SPEECH EMOTION RECOGNITION USING DEEP NEURAL NETWORK CONSIDERING VERBAL AND NONVERBAL SPEECH SOUNDS

Kun-Yi Huang, Chung-Hsien Wu, Qian-Bei Hong, Ming-Hsiang Su and Yi-Hsuan Chen

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan

ABSTRACT

Speech emotion recognition is becoming increasingly important for many applications. In real-life communication, non-verbal sounds within an utterance also play an important role for people to recognize emotion. In current studies, only few emotion recognition systems considered nonverbal sounds, such as laughter, cries or other emotion interjection, which naturally exists in our daily conversation. In this work, both verbal and nonverbal sounds within an utterance were thus considered for emotion recognition of real-life conversations. Firstly, an SVM-based verbal/nonverbal sound detector was developed. A Prosodic Phrase (PPh) auto-tagger was further employed to extract the verbal/nonverbal segments. For each segment, the emotion and sound features were respectively extracted based on convolutional neural networks (CNNs) and then concatenated to form a CNN-based generic feature vector. Finally, a sequence of CNN-based feature vectors for an entire dialog turn was fed to an attentive long short-term memory (LSTM)-based sequence-to-sequence model to output an emotional sequence as recognition result. Experimental results on the recognition of seven emotional states in the NNIME (The NTHU-NTUA Chinese interactive multimodal emotion corpus) showed that the proposed method achieved a detection accuracy of 52.00% outperforming the traditional methods.

Index Terms—Speech emotion recognition, prosodic Phrase, nonverbal segment, convolutional neural network, long-short term memory, sequence-to-sequence model

1. INTRODUCTION

In recent years, rapid progression of technology makes smart devices more attractive in our daily life. Intelligent services such as chatbot, mental problem diagnosis assistant, smart health care, sales advertising, and smart entertainment, consider not only completing the services but also humanizing the communication between human and computer. How to design and implement an intelligent human-computer interface has become an important issue. With the growth in the number of mobile device users, speech is one of the most common ways for human-human, human-machine communication. As people can realize others' emotion by information involved in speech [1], speech emotion recognition is one of the most important technique for realizing computer intelligence [2-7]. In the past, research on speech emotion recognition mainly focused on discriminative emotion features and recognition models. However, the choice of emotional speech corpus was also important for developing a robust speech emotion recognizer [8-9]. For feature extraction, many studies have strived to look for appropriate speech duration or tried to extract suitable speech feature sets [10-11]. Other studies tested if particular discrete prosodic events provided significant discriminative power for emotion recognition [12]. In recent years, with the progression of neural network (NN), many researchers tried to extract discriminative features from speech for emotion recognition [13-15], such as spectral acoustic features, prosodic features, prosodic action unit, and NN-based emotion feature. With the research mentioned above, convolution neural network (CNN) was common for feature extraction from speech. Global and local features were extracted in the convolution step. More research showed that when extracting emotion features, using the raw signal as input achieved better performance than using spectrum.

Recent emotion recognition systems, limited by lack of emotion speech data resources, seldom focused on nonverbal part of speech in spontaneous communications. In real-life communication, nonverbal sounds within an utterance play an important role for people to recognize others' emotion [16]. When human brain analyzes emotional speech, nonverbal sounds could help gain the difference of emotion more clearly [17]. Therefore, this work expected to consider sound features of nonverbal parts in emotion recognition mechanism to improve the performance of emotion recognition.

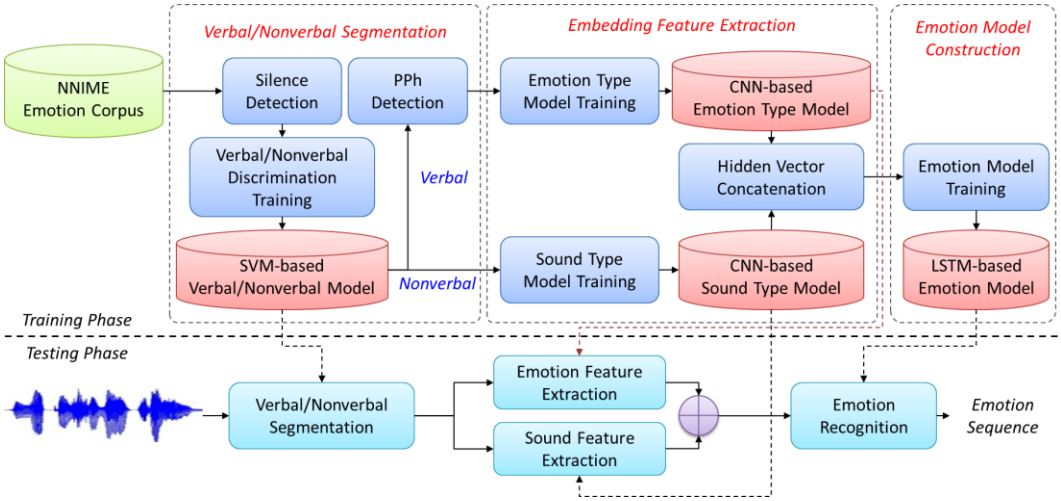

Figure 1. The framework of the proposed speech emotion recognition system

2. CORPUS AND ANNOTATION

In this study, an emotion speech corpus consisting of verbal and nonverbal sounds is desirable. Different from other scripted emotion speech database or segmented audio from dramas, NTHU-NTUA Chinese Interactive Multimodal Emotion Corpus (NNIME) [18], one of the few large-scale Chinese affective dyadic interaction database which contains various emotional nonverbal sounds, such as laughter, sobbing, and sigh in speech, is a suitable corpus for this work. The NNIME corpus was designed with three major elements: (1) Adopt the dyadic interactions for natural elicitation of affective behaviors, (2) Collect both external behaviors and internal physiology of the dyads, and (3) Annotate emotion attributes of the interacting dyads from multiple perspectives.

The NNIME corpus included recording of 44 subjects who were recruited from the Department of Drama at the National Taiwan University of Arts (NTUA) with 22 females and 20 males. The database contains audio, video and electrocardiogram recordings, in which the audio data are composed of 102 dyadic interaction sessions with roughly 11 hours (the duration of each session is μ=195.35s, σ=73.26s). Totally, there are 4766 single-speaker dialogue turns in the database. Table 1 shows the distribution of seven emotion types for verbal segments and Table 2 shows the distribution of sound types of nonverbal segments.

Table 1. Distribution of emotion types of verbal segments

Emotion type	Number of segments
Anger	863
Anxiety	1032
Boredom	491
Happiness	533
Neutrality	5080
Sadness	317
Surprise	1068
Total	**9384**

Table 2. Distribution of emotion types of nonverbal segments

Sound type	Number of segments
Laugh	183
Breath	409
Shout	67
Silence	4593
Total	**5252**

3. PROPOSED METHOD

Figure 1 shows the proposed system framework. In the training phase, there were three main stages including verbal/nonverbal sound segmentation, embedding feature extraction and emotion model construction. First, speech signals went through silence detection, verbal/nonverbal segment detection and prosodic-phrase segmentation procedures to obtain sound/speech segments. Secondly, sound/speech segments of verbal and nonverbal sounds were used to train the corresponding CNN-based models for extracting the generic features of emotion and sound. CNN models which were trained for emotion/sound type classification were used as feature extractors by removing the output layer. Finally, the CNN features of sound and emotion are concatenated as the representative feature vector of each segment. The feature vector sequence was used for emotion recognition to obtain the emotion recognition results of each segment considering emotional changes in a temporal context.

3.1. Verbal/Nonverbal Sound Segmentation

First, this work detected silence segments using the **praat** silence detection tool [21]. The algorithm in **praat** mainly considered three parameters: silence threshold (dB), minimum silence interval (second), and minimum sounding interval (second). The algorithm first evaluated intensity contour and marked sound and silence intervals where the values were below the silence threshold. It then removed sound intervals with durations shorter than the minimum sound interval duration. This step was followed by joining

the neighboring silence intervals. Finally, silence intervals with durations smaller than a pre-defined minimum silence interval duration were removed. After these steps, long silence sections were detected.

Then the verbal segments were obtained from the subsegments using the Prosodic Phrase (PPh) auto-tagger [22]. The PPh auto-tagger determined the boundaries for the verbal segments with pauses, final rising intonations, lengthening in the last word, and sharp falls in intensity. Finally, this work used support vector machine (SVM) to determine the verbal and the nonverbal intervals. OpenSMILE [20] was used to extract low-level descriptors (LLDs) with functionals proposed in [21] with a frame size of 100ms and a frame shift of 50ms for each segment.

Table 3. Low-level descriptors with functionals

LLDs (16x2)	Functionals (12)
(Δ)ZCR (Δ)RMS Energy (Δ)F0 (Δ)HNR (Δ)MFCC 1-12	Max, min, range, max Position, min Position Mean, slope, offset quadratic error standard deviation kurtosis, skewness

3.2. Feature Extraction for Emotion and Sound

With previous studies [16-17], nonverbal sounds in a dialogue were helpful for recognizing emotion. Both verbal and nonverbal segments existed in an utterance concurrently. For example, loudly shouting out a sentence or talking and laughing at the same time is likely to happen in an utterance. To imitate the mechanism in which the human brain discriminates emotion with sounds [17], an emotion type model and a sound type model using CNN were trained with different targets to extract generic emotion features and sound features. For example, sobbing usually expresses sad emotion and when talking some joyful things, the speech may contain laughter. Based on the knowledge from previous research, verbal intervals contain emotion features and nonverbal intervals are additional information for emotion recognition. For the reason mentioned above, verbal segments were used to train the CNN-based emotion model for emotion feature extraction, while the nonverbal segments were used for training the CNN-based sound type model for sound feature extraction. Figure 2 shows the CNN-based models for classification of emotions and sounds, respectively. The convolution layer can accept different lengths of input and compress them to the same length by the adaptive mean pooling layer. The proposed method regarded emotion types and sound types as targets of two CNN-based models for feature extraction and expected to project speech signal to these target feature space. The values of the last hidden layers of two CNN models (by discarding the output layer) were concatenated as the generic feature vector for each speech segment. The concatenated CNN features are used as the input of the LSTM-based sequence-to-sequence emotion recognition model to obtain the emotion recognition results considering the emotional change.

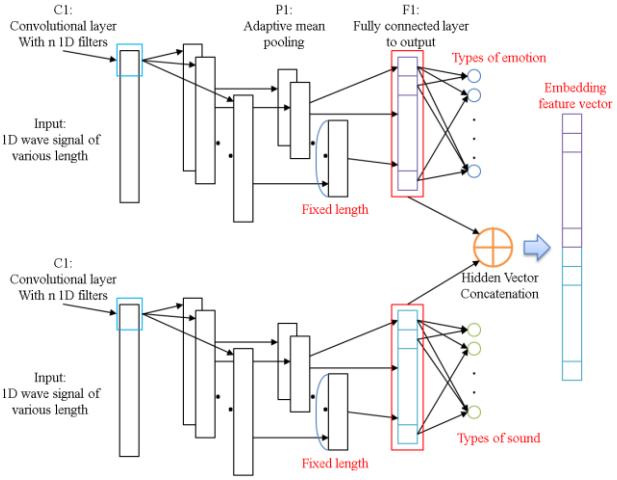

Figure 2. CNN-based emotion and sound feature extraction

3.3. Sequence-to-Sequence Emotion Recognition

As speech signals in a dialog turn in the NNIME corpus were composed of a sequence of verbal/nonverbal segments, emotion recognition should consider the sequential information of the emotional change. The LSTM-based sequence-to-sequence model with an attention mechanism was thus selected for emotion recognition in this study.

The LSTM and attention mechanism for developing a sequence-to-sequence emotion recognition model contained a bidirectional LSTM (Bi-LSTM) as the encoder for attention mechanism and a uni-directional LSTM as the decoder for emotional sequence output. The structure of the sequence-to-sequence emotion recognition model is shown in Figure 3.

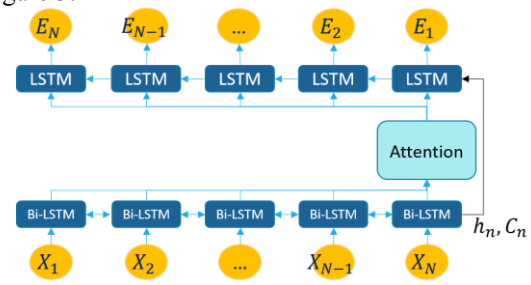

Figure 3. Structure of the LSTM-based emotion recognition model

The input feature vector sequence $X_1, X_2, \ldots, X_N$ is produced by concatenating the feature vectors extracted from the CNN-based emotion model and the CNN-based sound model. The feature vector sequence went through the Bi-LSTM based encoder and focused on some of the segments by the attention mechanism with weights calculated from the encoder output. The LSTM-based decoder then outputted the hidden state and cell state from the encoder and took input sequences obtained from the attention procedure to output the final emotion as the result of each segment within the entire dialog turn.

4. EXPERIMENTAL RESULTS

4.1 Boundary Detection of Verbal and Nonverbal Intervals

For evaluation on the performance of verbal/nonverbal segmentation, this work selected 300 dialog turns from each pre-specified emotion and duration range to annotate boundaries for evaluation on boundary detection. The features with dimensionalities of 32 and 384 was selected with window sizes of 100ms and 200ms and a shift size of 50ms, and the thresholds for boundary detection were selected as 0.6, 0.8, 1, and 1.2. Table 4 shows the precision, recall, F1 score of boundary detection. F is feature dimension, W is the window size and S is the shift size. FM is full match and PM is partial match. The tolerance is 200ms for the boundaries in PM.

Table 4. Performance of boundary detection

		F=32 W=100 S=50		F=32 W=200 S=50		F=384 W=100 S=50		F=384 W=200 S=50	
		FM	PM	FM	PM	FM	PM	FM	PM
0.6	Pre	0.24	0.46	0.23	0.47	0.34	0.62	0.31	0.57
	Rec	0.31	0.60	0.29	0.58	0.37	0.65	0.36	0.66
	F1	0.27	0.52	0.25	0.51	0.35	0.63	0.33	0.61
0.8	Pre	0.24	0.46	0.23	0.47	0.37	0.66	0.32	0.53
	Rec	0.31	0.60	0.28	0.57	0.37	0.64	0.36	0.64
	F1	0.27	0.52	0.25	0.51	0.37	0.64	0.34	0.61
1	Pre	0.25	0.48	0.23	0.49	0.38	0.67	0.33	0.59
	Rec	0.30	0.59	0.27	0.56	0.35	0.60	0.35	0.62
	F1	0.27	0.53	0.25	0.51	0.36	0.63	0.34	0.60
1.2	Pre	0.26	0.50	0.23	0.50	0.41	0.69	0.35	0.61
	Rec	0.30	0.58	0.26	0.55	0.32	0.54	0.34	0.58
	F1	0.28	0.54	0.24	0.52	0.36	0.61	0.34	0.59

4.2 Evaluation of Feature Extraction

For evaluation on feature extraction, this work selected a number of filters and different sizes of adaptive pooling layer based on the accuracy of emotion classification. The results of comparison between the methods using raw speech signal and extracted acoustic feature sets were obtained. The best parameters that achieved the highest accuracy for CNN models were used. The last hidden layer outputs of the CNN emotion/sound models were concatenated and fed to the LSTM-based sequence-to-sequence model for emotion recognition. Table 5 shows the performance of emotion type classification. Table 6 shows the performance of sound type classification.

Table 5. Accuracy of emotion type classification

Input	Best parameters	Accuracy
Speech signal	Filter number = 100, Kernel size = 512, step = 256, pooling = 2	30.10%
32-dim LLDs	Filter number = 150, Kernel size = 2, step = 1, pooling = 2	26.10%
32-dim LLDs with 12 functionals	Filter number = 100, Kernel size = 2, step = 1, pooling = 10	21.20%

Table 6. Accuracy of sound type classification

Input	Best parameters	Accuracy
Speech signal	Filter number = 100, Kernel size = 512, step = 256, pooling = 2	54.90%
32-dim LLDs	Filter number = 100, Kernel size = 2, step = 1, pooling = 2	53.63%
32-dim LLDs with 12 functionals	Filter number = 250, Kernel size = 2, step = 1, pooling = 10	47.95%

4.3 Evaluation of Emotion Recognition

For performance evaluation on emotion recognition, hidden layer sizes of the LSTM were selected from 32, 64, 128, 256, and 512 to achieve the highest accuracy of emotion recognition. Experimental results showed that the proposed method achieved 52.00% when the hidden size of the LSTM was set to 128.

This work compared the performance of the proposed method with traditional emotion recognition models with frame-based acoustic features or raw speech signal as input. As shown in Table 7, the proposed method considering the effect of short duration segments and the difference of background noise, the accuracy of the proposed method outperformed the traditional model. The result presented that the proposed method considering previous segments could help emotion recognition of the dialogue speech with verbal and nonverbal sounds.

Table 7. Comparisons with the traditional model

	Input	Best parameters	Accuracy
Proposed method	CNN-based feature extraction	Hidden size = 128	52.00%
LSTM	32-dim LLDs	Hidden size = 256	44.30%
CNN	Speech signal	Pooling = 2, filter number = 100	30.10%

5. CONCLUSIONS

This work proposed an approach to spontaneous speech emotion recognition considering verbal and nonverbal speech sounds. A segmentation method was present to decompose input speech into prosodic phrases, nonverbal intervals, and silence sections. Then, each segment is represented as feature vectors produced by concatenating emotion features and sound features extracted from the CNN-based models. Given a sequence of feature vectors, the LSTM-based sequence-to-sequence model outputted the emotional sequence as the result. Experimental results show that considering nonverbal speech interval achieved a better performance and sound feature also proved its effectiveness on emotion recognition. Finally, compared with traditional frame-based emotion recognition method, the results showed that CNN-based features and LSTM-based emotion model are beneficial for emotion recognition.

REFERENCES

[1] S. Blanton, "The voice and the emotions," *Quarterly Journal of Speech*, vol. 1, no. 2, pp. 154-172, 1915.

[2] B. W. Schuller, "Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90-99, 2018.

[3] C. H. Wu, J. C. Lin, and W. L. Wei, "Two-Level Hierarchical Alignment for Semi-Coupled HMM-Based Audiovisual Emotion Recognition with Temporal Course," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1880-1895, 2013.

[4] J. C. Lin, C. H. Wu, and W. L. Wei, "Error Weighted Semi-Coupled Hidden Markov Model for Audio-Visual Emotion Recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp.142-156, 2012.

[5] C. H. Wu, and W. B. Liang, "Emotion Recognition of Affective Speech based on Multiple Classifiers using Acoustic-Prosodic Information and Semantic Labels," *IEEE Transactions on Affective Computing*, vol. 2, no. 1, pp. 10-21, 2011.

[6] K. Y. Huang, C. H. Wu, M. H, Su, and Y. T. Kuo, "Detecting Unipolar and Bipolar Depressive Disorders from Elicited Speech Responses Using Latent Affective Structure Model," *IEEE Transactions on Affective Computing*, DOI:10.1109/TAFFC.2018.2803178, 2018.

[7] K. Y. Huang, C. H. Wu, and M.-H, Su, "Attention-based convolutional neural network and long short-term memory for short-term detection of mood disorders based on elicited speech responses," *Pattern Recognition*, vol. 88, pp. 668-678, 2019.

[8] S. Lugović, I. Dunđer, and M. Horvat, "Techniques and applications of emotion recognition in speech," in Proc. International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1278-1283, 2016.

[9] X. Zhang, Y. Sun, and D. Shufei, "Progress in speech emotion recognition," in Proc. IEEE TENCON, pp. 1-6, 2015.

[10] E. Tzinis, and A. Potamianos, "Segment-based speech emotion recognition using recurrent neural networks," in Proc. International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 190-195, 2017.

[11] K. S. Rao, S. G. Koolagudi, and R. R. Vempada, "Emotion recognition from speech using global and local prosodic features," *International journal of speech technolog*y, vol. 16, no. 2, pp. 143-160, 2013.

[12] H. Cao, S. Benus, R. Gur, R. Verma, and A. Nenkova, "Prosodic cues for emotion: analysis with discrete characterization of intonation," Speech prosody, 2014.

[13] N. Anand, and P. Verma, "Convoluted feelings convolutional and recurrent nets for detecting emotion from audio data," in Technical Report: Stanford University, 2015.

[14] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5200-5204, 2016.

[15] L. Zhu, L. Chen, D. Zhao, J. Zhou, and W. Zhang, "Emotion Recognition from Chinese Speech for Smart Affective Services Using a Combination of SVM and DBN," *Sensors*, vol. 17, no. 7, pp. 1694, 2017.

[16] N. Campbell, "On the Use of NonVerbal Speech Sounds in Human Communication," in Verbal and Nonverbal Communication Behaviours, pp. 117-128, 2007.

[17] A. Schirmer, and T. C. Gunter, "Temporal signatures of processing voiceness and emotion in sound," *Social Cognitive and Affective Neuroscience*, vol. 12, no. 6, pp. 902-909, 2017.

[18] H. C. Chou, W. C. Lin, L. C. Chang, C. C. Li, H. P. Ma, and C. C. Lee, "NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus," in Proc. International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 292-298, 2017.

[19] P. W. Boersma, David, "Praat: doing phonetics by computer [Computer program]. Version 6.0.40."

[20] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in Proc. ACM international conference on Multimedia, pp. 1459-1462, 2010.

[21] E. Bozkurt, E. Erzin, Ç. E. Erdem, and A. T. Erdem, "INTERSPEECH 2009 Emotion Recognition Challenge evaluation," in Proc. IEEE Signal Processing and Communications Applications Conference, pp. 216-219, 2010.

[22] M. Domínguez Bajo, M. Farrús, and L. Wanner, "An automatic prosody tagger for spontaneous speech," in Proc. International Conference on Computational Linguistics (COLING): Technical Papers; pp. 377-387, 2016.