딥네트워크 기반 음성 감정인식 기술 동향

무스타킴*, 권순일*[1]
*세종대학교 소프트웨어학과
mustaqeemicp@gmail.com, skwon@sejong.edu

Speech Emotion Recognition Based on Deep Networks: A Review

Mustaqeem*, Soonil Kwon*[1]
*Dept. of Software, Sejong University

Abstract

In the latest eras, there has been a significant amount of development and research is done on the usage of Deep Learning (DL) for speech emotion recognition (SER) based on Convolutional Neural Network (CNN). These techniques are usually focused on utilizing CNN for an application associated with emotion recognition. Moreover, numerous mechanisms are deliberated that is based on deep learning, meanwhile, it's important in the SER-based human-computer interaction (HCI) applications. Associating with other methods, the methods created by DL are presenting quite motivating results in many fields including automatic speech recognition. Hence, it appeals to a lot of studies and investigations. In this article, a review with evaluations is illustrated on the improvements that happened in the SER domain though likewise arguing the existing studies that are existence SER based on DL and CNN methods.

Keywords: Affective computing, deep learning, convolutional neural networks, machine learning, speech recognition.

1. Introduction

Speech signals are the most dominant source of human communication, and they are the efficient method of human-computer interaction (HCI) using 5G technology. Emotions express human behaviors, which are recognized from various body expressions. They are evident in speech patterns, facial expressions, gestures, and brain signals [1, 2]. In the field of speech signal processing, speech emotion recognition (SER) is the most attractive area of research in this era. Speech signals play an important role to recognize the emotional state and the human behavior during his/her speech. In this technological era, the researchers have utilized neural networks and deep learning tools in order to search for an efficient way to extract the deep features that ensure the emotional state of a speaker in the speech data [3, 4].

Some researchers introduced hybrid techniques to evaluate the handcrafted features with CNN models in order to improve the recognition accuracy of the speech signals. The handcrafted features particularly ensure the accuracy, but this process is difficult due to the features engineering. The amount of time is required, and this is exclusive to manual selection, which is particularly contingent on expert knowledge [5]. The deep learning approaches, which include the 2D-CNN models, account for the visual data, such as the images and the videos in computer vision [3], but the researchers adopted these models in speech processing and achieved better results than the classical models [3]. In addition, the researchers achieved good performances with the emotion recognition from the speech signals that utilize the deep learning approaches, such as the deep belief networks (DBNs), 2D-CNNs, 1D-CNNs, and the long short-term memory (LSTM) network. The performance of the deep learning approaches is better than the traditional methods.

Hence, Fiore et al. [6] developed an SER for on-board system to detect and analyze the emotional conditions of the driver, which involved taking the appropriate actions in order to ensure the passenger's safety. Badshah et al. [7] introduced an SER system for the smart health care centers in order to analyze the customer emotions using a fine-tuned Alex Net model with rectangular kernels. Kwon et al. [8] proposed a novel deep stride CNN network for speech emotion recognition in order to improve the prediction accuracy and decrease the overall model complexity [4]. Min et al. [9] developed an SER technique in order to analyze the emotion type and the intensity from the arousal features and the violence features using the content analysis in movies. Miguel et al. [10] developed an SER method in order to recognize a speaker's privacy using a privacy-preserving-based hashing technique that used paralinguistic features.

The remaining article is divided into different sections, the concept of traditional SER methods is presented in Section 2 and the introduction of deep learning techniques for SER is illustrated in Section 3. The publically available datasets for SER is presented in Section 4 and Section 5 shows summarized SER overviews. The conclusion and direction toward future work is illustrated in Section 5.

2. Traditional SER Methods

A traditional SER techniques digitized into main three chunks: 1) speech signal preprocessing, 2) features extraction by specific algorithm, and 3) classifier to differentiate among classes or emotions. **Figure 1** illustrates a simple flow diagram of traditional system for emotion recognition based on low-level handcrafted features. In the initial stage, the system processes the speech signal to remove noises and background clutter, second stage consists of features

*[1] Corresponding author: Soonil Kwon

extraction and selection based on different analysis such as time and frequencies. In the final stage, a classifier is utilized to differentiate the learnt features into different classes by GMM, HMM, and SVM machines.

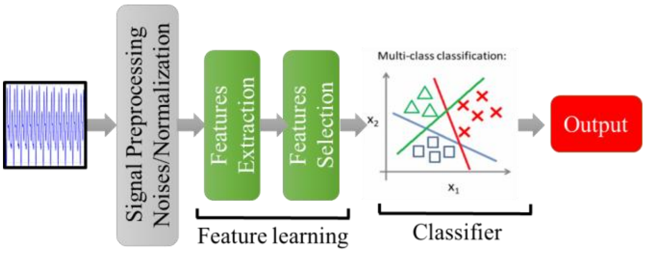


Figure 1: Traditional SER simplified architecture.

3. Deep Learning Approaches For SER

In this technological era, the deep learning approaches have become popular in all fields and specifically in the field of SER that utilizes the CNN models to extract the deep hidden cues from the spectrograms of the speech signals. Similarly, some researchers used the transfer learning techniques for the SER that utilizes spectrograms to train the pre-trained Alex Net [11]and VGG [12] models in order to identify the state of the speakers in term of the emotions. Furthermore, the researchers used the 2D-CNNs to extract the special information from the spectrograms and the LSTM, or the RNNs were utilized to extract the hidden sequential and temporal information from the speech signals.

Currently, the CNNs have increased the research interest of the SER. In this regard, [13] developed a new end-to-end method for an SER that utilizes a deep neural network (DNN) with the LSTM, which directly accepts the raw audio data and extracts the salient discriminative features rather than obtaining the handcrafted features. Most researchers used the joint CNNs with the LSTM and the RNNs for the SER to capture the special cues and the temporal cues from the speech data in order to recognize the emotional information. **Figure 2** shows the simple deep learning structure with hidden layer representation of simple DNN approach.

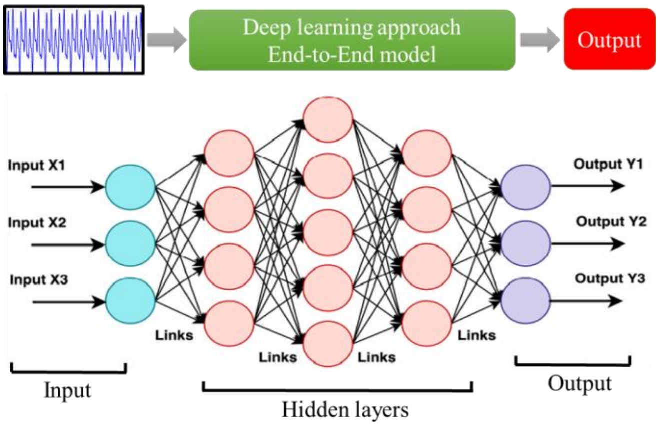


Figure 2: Deep learning simplified architecture with DNN.

Deep learning is a fast growing domain where researchers developed and explored various architectures due to its structure, layers, and efficient outcomes in all areas. Researchers were inspired by their convenient results and utilized it in digital audio processing field such as speech emotion recognition, speakers' recognition, pattern recognition, and natural language processing with different structures like DBMs, CNNs, DBNs, AEs, and RNNs, which are shown in **Figure 3**.

4. Databases Utilized For SER

For the system evaluations researchers utilized different public available emotional speech corpora to show the quality of the system and performance measured with baseline to show the importance. All emotional speech corpora are mainly categorized into three categories such as simulated, induced, and natural databases. Simulated corpus has been recorded by professional well trained actors, induced corpus has been collected by artificial emotional condition, and natural corpus has been recorded in most realistic environment to record natural emotions. The detail of public emotional speech corpora is illustrated in Table 1.

Table 1: detailed of free available speech corpora and its characteristics such as number of emotions, language, and acted or non-acted.

Dataset	Language	Emo	Source	Access
EMO-DB	German	7	Pro-actor	Free, public
RAVDESS	English	8	Actor	Free license
DEB	Danish	5	Npro-actor	Free license
IEMOCAP	English	8	Pro-actor	Free license
INT-FC05	Eng./Spa/Fch	5	Actor	Commercial
LDC	English	12	Pro-actor	Commercial

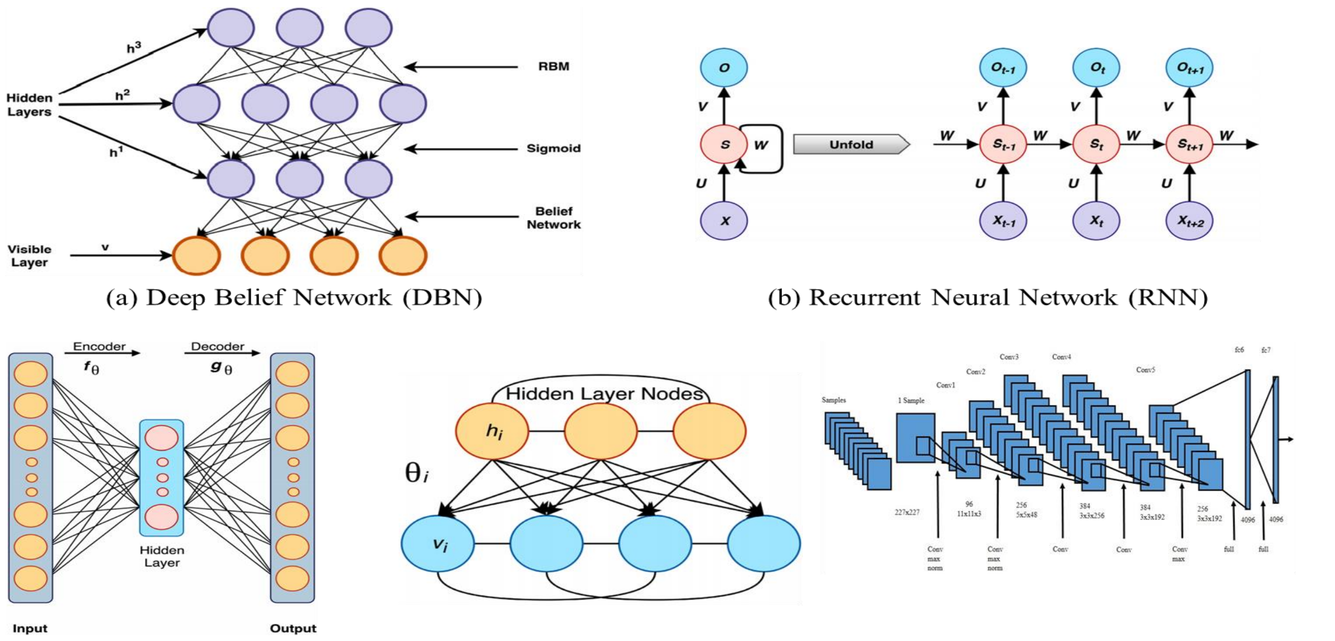
The above all datasets are publically available for experimentations to evaluate the significant of the proposed system. In contrast, there is a huge difference and variation among the corpora, emotions quantity, performers, and techniques. Speech emotions recognition technology needs a psychological studies depending on various situations to recognize desired emotion. The system hardly recognizes emotion in real-time environment because the data is so complex for automatic emotion recognition.

5. Summarized Literature of SER

In this section, we summarized the literatures for speech emotion recognition (SER) and discuss recent developed deep learning techniques. Deep learning is rapidly popular in all domain due to its hybrid nature by utilizing the key features in different applications. For example, SER, NLP, and sequential information processing, which is described with detailed in **Table 2**. The discriminative features selection is an important task for making an intelligent and efficient system.

Table 2: Summarized deep learning methods with descriptive key information.

DL Methods	Key Features Description
DBM	RBM based unsupervised learning with directionless connections
RCNN	Long and short term sequential info for SER and NLP
RVNN	Tree-like structure particularly for NLP
DBN	Unsupervised learning with directed connections
CNN	Basic architecture for image but extended to speech processing, and NLP
AE/SAE	Unsupervised learning with probabilistic models



(a) Deep Belief Network (DBN)

(b) Recurrent Neural Network (RNN)

(c) Auto-Encoder (AE)

(d) Deep Boltzmann Machine (DBM)

(e) Convolution Neural Network (CNN)

Figure 3: Graphical representation of different deep learning approaches.

Furthermore, we summarized the current existence deep learning based SER methods in order to recognize emotions, utilized dataset, used DL approach, contribution in term of accuracy for each dataset, and future direction. We compared different DL models for SER in order to show the improvement, the robustness, and the effectiveness of the system. **Table 3** shows the un-weighted evaluation matrix, which is mostly used in the literature to find the recognition rate of the system. The literature of the SER shows good improvements with the performance based on DL architectures.

We analyzed different one-dimensional and two-dimensional CNN models, which are shown in the **Table 3**. In [14, 15], the authors used 1D-CNN architectures with local features learning blocks as well as global features learning blocks by different learning strategies to easily recognize emotional features. Some methods were utilized sequential learning networks such as LSTM, gated recurrent units (GRUs) with simple, bi-direction, and multi-layer structures to extract the temporal cues in order to recognize the emotions. Researcher developed various novel strategy for the SER in order to recognize the emotional features in speech signals. **Table 3** illustrates a comparative summarized study of the models under the same conditions, such as datasets and an accuracy matrix. In **Table 3,** the researchers solved the SER problem by using different techniques in this era, but they mostly used the 2D-CNN, and 1D-CNN architectures to address the stated problem. Overall, the 2D-CNN strategies were built for visual data recognition and classification in the field of computer vision [16]. With the usage of traditional machine learning, we lost some paralinguistic cues in the speech signals, and we didn't achieve better accuracy for the emotion recognition. In order to address this limitation, researchers proposed deep learning models that can accept the direct speech data in order to extract the features and recognize the paralinguistic information, such as emotions. The deep learning models are able to predict emotions with a high accuracy rate compared to the other prior models that involve emotion recognition, which is shown in **Table 3**. In this short review paper, we shows the most recent and novel research in the field of speech emotion recognition. Additionally, we collected information from other research which has utilized similar datasets during system evaluations and increased the level of accuracy in order to address the limitation of the previous works.

Table 3: Summarized literatures of current deep learning based speech emotion recognition methods with discussion, datasets, utilized algorithm, and possible future directions for further improvements.

Refe #	**Dataset**	**Method**	**Accuracy (%)**	**Future-direction**
[7] (2019)	Emo-DB	CNN-SVM	EMO-DB=85.5	Authors shows interest for diagnostic system using audio-visual.

[17]
(2019)
	IEMOCAP
	EMO-DB
		SVM with Timber features and MFCC	IEMOCAP=63.03
			EMO-DB=96.49
				It can be tested in other platform and extended this work to deep learning method for high performance.

[8]
(2020)
	IEMOCAP
	RAVDESS
		Stride Net-CNN	IEMOCAP=**81.75**
			RAVDESS=**79.5**
				This comparison can be modeled and extended toward multi-modality.

[3]
(2020)
	IEMOCAP
	EMO-DB
	RAVDESS
		CNN-LSTM with k-mean	IEMOCAP=**72.25**
			EMO-DB=**85.57**
			RAVDESS=**77.02**
				Exploration and extraction more rationalized technique for sequences.

[4]
(2020)
	IEMOCAP
	EMO-DB
		CNN with modified kernel	IEMOCAP=**77.01**
			EMO-DB=**92.02**
				Model can be modified for time based features learning

[5]
(2020)
	IEMOCAP
	RAVDESS
		MLT using CNN-GRUs	IEMOCAP=**73.00**
			EMO-DB=**90.00**
				This model can be modified and beneficial for real-time results.

[14]
(2020)
	IEMOCAP
	EMO-DB
		ConvLSTM-GRU	IEMOCAP=**75.00**
			RAVDESS=**80.00**
				The model need more training with high computation devices.

[18]
(2020)
	IEMOCAP
	RAVDESS
		Attention based CNN	IEMOCAP=**78.00**
			RAVDESS=**81.00**
			EMO-DB=**93.00**
				The model can be train and evaluate with more natural data to achieve accurate decision for different emotions.

The above table is the pin-point analysis and partial overview of the current deep learning based SER methods with different learning strategies and improved results.

6. Conclusion

In this study, we have explored some research articles about speech emotion recognition (SER) based on Deep Learning (DL). Then we have studied a number of works that were experimented in the field and a comparison among them was discussed. It has been observed that the outcomes of the SER based on DL or CNN were much better than classical models. DL significantly increases the accuracy of the models and has less error rate. The outcomes of this exploration will enable potential scholars to create new and significant research ideas that have not yet been explored, as well as to highlight some of the flaws in existing studies.

Acknowledgement

This work was supported by the National Research Foundation of Korea funded by the Korean Government through the Ministry of Science and ICT under grant NRF-2020R1F1A1060659.

Reference

[1] R. A. Naqvi, M. Arsalan, A. Rehman, A. U. Rehman, W.-K. Loh, and A. Paul, "Deep Learning-Based Drivers Emotion Classification System in Time Series Data for Remote Applications," *Remote Sensing,* vol. 12, p. 587, 2020.

[2] S. Z. Bong, K. Wan, M. Murugappan, N. M. Ibrahim, Y. Rajamanickam, and K. Mohamad, "Implementation of wavelet packet transform and non linear analysis for emotion classification in stroke patient using brain signals," *Biomedical signal processing and control,* vol. 36, pp. 102-112, 2017.

[3] M. Sajjad and S. Kwon, "Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM," *IEEE Access,* vol. 8, pp. 79861-79875, 2020.

[4] S. Kwon, "MLT-DNet: Speech Emotion Recognition Using 1D Dilated CNN Based on Multi-Learning Trick Approach," *Expert Systems with Applications,* p. 114177, 2020.

[5] T. Anvarjon and S. Kwon, "Deep-Net: A Lightweight CNN-Based Speech Emotion Recognition System Using Deep Frequency Features," *Sensors,* vol. 20, p. 5212, 2020.

[6] U. Fiore, A. Florea, and G. Pérez Lechuga, "An Interdisciplinary Review of Smart Vehicular Traffic and Its Applications and Challenges," *Journal of Sensor and Actuator Networks,* vol. 8, p. 13, 2019.

[7] A. M. Badshah, N. Rahim, N. Ullah, J. Ahmad, K. Muhammad, M. Y. Lee, *et al.*, "Deep features-based speech emotion recognition for smart affective services," *Multimedia Tools and Applications,* vol. 78, pp. 5571-5589, 2019.

[8] S. Kwon, "A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition," *Sensors,* vol. 20, p. 183, 2020.

[9] S. Kang, D. Kim, and Y. Kim, "A visual-physiology multimodal system for detecting outlier behavior of participants in a reality TV show," *International Journal of Distributed Sensor Networks,* vol. 15, p. 1550147719864886, 2019.

[10] M. Dias, A. Abad, and I. Trancoso, "Exploring hashing and cryptonet based approaches for privacy-preserving speech emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2018, pp. 2057-2061.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems,* 2012, pp. 1097-1105.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,* 2014.

[13] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2015, pp. 4580-4584.

[14] S. Mustaqeem; Kwon, "CLSTM: Deep Feature-Based Speech Emotion Recognition Using the Hierarchical ConvLSTM Network. ," *Mathematics,* vol. 8, p. 2133, 2020.

[15] Mustaqeem and S. Kwon, "1D-CNN: Speech Emotion Recognition System Using a Stacked Network with Dilated CNN Features," *Computers, Materials \& Continua,* vol. 67, pp. 4039--4059, 2021.

[16] N. Khan, A. Ullah, I. U. Haq, V. G. Menon, and S. W. Baik, "SD-Net: Understanding overcrowded scenes in real-time via an efficient dilated convolutional neural network," *Journal of Real-Time Image Processing,* pp. 1-15, 2020.

[17] A. Tursunov, S. Kwon, and H.-S. Pang, "Discriminating Emotions in the Valence Dimension from Speech Using Timbre Features," *Applied Sciences,* vol. 9, p. 2470, 2019.

[18] S. Kwon, "Att-Net: Enhanced emotion recognition system using lightweight self-attention module," *Applied Soft Computing,* vol. 102, p. 107101, 2021.