A Multilingual Framework Based on Pre-training Model for Speech Emotion Recognition

Zhaohang Zhang*, Xiaohui Zhang†, Min Guo‡, Wei-Qiang Zhang‡[1], Ke Li§, Yukai Huang§
* Beihang University, Beijing, China
E-mail: sy1902323@buaa.edu.cn
† Beijing Jiaotong University, Beijing, China
E-mail: cecilehui@163.com
‡ Beijing National Research Center for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
E-mail: gm123@mail.tsinghua.edu.cn, wqzhang@tsinghua.edu.cn
§ Beijing Haitian Ruisheng Science Technology Ltd., Beijing 100083, China
E-mail: like@speechocean.com, huangyukai@speechocean.com

***Abstract*—Speech emotion recognition (SER) attracts much attention in recent years, especially under multilingual circumstances because of its potential in understanding human psychology and developing human-computer interaction. However, recent works in SER task mainly focus on developing fantastic structures to improve performance on monolingual datasets. Little attention is paid to promote the transfer performance on multilingual datasets. In this paper, we propose a multilingual SER framework that utilizes the pre-training model as an upstream to learn high-level speech representations and develop a hierarchical grained and feature model (HGFM) as a classifier. The proposed framework extracts speech representations based on a cross-lingual speech representations (XLSR) model and utilizes the HGFM structure to finish the classification task. We validate our framework on a multilingual dataset including IEMOCAP (English), EmoDB (German), TESS (English), SAVEE (English), EMA (English), and EMOVO (Italian). Experimental results show that features extracted by upstream model achieve an average weighted accuracy (WA) of 70.6% and unweighted accuracy (UA) of 73.4% in the downstream task, which outperforms not only manual features but other upstream structures. We also compare our results with the state-of-the-art and alternative methods to validate our framework and evaluate the performance of the structure in terms of F1-score.**

I. Introduction

Research in speech emotion recognition has attracted a lot of attention these years. Humans express their feelings through diverse modalities like images, voices, and emotions. Different attitudes correspond to various voice characteristics, so speech emotion recognition plays an important role in understanding speaker's psychology and responses. With recent advances in deep learning algorithms, acoustic signal embeddings have evolved from manual features [1] to high-level speech representations based on deep neural networks.

However, understanding emotions from multilingual speeches is still a challenge. Previous work [2] has shown a strong correlation between language and speech emotions. Success in SER is expected to mine the appropriate features that reflect the intrinsical character of emotion and are independent of languages. Besides, a downstream model is also needed to characterize the acoustic model as well as classifying high-level features.

The work to find appropriate speech representations has been investigated for a long period. Traditional approaches focused on hand-craft features [3] including Mel-frequency cepstrum coefficients, zero-crossing rate, and constant Q transform. To extract effective high-level acoustic features, [7] proposed a mixed CNN-LSTM architecture, where CNN refines local features and LSTM extracts long-term timing features. In [4], Yeh innovatively used the Automatic Speech Recognition (ASR) acoustic encoder layer as the feature extractor to finsih SER task and the result shows that the ASR encoder fine-tuned on the SER dataset can produce great gains for downstream emotion recognition tasks. Alexei introduced a new Wav2vec2 model in [5] and novelly utilized Connectionist Temporal Classification loss (CTC loss) which enabled the model to get the best performance in the ASR task. Pepino in [6] utilized the Wav2vec2 to extract acoustic features for SER and his experiments showed that fine-tuning the model with the labeled data improved the accuracy of SER by 3.4%. Conneau in [10] proposed XLSR model, which jointly learned a quantization of the latents shared across languages based on Wav2vec2. The experiment showed that the cross-lingual pre-training significantly outperformed the monolingual pre-training. Having developed the upstream feature extractor, an effective downstream classifier is needed. In [8], researchers compared the advantages and disadvantages of several traditional machine learning classifiers like SVM, KNN, and decision trees. [9] proposed the HGFM framework which modeled the frame-level and utterance-level structures of acoustic data.

Notwithstanding the progress made in previous works, multilingual SER still leaves room for improvement. For example, few works have been done to improve multilingual SER transfer performance. Most former works were evaluated on standard monolingual datasets like IEMOCAP, while they performed poor when the language of dataset was changed [2].In

[1]Corresponding author

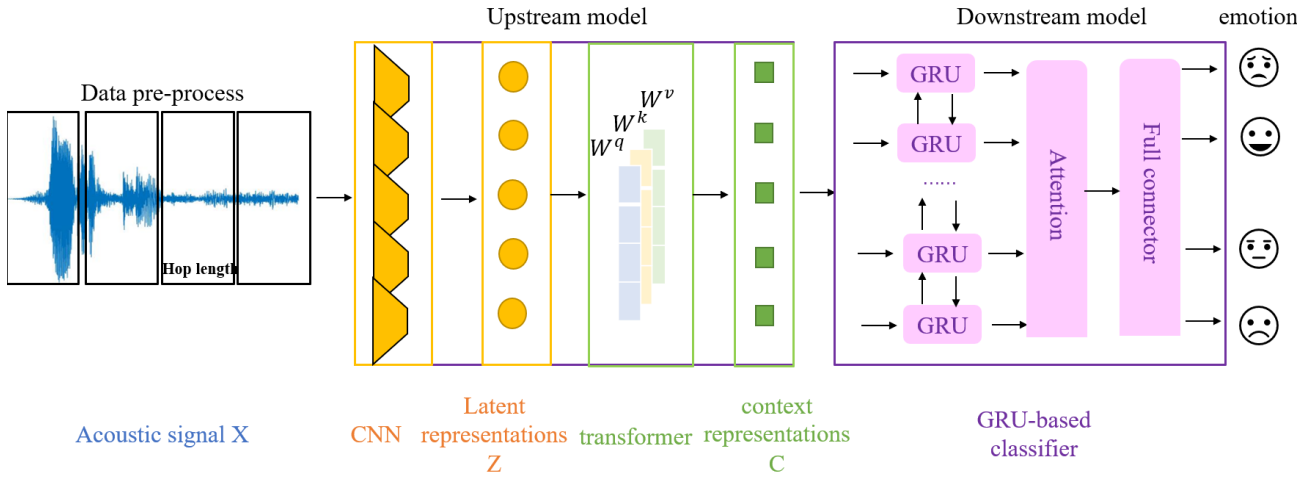


Fig. 1. Proposed framework for multilingual SER, with unsupervised pre-training model as upstream and acoustic model as downstream. Note how the caption is centered in the column.

this paper, we propose a framework aiming for extracting transferable representations of acoustic signals to enhance the multilingual SER performance. We verify the superiority of our framework's transfer performance on multilingual SER using a new dataset outside of the train set. The significant contributions of our work are as follows:

- We propose a framework including an upstream feature extractor and a downstream classifier which could improve multilingual SER performance.
- We compare several upstream models' performance and novelly utilize the cross-lingual speech representations (XLSR)[10] model as the feature extractor for multilingual SER task, which provides a 2.0% (absolute) improvement for WA compared with the state-of-the-art method. We utilize parts of HGFM[9] structure to form a classifier for high-level representations, which provides a 9.2% (absolute) improvement for WA compared with the state-of-the-art method.
- We analyze the effect of fine-tuning strategies on multilingual SER and verify the superior performance of our model on the new dataset.

The paper is structured as follows: Section 2 describes the proposed framework, Section 3 presents the experimental settings and results, and Section 4 concludes the script.

II. Proposed Framework

In this section, we present our proposed framework for multilingual SER, as shown in Fig. 1. We also discuss the advantages of our framework to improve transfer performance under multilingual circumstances. The proposed framework is shown in Fig. 1.

A. Upstream Model

As shown in Fig. 1, the preprocessed acoustic signals are fed to a convolutional feature encoder and the raw audio X are embedded to latent speech representations Z :

$$[z_1, z_2, ..., z_T]^T = \text{CNNblocks}([x_1, x_2, ..., x_T]^T) \quad (1)$$

The encoder consists of several convolutional blocks, each of which consists of three parts: one-dimension convolution layer, normalization layer, and GELU activation layer. Then the latent speech representations Z are fed to a context network that is formed by Transformer-like structures [17]. In this layer, network will output contextualized representations C, which are features for downstream input.

$$[c_1, c_2, ..., c_T]^T = \text{Transformer}([z_1, z_2, ..., z_T]^T) \quad (2)$$

In [10] a quantization module is applied to discretize Z to Q which can be regarded as targets in the self-supervised learning objective. Pre-training the upstream model can be time-consuming and cost significant computing resources thus we just download the pre-trained model and fine-tune it in the latter experiment. It has been proved in [10] that XLSR pre-trained model shares capacity across languages and particularly so with related languages. Consequently, we choose XLSR as the feature extractor of multilingual framework in hope of understanding intrinsical characters of emotion and those characters should have little correlation with languages.

B. Downstream Model

Having upstream model extracted representations of data, which can be denoted as $C^o \in \mathbb{R}^{\text{batchsize} \times 1 \times H}$, we build a downstream structure to further explore hidden features and classify different emotions. The output vector can be huge dimensions, which need a lot of computing resources for training. So the first layer of the downstream structure is a full connector to map the input features to lower dimensions, where $W_{fc} \in \mathbb{R}^{H \times L}$ denotes the paramters matrix of full

connector layer.

$$F_{\text{batchsize}}^{1\times L} = C_{\text{batchsize}}^{1\times H} \times W_{fc}^{H\times L} \tag{3}$$

Assume that the features extracted from the time series by the upstream model still have long-term dependence traits, we employ GRU [18] to encode the representations again. Where h_1 denotes the number of features in hidden state.

$$w_c^{1\times 2h_1} = \text{BiGRU}(F_{\text{batchsize}}^{1\times L}) \tag{4}$$

Then we utilize attention mechanism which could enable the downstream model to run fast, gain high parallelism, and digging long-distance dependence among extracted features [18]. We firstly divide features $w_c \in \mathbb{R}^{1\times 2h_1}$ into tow parts w_l and w_r, then we calculate the attention of two vectors separately. Here Q_i, K_i, V_i stand for the query vector, key vector, and value vector of input features. There are five forms of similarity function f :dot product, weighted dot product, concatenating, cosine similarity, and perceptron. Different from [9], the input features are upstream encoder embeddings that have higher-level representations instead of manual features. So we choose perceptron algorithm to learn the proper similarity function.

$$\begin{aligned} [w_l^{1\times h_1}, w_r^{1\times h_1}] &= \text{seperate}(w_c^{1\times 2h_1}) \\ similarity &= f(Q, K_i), i = 1, 2, ..., m \\ &= \tanh(WQ + UK_i) \end{aligned} \tag{5}$$

$$\begin{aligned} \alpha_i &= \text{softmax}(similarity) \\ attention &= \sum_{i=1}^{m} \alpha_i V_i \end{aligned}$$

Finally, we use the composite vector $v_{out} = [w_l, w_r, attention_l, attention_r, F_{batchsize}^{1\times L}]$ for classification. In the following formula, o denotes the emotion class numbers and $l_j \in \mathbb{R}^{1\times o}$ represents the probability vector of the j-th sample with respect to different emotion categories :

$$l_j = \text{softmax}(v_{\text{out},j}^{1\times(4h_1+L)} \times W_{fc}^{(4h_1+L)\times o}), j = 1, 2, ..., m \tag{6}$$

III. Experimental Setup

In this section, we present experimental setup including dataset preparation and training details.

A. Dataset

Our mixed dataset includes English database IEMOCAP, TESS, SAVEE, EMA, German database EmoDB, and Italian database EMOVO. In order to maintain data consistency, raw acoustic signals are resampled to 16kHz. In addition, the number of emotion categories contained in different datasets is different. This paper uses the intersection of emotion labels from different datasets, namely happy, sad, angry, neutral. The proportion of different emotions in our dataset is shown in the Fig. 2.

IEMOCAP [11] dataset is an acted, multimodal and multispeaker database, which is known for SER task. The corpus contains approximately 12 h of data consisting of five emotions including happiness, anger, sadness, frustration and neutral. Five female and five male actors were selected to record emotions. In our work, we only pick four of five emotions e.g. happiness, anger, sadness, and neutral in order to make consistent with other datasets. **TESS** database [12], in which a set of 200 target words were spoken in the carrier phrase "Say the word" by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral), containing 2800 stimuli in total. **SAVEE** database [13] consists of recordings from 4 male actors in 7 different emotions, 480 British English utterances in total and the data were recorded in a visual media lab with high quality audio-visual equipment, processed and labeled. **EMA** database [14] contains a total of 680 utterances spoken in four differenct target emotions spoken by three native speakers of American English: two females and one male. The two female talkers produced 10 sentences, and the male produced 14 sentences. Each sentence was repeated 5 times for each of the four different emotions. In **EmoDB** database [15] ten actors (5 female and 5 male) simulated the emotions, producing 10 German utterances (5 short and 5 longer sentences) which could be used in everyday communication. EmoDB database mainly divide actors' emotions into five categories: neutral (neutral), anger (Ärger), fear (Angst), joy (Freude), sadness (Trauer), disgust (Ekel) and boredom (Langeweile). Here we regard anger and disgust emotions as the same emotion and discard fear and boredom emotions in our mixed dataset. **EMOVO** [16] is the first database of emotional speech for the Italian language. Six actors were summoned, three males and three females with proven expertise, and have made them perform fourteen sentences (assertive, interrogative, lists) based on six basic emotional states (disgust, fear, anger, joy, surprise, sadness) plus the neutral state.

B. Training

Pre-process Because of the diversity of multilingual datasets, we deal with various datasets separately. All acoustic signals in the dataset are resampled to 16kHz and are padded to the same length. Then those data storage paths and corresponding tags are written in the JSON file in the form of a dictionary. All data is allocated to the train set and test set at

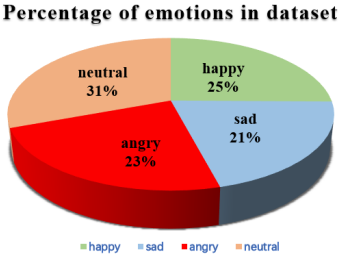


Fig. 2. The proportion of four emotions in mixed dataset

a ratio of 4:1. Then training data is divided into five sessions and the method of cross-validation is used to visualize the performance, where every four sessions are used for training and one session for evaluating.

Objective Cross-entropy loss is applied in our framework. It mainly describes the distance between the predicted output and the expected output, that is, the smaller the cross-entropy value is, the closer the two probability distributions are. Assuming that the vector $x = [x_1, x_2, x_3, x_4]^T$ is the output of our framework, the probability distribution p is the expected output, the probability distribution q is the actual output, and the cross-entropy loss can be expressed by the following formula:

$$q(x) = \log \frac{e^{x_i}}{\sum_j e^{x_j}} \quad H(p,q) = -\frac{1}{N}\sum_x p(x)q(x) \tag{7}$$

Fine-tuning We fine-tune our framework on two NVIDIA GeForce RTX 3090 GPUs with multilingual speech emotion data. AdamW optimizer is utilized to optimize our framework and the initial value of the learning rate is set to 2×10^{-4}. There exists a question that how to determine the best fine-tune strategy. The upstream model has several layers and the shallow layers tend to encode more general features compared with the deep layers. In the experiment, we compare the performance of different fine-tuning strategies.

IV. Results

As detailed in Section 2, the main novelty of our work is utilizing the unsupervised pre-training model to embed the acoustic signals and devising a framework to improve the multilingual SER performance. To evaluate the performance of our framework, the experiments are arranged as following steps. Firstly we compare the performance of various structures, and both upstream model and downstream model are considered. Then we discuss the fine-tuning strategies of the proposed structure. Finally the performance on another language dataset [22] is tested and comparision is made with the state-of-the-art method. Macro-F1 score is applied to evaluate the performance of the framework: $F_1 = \frac{1}{K}\sum_{k=1}^{K} 2 \times \frac{P_k \times R_k}{P_k + R_k} (k = 1, 2, 3, 4)$. where P_k denotes the precision of the k-th class and R_k denotes the recall of the k-th class.

For upstream encoder structures, we research six algorithms to embed acoustic signals. As shown in Table I, we first extract 13-dimension MFCC features and 80-dimension Fbank features (frame size 25ms, hop length 10ms with Hamming window) and feed these two traditional features into the downstream HGFM model separately. The result shows that these two traditional features have similar but poor performance in the multilingual SER task. Then our framework's performance is compared with other upstream structures based on deep learning algorithms. The proposed framework that uses XLSR as the upstream model to extract high-level emotion representations give us superior performance with WA of 68.8% and UA of 69.7%. Results in Table I indicate that the XLSR pre-trained model serves as the best upstream structure in our framework compared with other feature extractors. After analysis, the reason for the good performance of our model is that XLSR can extract higher-order features compared with traditional features, and absorb prior cross-lingual knowledge in the pre-training stage compared with other deep learning algorithms.

TABLE I
Comparison results on mixed-dataset using different multilingual-SER structures

Different structures	WA	UA	F1-score
Upstream structure research			
Fbank+HGFM[1]	62.3%	63.7%	62.8%
MFCC+HGFM[2]	59.7%	61.9%	60.1%
CNN-LSTM+HGFM[7]	62.7%	64.3%	63.1%
Tera+HGFM[21]	63.4%	64.9%	63.7%
Wav2vec2+HGFM[6]	66.8%	68.8%	67.1%
Downstream structure research			
XLSR+FCN[19]	59.6%	61.3%	59.9%
XLSR+CNN-Self-Attention-DNN[20]	60.4%	62.5%	60.7%
Proposed structure performance			
XLSR+HGFM(ours)	**68.8%**	**69.7%**	**69.2%**

TABLE II
Fine-tuning startegies of proposed framework

Fine-tuning strategies	WA	UA	F1-score
Upstream frozen(baseline)	68.8%	69.7%	69.2%
Upstream semi-fine-tuned	69.4%	70.9%	69.7%
Upstream completely fine-tuned	**70.6%**	**73.4%**	**70.8%**

For downstream classifiers, 3 different structures are considered to enchance the performance of SER. FCN, which performs well at semantic segmentation, is utilized to classify emotions based on features extracted from the upstream model. Then a CNN-attention network is applied to do the same work as FCN and we test its performance on SER. Results show that HGFM (a GRU-based model) performs better than these two models (CNN-based model) in SER task and shows an improvment of 4.3% absolute for WA and 4.1% for UA. That's because the acoustic signal feature has a strong time dependence and the HGFM model can make good use of this phenomenon.

The above work allows us to get the optimal architecture. It's believed that shallower layers of model tend to extract general features while deeper layers of model tend to extract task-target features, so we study the impact of fine-tuning strategies on the performance of SER tasks during the training process. In this experiment, we research three strategies to fine-tune the upstream structure: Fine-tune the whole upstream model, fine-tune the context network (as described in equation 2) only, and not fine-tune the upstream model. The results shown in Table II indicate that the completely fine-tuned upstream model gains the best performance in SER task with WA of 70.6% and UA of 73.4%.

In summary, we have determined the optimal model structure (XLSR+HGFM) and the best training method (fine-tune the whole upstream model with database). The training curve and the confusion matrix diagram are shown in Fig. 3.

Moreover, in order to test the transfer performance of our

TABLE III
THE COMPARASION OF MODELS' TRANSFER PERFORMANCE ON HAPPY, SAD, AND ANGRY EMOTIONS

SER framework	WA	UA	F1-score
Wav2vec2+HGFM(baseline)	50.2%	52.0%	48.5%
XLSR+HGFM(ours)	**70.7%**	**70.7%**	**71.0%**

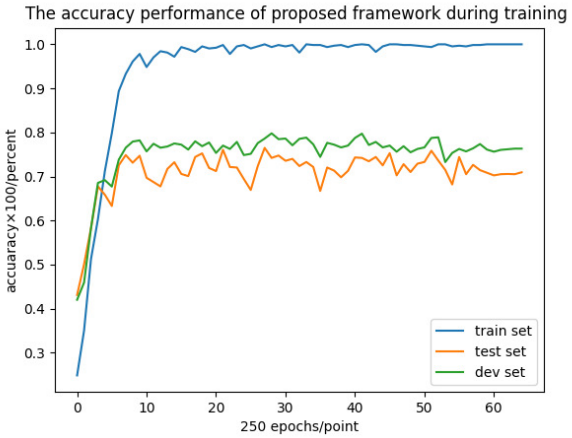
framework, we utilize another dataset [22] that is not in the hybrid database to test the model performance. There is no neutral emotion class in [22], thus we only consider the other three emotion classes to calculate WA, UA, and F1-score. Table III indicates that the proposed framework improves transfer performance of the multilingual SER task compared with the baseline.

V. CONCLUSION

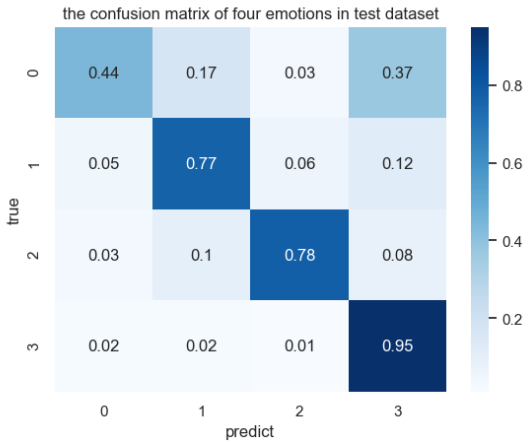
In this paper, we propose a XLSR-HGFM framework to deal with the cross-lingual SER problem. Experiments on multilingual datasets have shown the following two advantages of our framework. Firstly, the upstream model using the unsupervised pre-training model could better extract the essential feature representations of the acoustic signal. These feature representations can not only summarize the low-level features that can be manually extracted (e.g. MFCC in SER task) but also realize the high-level features that are difficult to manually extract. Secondly, the downstream model utilizes HGFM to model the acoustic signal so as to gain better performance in multilingual SER task. Our proposed framework also shows significant potential for improving the transfer performance on multilingual SER task.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant No. U1836219, and in part by the National Key R&D Program of China, and the Institute for Guo Qiang of Tsinghua University under Grant No. 2019GQG0001, and the Cross-Media Intelligent Technology Project of Beijing National Research Center for Information Science and Technology (BNRist) under Grant No. BNR2019TD01022.



(a) accuracy curve



(b) confusion matrix: 0-neutral, 1-happy, 2-angry, 3-sad

Fig. 3. The accuracy curve of training process(a) and the confusion matrix of prediction results(b)

REFERENCES

[1] A. Meftah, Y. A. Alotaibi and S. Selouani, "A Comparative Study of Different Speech Features for Arabic Phonemes Classification," 2016 European Modelling Symposium (EMS), 2016, pp. 47-52, doi: 10.1109/EMS.2016.018.

[2] A. H. Abo absa and M. Deriche, "A Two-Stage Hierarchical Multilingual Emotion Recognition System Using Hidden Markov Models and Neural Networks," 2017 9th IEEE-GCC Conference and Exhibition (GCCCE), 2017, pp. 1-6, doi: 10.1109/IEEEGCC.2017.8448155.

[3] Fayek HM, Lech M, Cavedon L. Evaluating deep learning architectures for Speech Emotion Recognition. Neural Netw. 2017 Aug;92:60-68. doi: 10.1016/j.neunet.2017.02.013. Epub 2017 Mar 21. PMID: 28396068.

[4] Yeh, S., Lin, Y., Lee, C. (2020) Speech Representation Learning for Emotion Recognition Using End-to-End ASR with Factorized Adaptation. Proc. Interspeech 2020, 536-540, DOI: 10.21437/Interspeech.2020-2524.

[5] Baevski A,Zhou H, Mohamed A. wav2vec 2.0:A framework for self-supervised learning of speech representations[J]. arXiv preprint arXiv:2006.11477, 2020.

[6] Pepino L, Riera P, Ferrer L. Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings[J]. arXiv preprint arXiv:2104.03502, 2021.

[7] Caroline Etienne, Guillaume Fidanza, Andrei Petrovskii, Laurence Devillers, Benoit Schmauch. CNN+LSTM Architecture for Speech Emotion Recognition with Data Augmentation.arXiv preprint arXiv:1802.05630,2018.

[8] M. Ghai, S. Lal, S. Duggal and S. Manik, "Emotion recognition on speech signals using machine learning," International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), no. 6, pp. 34-39, 2017.

[9] Y. Xu, H. Xu and J. Zou, "HGFM : A Hierarchical Grained and Feature Model for Acoustic Emotion Recognition," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6499-6503, doi: 10.1109/ICASSP40776.2020.9053039.

[10] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, Michael Auli. "Unsupervised Cross-lingual Representation Learning for Speech Recognition". arXiv preprint arXiv:2006.13979, 2020.

[11] Busso, C., Bulut, M., Lee, CC. et al. "IEMOCAP: interactive emotional dyadic motion capture database." Lang Resources and Evaluation 42, 335 (2008). https://doi.org/10.1007/s10579-008-9076-6

[12] Kate Dupuis, M. Kathleen Pichora-Fuller."Toronto emotional speech set (TESS) ".https://doi.org/10.5683/SP2/E8H2MF

[13] Kevin Lithgow, James Edge, Joe Kilner, Darren Cosker, Nataliya Nadtoka, Samia Smail, Idayat Salako, Affan Shaukat, Aftab Khan."Surrey audio-visual expressed emotion database". 2019. https://sail.usc.edu/iemocap/.

[14] Sungbok Lee, Serdar Yildirim, Abe Kazemzadeh, Shrikanth S. Narayana. "An articulatory study of emotional speech production," Proceedings of InterSpeech, pages 497-500, 2005

[15] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter Sendlmeier und Benjamin Weiss, "A Database of German Emotional Speech," Proceedings Interspeech 2005, Lissabon, Portugal

[16] G Costantini, I Iaderola, A Paoloni and M Todisco, "Emovo Corpus: an Italian Emotional Speech Database, " Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)

[17] Vaswani, Ashish, et al. "Attention is all you need."Advances in Neural Information Processing Systems. 2017.

[18] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio."Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation,"arXiv:1406.1078

[19] Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation. IEEE Trans Pattern Anal Mach Intell. 2017 Apr;39(4):640-651. doi: 10.1109/TPAMI.2016.2572683. Epub 2016 May 24. PMID: 27244717.

[20] C. Cai and D. Guo, "CNN-Self-Attention-DNN Architecture For Mandarin Recognition," 2020 Chinese Control And Decision Conference (CCDC), 2020, pp. 1190-1194, doi: 10.1109/CCDC49329.2020.9164333.

[21] Andy T. Liu, Shang-Wen Li, Hung-yi Lee, "TERA: Self-Supervised Learning of Transformer Encoder Representation for Speech," arXiv preprint arXiv:2007.06028.

[22] Martin, O., Kotsia, I., Macq, B., Pitas, I., 2006. The enterface'05 audio-visual emotion database. In: 22nd International Conference on Data Engineering Workshops(ICDEW'06). IEEE. 8–8.