DNN-BASED EMOTION RECOGNITION BASED ON BOTTLENECK ACOUSTIC FEATURES AND LEXICAL FEATURES

Eesung Kim[†] *and Jong Won Shin*

School of Electrical Engineering and Computer Science
Gwangju Institute of Science and Technology,
123 Cheomdan-gwagiro, Buk-gu, Gwangju, Korea

ABSTRACT

In this paper, we propose a novel emotion recognition method to reflect affect salient information using acoustic and lexical features. The acoustic features are extracted from the speech signal by applying statistical functionals of emotionally high-level features derived from Deep Neural Network (DNN). These acoustic features are early fused with two types of lexical features extracted from the text transcription of the speech signal, which are the distributed representation and affective lexicon-based dimensions. The fused features are fed to another DNN for utterance-level emotion classification. Experimental results on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) multimodal dataset showed 75.5% in unweighted accuracy recall, which outperformed the best results reported previously in the multimodal emotion recognition using acoustic and lexical features.

Index Terms— Multimodal emotion recognition, DNN-based emotion recognition, Acoustic feature, Lexical feature.

1. INTRODUCTION

Emotion recognition from speech signals has been one of the major topics in the field of affective computing. When people interact with others, they grasp emotion and content in the speech to understand actual intentions of the speakers. The majority of speech emotion recognition research rely on a single modality and exhibit limitations in recognition performance. In order to improve the performance, we focus on multimodal emotion recognition using the acoustic and lexical information extracted from the speech.

The biggest issue to successfully recognize human emotions from the speech and text modalities is the extraction of appropriate features that discriminate different emotions [1]. For capturing acoustic characteristics, the most popular features are low-level descriptors (LLD) with high-level statistical functions [2–5]. In [3], authors also suggested Gaussian supervectors and the bag-of-audio-words to model acoustic features. As for the linguistic information, bag-of-words (BOW) and their refinements were mainly used [2, 3]. In [3, 4], salience information weighting scheme is suggested to capture the emotional salience of both spoken content and verbal gestures using either word level or phoneme level transcripts. However, these traditional approaches are limited to represent high-level information to distinguish the emotions.

Recently, researchers have proposed several deep learning structures to extract high-level features from text and audio [5–8]. In [5], authors build a hybrid deep model structure that utilizes several types of features. As for the acoustic features, it uses a convolutional neural network (CNN)-long short-term memory (LSTM) model from Mel-frequency spectral coefficients (MFSC) energy maps, and the deep neural network (DNN) to learn high-level acoustic features from utterance-level LLD. Regarding the lexical features, CNN is used to extract textual features from word and Part-of-Speech embedding.

In [6], authors introduce attention mechanisms to focus the models on informative words and attentive audio frames for each modality. It extracts high-level informative textual and acoustic features through individual bidirectional gated recurrent units (GRU) and uses a multi-level attention mechanism to select the informative features in both the text and audio module. [7] uses LSTM with temporal pooling to obtain the acoustic features from LLD, and multi-resolution CNN to extract lexical features from word sequences. In [8], authors propose CNN-based emotion recognition system using spectrogram and phoneme embedding.

In this paper, we introduce a novel multimodal method to predict human emotions based on emotional salient information in both acoustic and lexical features. We extract DNN-based bottleneck segment-level acoustic features and then calculated statistical functionals of them for an utterance-level classification. In addition, we extract two types of word-level lexical features represented in the forms of the distributed

[†] Now with Kakao Corp.

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2016R1C1B1015291) and the Technology Innovation Program (10076583, Development of free-running speech recognition technologies for embedded robot system) funded by the MOTIE, Korea.

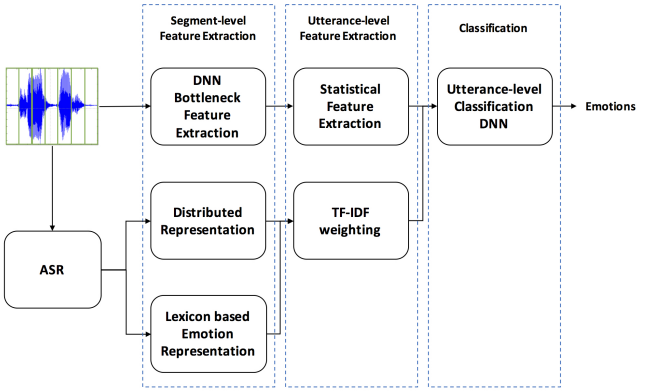

Fig. 1. Overview of the proposed multimodal emotion recognition method integrating the acoustic and lexical features.

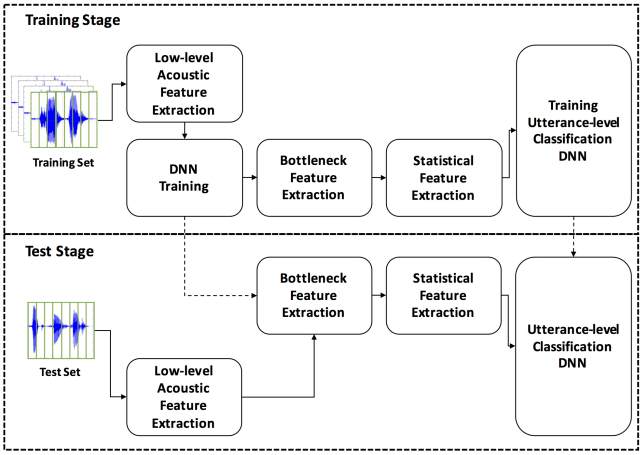

Fig. 2. Block diagram of the training and test stages of the speech emotion recognition using deep bottleneck features.

representation and affective lexicon-based dimensions, and obtain the utterance-level lexical features using appropriate weights that represent the importance of each word in an utterance. These acoustic and lexical features are early fused and then fed into an utterance-level classification DNN. The proposed method achieves 75.5% unweighted average recall (UAR) for four emotions on Interactive Emotional Dyadic Motion Capture (IEMOCAP) multimodal dataset [9]. The proposed method also demonstrates promising performances with acoustic or lexical features only.

2. DNN-BASED SPEECH EMOTION RECOGNITION INCORPORATING LEXICAL FEATURES

In this section, we propose a novel emotion recognition method using two types of features, which are acoustic and lexical features reflecting the emotional states. Figure 1 illustrates the proposed multimodal emotion recognition algorithm. We first extract the DNN bottleneck features to discriminate the emotions at the acoustic feature level. After that, statistical functions are applied to the features from multiple frames constituting an utterance. As for the linguistic information, two types of segment-level lexical features are extracted in the forms of the distributed representation and affective lexicon-based dimensions. These segment-level lexical features for an utterance are summarized with appropriate weights, which represent the importance of each word for various emotion classes. The feature-level early fusion technique is applied to reflect mutual correlations between acoustic and lexical features. Finally, the fused features are fed into another DNN, which is recognized as one of the most effective classifiers in emotion recognition [10].

2.1. Acoustic features based on DNN Bottleneck Representation

Bottleneck (BN) features have been shown to be effective in improving the accuracy of automatic speech recognition, speaker recognition, and acoustic event recognition [11–13]. They are generated by a multi-layer perceptron, in which one of the hidden layers has a small number of hidden units compared to the size of the other hidden layers. The special hidden layer acts as the bottleneck layer that generates a low dimensional representation reflecting target information of the neural network. Bottleneck features may represent salient features highly correlated with the emotional states which are the target of the network.

The detailed algorithm using bottleneck features is shown in Fig. 2. As the first step, we extract acoustic features such as fundamental frequency (F0), voicing probability, Mel-frequency cepstral coefficient (MFCC), and 40 mel-filterbank energies from all frames in an utterance. These features are converted into segments with context windows. With the segment-level acoustic features, we train DNN to predict the probabilities of each emotional states corresponding to the utterance. After that, we obtain high-level bottleneck features from trained DNN model. In order to model utterance-level features capturing details of the distribution of segment-level features, we apply statistical functions such as maximum, mean, median, minimum, standard deviation, percentile 10%, and percentile 90% to the bottleneck feature. Finally, another DNN estimates emotional states with the utterance-level features.

2.2. Distributed Representation of Lexical Information (word2vec)

Word2vec [14] is one of the most popular word embedding methods, which proposed the continuous bag-of-words

(CBOW) and skip-gram models to construct high-quality distributed vector representations efficiently. CBOW predicts a target word from the context words surrounding it across a fixed size context window, while the skip-gram model does the inverse and predicts the surrounding context words given the central target word. In the proposed method, we obtain the word embedding vectors with 300 dimensions for each word based on a pre-trained word embedding model, which use 100 million words from Google news. Word2vec representation relies on the distributed hypothesis that implies that words in the same context share semantic meanings. However, it does not provide any quantitative correlation of each word to a certain emotional state. Therefore, it is necessary to have an additional lexical feature representation method that takes emotion-related cues into account.

2.3. Affective Lexicon-based Lexical Features

Inspired by previous music information retrieval work [15] which uses psycholinguistic resources to identify emotional words in the utterances, we introduce affective lexicon-based emotional dimensions as lexical feature to capture the affective salient words from utterances. We use the affective lexicon as Affective Norms for English Words (ANEW) [16] lexicon. It contains 13,915 English words with scores in three affect-related dimensions: valence, arousal, and dominance (VAD) with a value from 1 to 10 evaluated by humans. It is well known that all emotions can also be represented as points in a 3-dimensional emotional VAD space [15]. These three scores are evaluated by the groups such as male, female, old, young, highly educated people, and low educated people. The means, standard deviations, and number of contributing ratings for every words in the ANEW lexicon are used as lexical features in addition to the word2vec word embedding vectors.

2.4. Construction of the Utterance-Level Lexical Features

To construct the utterance-level lexical features from the word-level word2vec and affective lexicon-based dimensions, the term frequency-inverse document frequency (tf-idf) [17] is introduced. The lexical feature vectors for a certain utterance s_j out of all utterances *S* in the given dataset can be represented through the following equations:

$$\mathbf{v}_{w2v}^{utt}(s_j) = \sum_{w_i \in s_j} tfidf(w_i; s_j)\mathbf{v}_{w2v}^{word}(w_i) \quad (1)$$

$$\mathbf{v}_{al}^{utt}(s_j) = \sum_{w_i \in s_j} tfidf(w_i; s_j)\mathbf{v}_{al}^{word}(w_i) \quad (2)$$

where w_i is a certain word contained in the utterance, and $\mathbf{v}_{w2v}^{word}(w_i)$ and $\mathbf{v}_{al}^{word}(w_i)$ are the word2vec word embedding vector and the affective lexicon-related dimensions for the word w_i, respectively. The *tf-idf* weight is given as

$$tfidf(w_i; s_j) = \frac{n_{i,j}}{\Sigma_k n_{k,j}} \log \frac{|S|}{|\{s_l : w_i \in s_l\}|}. \quad (3)$$

where $n_{i,j}$ is the number of occurrences of the word w_i in utterance s_j, and $|\cdot|$ is the cardinality of a set. The *tf-idf* weighting scheme consists of two components: *tf* and *idf*. The *tf*, $n_{i,j}/\Sigma_k n_{k,j}$, indicates how frequently the word w_i is used in the utterance s_j. The *idf*, $\log\{|S|/|\{s_l : w_i \in s_l\}|\}$, penalizes common words that appear in many documents. With *tf-idf* weights, the utterance-level lexical feature vectors may represent important contents of the utterance. The two utterance-level lexical feature vectors are used as the inputs of the utterance-level emotion classification DNN along with the acoustic features.

3. EXPERIMENTS

3.1. Experimental Setup

We conduct our all experiments on IEMOCAP dataset to demonstrate the efficiency of our approach. It consists of about 12 hours of audiovisual data (speech, video, facial motion capture) which consists of dyadic interaction between professional female and male actors. Each interaction can be divided into two recording scenarios: scripted play and improvised speech. Three evaluators annotated each utterance in the dataset with the categorical emotion labels. In order to match experimental condition with the previous studies [2–4, 6], we use emotional utterances for which at least two out of three annotators gave the same emotion label among angry, happy, neutral, and sad. The happy and excitement classes are merged into the happy class to balance data distribution between classes. Final experiment dataset consists of 1103 angry, 1636 happy, 1708 neutral, and 1084 sad utterances.

The experiments were conducted in a leave-one-speaker-out cross-validation scheme to focus on speaker-independent emotion recognition. The emotion recognition performance was assessed using the weighted average recall (WAR) and

Feature Set	WAR	UAR
ACO+MFCC$_{modified}$[2]	59.3	60.2
Cepstrum+GSV$_{mean}$[3]	55.4	-
GeMAPS	52.9	53.3
eGeMAPS	54.7	55.3
IS09	56.4	57.5
IS10	57.2	59.3
IS13	57.3	58.6
BN (Proposed)	**59.7**	**61.4**

Table 1: Accuracies for different types of acoustic features.

Method	WAR	UAR
LLD+MMFCC+BOW$_{Lexicon}$[2]	69.5	70.1
LLD+BOW$_{Cepstral}$+GSV$_{mean}$+BOW+eVector [3]	69.2	-
LLD+mLRF [4]	67.2	67.3
Hierarchical Attention Fusion Model [6]	72.7	72.7
BN+$\mathbf{v}_{w2v}^{utt}$+$\mathbf{v}_{al}^{utt}$ (Proposed)	**73.7**	**75.5**

Table 3: Accuracies for the multimodal emotion recognition methods.

Feature Set	WAR	UAR
BOW$_{Lexicon}$[2]	56	55.3
eVector+BOW [3]	58.5	-
mLRF [4]	63.8	64
$\mathbf{v}_{w2v}^{utt}$+$\mathbf{v}_{al}^{utt}$ (Proposed)	**64.8**	**65.7**

Table 2: Accuracies for different types of lexical features.

Method	WAR	UAR
LLD+MMFCC+BOW$_{Lexicon}$[2]	64.9	65.7
LLD+mLRF [4]	58.6	59.2
BN+$\mathbf{v}_{w2v}^{utt}$+$\mathbf{v}_{al}^{utt}$ (Proposed)	**66.6**	**68.7**

Table 4: Accuracies for the multimodal emotion recognition methods without transcript.

the UAR. The WAR is the ratio of the total number of correctly predicted test samples and the total number of test utterances, while the UAR is defined as the accuracy per class averaged over all classes so that the accuracy for each class has the same importance regardless of the number of test samples in the class.

The DNN structure for acoustic bottleneck feature extraction was 256-256-32-256, and leaky rectified linear units were used in each layer. Batch normalization technique was applied to learn weights well [18]. The transcripts were used to extract lexical information for the first two experiments to show the potential performances, and the automatic speech recognition (ASR) results were utilized in the last experiment. For modeling the lexical feature, we eliminate stop-words and apply stemming to remove common morphological and inflectional endings before extracting lexical features from the text. The utterance-level classification DNN has three hidden layers with 1024 units and is regularized using early stopping and dropout ratio of 0.5

3.2. Experimental Results

Firstly, we demonstrate the effectiveness of the proposed acoustic and lexical features each. Table 1 shows the WAR and UAR for the emotion recognition systems with only acoustic features. Compared acoustic features include IS09 [4], IS10 [3, 19], IS13 [20], GeMAPS and eGeMAPS [21] feature sets designed to standardize features used in affective computing. The acoustic features considered in [2] and [3] are also compared. The classifier for all the features were the DNN-based utterance-level emotion classifier trained for the given features. Table 2 compares the WAR and UAR for the systems with lexical features only. The lexical features considered in [2–4] are compared. We can see that the proposed acoustic and lexical features outperformed previously proposed acoustic and lexical features, respectively.

The performances of the emotion recognition systems that utilizes both the acoustic and lexical features [2–4,6] are compared with that of the proposed system in Table 3. We can see that both the WAR and UAR of the proposed system outperformed all of the previously reported performances for the 4-class classification on the IEMOCAP dataset by 1.0% and 2.8%, respectively. It is noted that [6] used 5398 sentences, while the proposed and other compared method used all the 5531 utterances.

Finally, the performance of the emotion recognition methods with a practical ASR system were demonstrated in Table 4. Instead of the transcripts provided with the IEMOCAP dataset, the transcripts obtained using the Googles speech recognition system were used to retrieve lexical features. We can confirm that the proposed method outperformed the previously reported methods with transcripts from the practical ASR, too, by 1.7% in WAR and 3.0% in UAR, respectively.

4. CONCLUSION

In this paper, we proposed a novel emotion recognition method that combines acoustic features and two types of lexical features to predict emotional states considering affect salient information. We extracted segment-level acoustic features based on DNN and derived utterance-level features through the statistical functionals. Meanwhile, we extracted word-level lexical features constituting word2vec and affective lexicon-based dimensions and constructed utterance-level lexical features through appropriate weighting scheme. These acoustic and lexical features were concatenated and then used to discriminate the emotional states through utterance-level classification DNN. Our experiments on the IEMOCAP dataset show that the proposed method outperformed other previously reported methods in both WAR and UAR.

5. REFERENCES

[1] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[2] V. Rozgić, S. Ananthakrishnan, S. Saleem, R. Kumar, A. N. Vembu, and R. Prasad, "Emotion recognition using acoustic and lexical features," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[3] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4749–4753, 2015.

[4] K. W. Gamage, V. Sethu, and E. Ambikairajah, "Salience based lexical features for emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5830–5834, 2017.

[5] Y. Gu, S. Chen, and I. Marsic, "Deep multimodal learning for emotion recognition in spoken language," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[6] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

[7] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," in *Proc. INTERSPEECH*, pp. 247–251, 2018.

[8] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech emotion recognition using spectrogram & phoneme embedding," in *Proc. INTERSPEECH*, pp. 3688–3692, 2018.

[9] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[10] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: raising the benchmarks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5688–5691, 2011.

[11] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *Twelfth annual conference of the international speech communication association*, 2011.

[12] Y. Song, B. Jiang, Y. Bao, S. Wei, and L.-R. Dai, "I-vector representation based on bottleneck features for language identification," *Electronics Letters*, vol. 49, no. 24, pp. 1569–1570, 2013.

[13] S. Mun, S. Shon, W. Kim, and H. Ko, "Deep neural network bottleneck features for acoustic event recognition.," in *Proc. INTERSPEECH*, pp. 2954–2957, 2016.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *International Conference on Learning Representations Workshop*, 2013.

[15] Y. Hu, X. Chen, and D. Yang, "Lyric-based song emotion detection with affective lexicon and fuzzy clustering method.," in *The International Society of Music Information Retrieval*, pp. 123–128, 2009.

[16] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior research methods*, vol. 45, no. 4, pp. 1191–1207, 2013.

[17] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[18] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *International Conference on Machine Learning*, pp. 448–456, 2015.

[19] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9-10, pp. 1062–1087, 2011.

[20] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proc. INTERSPEECH*, 2013.

[21] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.