Journal of Computer Technology and Software

ISSN:2998-2383

Vol. 3, No. 6, 2024

Speech Emotion Recognition with Dynamic CNN and Bi-LSTM

Elliot Finch

Department of Computer Science, Midwestern State University, Wichita Falls, USA
e.finch763@msu.edu

Abstract: This study presents a speech emotion recognition system that integrates a dynamic convolutional neural network with a bi-directional long short-term memory (Bi-LSTM) network. The dynamic convolutional kernel enables the neural network to capture global dynamic emotional patterns, enhancing model performance without significantly increasing computational demands. Simultaneously, the Bi-LSTM component allows for more efficient classification of emotional features by leveraging temporal information. The system was evaluated using three datasets: the CISIA Chinese speech emotion dataset, the EMO-DB German emotion corpus, and the IEMOCAP English corpus. The experimental results yielded average emotion recognition accuracies of 59.08%, 89.29%, and 71.25%, respectively. These results represent improvements of 1.17%, 1.36%, and 2.97% over the accuracy achieved by existing speech emotion recognition systems using mainstream models, demonstrating the effectiveness of the proposed approach**.**

Keywords: Speech emotion recognition; Dynamic convolutional kernel; Neural network.

1. Introduction

Speech is a multifaceted signal that conveys a wealth of information, including speaker semantics, emotions, and uses language as an information carrier [1]. Speech emotion recognition (SER) technology aims to extract the features from speech signals that characterize the speaker's emotional state and establish a mapping between these features and human emotions using methods such as machine learning [2]. The ultimate goal of SER is to enable machines to accurately recognize a speaker's emotional state, thereby facilitating intelligent and harmonious human-computer interactions. While speech emotion recognition has existed for decades, recent advancements in deep learning have spurred significant developments in SER technology, which holds great potential for applications in human-computer interaction, in-car navigation systems, teaching aids, medical therapy, robotics, and even video games [3]-[6]. As such, research in SER is highly valuable and holds significant application prospects.

Speech, as a continuous signal of varying lengths, transmits both the speaker's message and emotional expressions. Emotions can manifest in various signals, such as happiness, anger, sadness, calmness, boredom, disgust, and fear [7]. The emotional features embedded within speech signals can be extracted by classification models, and these features can be categorized into three primary groups: rhythmic features, phonological-related features, and spectral features [8]-[9].

Rhythmic features, such as intonation and rhythm, are perceptible by humans and are the most prominent features for conveying emotional content in SER [10]-[14]. Phonological features are related to the quality of sound, measuring attributes like clarity and recognizability. Spectral features, on the other hand, reflect the relationship between vocal tract shape variations and the speaker’s vocalization [15].

Traditional SER systems have commonly employed algorithms such as Hidden Markov Models (HMM) [16], Gaussian Mixture Models (GMM) [17], Support Vector Machines (SVM) [18], and Artificial Neural Networks (ANN) [19]. Additional approaches include decision trees (DT) [20] and k-nearest neighbors (KNN) [21]. HMM is well-suited for time-series sequence recognition but is sensitive to phonemic variations; GMM effectively fits data but is highly reliant on the training dataset; SVM works well for small sample sets but struggles with multi-class classification issues; ANN excels at approximating complex nonlinear relationships but tends to encounter issues like local minima and slow convergence.

In recent years, deep learning algorithms have surpassed traditional machine learning approaches, leading to a shift in research focus toward these methods. The most commonly used deep learning algorithms for SER are convolutional neural networks (CNN) and recurrent neural networks (RNN). CNNs are a specialized type of neural network designed to process grid-like data structures, including images and two-dimensional speech features [22]. By applying multiple convolutional filters, CNNs can effectively capture both temporal and spatial dependencies from the input data, reducing the computational complexity while maintaining feature integrity [23]. However, modern CNNs typically demand significant computational resources due to the extensive convolutional operations, resulting in considerable redundant calculations.

RNNs are capable of processing entire time-series data; however, they retain a strong memory of recent input signals, while earlier signals diminish in importance over time, thus limiting RNNs to short-term memory. By incorporating long-short-term memory (LSTM) structures, RNNs can overcome this limitation and capture long-term dependencies. LSTMs, a subclass of gated RNNs, are highly effective for addressing the long-term dependency issues in RNNs and are widely used in SER [24]. However, LSTMs only extract past information unidirectionally, which limits their ability to fully capture the rich emotional context of complex human speech, as they overlook the influence of later text on earlier information.

To address these challenges, this paper proposes a speech emotion recognition model based on dynamic convolutional neural networks (DyCNN) [25]-[28] and bidirectional long-short-term memory (Bi-LSTM). Dynamic convolution enhances algorithmic performance by reducing network redundancy and improving the flexibility of convolutional kernels, enabling more effective extraction of global emotional information. Additionally, an attention mechanism is employed to assign varying weights to different feature regions in the speech, allowing for better extraction of prominent emotional features in a sentence. The integration of Bi-LSTM addresses the limitations of traditional RNNs regarding long-term dependency and compensates for

LSTM's shortcomings in contextual information extraction, making more efficient use of temporal information. Simulation results demonstrate that the proposed model significantly improves the accuracy of speech emotion recognition systems.

2. DyCNN model combining Bi-LSTM

2.1. Model Framework

In this paper, we propose a network architecture that integrates an LSTM with a DyCNN, as depicted in Figure 1. The model begins by preprocessing the raw speech signal to generate a log-mel spectrogram. The global dynamic spectrogram features are then extracted through the DyCNN, followed by the Bi-LSTM, which captures temporal sentiment information in the context of the surrounding data. Finally, the Softmax layer is employed to perform sentiment classification. The functionality of each module is elaborated upon below.

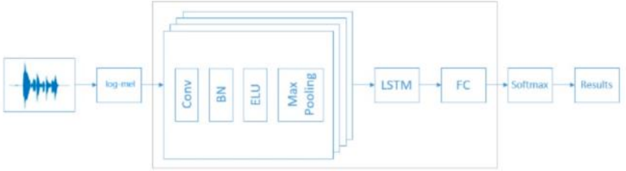

Figure 1. The network architecture

2.2. Feature extraction

The log-mel spectrogram is a technique that captures mood-related changes in speech signals, making it a suitable input for the network. In this paper, the raw speech signal is first acquired, followed by a series of preprocessing steps including pre-emphasis, framing, windowing, and short-time Fourier transformation. The resulting acoustic spectrogram is then passed through a mel filter bank to produce the mel spectrogram, after which the logarithm is applied to obtain the log-mel spectrogram. Finally, the extracted spectrogram features are fed into the subsequent network layers.

2.3. Dynamic Convolutional Neural Network

Traditional convolutional neural networks (CNNs) utilize static convolutional kernels, which share the same parameters across different input samples. However, in the context of speech emotion recognition, it is evident that dynamic convolution is more advantageous than static convolution, as different speakers and content benefit from adaptive processing. Therefore, this paper employs a dynamic convolutional neural network capable of adaptively adjusting its attention based on the input to construct a speech emotion recognition system.

In this work, dynamic convolution is introduced to enhance the convolutional component of the baseline network architecture. Rather than using a single convolutional kernel at each layer, dynamic convolution aggregates multiple parallel convolutional kernels based on attention mechanisms that are dependent on the input. These convolutional kernels are weighted and combined using a matrix of attention weight parameters from the previous layer, resulting in dynamic convolutional kernels that can adapt their attention to the input. These kernels then convolve with the input spectrogram to extract more emotionally relevant features. The structure of a dynamic convolutional layer is illustrated in Figure 2 below.

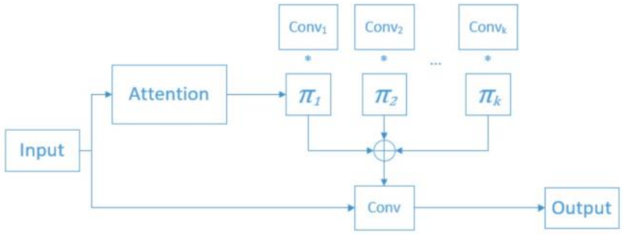

Figure 2. A dynamic convolutional layer

The attention mechanism employs an average pooling layer and two fully connected layers, maintaining low computational complexity and high efficiency. Softmax is applied to constrain the attention weights πk within the range of 0 to 1, allowing the model to learn deep features. For the feature map *xi* produced during the convolution process, several operations are performed to generate *K*= attention weight parameters πk, which sum to 1. These *K* convolution kernel parameters are then linearly combined, resulting in a convolutional kernel that adapts to changes in the input during inference.The dynamic convolution kernel model is calculated as follows.

$$y = \sigma\left(\sum_{k=1}^{K} \pi_k(x)\widetilde{W}_k \cdot x + \sum_{k=1}^{K} \pi_k(x)\tilde{b}_k\right)$$

2.4. Bidirectional long- and short-term memory network

The LSTM model builds upon the RNN architecture by incorporating input gates, forget gates, unit states, and output gates. During network training, the gate structure enables the addition or removal of information, allowing the model to decide which relevant information should be retained or discarded through the gating mechanism on the unit state. At time step *t*, the update of each gate state can be expressed as follows.

$$\begin{gathered} f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\ i_t = \sigma(W_i \cdot [h_{t-1} \cdot x_t] + b_i) \\ O_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ C_t = f_t * C_{t-1} + i_t * tanh(W_C \cdot [h_{t-1} \cdot x_t] + b_C) \\ h_t = o_t * tanh(C_t) \end{gathered}$$

Here, f_t represents the forget gate, i_t is the input gate, o_t is the output gate, and C_t is the cell state. W* denotes the weight matrix, while x_t and h_t refer to the input vector and the hidden state vector at time step *t*, respectively. The term b* is the bias, and σ is the activation function.

In this paper, we replace the LSTM component of the baseline network architecture with Bi-LSTM. Bi-LSTM combines both forward and backward LSTMs. Similar to LSTM, Bi-LSTM is frequently used to capture contextual information in natural language processing tasks. However, when using LSTM to model speech emotion signals, it is limited by its inability to encode information from back to front, a problem effectively addressed by Bi-LSTM. Figure 3 illustrates the structure of Bi-LSTM unfolded across time.

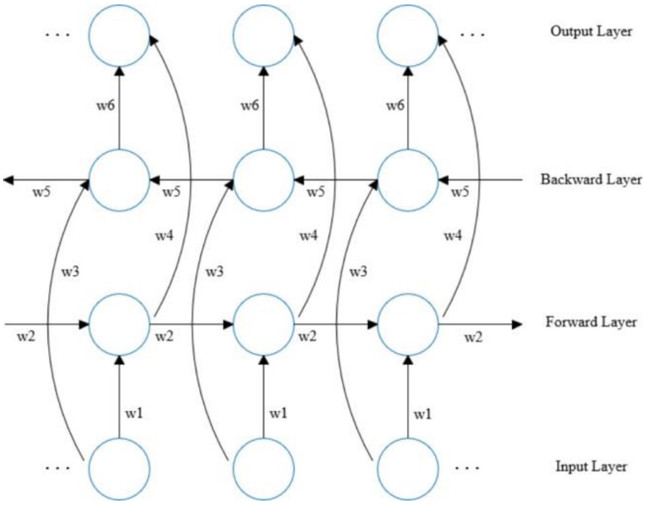

Figure 3. The structure of Bi-LSTM expanded along time

As illustrated in the figure, the Bi-LSTM consists of two unidirectional LSTMs: the forward LSTM and the backward LSTM. The forward LSTM calculates the forward contextual information, while the backward LSTM computes the backward contextual information. This bidirectional structure enables the model to capture the full context, thereby enhancing the recognition accuracy.

3. Experiments

3.1. Speech emotion dataset

Speech emotion recognition aims to identify the emotional state of a speaker during vocal communication. In everyday life, natural speech carries a wide range of emotions, and the actual environment is highly complex. As a result, capturing natural speech in real-world settings to create a speech emotion dataset is both challenging and intricate. Such a dataset must satisfy four key criteria: authenticity, continuity, interactivity, and diversity, while minimizing external interference as much as possible. Consequently, speech emotion recognition research typically relies on speech emotion corpora recorded in controlled, quiet environments like recording studios.

A speech emotion corpus forms the foundation of speech emotion recognition systems, and the performance of these systems heavily depends on having a large, diverse, and high-quality corpus. Discrete emotion datasets, which contain relatively distinct emotional states, are particularly useful for recognizing simple emotional speech signals. Below is an overview of several widely-used speech emotion corpora for the classification and recognition of discrete emotions.

1. The EmoDB corpus is a German discrete emotion corpus created in the Berlin laboratory. It consists of recordings from five male and five female actors, producing 535 sentences with seven emotional categories: happiness, anger, sadness, calmness, boredom, disgust, and fear.

2. The CASIA corpus is a Chinese discrete emotion corpus recorded by the Institute of Automation at the Chinese Academy of Sciences. It includes recordings from four professional speakers (two male and two female), with 9,600 utterances in total. These recordings, performed with six emotional states—anger, fear, happiness, neutrality, sadness, and surprise—contain 300 identical texts and 100 unique texts.

3. The IEMOCAP corpus, collected by the SAIL Lab at the University of Southern California, is one of the largest dimensional speech emotion datasets used for emotion recognition. It comprises approximately 12 hours of conversational speech from 10 actors, recorded in five separate sessions, with each conversation involving two speakers. The corpus is annotated by at least two annotators, with labels for emotions such as anger, happiness, sadness, and neutrality, as well as three emotional dimensions: arousal, valence, and dominance.

To evaluate the effectiveness of the model proposed in this paper for speech emotion recognition, experiments are conducted using the three corpora mentioned above. The weighted average accuracy (WA) is employed as the evaluation metric, and the results are compared with existing mainstream models. The WA is calculated using the following equation, where *n* denotes the number of correctly identified test samples, and *N* represents the total number of test samples.

$$WA = \frac{\mathrm{n}}{\mathrm{N}}$$

Next, the MFCC feature parameter extraction is performed, yielding a data dimension of 39. The extracted speech spectrogram is then normalized, and the resulting data is used as the input for the model.

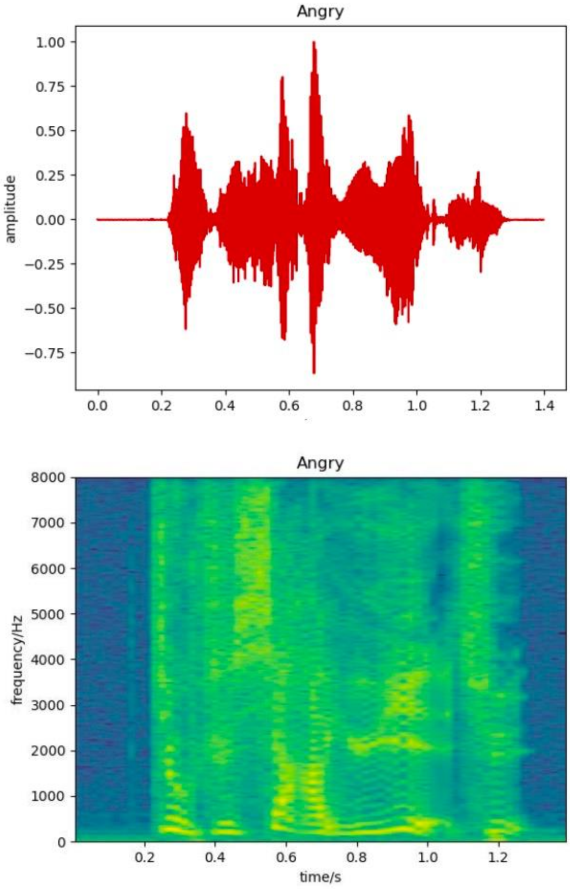

Figure 4. Original speech signal and speech spectrogram of the angersample

In this paper, the TensorFlow toolkit is employed to construct the network model and implement the training algorithm. The model parameters are optimized using the RMSProp algorithm, with an initial learning rate of 0.01. Cross-entropy serves as the loss function, with a training batch size of 200 and 1000 iterations.

4. Conclusion

This paper enhances the classification network model for speech emotion recognition by introducing a novel network model that employs dynamic convolutional neural networks (DyCNN) in place of traditional convolutional neural networks (CNN), combined with bidirectional long short-term memory networks (Bi-LSTM). The proposed model leverages dynamic convolution to address the data redundancy problem inherent in traditional CNNs, allowing for more flexible extraction of key emotional features while maintaining computational efficiency. The integration of Bi-LSTM enables the model to more comprehensively and effectively utilize the weight coefficients and temporal information of each emotion feature for emotion recognition, resulting in a significant improvement in the system's recognition accuracy.However, the proposed network has been tested on only three datasets, and there remains room for improvement in recognition accuracy when compared to other well-established speech emotion recognition methods. Future research will focus on selecting additional high-quality speech emotion datasets for further experimentation and on optimizing the network architecture to enhance recognition accuracy while maintaining computational power and processing speed.

References

[1] Akçay M B, Oğuz K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers[J]. Speech Communication, 2020, 116: 56-76.

[2] Abbaschian B J, Sierra-Sosa D, Elmaghraby A. Deep learning techniques for speech emotion recognition, from databases to models[J]. Sensors, 2021, 21(4): 1249.

[3] Pandey S K, Shekhawat H S, Prasanna S R M. Deep learning techniques for speech emotion recognition: A review[C]//2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA). IEEE, 2019: 1-6.

[4] Issa D, Demirci M F, Yazici A. Speech emotion recognition with deep convolutional neural networks[J]. Biomedical Signal Processing and Control, 2020, 59: 101894.

[5] Lee J, Tashev I. High-level feature representation using recurrent neural network for speech emotion recognition[C]//Interspeech 2015. 2015.

[6] Kim J, Saurous R A. Emotion Recognition from Human Speech Using Temporal Information and Deep Learning[C]//Interspeech. 2018: 937-940.

[7] El Ayadi M, Kamel M S, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases[J]. Pattern recognition, 2011, 44(3): 572-587.

[8] Fahad M S, Ranjan A, Yadav J, et al. A survey of speech emotion recognition in natural environment[J]. Digital Signal Processing, 2020: 102951.

[9] Roy T, Marwala T, Chakraverty S. A survey of classification techniques in speech emotion recognition[J]. Mathematical Methods in Interdisciplinary Sciences, 2020: 33-48.

[10] Reshma C V, Rajasree R. A survey on Speech Emotion Recognition[C]//2019 IEEE International Conference on Innovations in Communication, Computing and Instrumentation (ICCI). IEEE, 2019: 193-195.

[11] Ai X, Sheng V S, Fang W, et al. Ensemble learning with attention-integrated convolutional recurrent neural network for imbalanced speech emotion recognition[J]. IEEE Access, 2020, 8: 199909-199919.

[12] Hajarolasvadi N, Demirel H. 3D CNN-based speech emotion recognition using k-means clustering and spectrograms[J]. Entropy, 2019, 21(5): 479.

[13] Iqbal A, Barua K. A real-time emotion recognition from speech using gradient boosting[C]//2019 International Conference on Electrical, Computer and Communication Engineering (ECCE). IEEE, 2019: 1-5.

[14] Ringeval F, Eyben F, Kroupi E, et al. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data[J]. Pattern Recognition Letters, 2015, 66: 22-30.

[15] Garg U, Agarwal S, Gupta S, et al. Prediction of Emotions from the Audio Speech Signals using MFCC, MEL and Chroma[C]//2020 12th International Conference on Computational Intelligence and Communication Networks (CICN). IEEE, 2020: 87-91.

[16] Eddy S R. Profile hidden Markov models[J]. Bioinformatics (Oxford, England), 1998, 14(9): 755-763.

[17] Reynolds D A. Gaussian mixture models[J]. Encyclopedia of biometrics, 2009, 741(659-663).

[18] Hearst M A, Dumais S T, Osuna E, et al. Support vector machines[J]. IEEE Intelligent Systems and their applications, 1998, 13(4): 18-28.

[19] Jain A K, Mao J, Mohiuddin K M. Artificial neural networks: A tutorial[J]. Computer, 1996, 29(3): 31-44.

[20] Quinlan J R. Induction of decision trees[J]. Machine learning, 1986, 1: 81-106.

[21] Peterson L E. K-nearest neighbor[J]. Scholarpedia, 2009, 4(2): 1883.

[22] Kwon S. MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach[J]. Expert Systems with Applications, 2021, 167: 114177.

[23] Kwon S. A CNN-assisted enhanced audio signal processing for speech emotion recognition[J]. Sensors, 2020, 20(1): 183.

[24] Wang J, Xue M, Culhane R, et al. Speech emotion recognition with dual-sequence LSTM architecture[C]//ICASSP 2020- 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 6474-6478.

[25] Chen Y, Dai X, Liu M, et al. Dynamic convolution: Attention over convolution kernels[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11030-11039.

[26] Yang B, Bender G, Le Q V, et al. Condconv: Conditionally parameterized convolutions for efficient inference[J]. arXiv preprint arXiv:1904.04971, 2019.

[27] Zhang Y, Zhang J, Wang Q, et al. Dynet: Dynamic convolution for accelerating convolutional neural networks[J]. arXiv preprint arXiv:2004.10694, 2020.

[28] Wen H, You S, Fu Y. Cross-modal dynamic convolution for multi-modal emotion recognition[J]. Journal of Visual Communication and Image Representation, 2021: 103178.