Speech Emotion Recognition using Convolutional and Recurrent Neural Networks

Wootaek Lim, Daeyoung Jang and Taejin Lee
Audio and Acoustics Research Section, ETRI, Daejeon, Korea
E-mail: wtlim@etri.re.kr

***Abstract* — With rapid developments in the design of deep architecture models and learning algorithms, methods referred to as deep learning have come to be widely used in a variety of research areas such as pattern recognition, classification, and signal processing. Deep learning methods are being applied in various recognition tasks such as image, speech, and music recognition. Convolutional Neural Networks (CNNs) especially show remarkable recognition performance for computer vision tasks. In addition, Recurrent Neural Networks (RNNs) show considerable success in many sequential data processing tasks. In this study, we investigate the result of the Speech Emotion Recognition (SER) algorithm based on CNNs and RNNs trained using an emotional speech database. The main goal of our work is to propose a SER method based on concatenated CNNs and RNNs without using any traditional hand-crafted features. By applying the proposed methods to an emotional speech database, the classification result was verified to have better accuracy than that achieved using conventional classification methods.**

I. INTRODUCTION

Multimedia pattern recognition is an emerging technology that can extract and analyze large amounts of multimedia information from video and audio sources. In recent years, there has been a drastic growth in the application of machine learning technology using deep learning to solve various recognition problems. Speech Emotion Recognition (SER) is an especially significant task in understanding the characteristics of speech in media. However, recognizing emotions from speech is a very challenging problem because people express emotions in different ways, and the features are unclear to distinguish the emotions. Actually, the paralinguistic problem is challenging even for humans [1, 5].

Conventional techniques for solving this problem are extracting low-level descriptors and training the machine appropriately through learning those features. These methods have been accepted as state of the art for many years in machine learning. However, selecting good features to extract is difficult, and optimization is even more difficult, often being significantly time-consuming in research, development, and validation. Because of this, the traditional trend in speech/audio information retrieval is to focus on the use of powerful strategies for semantic analysis, often relying on model selection to optimize the results [2]. However, deep neural architectures can share low-level representations and naturally progress from low-level to high-level structures. Therefore, deep architectures automatically learn efficient features by stacking network layers.

II. RELATED WORKS

Researchers have developed various methods for emotion recognition in the speech signal. Traditional speech recognition methods are proposed in many research studies [3-5]. However, these conventional approaches have the drawback of using hand-crafted features. More recently, deep neural networks have taken center stage in speech and language analysis.

CNN-based image, video, speech, audio and music recognition methods have been proposed in several research studies [6-11]. From these studies, we already know that CNN-based analysis can be applied in one-dimensional signals such as speech and audio [12]. In particular, a CNN-based SER method has been proposed that learns salient features of SER using semi-CNNs [13]. This method, however, has the drawback of feeding the learned features into support vector machine classifier. The SER system using RNNs was proposed in [14], which accounts for long contextual effect in emotional speech and the uncertainty of emotional labels. In addition, audio/video based multimodal emotion recognition approaches were experimented in [15-16]. Furthermore, the cross-corpus emotion recognition method was explored in [17], which investigates the ability of an emotion recognition detector to be applied to other databases.

In this work, we propose the CNNs, RNNs, and time distributed CNNs-based SER method that acquires signals on a 2D domain speech signal representation. In particular, we propose a method for analyzing sequential audio data based on concatenated CNNs and RNNs. By applying the proposed method to a public emotional speech database, the recognition result was verified to be better than conventional methods [5, 13].

III. THE PROPOSED METHOD

The typical method for analyzing the audio and speech signals is 2D representation. Time-frequency analysis is commonly used in audio processing. We transform the speech signal to 2D representation using Short Time Fourier Transform (STFT) after pre-processing. Then the 2D

representation is analyzed through CNNs and Long Short-Term Memory (LSTM) architectures. Deep learning involves hierarchical representations with increasing levels of abstraction. By traversing sequentially constructed networks, the results corresponding to each selected audio frame are classified using a sum of probabilities. Figure 1 shows the detailed scheme of our proposed process.

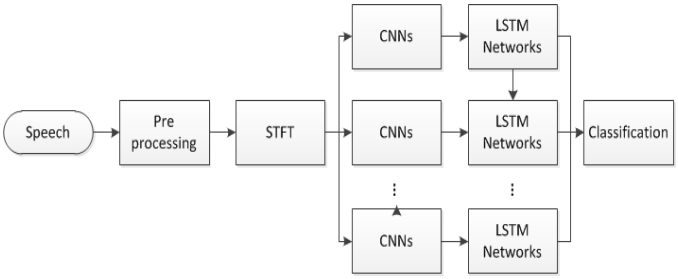

Fig. 1. Block diagram of our time distributed networks based SER method.

A. *Convolutional Neural Networks*

Typical CNN is designed to analyze data that come in the form of a multi-dimensional array [18]. An input array in which nearby data values are correlated is a good application for CNN. Therefore, it is deployed in many practical applications such as image, video and time-frequency representations of audio. The key idea of CNN is to take advantage of the properties of signals: local connectivity, weight sharing, pooling and use of many layers [19].

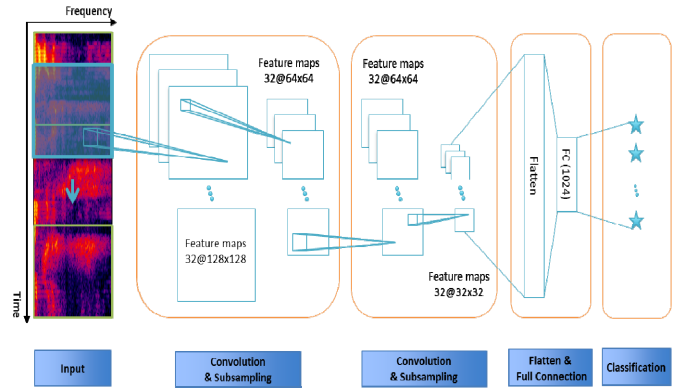

Fig. 2. The proposed CNN structure for emotion recognition in speech

Figure 2 shows the detailed scheme of the proposed CNN architecture. We used STFT for 2D representation of a speech signal with a frame size of 256 and 50% overlap. In addition, the look-ahead frame number was 128 time steps. Thus, the size of the input image was 128 x 128, as depicted in Figure 2. Subsequently, two convolution and max-pooling layers were stacked for learning representation. Finally, all values were flattened and fully connected with 1024 nodes.

B. *Recurrent Neural Networks*

The basic idea of RNN derives from the use of sequential data information. The other neural network techniques such as DNNs and CNNs assume that all input signals must be independent of each other. However, in many applications, we usually treat all types of time-distributed signals. Therefore, these sequential data analysis approaches are widely used not only for language modeling but also for machine translation [20-24].

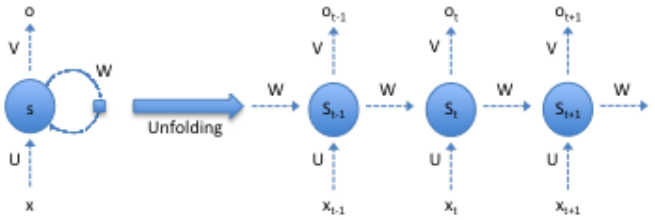

Fig. 3. A recurrent neural network and unfolding in time of the computation involved in its forward computation.

Figure 3 shows a basic concept of RNN and unfolding in time of the computation involved in its forward computation [19]. The traditional neural network uses different parameters at each layer. Unlike a traditional deep neural network, the RNN shares the same parameters (U, V, and W in Figure 3) across all steps. The hidden state formulas and variables are as follows:

$$s_t = f(Ux_t + Ws_{t-1}) \quad (1)$$

- x_t *is the input at time step t.*
- s_t *is the hidden state at time step t.*
- o_t *is the output at time step t.*
- *U, V, W are parameter matrices.*

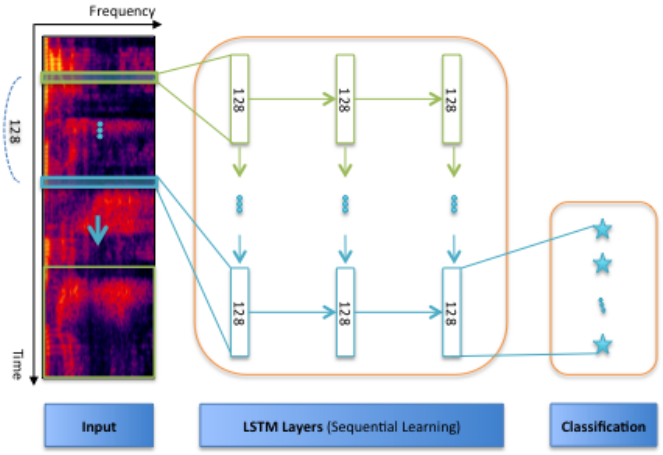

Fig. 4. The proposed LSTM structure for speech emotion recognition

Figure 4 shows the proposed LSTM network structure for comparing the classification accuracy with CNNs and sequential CNNs. Each time step is connected to an LSTM layer and three LSTM layers are stacked sequentially. The figure assumes that the input and output frame sizes are 1x128 each.

C. *Time Distributed CNNs*

One of the major goals of our work is combining a deep hierarchical CNNs feature extraction architecture with a

model that can learn to recognize and synthesize sequential dynamics in a speech signal. Because the temporal properties in speech signal provide important information about emotion, this idea is worth exploring. Therefore, the idea for this approach is introduced in [25]. Figure 5 depicts the time distributed CNNs model of our approach.

As previously described, we used the STFT for 2D representation with a frame size of 256 and 50% overlap. The size of the input image was 128 x 128, as depicted in Figures 2 and 5. Subsequently, two convolution and max-pooling layers were stacked for learning representation. After that, two additional sequentially stacked LSTM layers with 1024 nodes each, sequentially learn the CNNs features. We concatenate 4 CNNs with the LSTM networks, and the hopping size of time distributed CNNs is 8 time steps in each.

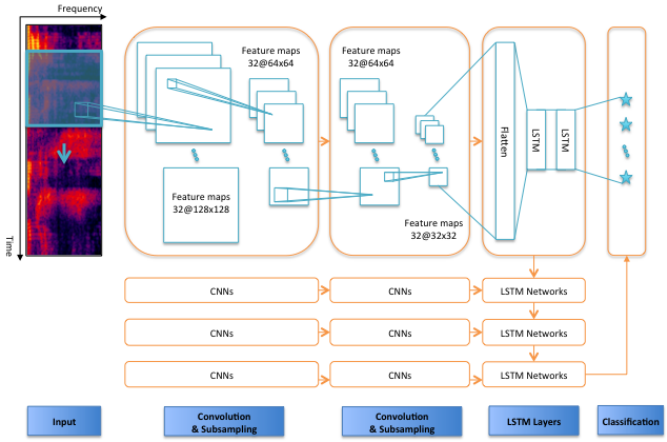

Fig. 5. The proposed Time Distributed CNNs structure for emotion recognition in speech.

IV. THE PERFORMANCE EVALUATION

In this section, we describe the experiment environment and report the recognition accuracy of using the proposed deep neural network structures on a public emotional speech database. We used Berlin database for network training and validation [26]. It consists of 535 utterances in German of 10 different statements by 10 actors expressing 7 emotions. All recorded wave files are approximately 2-3 s long.

A. Pre-Processing and Network Setting

We resampled the data to 16000Hz and randomly oversampled some datasets owing to the existence of a slight class imbalance problem in the training network. In Table 1, we have listed the hyper parameters and settings used in our experiments.

TABLE I. THE HYPER PARAMETERS AND SETTINGS OF PROPOSED NETWORK

Parameter	Value
Convolution filter size	3x3
Activation function	Relu
Dropout factor	0.25
Optimizer	Stochastic Gradient Descent
Learning rate	0.01
Decay	1e-6
Momentum	0.8

B. Experimental Results

The classification results of our research are listed in Tables 2–4. In Tables 2 and 3, the CNNs and LSTM network based SER results are provided as produced by methods depicted in Figures 2 and 4. The time distributed CNNs-based SER results are shown in Table 4. Comparing Table 4 to Tables 2 and 3, the proposed time distributed CNNs structure shows better results. Assuming that the speech signal is sequentially related, the result of the proposed scheme is a meaningful improvement.

The majority voting method was used for the final decision. All results from our experiment are validated using a 10-fold cross validation method and it was repeated 5 times.

TABLE II. RESULT OF EXPERIMENT 1 (CNNS)

Emotion	*Precision*	*recall*	*f-1 score*
Neutral	91.58	85.32	87.16
Anger	87.16	92.58	89.48
Fear	87.56	80.22	82.70
Disgust	87.76	90.30	88.48
Sadness	90.04	98.40	93.56
Boredom	89.20	88.54	87.78
Happy	80.88	68.86	73.26
Average (%)	**87.74**	**86.32**	**86.06 (±2.39)**

TABLE III. RESULT OF EXPERIMENT 2 (LSTM)

Emotion	*Precision*	*recall*	*f-1 score*
Neutral	75.28	76.36	74.58
Anger	84.82	89.04	86.46
Fear	84.90	76.60	79.76
Disgust	80.18	82.60	80.56
Sadness	85.62	94.08	88.96
Boredom	74.36	66.30	69.06
Happy	73.90	66.84	68.78
Average (%)	**79.87**	**78.83**	**78.31 (±4.59)**

TABLE IV. RESULT OF EXPERIMENT 3 (TIME DISTRIBUTED CNNS)

Emotion	*Precision*	*recall*	*f-1 score*
Neutral	93.80	88.16	90.02
Anger	86.88	91.42	88.64
Fear	82.94	82.18	81.56
Disgust	88.48	89.90	88.92
Sadness	90.68	99.66	94.58
Boredom	92.02	88.42	89.56

Emotion	*Precision*	*recall*	*f-1 score*
Happy	81.24	68.26	73.28
Average (%)	**88.01**	**86.86**	**86.65 (±1.73)**

V. CONCLUSION

The field of machine learning is sufficiently new to still be rapidly expanding, often from innovation in new formalizations of machine learning problems driven by practical applications. However, recognizing emotions from speech is still a challenging problem. In this paper, we proposed the CNNs, RNNs and time distributed CNNs based network without using any traditional hand-crafted features to classify emotional speech. For SER, we combined a deep hierarchical CNNs feature extraction architecture with LSTM network layers. Moreover, we investigated the recognition result by comparing with the basic CNNs and LSTM based emotion recognition results. We verified that CNNs-based time distributed networks show better results. This comparison of results provides a baseline for future research, and we expect that it can give a better result when using more concatenated CNNs. In future, we are planning to study the audio/video based multimodal emotion recognition task.

ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP). [R01261510340002003, Development of hybrid audio contents production and representation technology for supporting channel and object based audio]

REFERENCES

[1] Zeng, Zhihong, et al. "A survey of affect recognition methods: Audio, visual, and spontaneous expressions." Pattern Analysis and Machine Intelligence, IEEE Transactions on 31.1 (2009): 39-58.

[2] Humphrey, Eric J., Juan Pablo Bello, and Yann LeCun. "Moving Beyond Feature Design: Deep Architectures and Automatic Feature Learning in Music Informatics." ISMIR. 2012.

[3] Schuller, Björn, Gerhard Rigoll, and Manfred Lang. "Hidden Markov model-based speech emotion recognition." Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03). 2003 IEEE International Conference on. Vol. 2. IEEE, 2003.

[4] Nwe, Tin Lay, Say Wei Foo, and Liyanage C. De Silva. "Speech emotion recognition using hidden Markov models." Speech communication 41.4 (2003): 603-623.

[5] El Ayadi, Moataz, Mohamed S. Kamel, and Fakhri Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases." Pattern Recognition 44.3 (2011): 572-587.

[6] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[7] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[8] Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[9] Abdel-Hamid, Ossama, et al. "Convolutional neural networks for speech recognition." Audio, Speech, and Language Processing, IEEE/ACM Transactions on 22.10 (2014): 1533-1545.

[10] Sainath, Tara N., et al. "Deep convolutional neural networks for LVCSR." Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013.

[11] Schluter, Jan, and Sebastian Bock. "Improved musical onset detection with convolutional neural networks." Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014.

[12] Deng, Li, Geoffrey Hinton, and Brian Kingsbury. "New types of deep neural network learning for speech recognition and related applications: An overview." Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013.

[13] Huang, Zhengwei, et al. "Speech emotion recognition using CNN." Proceedings of the ACM International Conference on Multimedia. ACM, 2014.

[14] Lee, Jinkyu, and Ivan Tashev. "High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition." Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[15] Soleymani, Mohammad, Maja Pantic, and Thierry Pun. "Multimodal emotion recognition in response to videos." Affective Computing, IEEE Transactions on 3.2 (2012): 211-223.

[16] Kim, Yelin, Honglak Lee, and Emily Mower Provost. "Deep learning for robust feature generation in audiovisual emotion recognition." Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013.

[17] Lefter, Iulia, et al. "Emotion recognition from speech by combining databases and fusion of classifiers." Text, Speech and Dialogue. Springer Berlin Heidelberg, 2010.

[18] LeCun, Yann, et al. "Gradient-based learning applied to document recognition." Proceedings of the IEEE 86.11 (1998): 2278-2324.

[19] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." Nature 521.7553 (2015): 436-444.

[20] Mikolov, Tomas, et al. "Recurrent neural network based language model." INTERSPEECH. Vol. 2. 2010.

[21] Mikolov, Tomáš, et al. "Extensions of recurrent neural network language model." Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, 2011.

[22] Sutskever, Ilya, James Martens, and Geoffrey E. Hinton. "Generating text with recurrent neural networks." Proceedings of the 28th International Conference on Machine Learning (ICML-11). 2011.

[23] Liu, Shujie, et al. "A Recursive Recurrent Neural Network for Statistical Machine Translation." ACL (1). 2014.

[24] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems. 2014.

[25] Donahue, Jeffrey, et al. "Long-term recurrent convolutional networks for visual recognition and description." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015

[26] Burkhardt, Felix, et al. "A database of German emotional speech." Interspeech. Vol. 5. 2005.