ATTENTION BASED 3-D CONVOLUTION NEURAL NETWORK FOR SPEECH EMOTION RECOGNITION

Meiying Chen
Goergen Institute for Data Science
University of Rochester
meiying.chen@rochester.edu

ABSTRACT

Speech Emotion Recognition is important in many fields like human-machine interaction and health care. Recently, various deep neural networks are introduced to this domain and achieved great performance. Yet SER remains a difficult task due to several obstacles, e.g. the sparsity of emotion-relevant frames in a long speech utterance, the difficulties of bettering extracted acoustic feature quality, and the long-dependencies problem in RNN models. In this paper, we proposed a 3-D Convolutional Neural Network with attention mechanism, which helps to generate higher quality spectrogram features and mitigate the sparsity issue. The experimental results on the popular dataset IEMOCAP show the proposed approach outperforms the baseline models.

1. INTRODUCTION

Emotion state is considered an essential part of human communications. Recognition of emotions by machines greatly improves the human experience in human-machine interaction situations, like in robots, virtual reality, and games. It is also useful in medical diagnosis and intelligent assistant e.g. detecting clinical depression in speech during family interactions [7]. Speech is considered to be the most convenient and natural medium for human communication, which contains the implicit semantic information and intense affecting information [16]. Generally, it is hard for speakers to hide their emotions from their voices unless they are trained and intend to do so. This makes it feasible to detect emotion from speeches. Hence, automatic speech emotion recognition(SER) has become a popular research focus in the speech signal processing filed.

SER is a challenging task for the following reasons: 1) human emotion is abstract and subjective, which makes it hard to define accurately 2) human emotion is parse in audio records, which means it can only be detected in some specific moments during a long utterance, or, not every frame in a speech recording contains emotional information 3) the effectiveness of emotion-relevant features extracted from the raw speech data greatly affects the SER performance [6] 4) labeled data, especially fine-grained labeled data is insufficient [4].

Early research on speech emotion mainly concentrates on generating and selecting acoustic features that can represent different emotions effectively [11]. The most common procedure is first extracting long-term and short-term acoustic features from the speech data at different utterance levels, then use machine learning classifier or regression models e.g. Support Vector Machine, to map those features to the emotion categories that the speech contains. In such methods, it is still unclear which acoustic features have better representations of the speech emotion, which impairs the model's ability to process speech in other languages [12].

Recently, deep neural networks(DNN) have shown powerful leaning capacity in various tasks. Some previous researchers utilize DNN to automatically extract features from speech spectrogram or log-Mel spectrogram. Later, works with Convolutional Neural Networks(CNN) architectures generating hierarchical features showed great performance on several datasets [8]. Temporal information is then considered and extracted with Recurrent Neural Networks(RNN) [3]. To alleviate the longtime dependency problem, Long short-term memory networks(LSTM) [1] and its variant Bidirectional Long short-term memory Networks(Bi-LSTM) are introduced, with the later ones take both precedents and succeeding location information into modeling [14]. To make use of not only the hierarchical features but also the temporal attributes, CRNNs are proposed, in which the speech spectrogram is processed with several convolutional layers, and the result is then fed into an LSTM-RNN to do further high-level feature extraction [18]. In some cases, CNN and RNN are paralleled [19].

Not all time-frequency units, as described above, contribute equally to the whole utterance's emotional state. Attention mechanism avoids emphasis too much on the data point being close to one another by allowing the model to automatically search for parts of the source that are relevant to predicting the target, without fixing the length of the vectors used. Therefore, some researchers introduced the Attention Mechanism to extract the elements that are critical to the emotion of the utterance and aggregate those elements arrays to form an utterance emotion vector [17]. Attention layers are usually put after the feature extraction models with a softmax layer following. The attention mechanism is integrated with prior DNN models and thus formed multiple new ones like CRNN-Attention [5], LSTM-Attetion [15], Bi-LSTM-Attention [9], etc.

While former CNN models have achieved great performance on SER, there are still difficulties remain unsolved.

One is that CNNs use personalized features as inputs, which are greatly influenced by the style of speaking, and not robust to the variation of speakers, speak contents and environments [4]. Speech emotion features that directly reflect numerical values are called personalized features [8]. We use log-Mel, deltas, delta-deltas as 3-D inputs. Inspired by MFCC with deltas and delta-deltas and Chen et al. [8], we believe that measured deltas and delta-deltas are capable of representing the changing emotions and maintaining effective emotional knowledge while decreasing the impact of non-relevant emotional factors such as speakers, contents, and environment. Besides, existing models lack an efficient way of writing information to a long-term memory component and utilize these memories in the predicting process. Memory information in LSTMs/BLSTMs exists in the form of hidden layer and neuron weights, which is not ample for predicting, as what is the most relevant information that passed between the past and future remains unknown. Inspired by Transformer archetecture [13], we propose multi-attention layers to relace LSTM after 3-D convolution, to get long-distance memories and accelerate model training procedure.

The rest of this paper is arranged as follows. In section 2, we introduce our proposed model. In section 3, we report the experimental results. Finally, in section 4, we summarize our paper and discuss future works.

2. ALGORITHM DESCRIPTION

2.1 3-D Log-Mels Features

Chen et al. [4] found that 3-D convolution outperforms 2-D or 1-D convolution with limited data. So in this paper, we use the log-Mels with deltas and delta-deltas as the CNN input, where the deltas and delta-deltas reflect the process of emotional change.

The calculation process of 3-D input is as follows:

a) For a speech signal, zeros mean and unit variance to lower speaker difference. Then split the signal into short frames with Hamming windows of 25 ms and a shift of 10-ms.

b) Using discrete Fourier transform to calculate the power spectrum for every frame, feed the result into Mel-filter bank i to produce output p_i. d denotes delta. The number of Mel-bands used is 40.

c) The three-dimensional features are calculated as formula (1),(2) and (3), respectively. (1) simply calculate the log-Mels m_i by taking the logarithm of p_i. (2) calculates the delta features m_i^d, with a popular choice N=2. And (3) calculates the delta-deltas using the results of (2). Here dd denotes delta-delta. Finally, we obtain a 3-D presentation as the CNN layer input.

$$m_i = log(p_i) \tag{1}$$

$$m_i^d = \frac{\sum_{n=1}^{n} n(m_{i+n} - m_{i-n})}{2\sum_{n=1}^{n} n^2} \tag{2}$$

$$m_i^{dd} = \frac{\sum_{n=1}^{n} n(m_{i+n}^d - m_{i-n}^d)}{2\sum_{n=1}^{n} n^2} \tag{3}$$

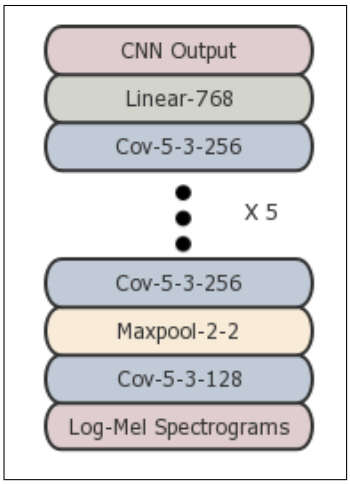

Figure 1. CNN blocks.

2.2 3-D CNN Layers

We use three-dimensional convolution layers to extract higher-level hierarchical attributes of the 3-D log-Mels inputs. The CNN model contains 7 layers, six of them are convolutional layers and the last one is a linear layer.

The first convolutional layer has 128 feature maps and the filter size is 5x3. A max-pooling layer is followed with the pooling size of 2x2. The remaining convolutional layers all have 256 feature maps with the same filter size 5x3, where 5 corresponds to the time axis and 3 corresponds to the frequency axis. The non-linear activation function of all convolution layers is leaky ReLU. After that, we perform a linear layer with outputs of 768.

2.3 Attention Layer

Instead of using LSTM layer in most previous architectures, in our proposed model, an attention layer is applied after the liner layer of the CNN model. The attention layer is worked to extract important elements in input sequences that are most emotion-relevant. Using only the attention layers rather than LSTM cells is considered to be more effective in handling long utterances [13], like those in our model. The attention layer is calculated by the following three formula:

$$e_i = u^T tanh(Wa_i + b) \tag{4}$$

$$\alpha_i = \frac{exp(\lambda e_i)}{\sum_{k=1}^{L} exp(\lambda e_i)} \tag{5}$$

$$c = \sum_{i=1}^{L} \alpha_i a_i \tag{6}$$

$A = \{a_1, a_2, ..., a_L\}$ is the output of the linear layer. In formula (4), we first calculate a new representation of a, by feed it into a multi-layer perceptron with tanh as the activation function. W is weights metrix and b is bias vector, which are both learnable. Then, the importance weight e_i is measured by the inner product between this new vector and a learnable vector u. Next, in equation (5), a softmax function is used to calculate the normalized importance

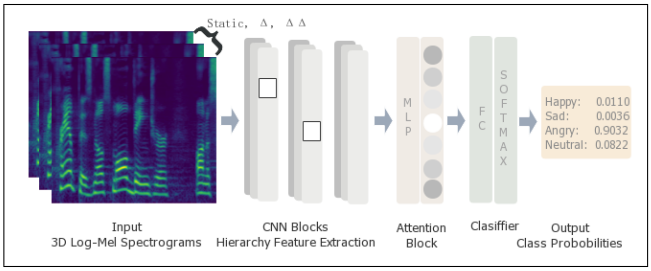

Figure 2. Overview of our proposed network architecture.

Session	Neutral	Happy	Sad	Angry	Sum
1	223	132	104	62	521
2	217	191	100	22	530
3	198	149	190	90	627
4	174	195	81	84	534
5	287	280	133	31	731
Sum	1099	947	608	289	2943

Table 1. Instance distribution for the IEMOCAP Dataset.

weight α_i. is a scale factor which controls the uniformity of the importance weights of the annotation vectors, which ranges from 0 to 1. The smaller is, the uniformer weights are. In this study, we choose = 0.3. Lastly, in (6) the utterance emotion vector c is calculated as the weighted sum of set a multiplied by important weights.

3. EXPERIMENTS

3.1 Dataset

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) [2] is one of the most widely used databases on speech emotion recognition task. The IEMOCAP contains five sessions, each of them includes audio-visual recordings of dialogues between two professional actors. So the dataset contains recordings from 10 different people. The corpus also contains two parts, improvise and script. The emotion state is divided into four categories: happy, angry, sad and neutral. Their distributions are as table 1. In this research, we only focus on improvised speech data and use all four categories. The average duration of IEMOCAP speech recordings is 4.5 s.

3.2 Baselines

We compare our approach with two baselines:

a) CNN in [10]. The method first extracts features from visual and textual modalities using deep convolutional neural networks and then feeds such features to a multiple kernel learning classifier.

b) CRNN model in [18]. This is a convolutional LSTM-RNN architecture.

c) CNN+LSTM+Attention Model with 2-D spectrogram features in [4]

Model	UAC
DNN in [10]	51.24%
CNN+LSTM in [18]	59.40%
CNN+LSTM+Attention in [4]	62.47%
3-D CNN+Attention(our proposed)	63.93%

Table 2. UAC of Comparative Models.

3.3 Experiment Result

As the dataset contains 10 speakers, we plan employ 10-fold cross-validation to assess our model. In our evaluating process, 8 of the 10 speakers are randomly chosen as training data, 1 of the rest speakers is taken as development data, and the remaining speaker is considered as the test data set. As the IEMOCAP dataset is imbalanced on different emotional classes, and to compare with the previous researches, we report the unweighted average recall(UAC) on the test set. The UAC is calculated as the mean of recall on all four classes. All parameter sets are chosen by optimizing UAC on the development set. Note that only 1-fold test is finally implemented and results reported here duo to time constrains.

4. CONCLUSION

In this paper, we proposed an attention-based CNN model on 3-D Mel-spectrograms for speech emotion recognition. We firstly form the 3-D spectrograms by calculating deltas and delta-deltas value of a 40-bands Mel-spectrogram. Then a 7-layer CNN is employed to extract hierarchical features from inputs. Later we applied an attention layer for our proposed model to concentrate on the emotional-irrelevant features and classify upon that output. This process helps to generate higher quality features and alleviate the emotion sparsity issue of SER. The experimental results show our proposed model outperforms the baseline models, and the model size is smaller than LSTM-based ones.

5. REFERENCES

[1] Shumin An, Zhenhua Ling, and Lirong Dai. Emotional statistical parametric speech synthesis using LSTM-RNNs. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1613–1616, Kuala Lumpur, December 2017. IEEE.

[2] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, December 2008.

[3] Chang-Hyun Park, Dong-Wook Lee, and Kwee-Bo Sim. Emotion recognition of speech based on RNN. In *Proceedings. International Conference on Machine*

Learning and Cybernetics, volume 4, pages 2210–2213, Beijing, China, 2002. IEEE.

[4] Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang. 3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition. *IEEE Signal Processing Letters*, 25(10):1440–1444, October 2018.

[5] Che-Wei Huang and Shrikanth Shri Narayanan. Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 583–588, Hong Kong, Hong Kong, July 2017. IEEE.

[6] Runnan Li, Zhiyong Wu, Jia Jia, Sheng Zhao, and Helen Meng. Dilated Residual Network with Multi-head Self-attention for Speech Emotion Recognition. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6675–6679, Brighton, United Kingdom, May 2019. IEEE.

[7] Lu-Shih Alex Low, M C Maddage, M Lech, L B Sheeber, and N B Allen. Detection of Clinical Depression in Adolescents' Speech During Family Interactions. *IEEE Transactions on Biomedical Engineering*, 58(3):574–586, March 2011.

[8] Qirong Mao, Ming Dong, Zhengwei Huang, and Yongzhao Zhan. Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks. *IEEE Transactions on Multimedia*, 16(8):2203–2213, December 2014.

[9] Hao Meng, Tianhao Yan, Fei Yuan, and Hongwei Wei. Speech Emotion Recognition From 3d Log-Mel Spectrograms With Deep Learning Network. *IEEE Access*, 7:125868–125881, 2019.

[10] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 439–448, Barcelona, Spain, December 2016. IEEE.

[11] Bjorn Schuller, Anton Batliner, Dino Seppi, Stefan Steidl, Thurid Vogt, Johannes Wagner, Laurence Devillers, Laurence Vidrascu, Noam Amir, Loic Kessous, and Vered Aharonson. The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals. page 4.

[12] Ting-Wei Sun and An-Yeu Andy Wu. Sparse Autoencoder with Attention Mechanism for Speech Emotion Recognition. In *2019 IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, pages 146–149, Hsinchu, Taiwan, March 2019. IEEE.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. page 11.

[14] Martin Wollmer, Angeliki Metallinou, Nassos Katsamanis, Bjorn Schuller, and Shrikanth Narayanan. Analyzing the memory of BLSTM Neural Networks for enhanced emotion classification in dyadic spoken interactions. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4157–4160, Kyoto, Japan, March 2012. IEEE.

[15] Yue Xie, Ruiyu Liang, Zhenlin Liang, Chengwei Huang, Cairong Zou, and Bjorn Schuller. Speech Emotion Classification Using Attention-Based LSTM. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(11):1675–1685, November 2019.

[16] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, and Wen Gao. Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. *IEEE Transactions on Multimedia*, 20(6):1576–1590, June 2018.

[17] Yuanyuan Zhang, Jun Du, Zirui Wang, and Jianshu Zhang. Attention Based Fully Convolutional Network for Speech Emotion Recognition. *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1771–1775, November 2018. arXiv: 1806.01506.

[18] Yue Zhao, Xingyu Jin, and Xiaolin Hu. Recurrent convolutional neural network for speech processing. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5300–5304, New Orleans, LA, March 2017. IEEE.

[19] Ziping Zhao, Zhongtian Bao, Yiqin Zhao, Zixing Zhang, Nicholas Cummins, Zhao Ren, and Bjorn Schuller. Exploring Deep Spectrum Representations via Attention-Based Recurrent and Convolutional Neural Networks for Speech Emotion Recognition. *IEEE Access*, 7:97515–97525, 2019.