Interspeech 2018
2-6 September 2018, Hyderabad

An End-to-End Deep Learning Framework with Speech Emotion Recognition of Atypical Individuals

Dengke Tang[2], *Junlin Zeng*[2], *Ming Li*[1]

[1]Data Science Research Center, Duke Kunshan University, Kunshan, China
[2]School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

ming.li369@dukekunshan.edu.cn

Abstract

The goal of the ongoing ComParE 2018 Atypical Affect subchallenge is to recognize the emotional states of atypical individuals. In this work, we present three modeling methods under the end-to-end learning framework, namely CNN combined with extended features, CNN+RNN and ResNet, respectively. Furthermore, we investigate multiple data augmentation, balancing and sampling methods to further enhance the system performance. The experimental results show that data balancing and augmentation increase the unweighted accuracy (UAR) by 10% absolutely. After score level fusion, our proposed system achieves 48.8% UAR on the develop dataset.

Index Terms: computational paralinguistics, atypical affect, end-to-end framework.

1. Introduction

Disability is defined as a physical condition that may affect an individuals ability to communicate, to interact with others, to learn or to function independently. According to the latest research by American Community Survey (ACS), it was estimated that, in the year of 2016, over 12.8 percent of the population in the United States reported a disability [1]. People with disabilities often have atypical behaviors, moods, feelings, and expressions. Accurately understanding their behaviors and interpreting their emotions are very crucial for the diagnosis and treatments.

In this paper, we aim to develop a speech-driven end-to-end framework to recognize and understand the emotions of people with disabilities. The task is proposed as part of the INTERSPEECH 2018 Computational Paralinguistics Challenge[2]. The data offered in this challenge is called the EmotAsS (EMOTional Sensitivity Assistance System for people with disabilities) dataset, which consists of the audio files recorded during the communication between the atypical individuals and the therapists. The labels of the data are their emotional states while communicating, including *Angry*, *Happy*, *Sad* and *Neutral*, annotated by volunteering annotators.

Many existing works have been performed on the speech emotion recognition (SER) task. Some acoustic features, such as the Geneva minimalistic acoustic feature set (GeMAPS)[3], Compare acoustic feature set[4] and Bagof-Audio-Words (BoAW) feature set[5], combined with traditional machine learning methods, have been proved to be effective for recognizing the human emotions from their speech signals. Recently, deep learning has made substantial achievements in the computational paralinguistics field. Among those general paralinguistic tasks, like speaker identification[6][7], language identification[8][9] and speech emotion recognition[10][11], deep learning methods have been widely used due to the superior performance.

However, in [2], the End-to-end Deep Learning framework, end2you network[12], did not achieve a better UAR compared to other three baseline systems on the test set. Moreover, the end2you framework clearly suffers from overfitting, its performance on the test set is 12% lower than the result on the development set. This might be because the scale of the dataset is not large enough to drive the deep learning networks to learn a good feature from the waveform directly and to make an accurate classification. In the EmotAsS database, there are only 9.2 hours of speech, collected from just 15 individuals in total. Furthermore, the labels of the data are severely unbalanced. More than 68% of the data in the training and development sets are labeled as *Neutral*. Therefore, some data augmentation and data balancing methods may need to be applied in the framework.

Accordingly, in this paper, we firstly investigate different types of features as input in the end-to-end framework, including the raw wave data, the constant Q transform (CQT) spectrogram[13] and the short-time Fourier transform (STFT) spectrogram[14]. And then we apply three different neural network setups in the end-to-end learning framework, including a traditional CNN+RNN, a ResNet and a CNN combined with extended features. The ResNet, according to the previous works, generally shows a strong performance in a variety of pattern recognition tasks[15], and may be a better solution to this task. And the CNN combined with extended features is another neural network structure we proposed, combining the convolutional neural network with extended feature set to enrich the features.

Furthermore, we propose our methods in data augmentation and data balancing, which achieve significant improvements in the end-to-end framework. Firstly, we have to augment the data as well as maintain the characteristic of emotion to enhance the performance of end-to-end framework. Then, in order to make the dataset labels balanced, we apply a sampling method to enrich the data labeled as *Angry*, *Happy* and *Sad*, which have less training samples. As a result, after performing data augmentation and balancing, our end-to-end framework works much better and can outperform the baseline on the development set.

The rest of this paper is organized as follows. In section 2, we describe the proposed methods in details. The experimental results are presented in Section 3 while conclusions and future works are provided in Section 4.

2. Methods

2.1. Data pre-processing

2.1.1. Data preparation

In order to feed the data into our neural network, we firstly need to chop the raw audio signal into multiple 100ms long segments. We also tried the 0.5s and 40ms long segments, but the results are almost same. Then we extract the spectrogram

10.21437/Interspeech.2018-2581

from audio data as the input of neural networks. Among various kinds of spectrograms, we mainly focus on the constant Q transform (CQT) spectrogram and the short-time Fourier transform (STFT) spectrogram.

2.1.2. Data Balancing

In the EmotAsS dataset, the imbalance of data distribution on different classes prevents the good performance of neural network[16]. In the development set, the number of samples in *Angry* class is only 1/56 of the ones in *Neutral* class. Furthermore, the duration of each segment is only 100ms. Thus the conventional balancing method, the time shift method[17], is not applicable in this scenario.

To address the data imbalance problem, we use the sampling method to augment the data labeled as *Angry*, *Happy* and *Sad*, which suffer from the data sparsity. To be specific, in every batch of data that we load from the disk to GPU to train the network, we use sampling without replacement to retrieve samples from the data labeled as *Neutral*, and perform sampling with replacement for other emotion states. In this way, we make sure all the data is used in training and maintain the diversity of training. Besides, we also develop some preference in the sampling, making the rate neutral: angry: sad: happy = 5:4:4:5. After each round of sampling, we can fetch a batch of data which is relevantly balanced among all the four classes, and we feed each batch of data into the end-to-end framework. In this way, we not only re-balance the data, but also augment the data set for training.

2.1.3. Data augmentation

To enrich the data and enhance the performance of our neural network model, we perform the data augmentation in the raw data set. In detail, we change the speed of the recorded audio to enrich our data set. In this way, we can create multiple data samples with slightly modified speed ratio. This may help in handling the overfitting issue.

2.2. Deep Neural Networks Structure

In our model, to predict the emotion of atypical individuals, we mainly used three end-to-end learning models, the traditional convolutional neural network (CNN) with recurrent neural network (RNN) model, the ResNet model and the new model, CNN combined with extended features.

2.2.1. CNN+RNN

The first one is the traditional CNN+RNN model, using a convolutional neural network (CNN) to extract features from the raw data and then a subsequent recurrent neural network (RNN) with Long Short-Term Memory (LSTM) to further analyze the features extracted from CNN, finally a full connected network with softmax function to classify the output from RNN. More specifically, our network structure is described in Tabel 1.

2.2.2. ResNet

The second model is the ResNet. We use the ResNet to extract the features from the spectrogram, and feed the outputs into the full-connected layers to perform classification. The detailed structure of ResNet is shown in Table 2.

Table 1: *CNN+RNN Network Structure*

Network	**Detail**
CNN	conv1: 16 5×5 kernels, 1 stride pooling: 3×3 pool, 1×2 stride conv2: 32 3×3 kernels, 1 stride pooling: 3×3 pool, 1×2 stride
Reshape	batch size×the time axis×rest
RNN	LSTM with 128 hidden units LSTM with 128 hidden units
Classification layer	softmax

Table 2: *Structure of Our ResNet Model*

block name	**configuration**
conv1 pooling	64 5×5 kernels, 1 stride 3×3 MaxPool, 2 stride
conv_block	64 5×5 kernels 64 5×5 kernels 256 5×5 kernels
identity_block×2	64 5×5 kernels 64 5×5 kernels 256 5×5 kernels
conv_block	128 5×5 kernels 128 5×5 kernels 512 5×5 kernels
identity_block×3	128 5×5 kernels 128 5×5 kernels 512 5×5 kernels
AveragePooling	3×3
Classification layer	full-connected 512 full-connected 256 softmax 4

2.2.3. CNN combined with extended features

In this model, we first use CNN to extract some features from the raw data, and then combining it with the extended acoustic feature set extracted by openSMILE[18], and at the end using the three-layer full connected network to classify the fused features. In this model, we aim to enrich the features and because in the baseline, COMPARE Acoustic Feature Set works well in the data set. The detailed structure is shown in Figure 1.

2.3. Different features with SVM classifier

According to the baseline[2], the SVM classifier with different features as input can all work better than the end-to-end framework in the test data set. Thus, in our SVM system, we select Compare acoustic feature and Bagof-Audio-Words (BoAW) features as the input of SVM, in order to enrich the methods of our system.

3. Experimental Result

In this section, we will evaluate our method using the unweighted average recall (UAR) on different modules, including

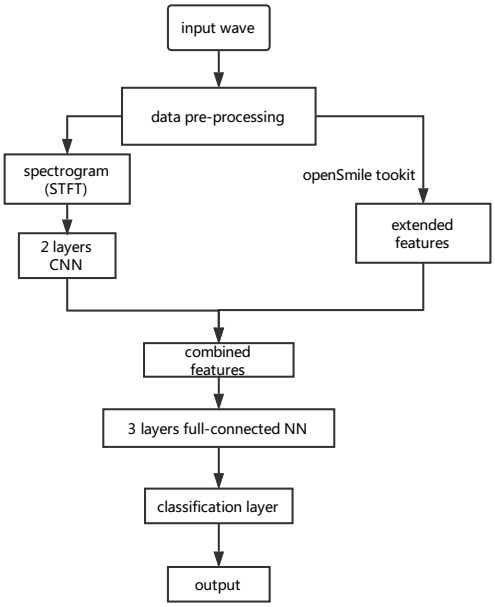

Figure 1: *Structure of CNN combined with extended features.*

the end2you and SVM module in the baseline[2].

3.1. End-to-end framework

Here we mainly discuss about the end-to-end learning framework, including the CNN+RNN model, ResNet model and the CNN combined with extended feature model. We focus on how different data preparation, augmentation and balancing influent our system.

3.1.1. Input features

We use three different features as the input of our end-to-end framework, the raw wave data, the constant Q transform (CQT) spectrogram and the short-time Fourier transform (STFT) spectrogram. The STFT spectrum achieves better performance against the CQT spectrum. Therefore, the rest of our experiments are based on the STFT spectrum. The results on the development set showed in Table 3 are base on the traditional CNN+RNN model without data augmentation or balance.

Table 3: *Performance of different input features on the development set*

Input type	UAR[%]
raw wave data	25.0
CQT spectrogram	26.18
STFT spectrogram	**27.46**

3.1.2. Data balancing

After we apply our data balancing approach in training, each model we adopt improves significantly in predicting the emotion. Especially, the CNN+RNN model achieves 45.01% UAR in the development set, 17.55% higher than the original imbalanced data. Moreover, although the CNN combined with Compare acoustic feature model achieves better performance than the CNN+RNN approach in the original imbalanced data, it cannot work as well as the CNN+RNN method in the balanced training data. The results with and without data balancing are shown in the Table 4.

Table 4: *Performance of different systems with or without data balancing on the development set*

UAR[%]	Imbalanced	Balanced
CNN+RNN	27.46	**45.12**
CNN combined with features	**29.56**	34.33
ResNet	28.94	37.78

3.1.3. Data augmentation

After we balance the data, we can find out from Table 4 that the CNN+RNN model achieves the best performance. Therefore we use the CNN+RNN system to evaluate different data augmentation methods.

As we discussed, we change the speed of speech to augment the training data, and keep the development set and test set unchanged. In our system, we perform 4 different speed ratios in modifying the speed of audio, namely 0.8, 0.9, 1.1 and 1.2. As a result, we find out that the difference of the results among different speed ratios is quite small, and slowing down the speed a little bit is relatively better than speeding up. Comparing with the data without been augmented, adding data with every ratio is useful. The detailed result is presented in Table 5.

Table 5: *Performance of different speed change ratios in data augmentation on the development set*

speed rate	UAR[%]
0.8	47.07
0.9	**47.76**
1.1	46.61
1.2	45.52
no augmentation	45.12

3.2. Fusion module

As shown in Table 5, we fused the scores of our end-to-end module with the scores of the SVM baseline together to further boost the performance.

Table 6: *Results of score level fusion on the development set*

ID	Modules	UAR[%]
1	COMPARE features + SVM	37.8
2	COMPARE BoAW + SVM	40.5
3	CNN+RNN	47.76
	Score level fusion(1+2+3)	**48.80**

Unfortunately, our fused result still suffers from overfitting. The final fusion result on the test set is 41.37%, a 7% degradation compared to the one on the development set. Regarding

to the confusion matrix, shown in Table 7, we can find out that Happy class is the most difficult one for our system on the development set while the unsatisfied performance on both Happy and Sad classes on the test set results in the performance degradation. This may be due to the overfitting in tuning the best configurations and parameters on the development set, and the mismatch between different partitions of the corpus.

Table 7: *The confusion matrix for the emotion recognition task with our final fusion model (upside) on the test set (below) on the development set*

	angry	happy	neutral	sad
angry	192	36	36	8
happy	326	92	184	48
neutral	272	101	1184	467
sad	23	4	92	34

	angry	happy	neutral	sad
angry	34	0	10	6
happy	147	101	335	382
neutral	168	207	1331	1136
sad	9	4	86	230

4. Discussion

As shown in [2], the baseline end-to-end framework suffers overfitting with limited and imbalanced data samples.

In our work, the overfitting issue has been mitigated at some certain level. Take the CNN+RNN model as an example, it achieves 45.12% on the development set, and 42.27% on the test set. Comparing with the end2you baseline, the overfitting problem is not that severe, but still challenging. However, the results from the ResNet and the CNN combined with features module, are not promising. One possible reason may be the lack of large scale training data.

As for the small scale and imbalanced training data, our methods generally achieved significant performance improvements on the development set. According to the experiments result presented above, the data balancing method increases the UAR by more than 10%. Besides, the data augmentation method we proposed also improves the performance.

To sum up, our proposed end-to-end module is 7% better than the best performance of the baseline systems on the development set. After score level fusion, the performance is further enhanced by 1% on the development set.

5. Conclusion

The computational paralinguistic task of recognizing the emotion states of atypical individuals is indeed a challenging problem. In this paper, we propose our data preprocessing methods and three neural network setups, namely the traditional CNN+RNN, the ResNet and the CNN combined with extended features. The CNN+RNN method achieves the best performance on the development set.

For example, the scale of database is not large enough to drive the neural network to learn features with good generalities. Our methods in data balancing and augmentation are effective to address this issue, but still cannot solve it. In the future, we will try some new ways in augmenting the data.

6. Acknowledgments

This research was funded in part by the National Natural Science Foundation of China (61401524,61773413), Natural Science Foundation of Guangdong Province (2014A030313123), Natural Science Foundation of Guangzhou City (201707010363), Science and Technology Development Foundation of Guangdong Province (2017B090901045), National Key Research and Development Program (2016YFC0103905).

7. References

[1] W. Erickson, C. Lee, and S. von Schrader, "Disability statistics from the american community survey (acs)." 2017. [Online]. Available: www.disabilitystatistics.org

[2] W. S. Bjrn, S. Stefan, B. Anton, B. M. Peter, B. Harald, D. Fengquan, H. Simone, P. Florian, R. Eva-Maria, D. B.-P. Katrin, E. Christa, Z. Dajie, B. Alice, A. Shahin, Q. Kun, R. Zhao, S. Maximilian, T. Panagiotis, and Z. Stefanos, "The interspeech 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats," in *Proc. of INTERSPEECH 2018, ISCA*, Hyderabad, India, 2018.

[3] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[4] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: what speech, music, and sound have in common," *Frontiers in psychology*, vol. 4, p. 292, 2013.

[5] M. Schmitt, F. Ringeval, and B. W. Schuller, "At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech." in *INTERSPEECH*, 2016, pp. 495–499.

[6] R. J. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition: A feature-based approach," *IEEE signal processing magazine*, vol. 13, no. 5, p. 58, 1996.

[7] J. S. Ren, Y. Hu, Y.-W. Tai, C. Wang, L. Xu, W. Sun, and Q. Yan, "Look, listen and learn-a multimodal lstm for speaker identification." in *AAAI*, 2016, pp. 3581–3587.

[8] J. Gonzalez-Dominguez, I. Lopez-Moreno, H. Sak, J. Gonzalez-Rodriguez, and P. J. Moreno, "Automatic language identification using long short-term memory recurrent neural networks," in *Proc. of INTERSPEECH*, 2014.

[9] S. Ganapathy, K. Han, S. Thomas, M. Omar, M. V. Segbroeck, and S. S. Narayanan, "Robust language identification using convolutional neural network features," in *Proc. INTERSPEECH*, 2014.

[10] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. of ICASSP*. IEEE, 2016, pp. 5200–5204.

[11] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using cnn," in *Proc. of ACM Multimedia*. ACM, 2014, pp. 801–804.

[12] P. Tzirakis, S. Zafeiriou, and B. W. Schuller, "End2you–the imperial toolkit for multimodal profiling by end-to-end learning," *arXiv preprint arXiv:1802.01115*, 2018.

[13] J. C. Brown, "Calculation of a constant q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[14] X. Zhu, G. T. Beauregard, and L. L. Wyse, "Real-time signal estimation from modified short-time fourier transform magnitude spectra," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1645–1653, 2007.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of CVPR*, 2016, pp. 770–778.

[16] E. DeRouin, J. Brown, H. Beck, L. Fausett, and M. Schneider, "Neural network training on unequally represented classes," *Intelligent engineering systems through artificial neural networks*, pp. 135–145, 1991.

[17] Z. Tianyan, X. Yixiang, L. Ming, and Z. Xiaobin, "An automated assessment framework for speech abnormalities related to autism spectrum disorder," *Affective Social Multimedia Computing*, 2017.

[18] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. of ACM Multimedia*. ACM, 2013, pp. 835–838.