IEMOCAP database, LLDs are also widely used for acoustic models (Metallinou *et al.* 2012; Wollmer *et al.* 2012). Therefore, we use the openSMILE toolkit (Eyben *et al.* 2010) to extract the acoustic features based on utterance and word segments with the baseline feature set of INTERSPEECH 2010 paralinguistic challenge (Schuller, Steidl, and Batliner 2010b) and Gemaps feature set (Eyben *et al.* 2016). The dimension of the emotional feature vector is 1582 and 62, respectively.

The normalization method has an effect on the experiment results. The goal of normalization is to eliminate speaker and recording variability while keeping the emotional discrimination. For our

https://doi.org/10.1017/S1351324920000339 Published online by Cambridge University Press

experiment, z-score normalization is implemented on all data to get a mean of 0 and a standard deviation of 1, meaning that our SEM is speaker independent.

4.2 Training recipe

A 10-fold leave-one-speaker-out cross-validation scheme (Schuller *et al.* 2010a) was employed in experiments using nine speakers as training data and one speaker as test data. As is standard practice in the field of automatic SER, the results are reported using UAR to reflect imbalanced classes. The equation is as follows:

$$UAR = \frac{1}{N}\sum_{N}^{i=1}\frac{c_i}{n_i} \times 100\% \tag{9}$$

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.

The classifier used in our experiment is SVM and CNN. SVMs complexity is from 0.1 to 1, and its kernel function is radial basis function. The architecture and configuration of CNN are as follows. We use one-dimensional convolution in CNN. We have two convolutional layers each followed with Batch Norm, Rectified Linear Unit, and a maximum pooling layer. The output layer utilizes a softmax nonlinearity instead of the nonlinear function used in previous layers. The base learning rate is set to 10^{-4} and the optimizer is Adam. The epoch is 20 and the training batch size is 32. The L2 regularization on weight is 0.01 added to the convolutional and dense layers. All filter sizes in the experiment are set to 10 and the max-pooling size is 2. We have four CNN with different topologies. In CNN classification, we present an in-depth exploration of various convolution architectures by changing the hidden neurons of fully connected layers and the number of filters.

4.3 Parameters setup

In our proposed method of SER, we need to verify three parameters, namely, the number of features selected $\boldsymbol{M}$, the number of discrete values $\boldsymbol{Q}$, and the number of emotion words N selected in AWED. We implement two experiments to choose the best parameters. The algorithm we used is SVM with radial basis function kernel and 1 complexity using word-level feature vector from AWED.

First, we decide the number of features $\boldsymbol{M}$ and the number of discrete values $\boldsymbol{Q}$ because they are the base to the rest experiments. If the value of M is too large, the emotion acoustic pattern would be overfit and the UAR will be lower. If it is too small, some information would be lost. Q is the every continuous features discrete value. If the discrete interval for value of the features is too wide, it would miss some emotion information. If the discrete interval is too narrow, it would overfit to lower the performance of classifier. It is extremely difficult to count each words frequency with continuous features. We try the value of $\boldsymbol{M}$ from 4 to 30 and $\boldsymbol{Q}$ from 2 to 6. The results are shown in Table 3 and Figure 4 for four discrete emotions, Table 4 and Figure 5 for activation dimension, Table 5 and Figure 6 for valence dominance, and Table 6 and Figure 7 for dominance, which indicate that when the dimensions of feature vector $\boldsymbol{M}$ are 4, 4, 6, and 4, respectively, and the numbers of feature value discretization $\boldsymbol{Q}$ are 3, 3, 4, and 2, respectively, the classification performance in four discrete emotions, as well as activation, valence, and dominance with AWED are the best.

Second, we tested the number of emotion words $\boldsymbol{N}$ in our AWED. Due to the fact that after selecting four features and dividing continuous feature into three discrete values in four discrete emotion and activation classification, the minimum number of acoustic words in each emotion is 53. So, we choose the number of emotion words ranging from 25 to 53 in four discrete emotion and activation classification, then the result is shown in Figures 7 and 8, respectively. The result

https://doi.org/10.1017/S1351324920000339 Published online by Cambridge University Press

Table 3. UAR for the given number of features ***M*** and the number of feature discretizations ***Q*** in four discrete emotion classifications.

	M					
Q	4	6	8	10	20	30
2	42.93	39.53	40.25	39.06	36.26	31.28
3	**43.99**	41.20	39.79	38.45	30.87	25.34
4	43.39	40.29	38.26	36.57	28.39	28.14
5	42.66	37.99	38.02	32.81	29.07	27.94
6	43.93	38.10	35.51	30.53	29.74	28.13

Table 4. UAR for the given number of features ***M*** and the number of feature discretizations ***Q*** in activation classification.

	M					
Q	4	6	8	10	20	30
2	51.45	50.80	51.89	50.10	43.05	34.13
3	**52.09**	50.91	50.59	47.91	33.94	35.68
4	50.71	50.20	49.87	45.96	35.41	36.47
5	50.16	49.17	48.60	37.64	36.37	35.70
6	51.13	48.18	39.95	39.49	37.96	34.67

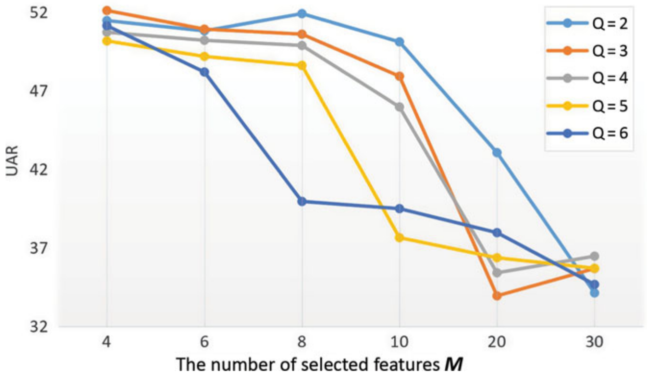

Figure 4. UAR of activation classification based on AWED for the given number of selected features ***M*** and the given number of feature value discretizations ***Q***

indicates that when the number of emotion words we choose in the emotion dictionary is 53, the classification performances in two classification tasks are the best. So, we choose 53 as the value of parameter ***N*** in four discrete emotion and activation classification. Due to the fact that after selecting six and four features and dividing continuous feature into four and two discrete values in valence and dominance classification, respectively, the minimum number of acoustic words in each emotion is 2000 and 16, respectively, so we choose the number of emotion words ranging from 500 to 2000 in valence classification and 6 to 16 in dominance classification, respectively, and

https://doi.org/10.1017/S1351324920000339 Published online by Cambridge University Press

Table 5. UAR of valence classification for the given number of features ***M*** and the number of feature discretizations ***Q*** in valence classification.

	M					
Q	4	6	8	10	20	30
2	35.27	34.95	34.87	34.79	34.62	33.54
3	35.91	35.70	33.58	33.25	32.99	32.90
4	34.98	**35.96**	33.22	34.57	33.22	33.16
5	35.58	34.76	33.87	33.84	33.63	32.74
6	34.81	34.88	34.04	33.81	33.58	33.35

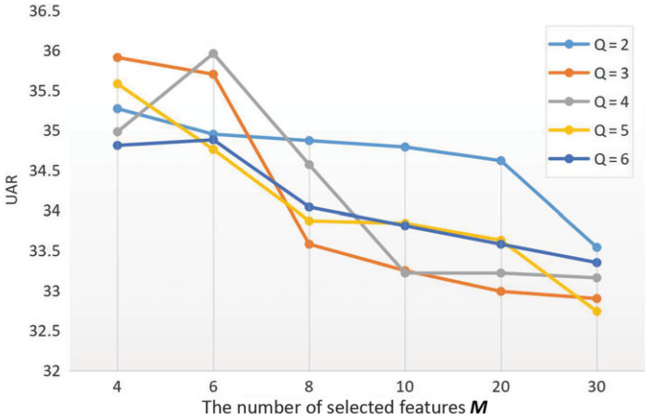

Figure 5. UAR of valence classification based on AWED for the given number of selected features ***M*** and the given number of feature value discretizations ***Q***.

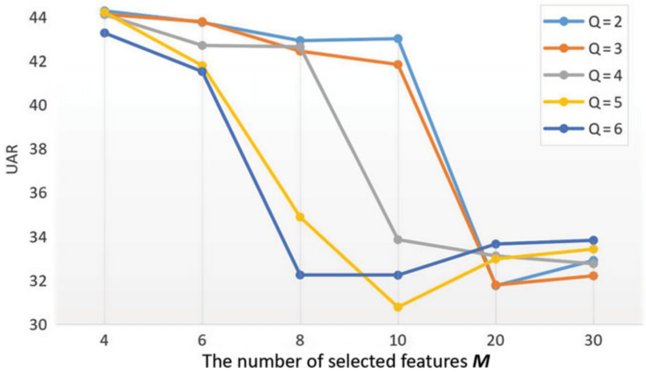

Figure 6. UAR of dominance classification based on AWED for the given number of selected features ***M*** and the given number of feature value discretizations ***Q***.

the result is shown in Figures 9 and 10, respectively. The result indicates that when the number of emotion words we choose in the emotion dictionary is 1800 and 9 in valence and dominance classification, the classification performances in two classification tasks are the best. So we choose 1800 and 9 as the value of parameter ***N*** in valence and dominance classification, respectively.

https://doi.org/10.1017/S1351324920000339 Published online by Cambridge University Press

Table 6. UAR of dominance classification for the given number of features ***M*** and the number of feature discretizations ***Q*** in dominance classification.

	M					
Q	4	6	8	10	20	30
2	**44.28**	43.76	42.93	43.02	31.76	32.91
3	44.12	43.79	42.45	41.84	31.79	32.21
4	44.11	42.71	42.65	33.86	33.12	32.76
5	44.22	41.79	34.89	30.78	32.98	33.43
6	43.28	41.52	32.25	32.24	33.66	33.83

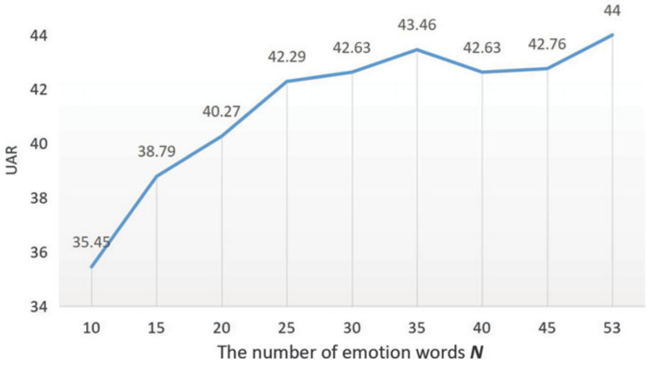

Figure 7. UAR of four discrete emotion classification based on AWED for the given emotion words size *N*.

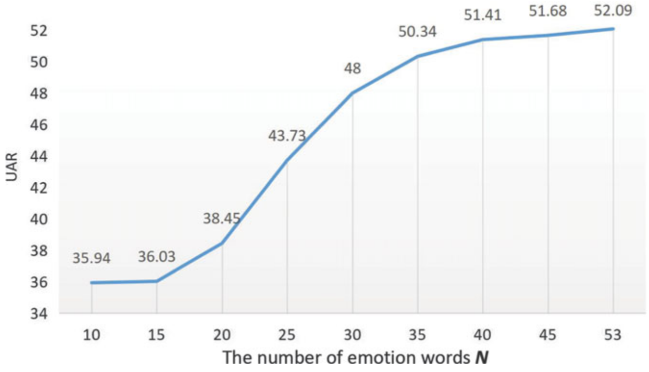

Figure 8. UAR of activation classification based on AWED for the given emotion words size *N*.

4.4 Results

The results of our proposed SER based on word level with AWED are shown in Table 7. In this experiment, we have decided the parameters of $\boldsymbol{M}$, $\boldsymbol{Q}$, and $\boldsymbol{N}$, namely, 4, 3, and 53 in four discrete emotion classification, 4, 3, and 53 in activation classification, 6, 4, and 1800 in valence activation, and 4, 2, and 9 in dominance classification. The baseline system is SVM and CNN with acoustic features from the utterance level. Overall, the proposed methods classification best results in four discrete emotions, activation, valence, and dominance are 61.70%, 57.47%, 49.38%, and 47.25%, respectively, where the results on four discrete emotion classification outperform the best results

https://doi.org/10.1017/S1351324920000339 Published online by Cambridge University Press

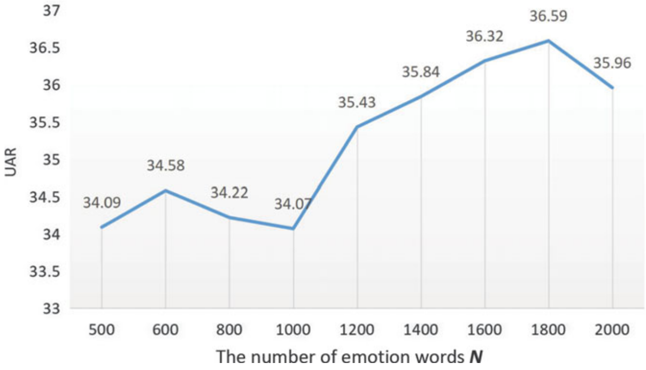

Figure 9. UAR of valence classification based on AWED for the given emotion words size *N*.

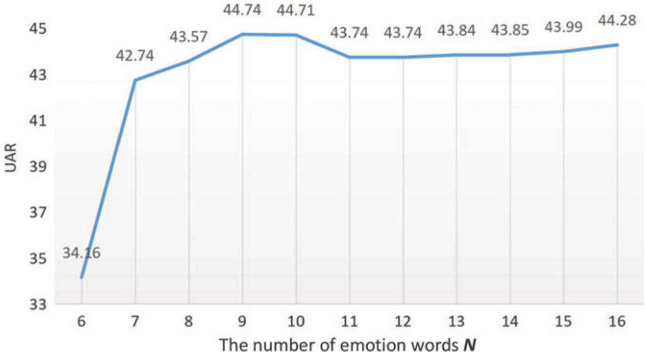

Figure 10. UAR of dominance classification based on the AWED for the given emotion words size *N*.

on the IEMOCAP database hence become the state of the art. The results show that our proposed method in four classification tasks provides 1.69%, 0.83%, 1.8%, and 2.14% improvement in UAR, respectively.

Although the recognition performance based on the word level with emotion dictionary is not as good as the performance based on the utterance level, combining word-level vector from the emotion dictionary and utterance-level acoustic features can provide improvement in accuracy. The results probably indicate that although the global features from the utterance level are more predictive than local words which are proved in Wollmer *et al.* (2012), word-level vector representation based on emotion dictionary can provide complementary information from local words to improve SER. In terms of classifiers, SVM with radial basis function kernel whose complexity is 1 performs best among all the classifiers. Compared with research in, they only use linear kernel in SVM whose result is much worse than ours. CNN outperforms the SVM in activation and valence classification.

We employed the t-test of independent samples to evaluate the statistical significance of SERs UAR between the U-AWED and utterance methods. First, we got 10 UAR by SVM classifier separately based on the U-AWED and utterance methods on the IEMOCAP database. Then, IBM SPSS 20, IBM company USA was used for independent samples t-test. The p-value for the categorical tags data is 0.046. The p-value for activation dimension is 0.036. The p-value for valence is 0.032. The p-value for dominance is 0.021. It indicates that the p-value for the categorical and every dimensional tags are all less than 0.05. These results suggest that the U-AWED improves the SER in a statistically significant manner.

https://doi.org/10.1017/S1351324920000339 Published online by Cambridge University Press

Table 7. Results of UAR on two classifiers: SVM with four different complexity (C) values and CNN with three different topologies. *CON1D*($m \times j$) denotes a one-dimensional convolutional layer of m and j filters, respectively, with size 10 and stride 2. *Softmax*(n_0) denotes a softmax output layer of n_0 units.

Emotions	Classifier	Utterance	AWED	U-AWED	UAR
Four discrete emotions	SVM ($C = 0.1$)	54.36	38.77	54.37	0.01
	SVM ($C = 0.5$)	59.84	43.86	60.05	0.21
	SVM ($C = 0.8$)	61.15	44.14	61.42	0.27
	SVM ($C = 1$)	61.50	44.00	61.70	0.2
	Conv1D (816)-Softmax(4)	58.22	39.52	59.91	1.69
	Conv1D (1632)-Softmax(4)	59.48	44.42	59.95	0.47
	Conv1D (3264)-Softmax(4)	59.26	45.97	60.05	0.79
Activation	SVM ($C = 0.1$)	54.74	51.21	54.78	0.04
	SVM ($C = 0.5$)	55.83	51.79	55.84	0.01
	SVM ($C = 0.8$)	55.92	51.70	55.99	0.07
	SVM ($C = 1$)	55.98	52.09	56.06	0.08
	Conv1D (816)-Softmax(4)	56.64	51.25	57.47	0.83
	Conv1D (1632)-Softmax(4)	55.55	51.52	56.05	0.5
	Conv1D (3264)-Softmax(4)	54.49	51.92	55.17	0.68
Valence	SVM ($\mathrm{C} = 0.1$)	44.56	34.47	44.59	0.03
	SVM ($\mathrm{C} = 0.5$)	46.70	35.98	46.76	0.06
	SVM ($\mathrm{C} = 0.8$)	47.16	36.36	47.32	0.16
	SVM ($\mathrm{C} = 1$)	47.81	36.59	47.88	0.07
	Conv1D (816)-Softmax(4)	47.59	35.78	48.85	1.26
	Conv1D (1632)-Softmax(4)	48.64	35.89	48.75	0.11
	Conv1D (3264)-Softmax(4)	47.58	36.29	49.38	1.8
Dominance	SVM ($\mathrm{C} = 0.1$)	44.67	45.05	45.05	0.38
	SVM ($\mathrm{C} = 0.5$)	45.53	44.63	45.63	0.1
	SVM ($\mathrm{C} = 0.8$)	44.74	44.74	46.88	2.14
	SVM ($\mathrm{C} = 1$)	47.20	44.74	47.25	0.05
	Conv1D (816)-Softmax(4)	45.75	43.83	46.88	1.13
	Conv1D (1632)-Softmax(4)	45.69	44.70	46.27	0.58
	Conv1D (3264)-Softmax(4)	45.06	44.93	45.77	0.71

5. Discussion

A U-AWED SER method is proposed based on word level by making AWED. This method combined with utterance-level features outperforms the state-of-the-art research in four discrete emotion classification. We make AWED to select emotionally salient words in each emotion

https://doi.org/10.1017/S1351324920000339 Published online by Cambridge University Press

classes so that we can highlight specific acoustic words weight in SER. Compared with the SER on the frame level (Fayek *et al.* 2017; Mirsamadi *et al.* 2017; Neumann and Vu 2017), our proposed method provides more knowledge about linguistic information which highlights the pattern of different acoustic words in expressing various emotions.

While the results of only word-level classification is not as good as utterance-level classification, the fused method provides good performance. It means probably the utterance-level classification includes more information of emotion expression than the word level. Or there are still much improving space for this word-level features extracting method.

Meanwhile, compared with research about SER with text and speech information on the same database (Shah, Chakrabarti, and Spanias 2014) whose final test UAR is 61.96% with replicated softmax models and SVM, with language UAR is 54.04%, our method based on AWED has not reached the same performance. The reasons are probably because that although we combine emotion words in each emotion classes, the input of our proposed method is still based on speech modality. Plus, when labeling the data, people integrate both speech and text information into emotion perception, which means that the data we used inherently contain emotion information in both sources. Thus, our proposed method is not as good as the method based on both speech and text. The limitation of this research is that the method of transforming continuous feature value into discrete values is too rough because different features have different distributions among each emotion, therefore just using the linear method to divide the continuous value is not correct.

In the future, we will test and compare the role of acoustic words and text words in emotion recognition and analyze the patterns of acoustic words in different emotions in terms of feature distribution. We will also use features that are useful in speech recognition, for example, *i*-vector feature (Tao et al. 2018), which are proved helpful in SER. We will explore the generalization of our proposed method and improve our methods robustness in the wild using different transfer learning methods.

Acknowledgments. Thanks for all anonymous reviewers and editors. This study was supported by the National Social Science Foundation of China (BCA150054).

Funding. The research is supported by National Education Science Fund of China "To construct contextualized evaluation environment of Chinese language Proficiency with Intelligent interactive technology" (BCA150054).

References

Busso C., Bulut M., Lee C.C., Kazemzadeh A., Mower E., Kim S., Jeannette N., Lee S. and Narayanan S.S. (2008). IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation* **42**(4), 335–359.

Cao H., Savran A. and Verma R. (2015). Acoustic and lexical representations for affect prediction in spontaneous conversations. *Computer Speech & Language* **29**(1), 203–217.

Cortes C. and Vapnik V. (1995). Support-vector networks. *Machine Learning* **20**(3), 273–297.

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

Eyben F., Wöllmer M. and Schuller B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *MM'10 - Proceedings of the ACM Multimedia 2010 International Conference*, pp. 1459–1462.

Eyben F., Scherer K.R., Truong K.P., Schuller B.W., Sundberg J. and Andre E. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing* **7**(2), 190–202.

Fayek H.M., Lech M., Cavedon L. and Wu H. (2017). Evaluating deep learning architectures for Speech Emotion Recognition. *Neural Networks* **92**(1), 60–68.

Fernandez R. (2004). A computational model for the automatic recognition of affect in speech. *Thesis Massachusetts Institute of Technology* **28**(1), 50–58.

Fernandez, R. and Picard R. (2011). Recognizing affect from speech prosody using hierarchical graphical models. *Speech Communication* **53**(9C10), 88–103.

Jin Q., Li C. and Chen S. (2015). Speech emotion recognition with acoustic and lexical features. pp. 4749–4753. doi:10.1109/ICASSP.2015.7178872.

https://doi.org/10.1017/S1351324920000339 Published online by Cambridge University Press

Keren G. and Schuller B. (2016). Convolutional RNN: an enhanced model for extracting features from sequential data. In *2016 International Joint Conference on Neural Networks (IJCNN) as part of the IEEE World Congress on Computational Intelligence (IEEE WCCI)*, Canada: Vancouver, pp. 3412–3419.

Lee C.C., Mower E., Busso C., Lee S. and Narayanan S. (2011). Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication* **53**(9¨C10), 1162–1171.

Litman D. and Forbes, K. (2003). Recognizing emotions from student speech in tutoring dialogues. *Automatic Speech Recognition and Understanding Workshop* **25**(3), 698–704.

Mao Q., Dong M., Huang Z. and Zhan Y. (2014). Learning ssalient features for speech emotion recognition using convolutional neural networks. *IEEE Transactions on Multimedia* **16**(8), 2203–2213.

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

Metallinou A., Wollmer M., Katsamanis A. and Eyben F. (2012). Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE Transactions on Affective Computing* **3**(2), 184–198.

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

Neumann M. and Vu N.T. (2017). Attentive convolutional neural network based speech emotion recognition: a study on the impact of input features, signal length, and acted speech. In *Interspeech*, Stockholm, Sweden, pp. 1263–1267.

Ozkan D., Scherer S. and Morency L.P. (2012). Step-wise emotion recognition using concatenated-HMM. In *14th ACM International Conference on Multimodal Interaction (ICMI)*, pp. 477–484.

Schuller B. and Rigoll G. (2006). Timing levels in segment-based speech emotion recognition. In *INTERSPEECH 2006, International Conference on Spoken Language Processing (ICSLP)*, pp. 1818–1821.

Schuller B., Steidl, S. and 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., Rigoll G. and Wendemuth A. (2010a). Acoustic emotion recognition: a benchmark comparison of performances. In *Automatic Speech Recognition & Understanding, ASRU 2009*, pp. 552–557.

Schuller B., Steidl S., Batliner A., Burkhardt F. and Narayanan S.S. (2010b). The INTERSPEECH 2010 paralinguistic challenge. In *INTERSPEECH 2010, Conference of the International Speech Communication Association*, Makuhari, Chiba, Japan, pp. 2794–2797.

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

Shami M.T. and Kamel M.S. (2005). Segment-based approach to the recognition of emotions in speech. In *IEEE International Conference on Multimedia and Expo*, Amsterdam, pp. 383–389.

Tian L., Moore J.D. and Lai C. (2015). Emotion recognition in spontaneous and acted dialogues. In *International Conference on Affective Computing and Intelligent Interaction*, Xi'an, pp. 698–704.

Trigeorgis G., Ringeval F., Brueckner R., Marchi E. and Zafeiriou S. (2016). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5089–5093.

Wollmer M., Metallinou A., Katsamanis N., Schuller B. and Narayanan S. (2012). Analyzing the memory of BLSTM Neural Networks for enhanced emotion classification in dyadic spoken interactions. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, pp. 4157–4160.

Yang N., Muraleedharan R., Kohl J., Demirkol I., Heinzelman W. and Sturge-Apple M. 2012. Speech-based emotion classification using multiclass SVM with hybrid kernel and thresholding fusion. In *IEEE Workshop on Spoken Language Technology*, Miami, FL, pp. 455–460.

Cite this article: Wei W, Cao X, Li H, Shen L, Feng Y and Watters PA (2021). Improving speech emotion recognition based on acoustic words emotion dictionary. *Natural Language Engineering* **27**, 747–761. https://doi.org/10.1017/S1351324920000339