Poster Abstract: Speech Emotion Recognition via Attention-based DNN from Multi-Task Learning

Fei Ma
Tsinghua-Berkeley Shenzhen Institute
Tsinghua University
mf17@mails.tsinghua.edu.cn

Weixi Gu
University of California, Berkeley
guweixigavin@gmail.com

Wei Zhang
Tsinghua-Berkeley Shenzhen Institute
Tsinghua University
wzhang17@mails.tsinghua.edu.cn

Shiguang Ni
Graduate School at Shenzhen
Tsinghua University
ni.shiguang@sz.tsinghua.edu.cn

Shao-Lun Huang
Tsinghua-Berkeley Shenzhen Institute
Tsinghua University
shaolun.huang@sz.tsinghua.edu.cn

Lin Zhang
Tsinghua-Berkeley Shenzhen Institute
Tsinghua University
linzhang@tsinghua.edu.cn

ABSTRACT

Speech unlocks the huge potentials in emotion recognition. High accurate and real-time understanding of human emotion via speech assists Human-Computer Interaction. Previous works are often limited in either coarse-grained emotion learning tasks or the low precisions on the emotion recognition. To solve these problems, we construct a real-world large-scale corpus composed of 4 common emotions (*i.e.*, anger, happiness, neutral and sadness). We also propose a multi-task attention-based DNN model (*i.e.*, MT-A-DNN) on the emotion learning. MT-A-DNN efficiently learns the high-order dependency and non-linear correlations underlying in the audio data. Extensive experiments show that MT-A-DNN outperforms conventional methods on the emotion recognition. It could take one step further on the real-time acoustic emotion recognition in many smart audio-devices.

CCS CONCEPTS

• **Human-centered computing** → *Human computer interaction (HCI)*; • **Computing methodologies** → *Machine learning*;

KEYWORDS

Speech Emotion Recognition, Multi-Task Learning

ACM Reference Format:
Fei Ma, Weixi Gu, Wei Zhang, Shiguang Ni, Shao-Lun Huang, and Lin Zhang. 2018. Poster Abstract: Speech Emotion Recognition via Attention-based DNN from Multi-Task Learning. In *The 16th ACM Conference on Embedded Networked Sensor Systems (SenSys '18), November 4–7, 2018, Shenzhen, China.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3274783.3275184

1 INTRODUCTION

Speech emotion recognition is crucial for achieving efficient human-computer interaction. Awareness of future emotional states not only allows people to have more friendly and natural communications with machines, but also can be used as adjuncts for many applications [6], for example, helps depression diagnosis and improves online learning efficiency and much more.

In the last few years, speech emotion recognition has made much progress [1]. Researchers proposed many models to predict speech emotion [6]. However, these models have the two major disadvantages: (1) The small training speech dataset made up of the actor's corpora often prevents the recognition ability of models from the fine-grained emotion recognition. (2) Plug-and-play models rarely capture the underlying multi-order dependency and sparseness evolved in the audio-data, resulting in the frustrating performance.

To cope with the above issues, we establish a large-scale real-world database from Chinese TV series and movies, which include anger, happiness, neutral and sadness, and also design a multi-task attention-based DNN model (*i.e.*, MT-A-DNN) for the fine-grained emotions recognition. The constructed database is composed of 33 Chinese TV series and 20 movies that reflect the real world, covering about 430 different speakers of different ages. The diverse scenarios of datasets provide us with rich emotion materials for model learning, guaranteeing the inference ability. Multi-task learning has been widely deployed in many scenarios [3, 4, 9], and achieves good performance. Hence we set the emotion classification as the main task and valence and activation classifications, describing the emotional states in other ways, as minor tasks. The hierarchical multi-task framework of MT-A-DNN shares the audio-data stream together, automatically studying the latent structures from distinct feature perspectives. MT-A-DNN also represents the inner temporal correlations of the audio-data with the attention mechanism [6], efficiently describes the underlying historical influence among the audio streams.

The extensive evaluation of 14364 audio-clips across 4 emotions shows that MT-A-DNN is able to recognize the speech emotion with 60.02% accuracy in general, superior to the state-of-arts. The outstanding performance and the easy implementation of MT-A-DNN could be widespread embedded in many kinds of audio-analytics platforms and Internet-of-Things devices.

2 MODEL DESIGN

In this paper, we propose MT-A-DNN which recognizes speech emotion effectively. Figure 1 depicts the architecture. It consists of

SenSys '18, November 4–7, 2018, Shenzhen, China

ACM ISBN 978-1-4503-5952-8/18/11…$15.00
https://doi.org/10.1145/3274783.3275184

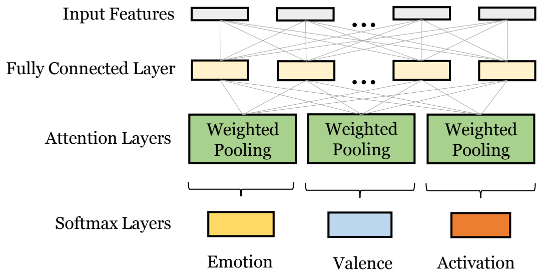

Figure 1: MT-A-DNN Architecture

three components: A DNN with one shared fully connected layer and three individual weighted pooling with attention layers and the corresponding softmax layers.

Input Features: Since all the audios come from different video materials, they are downsampled to 16KHz for consistency. We use the pyaudioanalysis toolkit [2] to extract the 34-dimensional acoustic features with a 25ms sliding window whose step size is 10ms for each audio clip, following [8], and the mean and variance for each dimensional feature are obtained as the model input, which is 68-dimensional.

Model Architecture: The first layer of MT-A-DNN is a fully connected layer which has 128 SELU nodes for shared features learning. We use three weighted-poolings with attention as the second individual attention layers to find the most salient features for each task after the first shared layer. The attention weights α_i of every task can be computed as $\alpha_i = \frac{exp(u \cdot y_i)}{\sum_j exp(u \cdot y_j)}$, where y_i is the i-th element of the first layer output, u is the weight to be learned for the corresponding task. Each attention layer has the same number of output parameters as the first shared layer and its i-th element, $z_i = \alpha_i \cdot y_i$. By learning u, we expect the network can simultaneously highlight pieces of salience and skip pieces of silence. The output from the attention layers are fed into three corresponding softmax layers for classifications. It is desired this framework can realize the fine-grained emotion recognition by effectively learning the non-linear correlations and high-order dependency underlying in the audio data.

Loss Function and Model Training: For multi-task learning, the four categorical emotions are mapped to binary valence and activation labels, like [7]. The final loss function is defined as the weighted sum of the individual cross entropy loss for emotion, valence and activation classification. Adam algorithm is used to optimize hyperparameters.

3 EVALUATION

Dataset: We evaluate the performance of MT-A-DNN on the dataset, resampled from our raw collected audio data through a sliding time window with a duration, which is 85% of original length, in a 0.03s time step. There are 3591 audio-clips for each emotion over about 212 different female speakers and 218 different male speakers. 80% of the dataset is randomly selected as the training set.

Performance: Table 1 shows the confusion matrix of MT-A-DNN inference performance. The results are averaged over the testing data. It indicates that "anger", "happiness", "neutral" and "sadness" can be recognized with accuracies of 52.03%, 67.73%, 60.18%, 60.25% respectively. In general, the overall accuracy is 60.02%.

Table 1: Confusion Matrix of MT-A-DNN

Intended	Predicted Emotion(%)			
Emotion	anger	happiness	neutral	sadness
anger	52.03	31.47	7.14	9.36
happiness	8.74	67.73	16.65	6.88
neutral	7.24	17.90	60.18	14.68
sadness	8.18	16.25	15.32	60.25

Comparison: Table 2 shows the average accuracy comparison results for four different methods over the corpus we collected. As can be seen, the MT-A-DNN far outperforms SVM, random forest, and the conventional DNN without attention mechanism and multi-task learning.

Table 2: Performance Comparison

Model	SVM	Random Forecast	DNN	MT-A-DNN
Accuracy(%)	55.46	52.61	57.79	60.02

4 CONCLUSIONS

Speech emotion recognition is advantageous for realizing effective human-computer interaction. In this work, we establish such a large-scale emotional speech database from Chinese TV series and movies. Besides, we design MT-A-DNN which efficiently learns the high-order dependency and non-linear correlations underlying in the audio streams. Experiments show MT-A-DNN is superior than the conventional methodologies in recognizing speech emotion on this corpus. In the future, we will adopt the transfer learning technologies [5] to extend our technology to different languages.

REFERENCES

[1] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition* 44, 3 (2011), 572–587.

[2] Theodoros Giannakopoulos. 2015. pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. *PloS one* 10, 12 (2015).

[3] Weixi Gu. 2017. PhD Forum Abstract: Non-intrusive Blood Glucose Monitor by Multi-task Deep Learning. In *Information Processing in Sensor Networks (IPSN), 2017 16th ACM/IEEE International Conference on.* IEEE, 249–250.

[4] Weixi Gu, Yuxun Zhou, Zimu Zhou, Xi Liu, Han Zou, Pei Zhang, Costas J Spanos, and Lin Zhang. 2017. Sugarmate: Non-intrusive blood glucose monitoring with smartphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 54.

[5] Weixi Gu, Zimu Zhou, Yuxun Zhou, Han Zou, Yunxin Liu, Costas J Spanos, and Lin Zhang. 2017. BikeMate: Bike Riding Behavior Monitoring with Smartphones. In *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, MobiQuitous 2017.* ACM.

[6] YAWEI MU, LUIS A HERNÁNDEZ GÓMEZ, ANTONIO CANO MONTES, CARLOS ALCARAZ MARTÍNEZ, XUETIAN WANG, and HONGMIN GAO. 2017. Speech Emotion Recognition Using Convolutional-Recurrent Neural Networks with Attention Model. *DEStech Transactions on Computer Science and Engineering* cii (2017).

[7] Bjorn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wollmer, Andre Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll. 2010. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing* 1, 2 (2010), 119–131.

[8] Fei Tao, Gang Liu, and Qingen Zhao. 2018. An Ensemble Framework of Voice-Based Emotion Recognition System for Films and TV Programs. *arXiv preprint arXiv:1803.01122* (2018).

[9] Rui Xia and Yang Liu. 2017. A multi-task learning framework for emotion recognition using 2D continuous space. *IEEE Transactions on Affective Computing* 1 (2017), 3–14.