International Journal of Engineering

Journal Homepage: www.ije.ir

Speech Emotion Recognition Using Scalogram Based Deep Structure

K. Aghajani, I. Esmaili Paeen Afrakoti*

Department of Engineering and Technology, University of Mazandaran, Babolsar, Iran

P A P E R I N F O

Paper history:
Received 27 October 2019
Received in revised form 16 January 2020
Accepted 17 January 2020

Keywords:
Continuous Wavelet Transform
Emotion Recognition
Convolutional Neural Network
Recurrent Network
Long-short Term Memory

A B S T R A C T

Speech Emotion Recognition (SER) is an important part of speech-based Human-Computer Interface (HCI) applications. Previous SER methods rely on the extraction of features and training an appropriate classifier. However, most of those features can be affected by emotionally irrelevant factors such as gender, speaking styles and environment. Here, an SER method has been proposed based on a concatenated Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN). The CNN can be used to learn local salient features from speech signals, images, and videos. Moreover, the RNNs have been used in many sequential data processing tasks in order to learn long-term dependencies between the local features. A combination of these two gives us the advantage of the strengths of both networks. In the proposed method, CNN has been applied directly to a scalogram of speech signals. Then, the attention-mechanism-based RNN model was used to learn long-term temporal relationships of the learned features. Experiments on various data such as RAVDESS, SAVEE, and Emo-DB demonstrate the effectiveness of the proposed SER method.

***doi**: 10.5829/ije.2020.33.02b.13*

1. INTRODUCTION

Emotions play an important role in many speech-based human-computer interface applications [1-3]. Therefore, in recent years, Speech Emotion Recognition (SER) has attracted increasing attention. However, in spite of improvements in this area, SER is still a challenging taskb ecause emotions can be expressed by people in different ways. Besides, emotion-irrelevant factors, such as gender and age can affect the speech signal in various ways.

In the conventional methods, distinguishable paralinguistic features should be extracted from the training data and then a machine learning algorithm should be learned using these features to estimate the emotion of the new input case [4-7]. These features should not depend on the lexical context or the speaker. Common features utilized in the field include energy-related, pitch, Linear Predictive Spectrum Coding (LPCC), Mel-Frequency Spectrum Coefficients (MFCC), Mel-Energy Spectrum Dynamic Coefficients (MEDC) and formant frequencies [8]. In addition, some classification methods such as K-Nearest Neighbors (KNN), Hidden Markov Model (HMM), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) have been utilized in many researches [8-12]. In [13], a feature combination of MFCC, MEDC, and energy was utilized along with the SVM method to recognize the emotion. In literature [14], short-time Log Frequency Power Coefficients (LFPC) were used as a feature vector and Hidden Markov Model (HMM) was used as the classifier. In [15], prosodic features were extracted using Continuous Wavelet Transform (CWT) coefficients and the SVM method was utilized for classification. In [16], hyper prosodic features were extracted, and a deep neural network was utilized for classification. Correa et al. extracted features using bionic wavelet transformation and utilized Gaussian Mixture Models (GMM) for classification [17].

Hardware improvement especially using the Graphical Processing Unit (GPU) for computation and also the development of robust learning algorithms become one of the significant causes of the spreading use of the Convolutional Neural Networks (CNN) and deep structures in many real-world applications [18-21]. CNN can be directly used for emotion recognition task using

*Corresponding Author Email: *i.esmaili.p@umz.ac.ir* (I. Esmaili Paeen Afrakoti)

Please cite this article as: K. Aghajani, I. Esmaili Paeen Afrakoti, Speech Emotion Recognition using Scalogram based Deep Structure, International Journal of Engineering (IJE), IJE TRANSACTIONS B: Applications Vol. 33, No. 2, (February 2020) 285-292

the image data [22-27]. But using these structures for speech input data has some difficulties. In the following, some significant works and ideas are reviewed about speech emotion recognition using CNN and deep structures.

In litrature [28], an end-to-end structure was used for SER based on a convolution operation as a feature step following a Long Short-Term Memory (LSTM) structure. After applying an FIR filter for reducing the noise, each 6-sec segment of the raw speech signal will be fed to multiple 1D convolution layer structures for extracting a high-level feature for LSTM. This system performed very well on some datasets with respect to state-of-the-art algorithms.

Short-Term Fourier Transform (STFT) is a famous method in this field for transforming the input speech signal into an image datum called a spectrogram. In litrature [29], a semi-CNN architecture consisting of an input layer, one convolutional layer, one fully-connected layer, and an SVM classifier was used. The spectrogram of the speech signal is given to the input of the structure as a 15×60 image datum. The proposed algorithm has shown high accuracy and robustness to speaker variation in the emotion recognition task. In litrature [30], an SER system is designed based on CNN networks with spectrogram and phoneme sequences as input data which shows 4 percent better accuracy compared to other state-of-the-art algorithms.

Jiang et al. introduced a parallelized convolutional recurrent neural network with spectral features [31]. Two parallel computing paths, consisting LSTM and CNN, form a unified strcuture applying on the frame level and log Mel-spectrogram features which uses a softmax classifier at the end. In [32], a CNN architecture followed by an LSTM structure is used for emotion recognition based on the raw spectrogram of the input speech signal using a data augmentation approach. The results of this paper wwere 64.5% for weighted accuracy and 61.7% for unweighted accuracy on the IEMOCAP dataset. A structure that contains two 1D and 2D CNN LSTM networks with speech and log-mel spectrogram inputs is proposed in litrature [33]. The results wwere 89.16 and 52.14 percent for speaker-dependent and speaker-independent experiments on IEMOCAP, respectively. The authors in litrature [34] splitted the input signal to overlapping segments, and then an 88-dimensional vector was extracted that contains features such as MFCC, pitch, and intensity for each frame . Using the K-means algorithm, the dimension of the data decreased and then it was encapsulated in a 3D tensor. The 3D tensor was fed to a 3D convolutional network which showed an acceptable result for the SER problem.

In this paper, an approach has been proposed to recognize emotion in speech signals using deep convolutional and recurrent networks with attention mechanisms. First, speech segments are converted into 3D scalograms. Given a 3D scalogram, convolutional layers are used to extract high-level features. To reflect the time-varying properties of the speech signal, an LSTM layer is used to extract long-term dependencies. Finally, an attention layer followed by fully connected and a softmax layer are used to make a final decision. In convolutional and max-pooling layers, rectangular kernels and square ones are utilized. The proposed method is evaluated on three widely used emotional speech databases with different languages. The experiments revealed that using scalogram as the input signal to the proposed model can improve the SER accuracy.

The rest of the paper is organized as follows. Section 2 presents the proposed method which is composed of generating the 3D scalogram as the input to the proposed CNN-LSTM architecture. In Section 3, some popular SER datasets are described and the experimental results are reported. Finally, concluding remarks are provided in Section 4.

2. MAIN IDEA AND THE PROPOSED METHOD

In this section, the proposed SER method is introduced. The audio signals are split into equal-length (2.5 seconds) segments. Then, the 3D scalogram for each segment is generated as the model input. The model consists of several 3D convolutional layers, an LSTM layer, and a fully connected layer followed by a softmax layer. Below we will discuss the details of the proposed method.

2. 1. 3D Scalogram Generation In order to reduce the variations between different speakers, the signals are normalized to zero mean with variance equal to 1. Then, a Voice Activity Detection (VAD) method is used to eliminate silence at the beginning and end of each signal. As described above, each speech signal has been split into segments. Each segment is also splitted into 25 *msec* frames with 10 *msec* overlap. Then, the scalogram of each frame is computed using CWT. To compute the CWT of a signal s(t), Equation (1) is used.

$$W(a,b)=\frac{1}{|a|^2}\int_{-\infty}^{+\infty} s(t)\bar{\psi}(\frac{t-b}{a})dt\,, \qquad (1)$$

where $a\in R^+$ and $b\in R$ are the scale and the transitional value, respectively; $\psi(t)$ is called the mother wavelet and $\bar{\psi}(t)$ is its complex conjugate.

Different daughter wavelets can be produced by changing the values of *a* and *b*. Among the available implemented wavelets, Morlet wavelet is a suitable wavelet for the application of speech processing [15]. Its complex-valued function is defined as Equation (2).

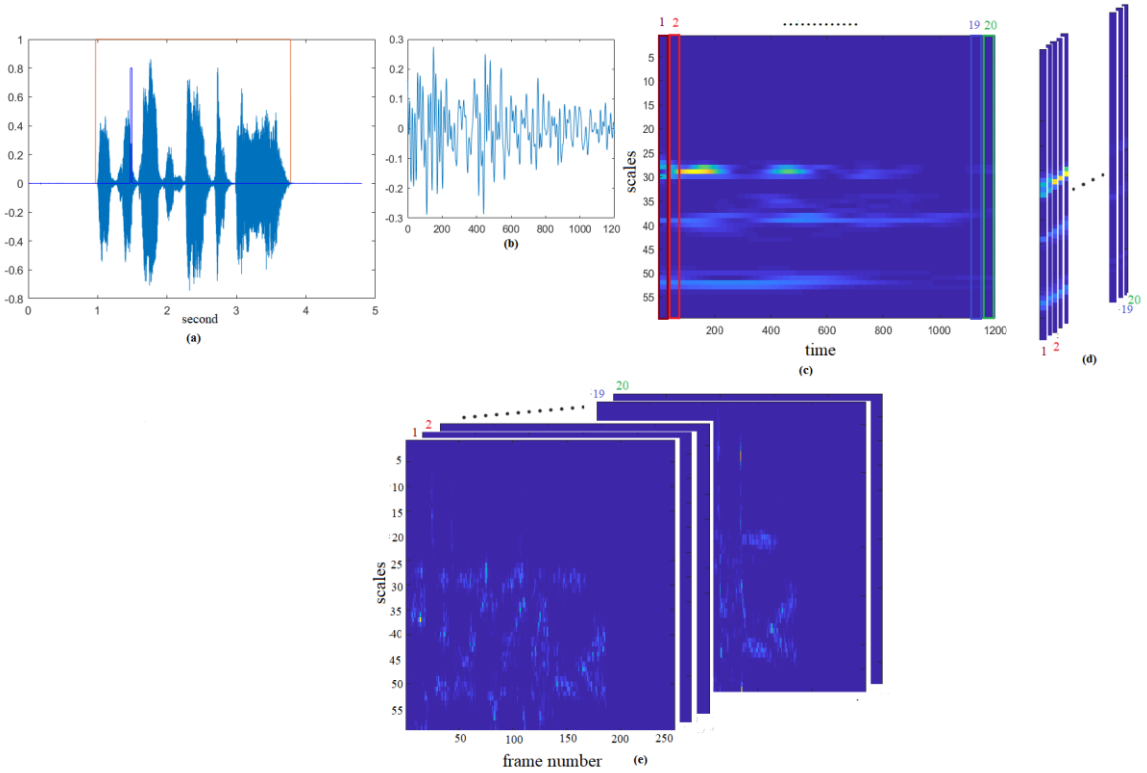

Figure 1. An example of generating a 3D scalogram of a speech segment: (a) The waveform of a sample speech signal. The active area is specified by a red box using a VAD method. (b) A sample frame that is specified by the blue box. (c) The scalogram of the specified frame using the 'Morlet' wavelet. (d) The obtained average amount in each stripe. (e) The final 3D scalogram which is generated by putting together the obtained vectors for all frames

$$\psi(t)=\exp\left(-\frac{t^2}{2\sigma^2}\right)\exp(i\xi t)\ , \qquad (2)$$

where σ is the width of the Gaussian and ξ can specify the time-frequency precision trade-off. Here, these parameters are set to 1 and 5, respectively [15].

The result of applying CWT on each frame is a scalogram. A typical sample frame with its scalogram is given in Figures 1.b and 1.c, respectively. To reduce the computational complexity, each scalogram is divided into strips of equal length (here 20 bars). Then, for each bar, the average values in each scale are computed. We put the vector obtained for each bar in depth (Figure 1.d). If we bring the obtained vectors (for the whole frames) together, we will obtain the data with dimensions of $n_f \times n_s \times n_b$ where n_f is the frame number, n_s is the number of the scales in CWT, and n_b is the number of bars. The whole process is described in Figure 1.

2. 2. CNN-LSTM Network The utilized deep neural network is depicted in Figure 2. As can be seen in this figure, two types of neural networks, namely, convolutional neural network and recurrent one have been used in the proposed architecture. In the following, the details of the proposed architecture are described.

2. 2. 1. Convolutional Neural Network One of the popular neural networks in image processing applications is CNN. Local connectivity and weight sharing in this model reduce the number of parameters to be learned. As described above, for each segment a 3D scalogram has been obtained. This image is a time-scale representation of the audio signal. Here, CNN can be learned to extract salient features from these images. The extracted features (the output of the CNN) can be viewed as a sequence of vectors, so it has been used as an input to the recurrent network. By considering the input of the CNN as a $n_f \times n_s \times n_b$ scalogram, the size of output will be $n_{f'} \times n_{s'} \times n_d$ where the values of $n_{f'}$, $n_{s'}$, and n_d depend on the network parameters such as filter sizes, strides, etc. Here, we can reshape the output of CNN to a $n_{f'} \times (n_{s'} \times n_d)$ datum and feed it into an LSTM network.

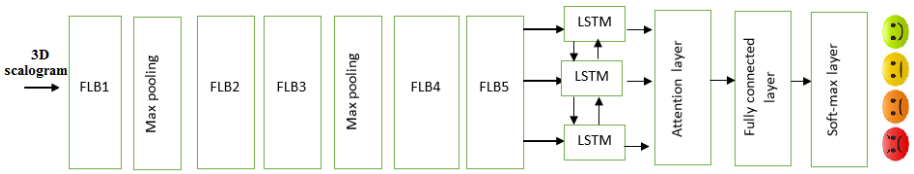

Figure 2. Proposed CNN LSTM network. Feature Learning Block (FLB) which is composed of CNN, batch normalization, and leaky Relu activation layers

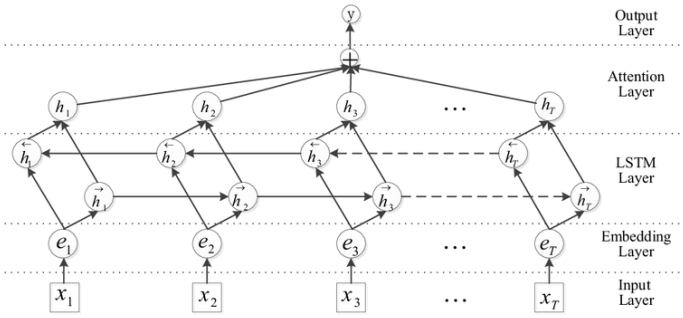

Figure 3. The block diagram of the LSTM layer with an attention layer [35].

2. 2. 2. Bidirectional Recurrent Network In emotion recognition, it seems that the past and the future audio information can be relevant. So sequence features obtained by the CNN part are fed into a bidirectional LSTM network. The block diagram of the utilized LSTM layer with an attention layer is depicted in Figure 3. As can be seen, there are two sub-networks for left and right sequence contex. The output of the ith vector is considered as $h_t = \left[\vec{h}_t, \overleftarrow{h}_t\right]$. Here x_i is the local extracted feature with size of $n_{s'} \times n_d$. We consider 128 cells in each direction. So the output of this network has 256-dimensional high-level features. In order to preserve and retire the most discriminative features for the final decision, an attention layer has been added to the model.

2. 2. 3. Attention Layer As can be seen in Figure 3, attention layer has been added after the LSTM layer to score the importance of the sequence of high-level features to the final decision [36]. By considering $h_t = \left[\vec{h}_t, \overleftarrow{h}_t\right]$ as the LSTM output at time step *t*, the normalized importance weight α_t is obtained as Equation (3).

$$\alpha_t = \frac{\exp(W.h_t)}{\sum_{r=1}^{T} \exp(W.h_t)} \quad (3)$$

The output of this layer is computed by performing a weighted sum on h_t according to the obtained weights α_t. The output of the attention layer is fed into a fully connected layer. It helps the softmax classifier to better map the audio signal into emotional categories.

3. EXPERIMENTAL RESULTS

Here, the proposed method is evaluated on three widely used emotional speech databases with different languages: the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), the Surrey Audio-Visual Expressed Emotion (SAVEE), and the Berlin Emotional Database (Emo-DB). In the following, a brief explanation of each dataset is provided first, and then experimental results are reported.

3. 1. Datasets **SAVEE** consists of 480 emotional speech utterances from four male speakers conveying seven emotions (anger, happiness, disgust, sadness, surprise, neutral, and fear) in English [37]. The average file length is 4 seconds. The sampling rate is 44.1 kHz.

Emo-DB contains 537 emotional speech utterances from 10 professional German actors (5 males and 5 females), with 7 emotions (anger, joy, sadness, neutral, boredom, disgust, and fear). The actors were asked to express 10 sentences with these 7 emotions. The audio files are 3 seconds long on average and the sampling rate is 16 kHz [38].

RAVDESS contains 24 American English actors (12 males and 12 females) speaking and singing with 8 different emotions (neutral, calm, happy, sad, angry, fear, surprise, and disgust) [39]. All emotion expressions except neutral are performed at two levels of intensity, namely normal and strong. The whole data is available in audio-visual, video-only, and audio-only (wave) formats. Here, only male speech signals were utilized for

evaluation. The audio files are 3 seconds long on average and the sampling rate is 16 kHz.

All experiments are performed on a computer system configured with Intel(R) Core (TM) i7-7700HQ CPU @ 2.80GHz, 16 GB RAM and a GPU Nvidia Geforce GTX1070 with 8 GB GDDR5. It is observed that by setting batch-size=40, it took approximately one hour to train the model.

4. SIMULATION RESULTS

In this section, the recognition accuracy on the three mentioned datasets are reported. The type and the size of the training samples should be considered in choosing the depth of the network. The experiments reveal that for a large database, 6 layers of FLB are sufficient, and for a smaller database, such as RAVDESS, taking more than 3 layers along with increasing complexity may lead to overfitting.

Here, for all databases, four emotional categories including happy, sad, neutral, and angry have been considered in the experiments. Each speech signal has been split into equal-length of 2.5-second segments, with a 1-second overlap. For the files that are less than 2.5 seconds in duration, zero-padding has been applied. The emotion label of each segment is assigned according to the label of the whole sentence. These segments with their labels are used in training and testing processes.

To better analyze the classification distribution of each emotion, the confusion matrixes during speaker-independent SER experiments for SAVEE, Emo-DB, and RAVDESS databases are shown in TABLEs 1, 2, and 3, respectively. The rows and the columns of each matrix correspond to the actual and predicted labels, respectively. The experiments performed on SAVEE data revealed that there is much confusion between happiness and anger. The reason is that both anger and happiness have high levels of energy and arousal. It can be seen that the proposed SER method can perform a prediction with average accuracy of above 77% for each emotion. Also, for Emo-DB data, the proposed SER method has performed a prediction with average accuracy of above 92% for anger, sadness, and neutral emotions. The neutral obtained the highest accuracy and happiness obtained the lowest rate. There are 35% happiness utterances detected as anger. However, there are only 7% anger utterances detected as happiness. The reason might be that a greater percentage of the utterances in the database is related to the angry emotion.

The experiments performed on RAVDESS data revealed that many neutral utterances are confused with sadness. Low arousal value for both of these two emotions can be the reason for this phenomenon. One possible reason for the great difference between the neutral/sadness and sadness/neutral confusion terms can be the different class distribution. Moreover, it is observed that 11% of happiness utterances are predicted as sadness. This is a little weird and should be investigated in future works in detail.

For speaker-independent evaluations, the k-fold cross validation approach has been used. In such a way, in any database, one speaker is selected as the testing data and the remaining ones are considered as the training data. Because of the effect of the initial value on the result, each evaluation has been repeated 5 times with different random initialization values, and the whole average of the results was subsequently reported.

For speaker-dependent experiments (only unweighted average recalls are reported in TABLE 4) for each dataset, the whole data was shuffled and randomly split into two disjoint sets; training (80%) and test (20%).

Generally, a direct comparison between SER methods is very difficult due to the differences in experimental setup or the choice of speech data. Here, for the sake of comparison, some well-established studies using common datasets are considered. The performances of the proposed method are evaluated using Unweighted Average Recall (UAR) and reported in Table 4. The UAR is a suitable metric for evaluation when the data are

TABLE 1. The confusion matrix of speaker-independent experiments on SAVEE database

	Anger	**Sadness**	**Happiness**	**Neutral**
Anger	77.5	6.3	8.1	8.1
Sadness	6.1	79.2	6.1	8.6
Happiness	16.7	1.7	81.6	0
Neutral	0.9	1.8	0	97.3

TABLE 2. The confusion matrix of speaker-independent experiments on Emo-DB database

	Anger	**Sadness**	**Happiness**	**Neutral**
Anger	92.9	0	7.1	0
Sadness	1.4	94.4	0	4.2
Happiness	35.5	0	61.3	3.2
Neutral	0	3.2	0	96.8

TABLE 3. The confusion matrix of speaker-independent experiments on RAVDESS database

	Anger	**Sadness**	**Happiness**	**Neutral**
Anger	67.2	7.8	25.0	0
Sadness	6.3	89.1	1.6	3.1
Happiness	9.4	10.9	79.7	0
Neutral	3.1	38.1	6.3	52.5

TABLE 4. Comparisons of the proposed method (PM) with some related methods according to the used datasets

Dataset	Method	Speaker dep.	Speaker indep.
SAVEE	PM	87.1	83.9
	[16]	-	**84.9**
	[31]	-	59.4
	[12]	-	42.3
	[17]	-	47.3
	[41]	75.4	73.6
Emo-DB	PM	89.1	86.4
	[16]	-	83.4
	[31]	-	84.5
	[12]	-	76.9
	[36]	-	82.8
	[17]	-	69.3
	[26]	-	79.6
	[40]	-	80.8
	[41]	86.4	82.9
RAVDESS	PM	85.1	72.1
	[15]	-	60.1
	[27]	-	70.0
	[42]	-	64.5

imbalanced, and is obtained by averaging the accuracy of all classes. In order to compare the performances of the proposed method with some conventional ones, the results reported in some references are tabulated in this table. The comparisons revealed that the proposed method conducted on a 3D scalogram achieves satisfactory accuracy. This indicates that effective emotional information can be retained using a 3D scalogram as an input.

5. CONCLUSION

In this paper, using the scalogram patches, a new SER model composed of the CNN and attention-based BLSTM has been proposed. The scalogram of each segment has been utilized as our model input. First, local features are learned using multiple FLBs in which each FLB consists of one convolutional layer and one batch normalization followed by a Relu layer. Then, to learn the contextual dependencies between locally learned features, they are reshaped and fed into an LSTM layer. Finally, a decision is made by utilizing an attention layer, and a fully connected layer followed by a softmax layer. The proposed SER model is evaluated on three widely used databases, namely RAVDESS, SAVEE, and Emo-DB. Speaker-independent and speaker-dependent SER experiments indicate that our model achieves more accurate results in terms of UAR compared with some state-of-the-art methods. According to the experimental results, it can be concluded thata significant improvement has been achieved by using the scalogram as an input to the CNN-LSTM model. However some degree of confusion between happiness and anger, and neutral and sad still exists which will be investigated in our future works.

6. REFERENCES

1. ediou, B., Krolak-Salmon, P., Saoud, M., Henaff, M. A., Burt, M., Dalery, J. and D'Amato, T., "Facial expression and sex recognition in schizophrenia and depression," ***The Canadian Journal of Psychiatry***, Vol .50, No. 9, (2005), 525-533.
2. Teixeira, T., Wedel, M. and Pieters, R., "Emotion-induced engagement in internet video advertisements", ***Journal of Marketing Research,*** .Vol49, No. 2, (2012), 144-159.
3. Liu, Z., Wu, M., Cao, W., Chen, L., Xu, J., Zhang, R., Zhou, M. and Mao, J., "A facial expression emotion recognition based human-robot interaction system ,"***IEEE/CAA Journal of Automatica Sinica,*** Vol 4, No 4, (2017).
4. El Ayadi, M., Kamel, M.S. and Karray, F., "Survey on speech emotion recognition: Features, classification schemes, and databases",. ***Pattern Recognition,*** . Vol44, No. 3, (2011), 572-587.
5. Kwon, O.W., Chan, K., Hao, J. and Lee, T.W., "Emotion recognition by speech signals", Eighth European Conference on Speech Communication and Technology, (2003).
6. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemz adeh, A., Lee, S., Neumann, U. and Narayanan, S., "Analysis of emotion recognition using facial expressions, speech and multimodal information", the 6th International Conference on Multimodal Interfaces, (2004), 205-211.
7. Esmaileyan, Z. and Marvi, H., "A database for automatic Persian speech emotion recognition: collection, processing and evaluation", ***International Journal of Engineering-Transactions A: Basics,*** Vol. 27, No. 1, (2014), pp.79-90.
8. Lin, Y.L., and Wei, G., "Speech emotion recognition based on HMM and SVM", IEEE International Conference on Machine Learning and Cybernetics. Vol. 8, (2005).
9. Hu, Hao, Ming-Xing Xu, and Wei Wu. "GMM supervector based SVM with spectral features for speech emotion recognition", IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07. Vol. 4. IEEE,)2007).
10. Chavhan, Yashpalsing, M. L. Dhore, and Pallavi Yesaware. "Speech emotion recognition using support vector machine." ***International Journal of Computer Applications*** 1.20 (2010), 6-9.
11. El Ayadi, Moataz, Mohamed S. Kamel, and Fakhri Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases." ***Pattern Recognition,*** Vol. 44, No. 3 (2011): 572-587.
12. Haider, F., Poolak, S., Albert, P., and Luz, S., "Emotion Recognition in Low-Resource Settings: An Evaluation of Automatic Feature Selection Methods" arXiv preprint arXiv: 1908.10623 (2019).

13. Pan, Yixiong, Peipei Shen, and Liping Shen. "Speech emotion recognition using support vector machine." ***International Journal of Smart Home,*** Vol. 6, No. 2, (2012), 101-108.
14. Nwe, Tin Lay, Say Wei Foo, and Liyanage C. De Silva. "Speech emotion recognition using hidden Markov models." ***Speech communication,*** Vol. 41, No. 4, (2003), 603-623.
15. Shegokar, Pankaj, and Pradip Sircar. "Continuous wavelet transform based speech emotion recognition." 2016 10th International Conference on Signal Processing and Communication Systems (ICSPCS). IEEE**,** 2016.
16. Jin, Bicheng, and Gang Liu. "Speech Emotion Recognition Based on Hyper-Prosodic Features." 2017 International Conference on Computer Technology, Electronics and Communication (ICCTEC). IEEE, 2017.
17. Vasquez-Correa, Juan Camilo, et al. "Wavelet-based time-frequency representations for automatic recognition of emotions from speech." Speech Communication; 12. ITG Symposium. VDE, 2016.
18. Gu, S., Holly, E., Lillicrap, T. and Levine, S., "Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates", IEEE International Conference on Robotics and Automation (ICRA), **(**2017), 3389-3396.
19. Ye, H., Li, G.Y. and Juang, B.H., "Power of deep learning for channel estimation and signal detection in OFDM systems",. ***IEEE Wireless Communications Letters,*** Vol. 7, No. 1, (2017), 114-117.
20. Zhang, F., Leitner, J., Milford, M., Upcroft, B. and Corke, P., "Towards vision-based deep reinforcement learning for robotic motion control" arXiv preprint arXiv:1511.03791, (2015).
21. Liu, X., Liu, W., Mei, T. and Ma, H., "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," In European Conference on Computer Vision, **(**2016), pp. 869-884,. Springer, Cham.
22. Yu, Z. and Zhang, C., "Image based static facial expression recognition with multiple deep network learning ," International Conference on Multimodal Interaction, (2015), 435-442.
23. Hu, M., Wang, H., Wang, X., Yang, J. and Wang, R., "Video facial emotion recognition based on local enhanced motion history image and CNN-CTSLSTM networks " ***Journal of Visual Communication and Image Representation,*** 59, (2019) ,176-185.
24. Chen, L., Zhou, M., Su, W., Wu, M., She, J. and Hirota, K., "Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction ," ***Information Sciences,*** 428,)2018) ,49-61.
25. Baber, J., Bakhtyar, M., Ahmed, K.U., Noor, W., Devi, V. and Sammad, A., "Facial Expression Recognition and Analysis of Interclass False Positives Using CNN", Future of Information and Communication Conference, (2019), 46-54.
26. Stolar, Melissa N., et al. "Real time speech emotion recognition using RGB image classification and transfer learning." 2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS). IEEE, 2017.
27. Gustav Sto.Tomas, "Speech Emotion Recognition using Convolutional Neural Networks."Thesis for M.S. in Audio Communication and Technology, Technische Universitt at Berlin, 2019.
28. Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou , M.A., Schuller, B. and Zafeiriou, S., "Adieu features, end-to-end speech emotion recognition using a deep convolutional recurrent network ,"IEEE international conference on acoustics, speech and signal processing (ICASSP), (2016), 5200-5204.
29. Huang, Z., Dong, M., Mao, Q. and Zhan, Y., "Speech emotion recognition using CNN", the 22nd ACM International Conference on Multimedia, (2014) ,801-804.
30. Yenigalla, P., Kumar, A., Tripathi, S., Singh, C., Kar, S. and Vepa, J., "Speech Emotion Recognition Using Spectrogram & Phoneme Embedding",***Interspeech,*** (2018), 3688-3692.
31. Jiang, Pengxu, et al. "Parallelized Convolutional Recurrent Neural Network With Spectral Features for Speech Emotion Recognition." ***IEEE Access*** 7 (2019), 90368-90377.
32. Etienne, C., Fidanza, G., Petrovskii, A., Devillers, L. and Schmauch, B., "CNN+ LSTM architecture for speech emotion recognition with data augmentation. arXiv preprint arXiv:1802.05630, (2018).
33. Zhao, J., Mao, X. and Chen, L., "Speech emotion recognition using deep 1D & 2D CNN LSTM networks ,"***Biomedical Signal Processing and Control,*** .Vol47, (2019) ,312-323.
34. Hajarolasvadi, N. and Demirel, H., "3D CNN-Based Speech Emotion Recognition Using K-Means Clustering and Spectrograms ,"***Entropy,*** .Vol21, No. 5, (2019), 479.
35. Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. , "Attention-based bidirectional long short-term memory networks for relation classification", the 54th Annual Meeting of the Association for Computational Linguistics**,** Vol. 2, (2016), 207-212.
36. Chen, Mingyi, et al. "3-D convolutional recurrent neural networks with attention model for speech emotion recognition." ***IEEE Signal Processing Letters*** 25.10 (2018), 1440-1444.
37. Haq, Sanaul, Philip JB Jackson, and J. Edge. "Speaker-dependent audio-visual emotion recognition." AVSP. 2009.
38. Burkhardt, Felix, et al. "A database of German emotional speech." Ninth European Conference on Speech Communication and Technology. 2005.
39. Livingstone, Steven R., Katlyn Peck, and Frank A. Russo. "Ravdess: The ryerson audio-visual database of emotional speech and song." 22nd Annual Meeting of the Canadian Society for Brain, Behaviour and Cognitive Science (CSBBCS). 2012.
40. Badshah, A., Rahim, N., Ullah, N., Ahmad, J., Muhammad, K., Lee, M. Y., Kwon, S., Baik, S. W., "Deep features-based speech emotion recognition for smart affective services." ***Multimedia Tools and Applications*** 78.5 (2019), 5571-5589.
41. Mao, Q., Dong, M., Huang, Z., Zhan, Y., "Learning salient features for speech emotion recognition using convolutional neural networks." ***IEEE Transactions on Multimedia*** 16.8 (2014), 2203-2213.
42. Zeng, Y., Mao, H., Peng, D., and Yi, Z., "Spectrogram based multi-task audio classification." ***Multimedia Tools and Applications*** 78.3 (2019), 3705-3722.

Speech Emotion Recognition Using Scalogram Based Deep Structure

K. Aghajani, I. Esmaili Paeen Afrakoti

Department of Engineering and Technology, University of Mazandaran, Babolsar, Iran

P A P E R I N F O

Paper history:
Received 27 October 2019
Received in revised form 16 January 2020
Accepted 17 January 2020

Keywords:
Continuous Wavelet Transform
Emotion Recognition
Convolutional Neural Network
Recurrent Network
Long-short Term Memory

چکیده

شناسایی احساس با استفاده از صدا به‌عنوان یکی از بخش‌های مهم در کاربردهای ارتباط صوتی بین انسان و کامپیوتر است. روش‌های قدیمی در این حوزه بر پایه‌ی استخراج ویژگی‌ها و سپس آموزش طبقه‌بند مناسب استوار هستند. از طرفی بسیاری از این ویژگی‌ها تحت تأثیر عواملی مستقل از احساس مانند جنسیت، طرز صحبت و محیط پیرامونی هستند. در این پژوهش روشی برای شناسایی احساس بر پایه‌ی شبکه‌ی عصبی کانولوشنال و بازگشتی ارائه شده است. شبکه‌ی کانولوشنال جهت یادگیری ویژگی‌های برجسته‌ی محلی در سیگنال‌های صوت، تصویر و ویدئو مورد استفاده قرار می‌گیرد. از طرفی شبکه‌ی بازگشتی در مسائل پردازش داده برای یادگیری رابطه‌ی بلندمدت بین ویژگی‌های محلی استفاده می‌شود. ترکیب این دو روش مزیت‌های هر دو روش را در اختیار ما می‌گذارد. در روش پیشنهادی شبکه‌ی کانولوشنال به‌طور مستقیم بر روی اسکیلوگرام سیگنال‌های صحبت اعمال می‌شود. سپس از شبکه‌ی بازگشتی مبتنی بر مکانیسم توجه برای یادگیری روابط بین ویژگی‌های یاد گرفته شده‌ی موقتی و محلی استفاده می‌شود. نتایج اعمال روش پیشنهادی بر روی داده های مختلفی مانند RAVDESS، SAVEE و Emo-DB کارآمدی این روش را در کاربرد شناسایی احساس بر پایه‌ی صوت نشان می‌دهد.

***doi**: 10.5829/ije.2020.33.02b.13*