Negative Emotion Recognition in Spoken Dialogs

Xiaodong Zhang[1], Houfeng Wang[1], Li Li[1], Maoxiang Zhao[2], and Quanzhong Li[2]

[1] Key Laboratory of Computational Linguistics (Peking University),
Ministry of Education, China
[2] Pachira Technology, Inc.
zxddavy@gmail.com, wanghf@pku.edu.cn, li.l@pku.edu.cn,
zmxiangde_88@163.com, quanzhong@hotmail.com

Abstract. Increasing attention has been directed to the study of the automatic emotion recognition in human speech recently. This paper presents an approach for recognizing negative emotions in spoken dialogs at the utterance level. Our approach mainly includes two parts. First, in addition to the traditional acoustic features, linguistic features based on distributed representation are extracted from the text transcribed by an automatic speech recognition (ASR) system. Second, we propose a novel deep learning model, multi-feature stacked denoising autoencoders (MSDA), which can fuse the high-level representations of the acoustic and linguistic features along with contexts to classify emotions. Experimental results demonstrate that our proposed method yields an absolute improvement over the traditional method by 5.2%.

Keywords: Emotion Recognition, Spoken Dialogs, MSDA

1 Introduction

Emotion recognition of speech signals aims to identify the emotional or physical states of a person by analyzing his or her voice [25]. The automatic recognition of emotions in human speech has drawn increasing attention over the past few years, mainly because of the growing number of applications that may benefit from this research field, e.g. call center, man-machine interaction system, and speech recognition, etc. Take call center as an example. By analyzing emotions in spoken dialogs between customers and agents, the managers can find problems in the customer service so as to reduce customer losses. Besides, it can serve as evidences for agent performance evaluation.

Emotion recognition of spoken dialogs is a challenging and cross-disciplinary research area. A variety of acoustic features have been explored by previous work [5, 24, 4]. However, these works have neglected linguistic features. Emotion in an utterance is expressed by not only how it is being said, but also what is being said. Recently, more attention has been paid to the integration of acoustic and linguistic information. For linguistic feature representation, the mainstream is the bag-of-words (BoW) and n-grams model [12, 19, 18]. In previous works, the different kinds of features are combined at input level [12, 18, 21] or at decision level [11, 16]. Both ways have drawbacks. How to represent and combine linguistic information for emotion recognition is worthy of further exploration.

In the past few years, a variety of deep neutral networks for emotion recognition have been studied [22, 27] and have achieved good performance. The deep learning method can learn abstract representation from the raw feature space, and can tolerate noises, making it suitable for spoken language processing.

According to Ayadi's survey on speech emotion recognition [2], most of the previous studies employed speech data recorded from actors who were asked to express the prescribed emotions. Besides, these utterances were produced in isolation without any conversational context. In this work, we focus on recognizing emotions in Chinese spoken dialogs recorded in a call center that serves actual customers. The emotion recognition is at the utterance level. We only consider two categories, i.e. *negative* and *non-negative*, rather than a large variety of emotions, which may be unnecessary for our application. The negative emotion can be used as a strategy to improve the quality of service.

Here we give a brief introduction of our proposed method. First, we extracted some classical acoustic features mentioned in the previous work. Then the speech was transcribed to text automatically by an ASR engine. We employed the distributed representation (embeddings) as linguistic features. Therefore, for each utterance, there are two kinds of features. The contextual information is based on the surrounding utterances. A novel deep learning model, referred to as MSDA, was proposed to fuse the high-level abstractions of acoustic and linguistic features to a unified representation and classify the utterances into two categories.

The rest of the paper is organized as follows. The related work is surveyed in Section 2. The proposed approach is presented in Section 3. The experimental results are detailed in Section 4. Lastly conclusions are given in Section 5.

2 Related Work

The early works on speech emotion recognition have been focused on acoustic features. Various frame-level descriptors have been explored. Banse examined vocal parameters for emotion expression using actors' portrayals of 14 emotions [3]. Pitch, energy, speech rate, and spectral information were used. McGilloway studied 22 different acoustic features for the classification of five emotion states [14]. However, using only acoustic features cannot guarantee a good result because it is just one side of the problem.

Recently, more attentions have been paid to combining acoustic features with other information, especially linguistic information. The BoW and n-grams representation are often used as linguistic features [12, 19, 18]. Raaijmakers compared n-grams at different level (word, character, and phoneme) and concluded that character-level features outperform other two levels [19]. Lee proposed emotional salience to measure how much information a word provides towards a certain emotion [11] and Metze extended it to include bi-grams and tri-grams [16].

Some of the previous works were based on manual transcripts [11, 12], while other studies relied on ASR [16, 21]. In real-word application, only the ASR approach is feasible. Some studies concluded that the recognition errors brought by ASR were consistent enough that it had little influence on the results [16, 13]. However, Rozgi demonstrated that the results based on ASR were much worse than the results based on manual transcription [21]. We believe the opposite conclusions are due to the different dataset and ASR system.

Some researchers employed discourse information for emotion recognition in human-computer interactive system and achieved good results [11, 12]. However, the discourse feature was manually labeled so that it is not feasible for real-world applications. Liscombe augmented standard lexical and prosodic features with contextual features [12]. The contextual features were defined on the difference between present utterance and previous two utterances. In our method, the previous and following utterances are both taken into consideration, and the relation is learned by the neural network, rather than the predefined difference.

For feature combination method, two ways are mainly employed in previous works. One way is to train separated classifiers for different kinds of features and then combine the results of these classifier to make the final classification [11, 16]. However, this way cannot learn the correlation of the different kinds of features and take full advantage of the complementation of them. The other way is to combine the different kinds of features at input level and train a unified classifier [12, 18, 21]. The acoustic and linguistic features generally have distinct statistical characteristics and the correlations between them are nonlinear. Consequently, joining the two kinds of features at low-level representation, e.g. the input layer, may not generate a good unified representation. Besides, this way may suffer from the dimensionality issues. Recently, Kiela proposed a multimodal representation method [10], which concatenated a skip-gram linguistic representation vector with a visual concept representation vector computed using a deep convolutional neural network. However, the different abstractions are just concatenated without learning their correlation. Our approach first learns the high-level abstraction of the acoustic and linguistic features separately and then fuses them to learn a unified high-level representation so that it can overcomes the shortcomings of the above methods.

3 The Proposed Approach

3.1 Features

Acoustic Features The acoustic features were automatically extracted from the speech signal of each utterance by the open source toolkit openSMILE[3]. At first, we computed 26 acoustic features (including MFCC, LSP, F0, Intensity, and MZCR) for each frame (25 ms) with their respective first derivatives. The F0 features contain F0, F0's slope, and the prior probability of voice frames. The

[3] http://www.audeering.com/research/opensmile

intensity features contain the absolute and relative amplitudes in time domain. MFCC and LSP contain the information about the formant and audio coding.

Based on these per-frame features, we computed the statistical features over a whole utterance using the statistics listed in Table 1. Hence, each utterance is represented as a 988-dimensional feature vector: $(1+\Delta)\times(12$ MFCC + 8 LSP + 3 F0 + 2 intensity + 1 MZCR$)\times(4$ regressions + 6 percentiles + 3 moments + 6 extremes). Without performing feature selection, we directly use all extracted features as input, because our model has inherent capability of dimensionality reduction.

Table 1. The statistics for global features

Statistics	**Number**	**Detail**
Regressions	4	two linear regression coefficients, absolute mean and variance of error
Percentiles	6	25%, 50%, 75%, 50%-25%, 75%-50%, 75%-25%
Moments	3	variance, skewness, kurtosis
Extremes	6	max, min, max-min, max position, min position, mean

Linguistic Features To extract linguistic features, we first transcribed the audio data into text via an ASR engine. Our ASR system is mainly composed of five components: feature extraction, acoustic model, language model, lexicon, and decoder. We used log filter-banks [6] with 40 dimensions as acoustic features. The acoustic model, language model, and lexicon were combined into a single weighted finite state transducers (WFST) as in [1]. The ASR system was measured on a dataset containing 40 hours of phone dialogs and the character error rate (CER) is 16%.

Word segmentation is the first step for Chinese text processing. The ICTCLAS[4] was utilized to segment our transcribed text into words. We did not remove any stop words on the consideration that some function words, especially tone words, can contribute to the emotion recognition.

We represented the text of each utterance by the distributed representation rather than the traditional BoW. First, word embeddings were trained by word2vec[5] on the combination of three corpus, namely Chinese Gigaword[6], Chinese Wikipedia[7], and SougouCA[8]. Next, we composed word embeddings to get the distributed representation of the utterance text. Due to the bad performance of a parser on text with ASR errors, composition methods based on a parser [23] are not suitable for our work. Following Hermann's work [8], we represent the text of an utterance by the average of its word embeddings. Formally,

$$f(x) = \sum_{i=1}^{n} x_i \Big/ n \tag{1}$$

[4] http://ictclas.nlpir.org/

[5] https://code.google.com/p/word2vec/

[6] https://catalog.ldc.upenn.edu/LDC2011T13

[7] http://download.wikipedia.com/zhwiki/

[8] http://www.sogou.com/labs/dl/ca.html

where x_i is the embedding of the i-th word in the utterance text x and $f(x)$ is the utterance text vector. As the variance of the length of our text is large, we used the averaged vectors of words or pairs rather than the sum to alleviate the influence of text length. The utterance text vectors served as the linguistic features.

For completely out of vocabulary utterance, the averaged vector of all utterances in the training set serve as the linguistic features. In this case, the classification mainly depends on the acoustic features.

Contextual Information Contextual information in dialogs is useful for emotion recognition. It is natural to use the surrounding utterances as additional evidence to help the emotion recognition of the present utterance.

The acoustic context is defined as the ordered concatenation of the acoustic feature vectors of utterances in a window. Formally,

$$w(t) = [x(t-s), ..., x(t), ..., x(t+s)] \tag{2}$$

where $x(t)$ is the acoustic features of t-th utterance, $w(t)$ is the acoustic features with context of t-th utterance, and the window size is $2 \times s + 1$.

The linguistic context is defined in the same way.

3.2 Emotion Classification

We propose a novel classification model, referred to as multi-feature stacked denoising autoencoder (MSDA). Here, the multi-feature means several kinds of features with different statistical characteristics and non-linear correlation. Figure 1 demonstrates the framework of the model, which mainly includes two parts. In the bottom part, the acoustic features and linguistic features with their respective contexts are employed as inputs to train two stacked denoising autoencoders (SDA) to learn the high-level abstractions independently. Subsequently, in the top part, the two high-level abstractions are fused to generate a unified high-level feature representation by another SDA. Finally, the unified representation serve as the input to a classifier to make the prediction. Next, we introduce the details of MSDA.

The basic building block for MSDA is the denoising autoencoder (DAE) [26], which is an extension of the classical autoencoder. The DAE is trained to reconstruct the input from a partially destroyed version of it, so as to force the hidden layer to discover more robust features. It can be stacked for building deep networks, i.e. the stack denoising autoencoders (SDA).

The unsupervised pretraining of MSDA is performed one layer at a time. Each layer is trained as a denoising autoencoder by minimizing the reconstruction of its input, which is the output of the previous layer. First, in the bottom part, the two SDAs are pretrained layer-wise from bottom to top. Then the outputs of the top layers of the two SDAs are joined together as the input of the top part. The top part is also pretrained layer-wise from bottom to top.

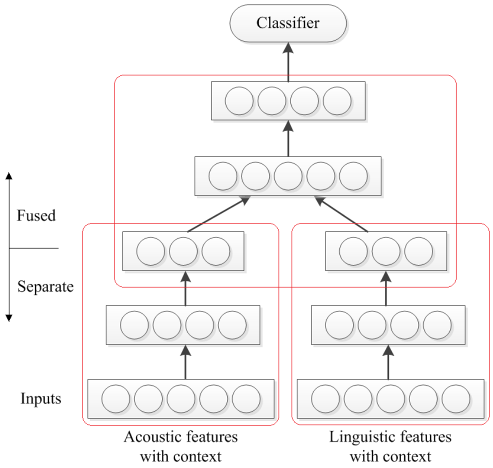
Classifier
Fused
Separate
Inputs
Acoustic features with context
Linguistic features with context

Fig. 1. The framework of MSDA

After pretraining, the MSDA are fine-tuned using labeled data. A classifier is put on the top of the network so as to be trained with the unified high-level representation. In our experiments, the logistic regression (LR) classifier was used. In our dataset, the number of non-negative utterances is far more than the negative utterances. The LR classifier does not perform well on imbalanced datasets. The most common solution of imbalanced learning is sampling [7], however this solution is not applicable to our approach because contextual information is used. Thinking differently, we made a modification on the loss function of LR.

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + \alpha(1-y^{(i)}) \log(1-h_\theta(x^{(i)}))\right] \tag{3}$$

$$h_\theta(x) = \frac{1}{1+\exp(-\theta^T x)} \tag{4}$$

where $x^{(i)}$ is the input, $y^{(i)} \in \{0,1\}$ is the label (1 for negative and 0 for non-negative), and θ are model parameters. Note that the second term of Equation 3 is multiplied by $\alpha \in (0,1]$, which is a penalty factor of non-negative data to alleviate the imbalanced learning problem.

Next we introduce the advantages of MSDA. In the previous work, the different kinds of features were fused at the input layer [12, 18, 21], which may not generate a good unified representation because of distinct statistical characteristics, or after the independent prediction [11, 16], which cannot take full advantage of the complementation of different features. Our approach fuses the acoustic and linguistic features at high-level abstraction and overcomes the shortcomings of the above two methods. It can be regarded as the trade-off of the two methods.

The utterance text was transcribed automatically by an ASR engine. In our dataset, some speakers use dialects and the Chinese dialects are diverse. Besides, the quality of our telephone recordings is not very good. Therefore, the transcription errors must be taken into account. In our model, the denoising autoencoders can discover robust features for classification so as to alleviate this problem.

It should be noted that although our model is introduced with two kinds of features, it can be easily extended to support more kinds of features. What needs to be done is to add more SDAs for more kinds of features and take the outputs of the SDAs as the input of the fused SDA.

4 Experiments and Results

4.1 Dataset and Evaluation Metrics

We collected 254 dialogs (8 kHz, 16 bit WAVE) from a call center, where actual customers are engaged in spoken dialog with human agents over the telephone. The dialogs were then segmented into utterances by a speaker diarization algorithm. The goal of speaker diarization is to segment an audio signal into several acoustic classes, each of which only contains the acoustic data from a single speaker [20]. The speaker diarization in our work mainly includes two processes: detecting speaker change points and unsupervised clustering [15]. To avoid overlong turns, each turn was segmented into one or several utterances based on the lengths of silences. Our speaker diarization algorithm was evaluated on a dataset containing 40 hours of phone dialogs and the diarization error rate (DER) was 7%. We got 34416 utterances after applying speaker diarization on the 254 dialogs.

Three annotators independently labeled each utterance as negative or non-negative. In our study, the negative emotion represents anger and discontent, etc., whereas the non-negative emotion represents the complement, i.e., neutral or positive emotions such as happiness or satisfaction, etc. The annotation was done after listening to the audio corresponding to an utterance. The majority label of the three annotators was taken as the label of an utterance when there were disagreements. After annotation, we got 2437 negative utterances and 31979 non-negative utterances. The constructed dataset is referred to as Emotional Utterances in Chinese Spoken Dialogs (EUCSD).

Most of the previous work is evaluated on a balanced dataset. For comparison, we sampled non-negative utterances from the EUCSD to acquire a subset having a comparable amount to the negative utterances. This balanced dataset is referred to as EUCSD-B. Note that EUCSD-B is just used for comparison experiments, and the results on EUCSD are the final results. For both datasets, we used 70% of the data for training and 30% for testing. The statistics of the two datasets are shown in Table 2.

Table 2. The statistics of the two datasets

Dataset	**Negative**	**Non-negative**	**Training**	**Testing**	**Total**
EUCSD	2437	31979	23993	10423	34416
EUCSD-B	2437	2664	3701	1400	5101

Since we focused on the recognition of the negative emotion, the precision, recall and F1 score for the negative emotion class were employed as the evaluation metrics.

4.2 Comparison between feature combination methods

In the first set of experiments, we compare our proposed approach with some baseline methods. Because the features used in the previous work are diverse, for simplicity we focus on the feature combination methods. Our acoustic and linguistic features are used for all of the following methods.

RG: We use random guess to describe the different difficulties of the two dataset. The results are calculated theoretically, not experimentally. The recall is 0.5, and the precision is the proportion of the negative emotion category.

SVM-A: An SVM classifier using only the acoustic features.

SVM-L: An SVM classifier using only the linguistic features.

SVM-AL-O: Two SVM classifiers are learned from the acoustic and linguistic features separately. The final prediction is combined at the decision level by taking the prediction with the larger posterior probability. This combination method has been employed in [11, 16].

SVM-AL-I: The acoustic and linguistic features are concatenated to unified features as the input of an SVM classifier. This feature combination method has been employed in [12, 18, 21].

MSDA-AL: Our proposed MSDA model is employed for classification but the contextual information is not used.

SVM-CRF: A two-stage SVM/CRF sequence classifier [9]. First, an SVM is trained to predict each individual sequence element. Second, a CRF is trained to predict the whole sequences using the prediction from the previously trained SVM as its input.

MSDA-ALC: Our proposed MSDA model is employed for classification and the contextual information is used. The contextual window size is 3.

In this paper, LIBSVM[9] and CRF++[10] are used for SVM and CRF implementation respectively, and MSDA is implemented using Theano[11]. As there are no unlabeled data in our datasets, we used the labeled data in the training set for unsupervised pretraining. In the pretraining and fine-tuning process, model parameters were optimized based on minibatch stochastic gradient descent and the batch size was 50. For experiments on EUCSD, the penalty factor of non-negative data was 0.2 for both the SVM and LR classifiers, while for experiments on EUCSD-B, the penalty factor was 1, i.e. no penalty. The destruction proportion of denoising autoencoders was 0.5.

The comparison results are shown in Table 3. The contextual information is not involved in Method 1-6, but considered in Method 7 and 8. Due to the

[9] https://github.com/cjlin1/libsvm

[10] http://taku910.github.io/crfpp/

[11] http://deeplearning.net/software/theano/

destruction of contexts caused by the sampling, Method 7 and 8 are not applicable on EUCSD-B. The results on EUCSD are much worse than the results on EUCSD-B, because it is harder to learn a classifier on an imbalanced dataset than a balanced one, especially for the data with lots of ASR errors.

Table 3. Comparison results of feature combination methods

#	**Methods**	**EUCSD-B**			**EUCSD**		
		precision	recall	F1	precision	recall	F1
1	RG	0.478	0.500	0.489	0.071	0.500	0.124
2	SVM-A	0.721	0.767	0.743	0.337	0.439	0.381
3	SVM-L	0.652	0.868	0.745	0.262	0.467	0.335
4	SVM-AL-O	0.717	0.853	0.779	0.390	0.396	0.393
5	SVM-AL-I	0.726	0.851	0.784	0.326	0.518	0.400
6	MSDA-AL	0.713	0.896	**0.794**	0.398	0.461	0.428
7	SVM-CRF	N/A	N/A	N/A	0.421	0.112	0.177
8	MSDA-ALC	N/A	N/A	N/A	0.396	0.478	**0.433**

First, we analyze the experimental results on EUCSD-B. The F1 scores of SVM-A and SVM-L are close, showing that the linguistic feature cannot improve the performance individually. However, SVM-AL-O, which makes the prediction based on two classifier, achieves a 3.6% higher F1 score than SVM-A. It demonstrates that the two kinds of features are complementary and the joint use can improve the emotion recognition. Furthermore, SVM-AL-I outperforms SVM-AL-O by 0.5%, which demonstrates that combing features at input level is better than at decision level. This is because the correlation of acoustic and linguistic features can be learned. MSDA-AL outperforms SVM-AL-I by 1.0% and SVM-A by 5.1%. The improvements are due to the high-level representation fusion and robust features extracted by denoising autoencoders.

For EUCSD, similar conclusions can be drawn. The only exception is that SVM-L performs much worse than SVM-A, because the introduction of more ASR errors does harm to the classification. Nevertheless, the combination of the acoustic and linguistic features can still help emotion recognition, with SVM-AL-O and SVM-AL-I outperforming SVM-A by 1.2% and 1.9% respectively. Furthermore, MSDA-AL outperforms SVM-AL-I by 2.8% and SVM-A by 4.7%. Next are two experiments with contexts taken into account. SVM-CRF modeled the task by sequence labeling model, however this method performs badly and may be because the CRF model needs a large number of data to learn the model parameters and the imbalanced learning problem is not concerned. This is why our proposed approach employs contextual information as input features rather than sequence labeling. One potential concern of our method is the curse of dimensionality, but it can be solved by the dimensionality reduction of DAE in our model. With contextual information, MSDA-ALC outperforms MSDA-AL by 0.5% and SVM-A by 5.2%. The absolute improvements on both datasets are highly consistent, demonstrating the effectiveness and reliability of the proposed approach.

4.3 Comparison between models for linguistic features

We compare our linguistic feature model with some baseline models. We only employ linguistic features for emotion classification and the experimental settings are as follows:

SVM-BOW: The inputs are the BoW representations and the classifier is SVM.

ES: The emotional salience [11] is used for emotion classification.

SVM-EMB: The inputs are the distributed representations and the classifier is SVM.

The dimensionality of the BoW representation is the vocabulary size, i.e. 13469 in our dataset, while the dimensionality of the distributed representation is 200, much lower and denser than BoW.

The EUCSD-B is used as the experimental dataset and the results are shown in Table 4. The F1 score of SVM-BoW is 0.728. The result of ES is even worse than SVM-BoW. This is because the classification method of emotional salience is simple. The SVM-EMB outperforms SVM-BoW, which demonstrates that the distributed representation is better than the traditional BoW representation.

Table 4. Comparison results of linguistic feature models

#	**Methods**	**Precision**	**Recall**	**F1**
1	SVM-BoW	0.740	0.718	0.728
2	ES	0.565	0.886	0.690
3	SVM-EMB	0.652	0.868	0.745

4.4 Comparisons between MSDA and SDA

We also compare the proposed MSDA with SDA. The experimental settings are as follows:

SDA-AL: The acoustic and linguistic features are concatenated to generate a unified feature representation, which is the input to an SDA.

MSDA-AL: The same as described in Section 4.2.

SDA-ALB: The same as SDA-AL except that the BoW representation is used as linguistic features.

MSDA-ALB: The same as MSDA-AL except that the BoW representation is used as linguistic features.

Tabel 5 shows the results of the above methods on EUCSD-B. Although MSDA-AL outperforms SDA-AL, the improvement is small, only 0.4%. This is because the linguistic feature is distributed representation and the same normalization methods are employed for the two kinds of features so that the statistical characteristics of acoustic and linguistic features are similar. The premise that the correlation of acoustic and linguistic features is non-linear does not hold. To prove the effectiveness of MSDA, two more experiments were conducted, in which the BoW representation was used as the linguistic feature. In this case, the correlation of acoustic and linguistic features is believed to be non-linear and MSDA-ALB outperforms SDA-ALB by 1.9%. Therefore, for the cases that there are different kinds of features with non-linear correlation, the MSDA is a good choice. Furthermore, before fusing different features, MSDA has fewer model parameters than SDA, because the two kinds of features are separated in low layers.

Table 5. Comparison between MSDA and SDA

#	**Methods**	**Precision**	**Recall**	**F1**
1	SDA-AL	0.717	0.879	0.790
2	MSDA-AL	0.713	0.896	0.794
3	SDA-ALB	0.688	0.844	0.758
4	MSDA-ALB	0.714	0.853	0.777

5 Conclusion and future work

In this paper, we propose a novel approach for emotion recognition in spoken dialogs. The utterances are transcribed to text by an ASR engine and then the distributed representations of the text are employed as linguistic features. The acoustic and linguistic features along with contextual information are provided to MSDA to learn the high-level representation, which are then fused to a unified feature representation for emotion classification. To evaluate the effectiveness of the proposed approach, we constructed a dataset based on dialogs from a call center. The experimental results demonstrate that our proposed approach outperforms other comparative methods.

As to future work, we plan to study other approaches for leveraging contextual information. Additionally, we will explore our MSDA model to other tasks.

Acknowledgments

Our work is supported by National High Technology Research and Development Program of China (863 Program) (No. 2015AA015402), National Natural Science Foundation of China(No.61370117 & No.61433015) and Major National Social Science Fund of China(No.12&ZD227).

References

1. C. Allauzen, M. Mohri, M. Riley, and B. Roark. 2004. A generalized construction of integrated speech recognition transducers. *ICASSP*, volume 1, pages 761-764.
2. M. E. Ayadi, M. S. Kamel, and F. Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572-587.
3. R. Banse and K. R. Scherer. 1996. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3):614-636.
4. D. Bitouk, R. Verma, and A. Nenkova. 2010. Class-level spectral features for emotion recognition. *Speech communication*, 52(7):613-625.

5. F. Dellaert, T. Polzin, and A. Waibel. 1996. Recognizing emotion in speech. *ICSLP*, volume 3, pages 1970-1973.
6. L. Deng, J. Li, J. T. Huang, K. Yao, D. Yu, F. Seide, M. L. Seltzer, G. Zweig, X. He, J. Williams, Y. Gong, and A. Acero. 2013. Recent advances in deep learning for speech research at Microsoft. *ICASSP*, pages 8604-8608.
7. H. He and E. A. Garcia. 2009. Learning from imbalanced data. *TKDE*, 21(9):1263-1284.
8. K. M. Hermann and P. Blunsom. 2014. Multilingual models for compositional distributed semantics. *ACL*.
9. G. Hoefel and C. Elkan. 2008. Learning a two-stage SVM/CRF sequence classifier. *CIKM*, pages 271-278.
10. D. Kiela and L. Bottou. 2008. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. *EMNLP*, pages 36-45.
11. C. M. Lee and S. S. Narayanan. 2005. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2):293-303.
12. J. Liscombe, G. Riccardi, and D. Hakkani-Tür. 2005. Using Context to Improve Emotion Detection in Spoken Dialog Systems. *EUROSPEECH*.
13. D. J. Litman and K. Forbes-Riley. 2004. Predicting student emotions in computer-human tutoring dialogues. *ACL*.
14. S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk, and S. Stroeve. 2000. Approaching automatic recognition of emotion from voice: a rough benchmark. *ITRW*.
15. S. Meignier, D. Moraru, C. Fredouille, J. F. Bonastre, and L. Besacier. 2006. Step-by-step and integrated approaches in broadcast news speaker diarization. *Computer Speech & Language*, 20(2):303-330.
16. F. Metze, T. Polzehl, and M. Wagner. 2009. Fusion of acoustic and linguistic features for emotion detection. *ICSC*, pages 153-160.
17. D. Morrison, R. Wang, and L. C. De Silva. 2007. Ensemble methods for spoken emotion recognition in call-centres. *Speech communication*, 49(2):98-112.
18. V. Pérez-Rosas, R. Mihalcea, and L. P. Morency. 2013. Utterance-Level Multimodal Sentiment Analysis. *ACL*, pages 973-982.
19. S. Raaijmakers, K. Truong, and T. Wilson. 2008. Multimodal subjectivity analysis of multiparty conversation. *EMNLP*, pages 466-474.
20. D. A. Reynolds and P. Torres-Carrasquillo. 2005. Approaches and applications of audio diarization. *ICASSP*, volume 5, pages 953-956.
21. V. Rozgić, S. Ananthakrishnan, S. Saleem, R. Kumar, A. N. Vembu, and R. Prasad. 2012. Emotion recognition using acoustic and lexical features. *INTERSPEECH*.
22. M. E. Sánchez-Gutiérrez, E. M. Albornoz, F. Martinez-Licona, H. L. Rufiner, and J. Goddard. 2014. Deep learning for emotional speech recognition. *Pattern Recognition*, pages 311-320.
23. R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. *EMNLP*.
24. R. Tato, R. Santos, R. Kompe, and J. M. Pardo. 2002. Emotional space improves emotion recognition. *INTERSPEECH*.
25. D. Ververidis and C. Kotropoulos. 2006. Emotional speech recognition: Resources, features, and methods. *Speech communication*, 48(9):1162-1181.
26. P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. *ICML*, pages 1096-1103.
27. G. Zhou, T. He, and J. Zhao. 2014. Bridging the Language Gap: Learning Distributed Semantics for Cross-Lingual Sentiment Classification. *NLPCC*