ISSN 2616-5775 Vol. 6, Issue 13: 169-177, DOI: 10.25236/AJCIS.2023.061324

Feature Selection and Fusion in Cantonese Speech Emotion Analysis

Jianghao Luo[a], Zhenhua Tang[b,*]

School of Physics & Optoelectronic Engineering, Guangdong University of Technology, Guangzhou, China
[a]2112115018@mail2.gdut.edu.cn, [b]tangzh@gdut.edu.cn
**Corresponding author*

***Abstract*:** *This work addresses the scarcity of Cantonese speech emotion datasets by introducing a dedicated dataset and employing innovative methodologies. A tailored feature set, specifically designed for Cantonese, captures intricate emotional expressions. Enhanced efficiency in Cantonese speech emotion recognition is showcased through the utilization of a self-normalization network-based model. With an impressive accuracy of 92.3% on the Cantonese dataset, the model demonstrates robust generalization capabilities across diverse Chinese and English datasets. The obtained results underscore the potential applications of this research in various domains, including Cantonese language education, psychological counseling, and voice assistants. Understanding of Cantonese emotional expressions is advanced, contributing to the preservation of linguistic and cultural heritage. Despite the notable achievements, limitations in dataset coverage and emotion variety are acknowledged. Future endeavors will prioritize expanding the dataset's breadth and incorporating a wider range of emotional expressions. Additionally, the exploration of more comprehensive Cantonese emotion recognition will involve the investigation of multimodal approaches, where audio, visual, and textual cues are combined. These efforts are aimed at addressing current limitations and pushing the field toward a more nuanced understanding of Cantonese emotional communication.*

Keywords: *Cantonese emotion recognition, feature set, self-normalizing neural network, Multimodal*

1. Introduction

With the development of the Great Bay Area, an increasing number of migrants are flocking to Guangdong in search of opportunities, leading to the convergence of diverse regional cultures. Cantonese, as a representative of the Cantonese culture, has sparked a craze for learning. However, compared to Mandarin and English, Cantonese's unique high-spot, rhythm, and pronunciation differences can manifest in distinct emotional expressions for the same sentence, leading to potential misunderstandings. With its nine tones and six intonations, Cantonese inherently carries multiple emotions. It continues the rhythmic characteristics of ancient Chinese, and there is ongoing research on the relationship between Tang Dynasty poetry and Cantonese phonology [1]. Moreover, Cantonese songs and films have attracted a large following, contributing to the spread of Cantonese culture. Cantonese, with its rich emotions, has a large and widespread user base. Research on Cantonese emotion recognition not only contributes to the preservation and inheritance of the language but also provides more efficient services for Cantonese users.

However, based on existing research, there is a deficiency in the exploration of Cantonese speech emotion recognition due to the uniqueness of Cantonese phonology and a shortage of relevant datasets. Therefore, our research is highly practical and provides insights for researchers in Cantonese feature extraction and emotion classification. We have established a Cantonese speech emotion dataset with performances by 10 actors portraying four emotions (joy, sadness, anger, neutrality), with 200 samples for each emotion and gender. We conducted training with three native Cantonese speakers and, with their assistance, achieved annotation consistency of 0.76, making the dataset usable. The prevailing trend in speech emotion recognition involves the analysis of Mel-Frequency Cepstral Coefficients (MFCC) features. For instance, Siba P. M. et al. applied MFCC, mel-spectrogram, approximate entropy (ApEn), and permutation entropy (PrEn) for downstream tasks [2]. S Jothimani et al. also incorporated Zero Crossing Rate (ZCR) and Root Mean Square (RMS) into their network [3]. Given Cantonese's emphasis on tonal patterns and rhyming [4], we conducted manual extraction of various low-level descriptors,

ISSN 2616-5775 Vol. 6, Issue 13: 169-177, DOI: 10.25236/AJCIS.2023.061324

including MFCC, ZCR, RMS, fundamental frequency, spectral frequency, mel-frequency, spectral contrast, and spectral flatness features. To enhance the feature set, high-level statistical features (HSF) of these low-level descriptors (LLD) were computed and used as part of the overall feature set.

With the advancement of deep learning, more researchers are focusing on acoustic models. In the research performed by C. Hema et al. on speech emotion classification, CNN demonstrated stronger advantages over traditional machine learning algorithms such as support vector machines, radial basis functions, and backpropagation networks, achieving an emotion classification accuracy of 78% on their self-constructed dataset [5]. Xinlei Xu et al. designed a three channel model, integrating CNN-extracted spectrogram features, DNN-extracted low-level descriptors (LLDs) and high-level spectral features (HSFs), and LSTM-extracted MFCC features, achieving scores of 91.25% and 72.02% on EMO-DB and IEMOCAP, respectively [6]. However, these algorithm models have the following drawbacks: relatively simple structures and relatively weak generalization capabilities. Therefore, our focus will be on using convolutional neural networks to construct an algorithm model for Cantonese emotion recognition, aiming to achieve higher accuracy and better generalization capabilities.

2. Dataset Preparation and Feature Selection

To address the scarcity of the dataset, we collected a Cantonese dataset featuring performances by 10 actors expressing four emotions (joy, sadness, anger, neutrality), with 200 samples for each emotion from both male and female participants. Three native Cantonese speakers were enlisted and underwent emotion annotation training, achieving a consistency score of 0.76 with their assistance, validating the dataset.

Comparative experiments were conducted on two English datasets MELD [7], IEMOCAP [8], and a Chinese dataset, CH-SIMS [9], for detailed analysis and comparison. Our dataset is illustrated in Fig. 1.

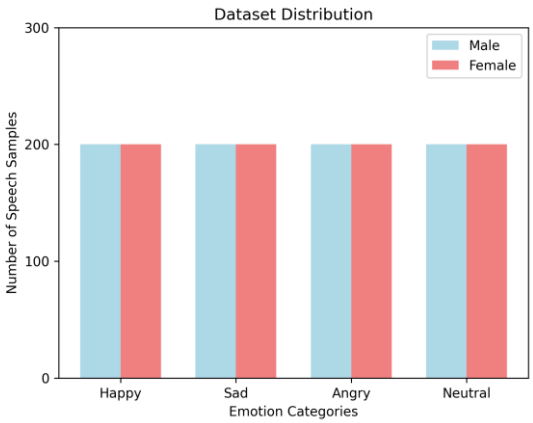

Fig. 1 Emotion categories and gender distribution in the dataset

2.1. Related Datasets

For comparing model accuracy and generalization capabilities, the following three datasets were employed in the experiments.

MELD: the Multimodal EmotionLines Dataset (MELD) introduces audio and visual modalities alongside text in its comprehensive collection. Derived from dialogues found in the Friends TV series, MELD consists of over 1400 dialogues and 13,000 utterances, with participation from multiple speakers. Each utterance is meticulously labeled with one of seven emotions: Anger, Disgust, Sadness, Joy, Neutral, Surprise, and Fear.

IEMOCAP: the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset includes performances by 10 actors in 10,039 dialogue instances, encompassing 5,255 scripted sentences and 4,784 spontaneous samples. The dataset encompasses nine emotions: angry, excited, fear, sad, surprised, frustrated, happy, disappointed, and neutral.

CH-SIMS: the Chinese Single and Multimodal Sentiment Analysis (CH-SIMS) dataset offers a comprehensive collection of 2,281 refined video segments. Sourced from movies, TV shows, and variety

ISSN 2616-5775 Vol. 6, Issue 13: 169-177, DOI: 10.25236/AJCIS.2023.061324

programs, these segments are curated with frame-level precision using Adobe Premiere Pro. CH-SIMS categorizes sentiments into five classes: weakly negative, negative, weakly positive, positive, and neutral.

2.2. Data Augmentation

As advances in emotion recognition hinge on the development of sophisticated neural networks, the scarcity of diverse datasets becomes a bottleneck, hindering the training and generalization capabilities of these models. Data augmentation emerges as a pivotal solution, addressing the constraints of limited datasets in both training and testing phases [10]. By synthetically expanding the dataset during training, augmentation ensures the neural network encounters a more comprehensive array of emotional variations. This, in turn, equips the model to better generalize during the testing phase [11].

In Fig. 2, we illustrate three key techniques employed in our data augmentation approach: (a) represents the original audio, providing a baseline for comparison. Next, (b) introduces noise to simulate real-world environmental variations. Moving on to (c), time-stretching alters the temporal dynamics, while (d) showcases pitch-shifting, mimicking diverse vocal characteristics. This not only enriches the training dataset but also ensures that the model becomes more robust in handling a variety of audio inputs during the testing phase.

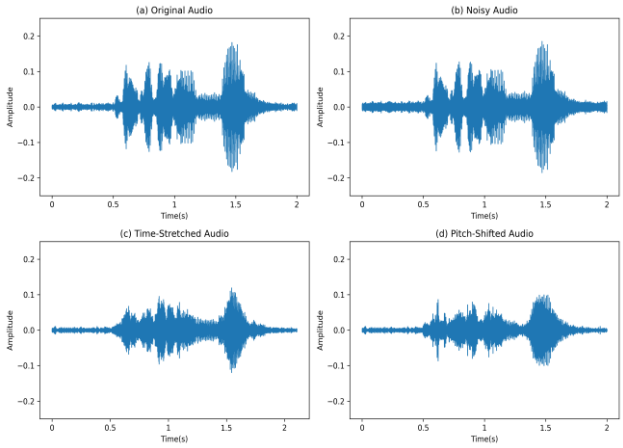

Fig. 2 Comparison of audio waveforms with three data augmentation methods (a) original audio, (b) noisy audio, (c) stretched audio, (d) pitch-shifted audio.

2.3. Feature Selection

For a more effective analysis of Cantonese emotions, we chose 8 low-level descriptors. Through statistical analysis, we derived 16 high-level descriptors. The following outlines the characteristics of the extracted features.

MFCC: the Mel-Frequency Cepstral Coefficients, are computed through a series of steps. First, the audio signal is divided into short frames, and for each frame, the power spectrum is calculated. Then, a Mel-filterbank is applied to the power spectrum, and logarithms are taken to obtain the cepstrum. Finally, discrete cosine transform (DCT) in (1) is applied to the cepstrum, forming the MFCCs, which represent the spectral characteristics of the audio signal.

$$X(n)=\sqrt{\frac{2}{N}}\sum_{m=0}^{N-1}X(m)\cos\left(\frac{(2m+1)n\pi}{2N}\right)\quad 0\le n\le N-1 \tag{1}$$

Here, $X(m)$ is the m-th sample point of the input signal, n is the index of the the DCT output coefficient, and N is the length of the input signal, indicating there are N sample points.

Fundamental Frequency (F0): F0 represents the lowest frequency component in a periodic signal, often associated with the pitch of the audio. F0 is crucial for pitch-related analysis, offering insights into the tonal characteristics of the audio signal.

Spectral Centroid (SC): SC indicates the “center of mass” of the spectrum, reflecting the average frequency of the signal, helps identify the dominant frequency and spectral balance in the audio signal.

ISSN 2616-5775 Vol. 6, Issue 13: 169-177, DOI: 10.25236/AJCIS.2023.061324

Root Mean Square (RMS): RMS measures the average energy in an audio signal. Calculated as the square root of the mean of the squared amplitude values, it provides information about the overall amplitude and energy distribution in the signal, aiding in identifying the signal's loudness.

ZCR: Zero Crossing Rate measures the rate at which a signal changes its sign, indicating the number of times it crosses zero. This feature is beneficial for identifying aspects of the audio signal related to pitch and noisiness. ZCR is defined in (2). Here, N represents the total number of samples in the audio signal, $s(n)$ is the signal amplitude at time n, and I is the indicator function.

$$ZCR = \frac{1}{N-1}\sum_{n=1}^{N-1} I\{s(n)s(n-1) < 0\} \tag{2}$$

Spectral Flatness (SF): SF quantifies the flatness or peakedness of the spectrum, providing information about the tonality of the signal, distinguishes between noise-like and tonal components, aiding in the characterization of signal coloration.

Mel Frequency: Mel Frequency represents the perceived frequency scale, emphasizing the human auditory system's sensitivity. It provides a more perceptually relevant frequency representation, supporting tasks like speech and audio recognition.

Spectral Contrast: Spectral Contrast measures the difference in amplitude between peaks and valleys across different frequency bands. It highlights spectral texture and is useful for distinguishing between harmonic and non-harmonic components.

In our analysis, we leverage low-level descriptors for statistical examination, extracting high-level statistical features such as mean, maximum, and standard deviation. This process culminates in the creation of a final feature vector. Table 1 provides pertinent details, with F denoting the number of speech frames. Throughout the experiment, we adopt a sampling rate of 22050, a frame length of 2048, a frame shift of 512, and a sampling time of 3 seconds. Equation (3) illustrates the calculation of F. Through calculation, we obtain the final feature vector dimension as $200\times F$.

$$F = \left\lfloor \frac{sr \times st}{w} \right\rfloor + 1 \tag{3}$$

Here, sr represents the sampling rate, st stands for sampling time, w is the frame shift, and $\lfloor\cdot\rfloor$ denotes the floor operation.

Table 1 Audio Feature Summary

LLDs	***HSFs***	***Shape***
MFCC	mean,max,std	(3,20,F)
F0	mean,max,std	(3,F)
SC	mean,max,std	(3,F)
RMS	mean,max,std	(3,F)
ZCR	mean	(1,F)
SF	mean	(1,F)
Mel-frequency	mean	(128,F)
Spectral Contrast	mean	(1,F)

3. Model Architecture

In addressing the challenges of Cantonese emotion classification, our research introduces a comprehensive solution designed to enhance accuracy and efficiency in emotion analysis. Our modular approach is structured with three key components: the Self-Normalization Networks (SNNs) Block, Feature Learning Layer, and Classifier, as illustrated in Fig. 3. Each module plays a crucial role in extracting relevant features, optimizing normalization, and facilitating accurate emotion classification in Cantonese language data.

ISSN 2616-5775 Vol. 6, Issue 13: 169-177, DOI: 10.25236/AJCIS.2023.061324

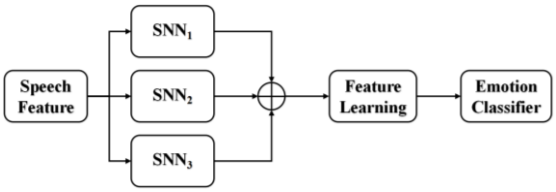

Fig. 3 The flowchart of the proposed model

3.1. Self-Normalization Networks Block

Self-Normalization Networks (SNNs), introduced by Klambauer et al., present a novel and effective architecture that addresses challenges associated with gradient vanishing and exploding in neural networks [12]. Junyi Li achieved successful utilization of Self-Normalization Networks (SNNs) in the analysis of copy number variation data, which classified four pan-cancer types. Furthermore, its versatility in extracting molecular-level features across diverse contexts is demonstrated by this approach [13]. Yao Lu et al. discovered that with ample width in neural networks, the issue of vanishing or exploding gradients is likely to vanish under mild conditions [14]. Our SNNs block consist primarily of convolutional layers, max pooling, SeLU activation functions, and group normalization. We concatenate the features extracted from three SNNs blocks for additional learning purposes. We will delve into more details below.

Convolutional Pooling Layer: three distinct SNNs utilize convolutional kernels with sizes 3, 5, and 7, each having a depth of 64 and a stride of 1 for extracting features at different granularities. The pooling layer employs max pooling with a size of 2 and a stride of 2.

SeLU: the Scaled Exponential Linear Unit is an activation function and it is known for its self-normalizing properties. It is mathematically defined in (4).

$$SeLU(x) = \lambda \begin{cases} x & x > 0 \\ \alpha(\exp(x) - 1) & x \le 0 \end{cases} \tag{4}$$

Here, λ controlling the scaling for positive values and α determining the rate of exponential decay for negative values. λ and α are approximately 1.05 and 1.67, respectively.

GN: Group Normalization is a normalization technique introduced as an alternative to Batch Normalization. It is designed to address challenges related to training deep neural networks by normalizing activations within smaller groups instead of the entire batch. GN does not impact cross-domain transfer learning and can save memory [15]. The mathematical formulation of Group Normalization is given by:

$$GN(x) = \gamma_g \times \frac{x - \mu_g}{\sigma_g} + \beta_g \tag{5}$$

Here, γ_g and β_g are learnable scale and shift parameters, μ_g and σ_g are the mean and standard deviation computed within the group.

By individually processing the input feature vector with three different SNNs, the resulting feature vectors are concatenated to obtain the output vector F_{out}. The specific calculation is shown in (6) and (7):

$$SNN_i = GN[SeLU[M_{(2,2)}(F_{in} \otimes K_{(k_i, d_i, s_i)})]] \tag{6}$$

$$F_{out} = \sum_{i=1}^{i=3} SNN_i \tag{7}$$

Here, F_{in} represents the input feature vector, $K_{(k_i, d_i, s_i)}$ is the convolutional kernel with parameters k_i (kernel size), d_i (depth), and s_i (stride), $M_{(2,2)}$ refers to max pooling with a pool size of 2 and a stride of 2. The symbol $\otimes$ denotes the convolution operation.

3.2. Feature Learning Layer

To capture more nuanced features of Cantonese, the Feature Learning Layer employs three layers of convolutional operations followed by max-pooling. The convolutional layers have kernel sizes of 3, with depths of 128, 64, and 32, respectively. The pooling layers have a size of 2 and a stride of 2. Finally, the

ISSN 2616-5775 Vol. 6, Issue 13: 169-177, DOI: 10.25236/AJCIS.2023.061324

features are processed through ReLU and Batch Normalization before being input into the classifier for classification.

ReLU: Rectified Linear Unit is an activation function commonly used in neural networks. ReLU speeds up training and effectively mitigates the problem of vanishing gradients [16], as defined in (8).

$$ReLU(x) = \begin{cases} 0 & x < 0 \\ x & x > 0 \end{cases} \tag{8}$$

Batch Normalization: Batch normalization (BN) is commonly used to standardize data at the input layer for ease of training. It stabilizes the numerical distribution of activation functions and improves model performance [17]. To alleviate overfitting, we added a dropout of 0.2 after the BN layer.

The Equation is given as follows:

$$BN(x) = \gamma_b \times \frac{x - \mu_b}{\sqrt{\sigma_b^{\,2} + \varepsilon}} + \beta_b \tag{9}$$

Here, γ_b and β_b are learnable scale and shift parameters, μ_b and σ_b are the mean and standard deviation computed within the batch, ε is a small constant.

Next, we further obtained feature F_{LL}, with higher granularity, calculated as follows:

$$u = M_{(2,2)}(F_{in} \otimes K_{(3,128,1)}) \tag{10}$$

$$v = M_{(2,2)}(u \otimes K_{(3,64,1)}) \tag{11}$$

$$F_{LL} = BN[LeLU[M_{(2,2)}(v \otimes K_{(3,32,1)})]] \tag{12}$$

Here, u and v denote intermediate layers.

3.3. Emotion Classifier

The Classifier consists of two fully connected layers. The first layer takes the flattened input feature F_{LL}. The ReLU activation and Batch Normalization follow and we applied a dropout of 0.2 to prevent overfitting. The second layer outputs the final classification result with N classes. The softmax function in (13) is applied to the output for probability normalization.

$$\text{Softmax}(z_i) = \frac{\exp(z_i)}{\sum_j \exp(z_j)} \tag{13}$$

Here, z_i represents the input to the softmax function for class i, and the final output is a probability distribution over the classes.

4. Experimental Methodology and Results

We conducted three groups of experiments. The first group served as a control experiment for self-comparison, evaluating the performance of our model on the proposed Cantonese dataset. In the second group, various feature sets were employed to compare our model with baseline models such as CNN [18], LSTM [19], and CNN-LSTM [20]. The third group of experiments involved comparing our model with several baseline models on four different datasets. The evaluation metric used in the experiments is accuracy, defined in (14). Our experimental environment consists of the Windows 10 operating system, a Tesla V100 32GB GPU, a 4-core 32GB CPU, and is implemented using PyTorch.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

Here, TP represents the number of true positives, TN represents true negatives, FP is the count of false positives, and FN is the count of false negatives.

In this experiment, we maintained a training set to test set ratio of 7:3.

ISSN 2616-5775 Vol. 6, Issue 13: 169-177, DOI: 10.25236/AJCIS.2023.061324

4.1. Control group with self-comparison

We tested the performance of our model on the proposed Cantonese dataset, which includes four emotions (happy, sad, angry, neutral). After 100 rounds of training, the average accuracy reached 92.3%. Fig. 4 shows the precision of the training and testing sets changing with epochs, and Fig. 5 displays the loss variation, which stabilizes around 0.35.

4.2. Feature Set & Baseline Model Comparison

We selected three feature sets: the first is the proposed feature set as set 1, the second consists of commonly used features, including MFCC and zero-crossing rate, as set 2 and the third includes MFCC, zero-crossing rate, mel-frequency, and RMS as set 3. The test results are shown in Table 2 indicating that our model performs best on the proposed feature set and outperforms the baseline models on the other two feature sets as well.

Table 2 Performance Comparison across Feature Sets

Model	***Set 1***	***Set 2***	***Set 3***
CNN	78.3%	77.4%	78.1%
LSTM	79.1%	74.6%	76.3%
CNN-LSTM	82.4%	81.7%	83.2%
Proposed	92.3%	88.7%	89.3%

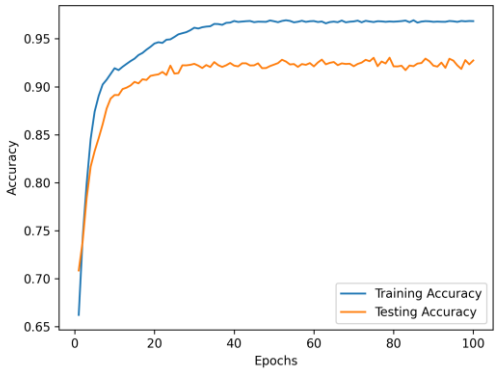

Fig. 4 Model accuracy in training vs. testing

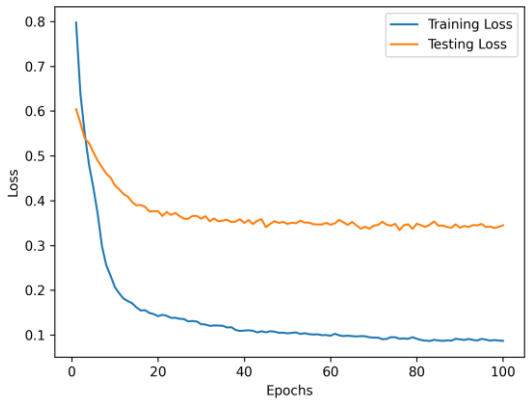

Fig. 5 Model loss in training vs. testing

4.3. Model Evaluation on Diverse Datasets

In this section, we present the performance of our proposed model compared to baseline models on four different datasets. The results in Table 3 demonstrate the model's effectiveness not only on the

ISSN 2616-5775 Vol. 6, Issue 13: 169-177, DOI: 10.25236/AJCIS.2023.061324

Cantonese dataset but also its strong generalization capability on English and Chinese datasets.

Table 3 Model Performance on Four Datasets

Model	***Proposed***	***IEMOCAP***	***MELD***	***CH-SIMS***
CNN	78.3%	69.6%	73.2%	77.3%
LSTM	79.1%	75.1%	73.8%	76.7%
CNN-LSTM	82.4%	77.3%	78.1%	79.7%
Proposed	92.3%	84.5%	83.7%	87.6%

5. Conclusion

In response to the scarcity of Cantonese speech emotion datasets, this study introduces a dedicated Cantonese dataset. Addressing the specific features of Cantonese, a tailored feature set is developed, and a model based on self-normalization networks is constructed to enhance the efficiency of Cantonese speech emotion recognition. We further improve the model's performance and optimize the training process. In the experimental phase, we achieve an impressive accuracy of 92.3% on the Cantonese dataset and demonstrate strong generalization capabilities on several Chinese and English datasets. Overall, the results are satisfactory, showcasing the potential applications of this research in various Cantonese speech emotion recognition domains, such as Cantonese language education, psychological counseling, and voice assistants. This not only facilitates Cantonese speakers but also contributes to the preservation of linguistic and cultural heritage.

However, due to limitations in experimental conditions, the collected dataset may not be comprehensive, and the range of emotions is relatively limited. Future efforts will focus on expanding the dataset in terms of breadth and emotion variety. Additionally, our research will explore the integration of multimodal approaches with Cantonese emotion recognition, representing our future research direction.

Acknowledgment

This work received financial support from the National Natural Science Foundation of China (NSFC) under grant numbers 51702055, 12172093, 62073084, 11904056, and 11704079. Additional support was provided by the Guangdong Provincial Natural Science Foundation of China (grant no. 2023A1515011599), the Guangzhou Basic and Applied Basic Research Foundation (grant no. 202102021035), the Open Foundation of Guangdong Provincial Key Laboratory of Electronic Functional Materials and Devices (grant no. EFMD2021008M), and the Special Funds for the Cultivation of Guangdong College Students' Scientific and Technological Innovation (Climbing Program Special Funds, grant no. pdjh2020a0174).

References

[1] C. Chunlan, "An Exploration of the Application of Cantonese Chanting in the Teaching of Tang Poetry," Art and Literature for the Masses, no. 20, pp. 200-202, 2023 (in Chinese), doi: 10.20112/j.cnki.ISSN1007-5828.2023.20.066.

[2] S. P. Mishra, P. Warule, and S. Deb, "Variational mode decomposition based acoustic and entropy features for speech emotion recognition," Applied Acoustics, vol. 212, p. 109578, 2023.

[3] S. Jothimani and K. Premalatha, "MFF-SAug: Multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network," Chaos, Solitons & Fractals, vol. 162, p. 112512, 2022.

[4] Y. Pan, "Integrating Cantonese nursery rhymes into early childhood music classrooms: A lesson for learning music, language, and culture," Journal of General Music Education, vol. 35, no. 1, pp. 34-45, 2021.

[5] C. Hema and F. P. G. Marquez, "Emotional speech recognition using cnn and deep learning techniques," Applied Acoustics, vol. 211, p. 109492, 2023.

[6] X. Xu, D. Li, Y. Zhou, and Z. Wang, "Multi-type features separating fusion learning for Speech Emotion Recognition," Applied Soft Computing, vol. 130, p. 109648, 2022.

[7] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," arXiv preprint arXiv:1810.02508, 2018.

[8] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," Language

ISSN 2616-5775 Vol. 6, Issue 13: 169-177, DOI: 10.25236/AJCIS.2023.061324

resources and evaluation, vol. 42, pp. 335-359, 2008.
[9] W. Yu et al., "Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 3718-3727.
[10] K. Nugroho and E. Noersasongko, "Enhanced Indonesian ethnic speaker recognition using data augmentation deep neural network," Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 7, pp. 4375-4384, 2022.
[11] L. Trinh Van, T. Dao Thi Le, T. Le Xuan, and E. Castelli, "Emotional speech recognition using deep neural networks," Sensors, vol. 22, no. 4, p. 1414, 2022.
[12] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," Advances in neural information processing systems, vol. 30, 2017.
[13] J. Li, Q. Xu, M. Wu, T. Huang, and Y. Wang, "Pan-cancer classification based on self-normalizing neural networks and feature selection," Frontiers in Bioengineering and Biotechnology, vol. 8, p. 766, 2020.
[14] Y. Lu, S. Gould, and T. Ajanthan, "Bidirectionally self-normalizing neural networks," Neural Networks, vol. 167, pp. 283-291, 2023.
[15] X. Cai, Z. Wu, K. Zhong, B. Su, D. Dai, and H. Meng, "Unsupervised cross-lingual speech emotion recognition using domain adversarial neural network," in 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2021: IEEE, pp. 1-5.
[16] A. S. Alluhaidan, O. Saidani, R. Jahangir, M. A. Nauman, and O. S. Neffati, "Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network," Applied Sciences, vol. 13, no. 8, p. 4750, 2023.
[17] K. Mountzouris, I. Perikos, and I. Hatzilygeroudis, "Speech Emotion Recognition Using Convolutional Neural Networks with Attention Mechanism," Electronics, vol. 12, no. 20, p. 4376, 2023.
[18] A. B. A. Qayyum, A. Arefeen, and C. Shahnaz, "Convolutional neural network (CNN) based speech-emotion recognition," in 2019 IEEE international conference on signal processing, information, communication & systems (SPICSCON), 2019: IEEE, pp. 122-125.
[19] A. Jadhav, V. Kadam, S. Prasad, N. Waghmare, and S. Dhule, "An Emotion Recognition from Speech using LSTM," in 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), 2023: IEEE, pp. 834-842.
[20] A. A. Abdelhamid et al., "Robust speech emotion recognition using CNN+ LSTM based on stochastic fractal search optimization algorithm," IEEE Access, vol. 10, pp. 49265-49284, 2022.