Biomedical Signal Processing and Control 47 (2019) 312–323

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Speech emotion recognition using deep 1D & 2D CNN LSTM networks

Jianfeng Zhao[a,b], Xia Mao[a], Lijiang Chen[a,*]

[a] School of Electronics and Information Engineering, Beihang University, 100083, Beijing, China
[b] School of Information Engineering, Inner Mongolia University of Science & Technology, 014010, Baotou, China

ARTICLE INFO

Article history:
Received 12 July 2017
Received in revised form 26 July 2018
Accepted 27 August 2018

Keywords:
Speech emotion recognition
CNN LSTM network
Raw audio clips
Log-mel spectrograms

ABSTRACT

We aimed at learning deep emotion features to recognize speech emotion. Two convolutional neural network and long short-term memory (CNN LSTM) networks, one 1D CNN LSTM network and one 2D CNN LSTM network, were constructed to learn local and global emotion-related features from speech and log-mel spectrogram respectively. The two networks have the similar architecture, both consisting of four local feature learning blocks (LFLBs) and one long short-term memory (LSTM) layer. LFLB, which mainly contains one convolutional layer and one max-pooling layer, is built for learning local correlations along with extracting hierarchical correlations. LSTM layer is adopted to learn long-term dependencies from the learned local features. The designed networks, combinations of the convolutional neural network (CNN) and LSTM, can take advantage of the strengths of both networks and overcome the shortcomings of them, and are evaluated on two benchmark databases. The experimental results show that the designed networks achieve excellent performance on the task of recognizing speech emotion, especially the 2D CNN LSTM network outperforms the traditional approaches, Deep Belief Network (DBN) and CNN on the selected databases. The 2D CNN LSTM network achieves recognition accuracies of 95.33% and 95.89% on Berlin EmoDB of speaker-dependent and speaker-independent experiments respectively, which compare favourably to the accuracy of 91.6% and 92.9% obtained by traditional approaches; and also yields recognition accuracies of 89.16% and 52.14% on IEMOCAP database of speaker-dependent and speaker-independent experiments, which are much higher than the accuracy of 73.78% and 40.02% obtained by DBN and CNN.

1. Introduction

Speech emotion recognition has attracted much attention in the last decades. Emotions are specific and intense mental activities, which can be signed outward by many expressive behaviors. Speech, facial expression, body gesture, and brain signals etc., are the cues of the whole-body emotional phenomena [1–3]. Speech is a fast, efficient and essential pathway of human's communication. So, recognizing speech emotion is one of the important research directions in emotion detection and recognition naturally [4,5].

In order to recognize the emotional state of the speaker, distinguishing paralinguistic features which do not depend on the speaker or the lexical content need to be extracted from the speech. In general, there are two types of information in speech: linguistic information, and paralinguistic information. The linguistic information always refers to the context or the meaning of the speech. The paralinguistic information comes to mean the implicit messages such as the emotion contained in the speech [4,6–8].

There are many distinguishing acoustic features usually used into recognizing the speech emotion: continuous features, qualitative features, and spectral features [9–13]. Many features have been investigated to recognize speech emotion. Some researchers weighted the pros and cons of each feature, but no one can identify which category is the best one until now [4,6,14,15].

In order to learn high-level features from emotion utterances and form a hierarchical representation of the speech, many deep learning architectures have been introduced in speech emotion recognition. The classification accuracy of handcrafted features extracted from certain emotion utterances is relatively high, but the extraction of handcrafted features always consumes expensive manual labor and depend on professional knowledge [6,16,17]. The handcrafted features extraction normally overlooks the high-level features, which are derived from lower level features. So, hierarchical learning, also known as deep learning, is introduced to model high-level abstractions of the data.

Speech signal processing has been revolutionized by deep learning. More and more researcher achieved excellent results

* Corresponding author at: School of Electronics and Information Engineering, Beihang University, Mailbox 206, 37 XueYuan Road, Beijing, China.
E-mail address: chenlijiang@buaa.edu.cn (L. Chen).

https://doi.org/10.1016/j.bspc.2018.08.035
1746-8094/

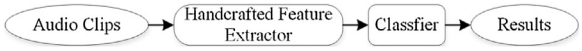

Fig. 1. A general flow chart of traditional speech emotion recognition approach. The handcrafted features are extracted from raw data.
(a) The deep features are extracted from raw data.
(b) The deep features are learned from handcrafted features.

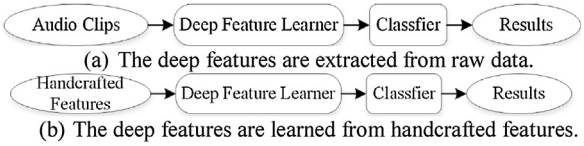

Fig. 2. Two flow charts of the speech emotion recognition approaches adopted in this paper.

in certain applications using deep belief networks (DBNs), convolutional neural networks (CNNs) and long short-term memory (LSTM) [18–20,32]. Deep neural networks are typical "black box" approaches, because it is extremely difficult to understand how the final output is arrived at. There are two models or methods have been introduced to study relevant problems or coincidences. Compared to the "data model" used largely by statisticians, deep networks focus on finding an algorithm to do prediction, so they are called "algorithmic model" [55], [56]. The interpretability of how the highly abstracted features are learned by deep neural networks (DNNs) is poor [57]. But deep neural networks perform dramatically better than traditional approaches (see Fig. 1) in some experiments [21,22].

We constructed two convolutional neural network and long short-term memory (CNN LSTM) networks by stacking four designed local feature learning blocks (LFLBs) and other building layers to extract emotional features. The speech signal is a time-varying signal which needs special processing to reflect time-varying properties. Therefore, LSTM layer is introduced to extract long-term contextual dependencies. The 1D CNN LSTM network is intended to recognize speech emotion from audio clips (see Fig. 2a); the 2D CNN LSTM network mainly focuses on learning global contextual information from the handcrafted features (see Fig. 2b). Most of the traditional features extraction algorithms can reduce data dimension dramatically. The amount of extracted low-level features, such as the spectrum features [23,24], is smaller than that of the raw data. A significant advantage of the learning from a small amount of the low-level features is the decreasing of the training time. The experimental results show that the designed CNN LSTM networks can recognize the speech emotion effectively. Moreover, the designed 2D CNN LSTM network does not only achieve high emotion recognition accuracies but also has better generalization ability. High recognition rate and good generalization ability can provide a guarantee for the application of the designed networks in disease prevention, health care, medical diagnosis, social intercourse etc.

Our original contributions of the work are as follows: 1) a local feature learning block (LFLB), which consists of one convolutional layer, one batch normalization (BN) layer, one exponential linear unit layer, and one max-pooling layer, is designed to extract local features; 2) to learn long-term dependencies from a sequence of local features, LSTM layer is introduced to build CNN LSTM networks following the LFLB; 3) it is proved experimentally that 1D CNN LSTM network can learn lots of emotional features from raw audio utterances for the first time. In our experiments, 2D CNN LSTM network achieves better results. 2D CNN LSTM network focuses on capturing both local correlations and global contextual information from log-mel spectrogram, which is a representation of how the frequency content of a signal changes with time. When log-mel spectrogram is considered as a grid or a sequence, it can be processed by LFLB or LSTM layer.

2. Related work

Distinguishing features are essential for recognizing the speech emotion. Among many paralinguistic features, spectrum features are widely used in speech emotion recognition. AB Kandali et al. presented a method based on MFCCs as features and Gaussian mixture model classifier to recognize emotion from Assamese speeches [25]. Milton, A. et al. used a 3-stage Support Vector Machine classifier to classify seven different emotions present in the Berlin Emotional Database (Berlin EmoDB) [26]. VB Waghmare et al. adopted MFCCs to analyze and recognize speech emotion from artificial emotional Marathi speech database [27]. Demircan, S. et al. used a k-NN algorithm to classify the speech emotion after extracting MFCCs from the audio clips of the Berlin EmoDB [28]. Nalini, N. J. et al. developed a speech emotion recognition system using the residual phase and MFCCs features with the autoassociative neural network (AANN) [29]. Chenchah, Farah et al. used a Hidden Markov Model (HMM) and Support Vector Machine (SVM) to classify the spectral features extracted from audio characteristics of emotional speech [30]. Nalini, N. J. et al. combined the evidence from MFCCs and residual phase (RP) features to recognize emotion in music using AANN, SVM, RBFNN, respectively [31]. Though handcrafted features are very effective to distinguish emotions in speech, most of them are low-level features.

With numerous successful applications of DNNs, more and more researchers began to focus on the learning of deep emotional features. Andre Stuhlsatz and collaborators introduced a generalized discriminant analysis (GerDA) DNNs stacked by several restricted Boltzmann machines (RBMs) to recognize the speech emotion and obtained a highly significant improvement over the previously reported baselines by SVMs [33]. Erik M. Schmidt et al. employed a regression-based deep belief network which was configured with three hidden layers to learn features directly from magnitude spectra and recognize music emotion [34]. Duc Le et al. proposed and evaluated a set of hybrid classifiers based on hidden Markov models and deep belief networks and achieved state-of-the-art results on FAU Aibo [35]. Kun Han et al. proposed to utilize deep neural networks (DNNs) to recognize utterance-level emotions, and obtained 20% relative accuracy improvement compared to the traditional state-of-the-art approaches [17]. Qirong Mao et al. introduced a semi-CNN architecture with a linear SVM to recognize speech emotion and achieved a stable and robust recognition performance in complex scenes [36]. W. Q. Zheng et al. also constructed a CNN architecture to implement emotion recognition on labelled audio data, the preliminary experimental results showed that this approach outperformed the SVM-based classification [21].

Our work differs from the work mentioned above. The designed 1D & 2D CNN LSTM networks learn hierarchical local and global features to recognize speech emotion. Whereas most of the data models can only extract low-level features to classify emotion, and most of the previous DBN-based or CNN-based algorithmic models can only learn one type of emotion-related features to recognize emotion.

3. Methods and materials

Extracting more distinguishing emotion features is one of the main tasks for researchers to recognize speech emotion. According to the difference of feature extraction methods, speech features can be classified as handcrafted features and learned features. Most of the extraction of handcrafted features are carefully designed using ingenious strategies and can be explained in more detail how it