DCRNNX: Dual-channel Recurrent Neural Network with Xgboost for Emotion Identification using Nonspeech Vocalizations

Xingwei Liang[1], You Zou[1], Tian Xie[2], and Qi Zhou[2]

[1] Konka Corporation, Shenzhen, China
[2] Harbin Instituite of Technology(Shenzhen), Shenzhen, China

Abstract. The human voice, especially nonspeech vocalizations, inherently convey emotions. However, existing efforts have ignored such emotional expressions for a long time. Based on this, we propose a Dual-channel Recurrent Neural Network with Xgboost (DCRNNX) to solve emotion recognition using nonspeech vocalizations. The DCRNNX mainly combines two Backbone models. The first model is a two-channel neural network model based on the Deep Neural Network (DNN) and Channel Recurrent Neural Network (CRNN). Channel 1 is constructed by CRNN, and the other model is constructed by Xgboost. Additionally, we employ a smoothing mechanism to integrate the outputs of the two classifiers to promote our DCRNNX. Compared with the baselines, DCRNNX combines not only multiple features but also combines multiple models, which ensures the generalization performance of DCRNNX. Experimental results show that our method achieves 45% and 42% UAR (Unweighted Average Recall), on the development dataset. After model fusion, DCRNNX achieves 46.89% UAR and 37.0% UAR on development and test datasets, respectively. The performance of our method on the development dataset is nearly 6% better than the baselines. Especially, there is a considerable gap between the performance of DCRNNX on the development and the test set. It may be the reason for the differences in emotional characteristics of the male and female voices.

Keywords: Speech Emotion Recognition, Convolutional Recurrent Neural Network, eXtreme Gradient Boosting, Model Fusion

1 Introduction

Humans express emotions through a variety of vocal channels, including verbal expressions that contain words, sentences, and non-verbal vocal expressions such as interjections. At present, many works on emotion recognition in acoustics have been devoted to the study of emotions expressed in language, while the research on emotion recognition of non-verbal expressions has yet to be developed.

However, Non-verbal vocal expressions play an important role in many applications, especially in today's increasingly common human-computer interaction systems. Correctly understanding the emotions embedded in non-verbal vocal

expressions is essential for the purposes such as intelligent healthcare and psychological counseling.

In this paper, we built a speech emotion recognition system to predict emotion contained in non-verbal speech. The task is first proposed as part of the tasks of the ACM Multimedia 2022 Computational Paralinguistics Challenge [1]. In the field of speech emotion recognition, the Channel Recurrent Neural Network (CRNN) and its variants have been shown to be effective in the field of computational paralinguistics, including keyword discovery [2], speaker recognition [3], and speech emotion recognition [4, 5]. However, most of them only focus on using a certain feature as a single-channel input instead of a multi-channel model. Therefore they lack the ability to collect speech emotion information from multiple feature sets. To solve this problem, we employ a dual-channel model. After experiments, we found that using the combination of Mel-Frequency Cepstral Coefficients (MFCCs) and Bag-of-Audio-Words (BoAWs) feature set extracted using Open-Source Crossmodal Bag-of-Words (openXBOW)[3] [6] tool as the model's input obtained the best recall rate on the development dataset. In addition to deep learning methods, we also employed the traditional machine learning methods for classification, such as the Xgboost classifier using the ComParE Acoustic Feature Set (ComParE) manual features extracted by opensmile[4] [7]. The experimental results show that while Xgboost works well in most machine learning classification tasks, but using Xgboost alone results in lower recall for most labels than the CRNN. Instead of using the traditional voting method, we smooth the prediction sequences of the two models using the L2 norm to perform a model fusion. Data augmentation methods such as changing the speed and volume of the sound are used to enhance robustness.The main contribution of the paper can be summarized as below:

- We propose a Dual-channel Recurrent Neural Network with Xgboost to solve emotion recognition using nonspeech vocalizations.
- We employ a smoothing mechanism to integrate the two-channel neural network classifier and the xgboost classifier.
- We designed a transformer convolutional recurrent network as a substitute design to the traditional CRNN structure.

The rest of this article is organized as follows. In Section 2, related work is briefly described. In Section 3, we illustrate the proposed method in detail. In Section 4, the experimental results are presented. Conclusions and future work are provided in Section 5.

2 Related Works

In the past decade, the advancement of deep learning technologies has achieved increased accuracy of artificial intelligence speech emotion recognition. In Han's

[3] https://github.com/openXBOW/openXBOW
[4] https://github.com/audeering/opensmile

work [8], the Deep Neural Networks (DNNs) are used to extract effective emotional features from short-term low-level descriptors, which in turn fed into other classifiers for emotion recognition. Latif [9] proposed to use the most primitive time-frequency band speech data for convolution to complete the emotion recognition work. Compared with the most original band data as input, the mainstream method is to use various spectrograms combined with CRNN to complete the Speech Emotion Recognition (SER). By observing the spectrogram, we can better distinguish the phoneme attributes and formant attributes to better identify the sound. CRNN was first proposed by Baoguang and Xiang [10] to solve text classification tasks. Later on, researchers applied it to the emotion recognition tasks. It achieved good experimental results. Satt [5] and Sainath [11] proposed to use the spectrogram as input to capture the speech emotion information contained in the spectrogram. They use the convolutional neural network and Long short-term memory (LSTM) to perform a discrete emotion prediction and classification task. Coincidentally, this two studies [12, 13] performed good sentiment classification on the RECOLA database using an end-to-end model combining CNN and LSTM layers. People use different representations of features in the SER tasks. For example, in [14], two convolution kernels of different sizes are used to perform spatial convolution, and temporal convolution on input MFCCs features, while [15, 16] use manual feature sets for SER tasks. In addition, studies have shown that the introduction of transformer blocks [17] can effectively improve the performance of the model for multi-channel sentiment analysis and emotion recognition [18].

Overall, Researchers have been exploring various features and models to improve performance for SER tasks. Our solution to this problem is to combine hand-crafted features, i.e., deep learning and machine learning to achieve better model performance with ensemble learning.

3 Proposed Method

In this work, we combined the dual-channel neural network model and the XGBoost classifier as a base structure to explore the possibility of discovering emotional information from non-vocal speech. To benefit from the high concentration of MFCCs and the complementarity of xbow features, we use dual channels on DNN-based structures, CNN+LSTM based-structure, and Transformer based-structure. Then the two classification prediction sequences are integrated through the L2 norm, and it turns out that after smoothing, a more reasonable and effective final prediction result can be obtained, thereby achieving a higher UAR.

3.1 Dual-channel Neural Network Model

We refer to the Luo's design [19] to establish the two Channel neural network model, and its overall framework is shown in Fig. 1. We put MFCCs and BoAWs manual features into two parallel channels. The CRNN channel uses MFCCs as

input, while the DNN channel uses BoAWs as input. The outputs of each channel are mapped to the same feature space. It is then concatenated as the input to the fully connected layer. Finally, We use the Classification Block to perform the sentiment classification.

Fig. 2 shows Channel 1's detailed design. We extract the speech signal into 40-dimensional MFCC features as the input to the first channel. We also tried 20-dimensional, 60-dimensional, and 80-dimensional MFCC features and found that feature size 40 works best. Then the convolution is performed through two convolutional layers. The size of the convolution kernel is 3×3. The number of feature maps in the first layer is 16, and the number of feature maps in the second layer is 32. After each convolutional layer is Batch Normalization and Relu activation unit, they enhance the generalization ability and expressive ability of the model. We use 2×2 max-pooling to downsample the feature map, therefore, reducing the dimension and the model size. It further reduces the risk of overfitting. The output of the previous layers is then processed by a 2-layer LSTM network. The LSTM network can solve the RNN's long-term dependency problem and has better performance in the time series prediction. Each LSTM network has 128 units. We use 0.5 as the dropout rate. The outputs of the LSTM's last time series for each layer are then concatenated. It is then fed to a fully connected layer to produce a 64-dimensional output.

The detailed design of channel 2 is shown in Fig. 3. We use the BoAWs feature set, which has 2000 feature dimensions. We also tried other feature sets offered by organizers, such as ComParE, DeepSpectrum, and auDeep. We found that none of them worked as well as BoAWs. The DNN model we use consists of three fully connected layers and produce a 64-dimensional output.

In the classification block, the outputs of the two channels are concatenated and classified into one of 6 emotion categories through a fully connected layer with softmax activation units.

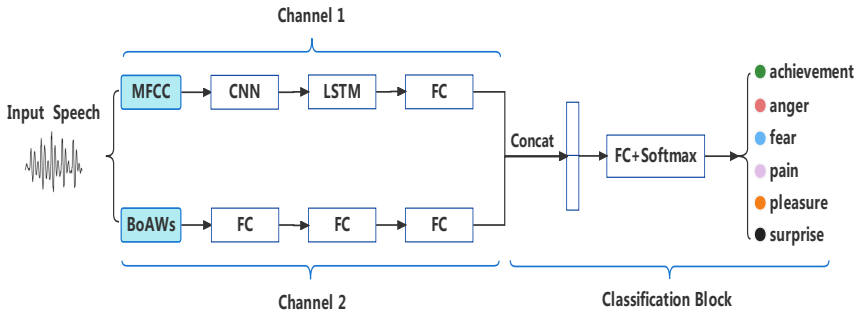

Fig. 1. Schematic diagram of proposed dual-model

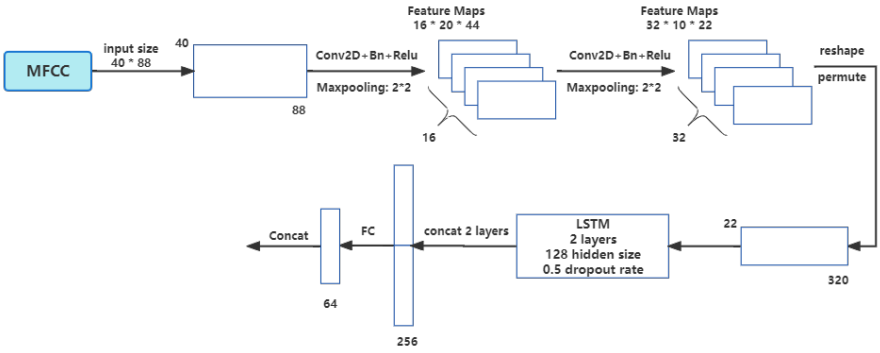

Fig. 2. Channel 1 detailed design

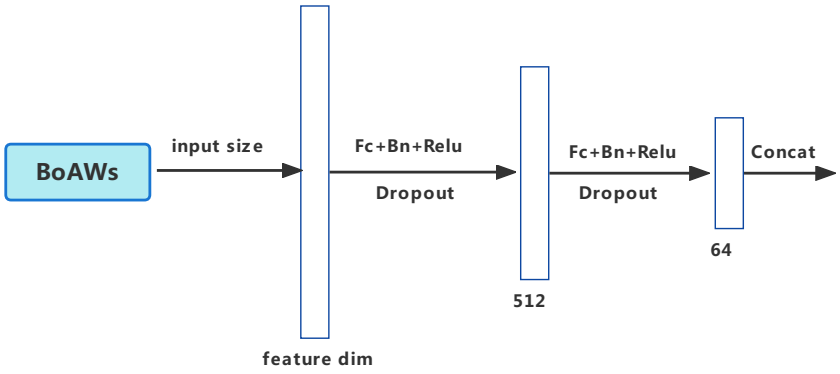

Fig. 3. Channel 2 detailed design

3.2 Introducing Attention Mechanism In Two-channel Model

Transformer were originally designed for Language Translation. Transformer enables modeling long dependencies between input sequence elements and supports parallel processing of sequence as compared to recurrent networks e.g., LSTM. As an alternative solution to the CRNN channel in our dual-channel Neural Network Model, we designed a transformer based-structure, which is a feedforward network based on multi-head attention in transformer [17]. We call this structure a transformer convolutional recurrent network (TCRN), as shown in Fig. 4.

The design of TCRN is based on an encoder-decoder structure. The audio feature MFCC is processed by preprocessing the convolution layer and then sent to four encoder layers. The structure of the encoder is the same as the encoder layer in transformer [17]. In the Multi-head Attention block, we used

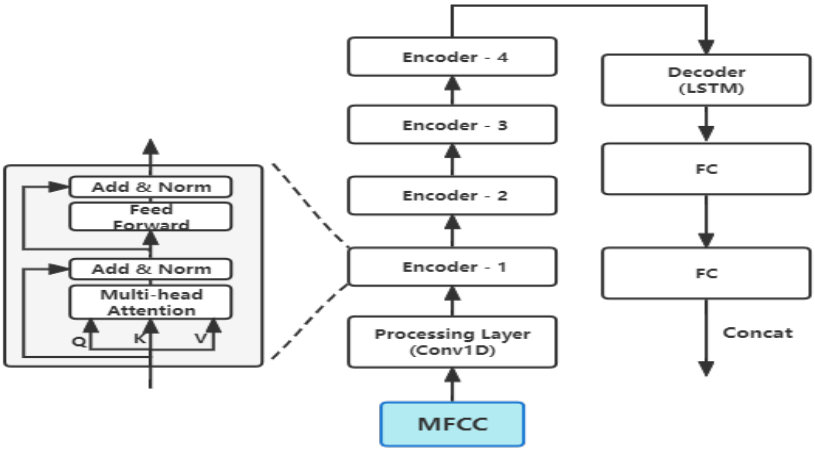

Fig. 4. Transformer convolution recurrent network

four attention heads. The multi-head attention network maps a plurality of heads in different spaces to extract cross-location information.

We use LSTM as our decoder in TCRN. LSTM is a neural network with a cyclic structure, which is composed of multiple network units (Cell). The parameters are shared among different network units. After receiving the input of the current time at time t, the network unit calculates the value of the hidden layer and then outputs the result of the current time. Unlike the network structure without timing, the hidden layer value not only depends on the input at the current moment but also is related to the hidden layer state at the previous moment. This connectivity property allows such networks to retain the memory of previous inputs, enabling them to process time series. The input gate, forget gate, and output gate is referenced in LSTM, which can effectively memorize information in a longer time dimension.

We use a two-layer LSTM for decoding, Each hidden layer has 128 units, and 0.5 is the dropout rate. Finally, input the decoding information of LSTM into two fully connected layers. It produces a 64-dimension output. Concatenated with the output from Channel 2, We have the final prediction with the Softmax function.

3.3 XGBoost Classifier

Since the dual-channel neural network model performance on some labels is not satisfying, our team considered various machine learning methods, specifically boosting methods.

XGBoost[5] is an open source software library, an implementation of the Xgboost algorithm, which provides a regularized gradient boosting framework. Through its optimization step, it is able to turn a large sum of weak learners into strong learners. Its regularization step helps prevent overfitting problems. It also comes with a customized loss function. In the training process, we use ComParE Acoustic Feature Set as our input feature set since this dataset performs the best compared with the other datasets. Furthermore, we use gridsearchCV for parameter tuning. For the parameter tuning process, we use the entire development set and the training set as input. When we got the best parameters for the model, we used the training set for training and the development set to calculate the UAR. As a result, the UAR on the development set reached 42.01%. The overall UAR does not perform better than the two-channel model. However, we found from the results that Xgboost performs much better in the recall rate in some labels, especially in the "achievement" label, where the recall rate of Xgboost is as high as 44.4%. It is the best performance of all the models we used for the achievement label. Therefore, we do not simply discard Xgboost, instead, we take its result and smooth it together with the dual-channel model's result. The specific label recall will be displayed in Fig. 5.

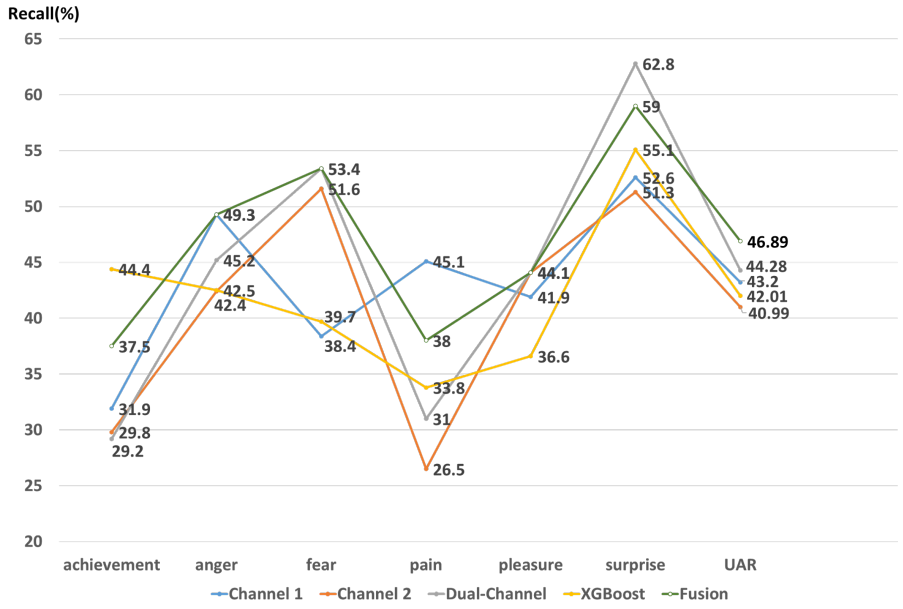

Fig. 5. Comparison of recall rate of each label in different models on the development set.

[5] https://github.com/dmlc/xgboost

3.4 Model Fusion Using L2 Norm

Assuming

$$\boldsymbol{A_j} = (a_{j0}, a_{j1}, ..., a_{j5}) \tag{1}$$

is the prediction sequence of j sample of the two-channel model and

$$\boldsymbol{B_j} = (b_{j0}, b_{j1}, ...b_{j5}) \tag{2}$$

is the prediction sequence of j sample of the Xgboost model, we combine the array of corresponding positions into a two-dimensional vector,

$$\boldsymbol{C_{ji}} = (a_{ji}, b_{ji}) \tag{3}$$

then calculate the L2 norm of this two-dimensional vector as the final predicted probability of the *i_th* label of the *j_th* sample

$$p_{ji} = \sqrt{a_{ji}^2 + b_{ji}^2}, i \in [0, 6] \tag{4}$$

4 Experiments

4.1 Datasets Used

We use Vocalisation Corpus VOC-C as our dataset, which is provided by Natalie Holz, MPI. Frankfurt is Main, features vocalizations (affect bursts) such as laughter, cries, moans, or screams, which have different affective intensities, and indicate different emotions [1]. The data set consists of 625 Train sets and 460 Development sets, of which are all female voices. The test set has 276 male voices. These data sets have 6 labels, including Achievement, Anger, Fear, Pain, Pleasure, and Surprise. The duration of each sample is about 1 second. The number of each type of label in this database is relatively balanced, so no compensation measures are needed.

4.2 Experimental Setup

In channel 1, we extract the MFCC features of the speech as the input to CRNN or TCRN model. First, we align the input speech data to a one-second length via truncation or zero-padding. We then call the librosa[6] library to generate 40×88 dimension MFCC features.

The input feature of channel 2 is the 2000 dimension BoAWs. We select DNN as the channel 2 standard model. Xgboost selects the entire data of the ComParE feature set as input.

For each training process, the two-channel model was trained for 80 epochs with a batch size 32. The cross-entropy criterion is used as the loss function. Using Adam as an optimizer, the weight decay rate is set to 2×10^{-5}. The initial learning rate is 10^{-3}, and its decay rate is 0.97. In the Xgboost experiment, we set the learning rate to 0.3, min-child-weight to 1, and max-depth to 6. We use the UAR metric to evaluate SER performance. UAR index can better deal with the uneven distribution of sample labels.

[6] https://github.com/librosa/librosa

4.3 Experimental Results

In this experiment, channel 1 and channel 2 of the dual-channel Neural network model were optimized and predicted, respectively, and then the dual-channel joint prediction classification was carried out after they were adjusted to the optimal parameter configuration. The classification effect of each model is shown in Fig. 5.

Fig. 5 shows UAR results of channel 1 and channel 2, which are 43.20% and 40.99%. The dual-channel Neural Network model has better results with UAR reaching 44.28%, indicating the model's effectiveness. XGBoost's classification effect is shown in the XGBoost branch in Fig. 5 and its overall UAR is 42.01%. After model fusion, its final UAR was 46.89%. All UAR results are shown in Table 1. The confusion matrix results of each emotion label are shown in Fig. 6.

From the above results, we can see that Channel 1 based on CRNN and Channel 2 based on full connection layer. They have already had good classification results. The convolutional layer in Channel 1 captures high-level abstraction, followed by the LSTM layer which performs the long-term temporal modelling. The fully connected layer in both channel performs discriminative representations which improve the classification results of the model. At the same time, the classificaiton block results are fused with the XGBoost classification reuslts, and they complement each other, such achieve better final classification results.

Table 1. UAR results for each channel in Devel

Channel	UAR(%)
Channel 1	43.20
Channel 2	40.99
Dual-channel Neural Network Model	44.28
XGBoost	42.01
Our result	**46.89**
Baseline	39.8

4.4 Introduce Attention Mechanism

In this experiment, we introduce the attention mechanism into the experiment of emotion classification and encode the input feature information through the stack of multi-head attention and feed-forward convolution. Experiments show that the introduction of the attention mechanism has a good improvement in emotion recognition. The model converges quickly during training, and only 9 epoch converges to the highest UAR: 0.44, as shown in Fig. 7. where The horizontal axis is the training epoch, and the vertical axis is the UAR on the validation set. It can be seen that the model reaches convergence faster and maintains a good validation set UAR.

Fig. 6. Confusion Matrix of Fusion result on the development set.

We use TCRN to replace the Channel 1 CRNN in the two-channel model for feature learning of MFCC features. The experimental results show that our emotion recognition UAR results are similar to CRNN, but the convergence speed of the model during training and its stability performance, robustness, and generalization ability is further enhanced, which may be due to the attention mechanism Dropout [20], and Batch Normalization [21] operations introduced in the network layers.

4.5 Data Augmentation

The dual-channel neural network model was used to evaluate the impacts of multiple data augmentation methods. All audio processing is done by calling the Libros library.

We changed the speed of sound and volume in the training dataset, the results are shown in Table 2. The two data enhancements did not bring an improvement in performance or even a decrease.

In the Vocalisation Corpus VOC-C, both the training and development sets have female vocalizations only, while the test set has male vocalizations only. Gender-specific differences in vocal annotation necessitate audio pitch transformation in the dataset. Male and female sound conversion is essential to ensure that the system has a better performance. Without the male-to-female voice conversion, the system only achieved a UAR of 46.89% on the training set and 30.4% on the test set.

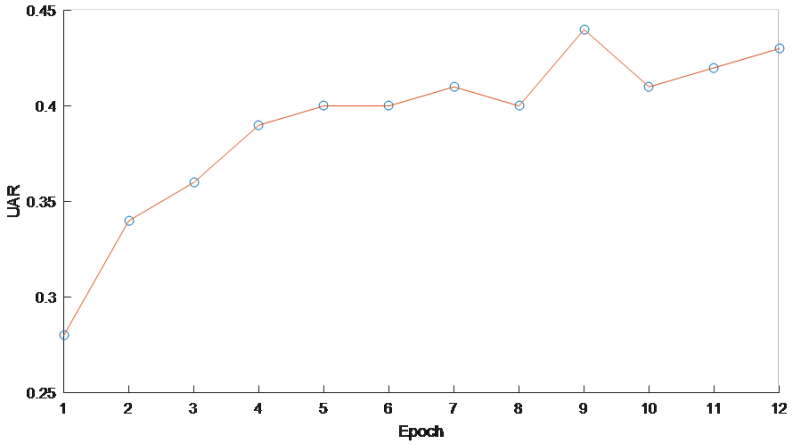

Fig. 7. The training result of TCRN

Table 2. Performance of different data augmentation methods on development set

Method	UAR
change speed	44.15%
change volume	44.02%
raw data	44.28%

To achieve gender-specific sound transitions, we use Parselmouth[7], a python library for the Praat software, to achieve pitch shift without affecting factors such as speed of sound. In the conversion process, we tried two different conversions, i.e., male to female in the test set and female to male in the training and development sets. After evaluation, we found that converting the training and development sets to male voices worked better. It produced a final result of 37.0% UAR on the test set.

Table 3. Performance of different speech gender conversion method

Method	UAR
Without Conversion	30.4%
Male to Female	36.3%
Female to Male	37.0%

[7] https://github.com/YannickJadoul/Parselmouth

5 Conclusion

In this paper, we propose a dual-channel neural network model using CRNN and DNN and achieve a UAR 4% higher than the baseline on the development set. Moreover, we replace the CRNN with the TCRN model using the attention mechanism to improve the stability and robustness of the SER system. And We use Xgboost to work on the dataset and achieved a UAR 2% higher than the baseline on the development set. Then, we design a smoothing method to integrate the two widely used models, which consequently improves the UAR by 6%. Additionally, Data augmentation strategies are also discussed. In particular, the differences between male and female voices result in a considerable gap between performance on the test dataset and the development dataset. In future work, we will try to find a way to reduce the model's sensitivity to the characteristics caused by individual differences.

References

1. Schuller, B.W., Batliner, A., Amiriparian, S., Bergler, C., Gerczuk, M., Holz, N., Larrouy-Maestri, P., Bayerl, S.P., Riedhammer, K., Mallol-Ragolta, A., Pateraki, M., Coppock, H., Kiskin, I., Sinka, M., Roberts, S.: The ACM Multimedia 2022 Computational Paralinguistics Challenge: Vocalisations, Stuttering, Activity, & Mosquitos. In: Proceedings ACM Multimedia 2022, Lisbon, Portugal, ISCA (October 2022) to appear.
2. Yan, H., He, Q., Xie, W.: Crnn-ctc based mandarin keywords spotting. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). (2020) 7489–7493
3. Meftah, A.H., Mathkour, H., Kerrache, S., Alotaibi, Y.A.: Speaker identification in different emotional states in arabic and english. IEEE Access **8** (2020) 60070–60083
4. Ma, X., Wu, Z., Jia, J., Xu, M., Meng, H., Cai, L.: Emotion recognition from variable-length speech segments using deep learning on spectrograms. (09 2018) 3683–3687
5. Satt, A., Rozenberg, S., Hoory, R.: Efficient emotion recognition from speech using deep learning on spectrograms. (08 2017) 1089–1093
6. Schmitt, M., Schuller, B.: openxbow - introducing the passau open-source crossmodal bag-of-words toolkit. Journal of Machine Learning Research **18** (10 2017) 1–5
7. Eyben, F., Wöllmer, M., Schuller, B.: opensmile – the munich versatile and fast open-source audio feature extractor. (01 2010) 1459–1462
8. Han, K., Yu, D., Tashev, I.: Speech emotion recognition using deep neural network and extreme learning machine. (09 2014)
9. Latif, S., Rana, R., Khalifa, S., Jurdak, R., Epps, J.: Direct modelling of speech emotion from raw speech (2019)
10. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(11) (2017) 2298–2304
11. Sainath, T.N., Vinyals, O., Senior, A., Sak, H.: Convolutional, long short-term memory, fully connected deep neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). (2015) 4580–4584

12. Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Zafeiriou, S.: Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In: IEEE International Conference on Acoustics. (2016)
13. Tzirakis, P., Zhang, J., Schuller, B.: End-to-end speech emotion recognition using deep neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). (2018)
14. Zhu, W., Li, X.: Speech emotion recognition with global-aware fusion on multi-scale feature representation. (2022)
15. Kim, J., Saurous, R.A.: Emotion recognition from human speech using temporal information and deep learning. In: Interspeech 2018. (2018)
16. Jian, H., Li, Y., Tao, J., Zheng, L.: Speech emotion recognition from variable-length inputs with triplet loss function. In: Interspeech 2018. (2018)
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: arXiv. (2017)
18. Zadeh A, Liang PP, P.S.V.P.C.E.M.L.: Multi-attention recurrent network for human communication comprehension. Proc Conf AAAI Artif Intell. (2018) 5642–5649
19. Luo, D., Zou, Y., Huang, D.: Investigation on joint representation learning for robust feature extraction in speech emotion recognition. In: Interspeech 2018. (2018)
20. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**(1) (2014) 1929–1958
21. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: JMLR.org. (2015)