SER
Raw Research Paper

1. Introduction

Speech emotion recognition (SER), which is the process that aims to classify the emotional state of the speaker, has currently drawn a great deal of interest among many researchers because of the increasing demand in many applications. SER is one of the vital tasks which can improve human-computer interaction. Examples of this usage are call center services, medical diagnosis, mental state detection, or stressed emotion detection of a car driver. From the viewpoint of those real-world applications, the SER is required to be implemented in embedded systems, which have limited resources and computational power. Several techniques are used in SER systems, such as the Gaussian mixture model (GMM), hidden Markov model (HMM), nearest-neighbor (k-NN), and artificial neural network (ANN) have been proposed to achieve high recognition rates. Recently, Iliou and Anagnostopoulos showed that a method based on deep learning outperforms other classifiers in terms of recognition rate. However, a critical issue in deep learning methods is that deep learning models are large and consist of over a million parameters. Consequently, the deep neural network-based SER systems require high computational complexity and are not suitable for implementing in embedded systems.

There are many speech features proposed and used for the SER, such as pitch, Linear Predictive Cepstral Coefficient (LPCC), Teager energy operator (TEO), and MelFrequency Cepstral Coefficient (MFCC). Some researchers claim that's MFCC might not be the best performance feature. However, MFCC is used in this work because of the promising results of the previous works in emotion classification, and it becomes the most widely used feature in SER. There are several pieces of research based on the implementation of MFCCs. From the literature review, it can be found that the dimension of MFCCs can be reduced by using the average values of MFCC]. Representing those MFCCs by their average values has been adopted in various recognition applications. For example, Badlaoui and Hammouch used mean values of MFCCs to achieve 84.21% recognition rate in their heart sound classification.

Xu et al. used the average value of Weighted-MFCCs in the voice of the Parkinson patient's Classification to achieve an 89.5% recognition rate. Naini A. Mel-Frequency Cepstral Coefficients (MFCC) Mel-Frequency Cepstral Coefficients (MFCC) represents the speech signals' spectral property. The vocal cords, the tongue, and the teeth are all the elements that filter the sound and make it unique for each speaker. The shape of these elements determines the voice. The shape manifests in the envelope of the short-time power spectrum, and MFCCs represent this envelope. MFCC extraction process consists of five-steps.

First, the signal is segmented into 20-40 frames, and the frame step is set, which allows some 5-15 ms to overlap to the frames. Second, the discrete Fourier transform (DFT) is performed to each frame, and the absolute value of the complex Fourier transform is determined, and the result is squared. Third, the mel-spaced filterbank is computed. This is a set of 20-40 triangular filters. Fourth, the log of energies is calculated to obtain log filterbank energies. Finally, the discrete cosine transform (DCT) is performed to the log filterbank energies to get cepstral coefficients.

2. MFCC

In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.

Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC.[1] They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound, for example, in audio compression.

MFCCs are commonly derived as follows

1. Take the Fourier transform of (a windowed excerpt of) a signal.
2. Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
3. Take the logs of the powers at each of the mel frequencies.
4. Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
5. The MFCCs are the amplitudes of the resulting spectrum.

There can be variations on this process, for example: differences in the shape or spacing of the windows used to map the scale,[3] or addition of dynamics features such as "delta" and "delta-delta" (first- and second-order frame-to-frame difference) coefficients

3. Artificial Neural Network (ANN)

The artificial neural network (ANN) is a machine learning algorithm that mimics the biological neuron’s action. It has a flexible way to handle the data without specifying any relationships between the input and output.

Usually, it consists of 3 layer types: the input layer, hidden layers, and the output layer. Each layer contains several neurons or nodes connected to each node in the previous layer by weighted

links. The node's output response is calculated by applying an activation function to the sum of the weighted inputs. This work aims to improve the performance of SER using fewer parameters. In a deep learning model, The number of parameters in the layer is the count of the learnable element for a filter in that layer. In the convolutional layer, it can be computed by multiply the shape of filter width, the shape of filter height plus 1, and all multiply by the number of filters in the layer. Bias term is also included; one is added because of it. In a fully-connected layer, all input has a separate weight to each output. The number of parameters used in a fully-connected layer can be computed by the number of input plus 1, which is the bias term, then multiply by the number of output.

Recently, there are many deep convolutional neural network methods for SER that have been proposed. Zhang et al. proposed the transfer learning technique on AlexNet [6] to and Hammounch used mean values of MFCCs concatenated with jitters and shimmers features, which can reduce the error rate to be 13.5% in speaker sex identification. It can be seen that the recognition rates obtained from using the mean values of MFCCs from those applications are excellent. It is our understanding that no research group directly uses the mean values of MFCCs in SER. Therefore, the goal of this work is to investigate the possibility of using the mean.

4. Mel-Frequency Cepstral Coefficients (MFCC)

Mel-Frequency Cepstral Coefficients (MFCC) represents the spectral property of the speech signals. The vocal cords, the tongue, and the teeth are all the elements that filter the sound and make it unique for each speaker. The shape of these elements determines the voice. The shape manifests in the envelope of the short-time power spectrum, and MFCCs represent this envelope. MFCC extraction process consists of five steps.

First, the signal is segmented into 20-40 frames, and the frame step is set, which allows some 5-15 ms to overlap to the frames. Second, the discrete Fourier transform (DFT) is performed to each frame, and the absolute value of the complex Fourier transform is determined, and the result is squared. Third, the mel-spaced filterbank is computed. This is a set of 20-40 triangular filters. Fourth, the log of energies is calculated to obtain log filterbank energies. Finally, the discrete cosine transform (DCT) is performed to the log filterbank energies to get cepstral coefficients.

5. Convolutional Neural Network (CNN)

The convolutional neural network (CNN) is a machine learning algorithm that mimics the biological neuron's action. It has a flexible way to handle the data without the need to specify any relationships between the input and output. Normally, it consists of 3 layer types, which are the input layer, hidden layers, and output layer, and each layer contains several neurons or nodes that are connected to each of the nodes in the previous layer by weighted links. The output response from the node is calculated by applying an activation function to the sum of the weighted inputs.

The aim of this work is to improve the performance of SER using fewer parameters. In a deep learning model, The number of parameters in the layer is the count of the learnable element for a filter in that layer. In the convolutional layer, it can be

computed by multiply the shape of filter width, the shape of filter height plus 1, and all multiply by the number of filters in the layer. Bias term is also included, 1 is added because of it. In a fully-connected layer, all input has a separate weight to each output. The number of parameters used in a fully-connected layer can be computed by the number of input plus 1 which is the bias term then multiply by the number of output. Recently, there are many deep convolutional neural network methods for SER that have been proposed.

Zhang et al. proposed the transfer learning technique on AlexNet to achieve a high recognition rate. However, this model consists of about 60 million parameters. Pandey et al. also proposed a CNN model in which the number of parameters is also over a million. The more layers are added, the more parameters are used, the more resources and computing units are needed. This is the main reason that the computational complexity of the model needs to be reduced to make the model more lightweight and consume fewer resources by simply use only fully-connected layers instead of using convolutional layers.