applied sciences

MDPI

*Article*

# Electroglottograph-Based Speech Emotion Recognition via Cross-Modal Distillation

**Lijiang Chen , Jie Ren, Xia Mao and Qi Zhao ***

School of Electronic and Information Engineering, Beihang University, No. 37 Xueyuan Road, Haidian District, Beijing 100191, China; chenlijiang@buaa.edu.cn (L.C.); rj980728@buaa.edu.cn (J.R.); moukyou@buaa.edu.cn (X.M.)

* Correspondence: zhaoqi@buaa.edu.cn; Tel.: +86-010-8231-6739

**Abstract:** Speech emotion recognition (SER) is an important component of emotion computation and signal processing. Recently, many works have applied abundant acoustic features and complex model architectures to enhance the model's performance, but these works sacrifice the portability of the model. To address this problem, we propose a model utilizing only the fundamental frequency from electroglottograph (EGG) signals. EGG signals are a sort of physiological signal that can directly reflect the movement of the vocal cord. Under the assumption that different acoustic features share similar representations in the internal emotional state, we propose cross-modal emotion distillation (CMED) to train the EGG-based SER model by transferring robust speech emotion representations from the log-Mel-spectrogram-based model. Utilizing the cross-modal emotion distillation, we achieve an increase of recognition accuracy from 58.98% to 66.80% on the S70 subset of the Chinese Dual-mode Emotional Speech Database (CDESD 7-classes) and 32.29% to 42.71% on the EMO-DB (7-classes) dataset, which shows that our proposed method achieves a comparable result with the human subjective experiment and realizes a trade-off between model complexity and performance.

**Keywords:** speech emotion recognition; knowledge distillation; cross-modal transfer; electroglottograph

check for updates

**Citation:** Chen, L.; Ren, J.; Mao, X.; Zhao, Q. Electroglottograph-Based Speech Emotion Recognition via Cross-Modal Distillation. *Appl. Sci.* **2022**, *12*, 4338. https://doi.org/10.3390/app12094338

Academic Editor: Douglas O'Shaughnessy

Received: 31 March 2022
Accepted: 22 April 2022
Published: 25 April 2022

## 1. Introduction

Speech is an effective medium to express emotions and attitudes through language. Applications of emotion recognition in speech can be found in many areas [1,2]. Extracting and recognizing emotional information from speech signals is an important subject to realize more natural human-computer interaction.

Speech emotion recognition with deep learning methods aims to extract deep emotion features through artificial neural networks. The majority of speech emotion recognition architectures utilize neural networks such as convolutional neural networks (CNN), recurrent neural networks (RNN), long-short term memory (LSTM), or their combinations [3–10]. In recent years, in order to obtain higher recognition accuracy, most research has adopted two strategies to enrich the emotion information that one model can obtain.

One strategy is to design and apply more complex architectures, such as deep neural networks (DNN). In 2018, Tzirakis [11] proposed an end-to-end continuous speech emotion recognition model to extract features from the raw speech signal based on DNN, and stack a 2-layer long short-term memory (LSTM) to consider the contextual information in the data. Furthermore, in a model also based on DNN, Sarma [12] investigated the choices of inputs and two different strategies of giving labels and applied the best combination to the IEMOCAP database.

Another approach to improve accuracy is to consider abundant speech features to model the emotion space. In 2020, Yu [13] proposed a speech emotion recognition model with an "attention long-term short-term memory (LSTM)-attention" structure, which combined IS09 and Mel-scaled spectrograms. Issa et al. [14] took Mel-frequency Cepstral Coefficients (MFCCs), chromagrams, Mel-scaled spectrograms, Tonnetz representations and spectral contrast features extracted from speech as inputs and achieved 86.1% recognition accuracy on a 7-class EMO-DB dataset using a deep CNN network.

Although the above research obtained good performance in the speech emotion recognition task, it must be pointed out that such performance is acquired at the sacrifice of the model's portability. There have been attempts to overcome the problems of huge model size and feature redundancy. In 2021, Muppidi [15] jumped out of the traditional method based on machine learning, focusing on high-level features in real value space, and proposed a unique method of feature and network coding using quaternion structure model (QCNN), which not only ensures good accuracy of speech emotion recognition, but also greatly reduces the size of the model. In order to overcome the problem of feature redundancy in speech emotion recognition, Bandela [16] applied unsupervised feature selection to a combination of INTERSPEECH 2010 paralinguistic features, Gammatone Cepstral Coefficients (GTCC) and Power Normalized Cepstral Coefficients (PNCC). The Feature Selection with Adaptive Structure Learning (FSASL), Unsupervised Feature Selection with Ordinal Locality (UFSOL) and the novel Subset Feature Selection (SuFS) algorithms were used to reduce the feature dimension of input features to obtain better SER performance. Although these research works have explored methods to simplify the model, they still depend on acoustic features extracted from speech. In other words, clear speech signals are still necessary.

However, when it comes to applications in real life, what we need is an efficient model to deal with much more complex scenes. For example, clear speech may be hard to obtain. Aiming at facing these problems and improving the practicability of the model, we seek to design an efficient speech emotion recognition model to realize a comparable performance with less and steadier input (just one) and simpler model architecture.

Electroglottograph (EGG) is a signal which can reflect vocal cord movement through recording electrical impedance in the glottis collected by electrodes situated on the throat [17]. The procedure of generating speech can be abstracted as the source-filter model, shown in Figure 1, set up by Fant [18]. It represents speech signals as the combination of a source and a linear acoustic filter, corresponding to the vocal cords and the vocal tract (soft palate, tongue, nasal cavity, oral cavity, etc.), respectively. As an aspect of the source-filter model, EGG is a credible resource to acquire the periodic source information exactly. Additionally, considering the special acquisition of the EGG signal, it is not affected by mechanical vibrations and noise, which makes it suitable for applications in the real life.

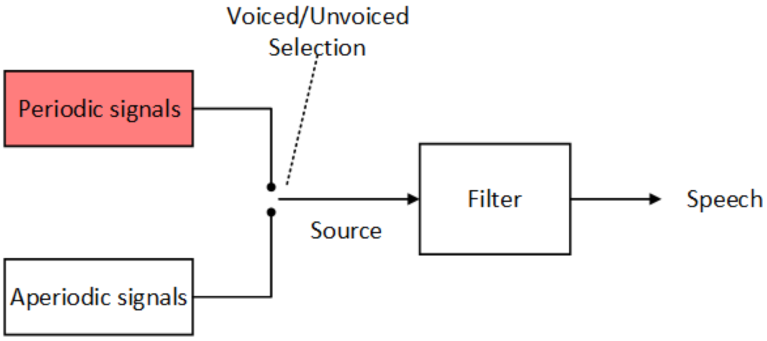

**Figure 1.** The procedure of generating speech based on the source-filter model. The red part indicates what EGG signals record.

In 2017, Sunil Kumar [19] proved that using the phase of the EGG signal can detect the glottal closure instant (GCI) and glottal opening instant (GOI) within a glottal cycle accurately and robustly, which indicated the advanced performance of the EGG signal in exactly extracting excitation source information of speech signals. In 2016, we realized a text-independent phoneme segmentation combining EGG and speech data, which reflected the superiority of EGG signals regarding robustness to noise [20].

As the EGG signal highly corresponds to speaking, a multitude of research has been carried out regarding EGG and its application in the task corresponding to speech [21–27]. As for figuring out the relationship between EGG and emotions, there are studies that have found a strong base to utilizing EGG signals in speech emotion recognition tasks. In 2015, Lu [28] found that EGG can actually serve to identify emotions between neutral, happy and sad. Based on traditional methods, Chen extracted two classes of speech emotional features from EGG and speech, which were the power-law distribution coefficients (PLDC) and the real discrete cosine transform coefficients of the normalized spectrum of EGG and speech signals [29].

In particular, EGG signals have been utilized to help extract emotion features from speech. In 2010, Prasanna et al. [30] analyzed changes in incentive source characteristics between different emotions and observed that the fundamental frequency and incentive intensity were related to emotions. Taking the fundamental frequency extracted from the electroglottograph as the ground truth, the features extracted from EGG and speech are compared to verify the effectiveness of extracting excitation source features from speech. Based on this conclusion, Pravena et al. [31] studied and proved the effectiveness of incentive intensity in identifying emotions in 2017. Incentive intensity explores and introduces the excitation parameters related to emotion (strength of excitation, SoE, and instantaneous fundamental frequency, $F_0$). Combined with the MFCC and GMM models, it realizes an emotion recognition model based on speech and electroglottograph signals. However, although Prevena proved that EGG signals can help improve performance in the SER task, it still relies on the information from speech signals and cannot realize recognition based on only EGG signals.

Cross-modal distillation aims to improve model performance by transferring supervision and knowledge from different modalities. It normally adopts a teacher-student learning mechanism, where the teacher model is usually pre-trained on one modality and then guides the student model on another modality to obtain a similar distribution. The distillation methods usually involve the traditional response-level knowledge distillation [32–34], which uses the logits as the supervision, and feature-level distillation [35,36], which encourages the student network to learn and imitate the intermediate representations of the teacher network. For the speech emotion recognition task, in 2018, Albanie et al. [37] proposed a method of training a speech emotion recognition model with unlabeled speech data via response-level distillation from a pre-trained facial emotion recognition model given visual-audio pairs. Li et al. [38] proposed a method of training the speech emotion recognition model without any labeled speech emotion data with the help of emotion knowledge from a pre-trained text emotion model. These studies apply cross-modal distillation to speech emotion recognition, which inspired our study. However, it must be highlighted that our paper aims to extract emotion information from EGG signals, not speech itself, which is a fundamental difference compared to the above research.

In the present paper, to face up with the reality of poor anti-noise performance of the inputs extracted from speech in the speech emotion recognition task, we propose an EGG-based speech emotion recognition model. Furthermore, to cover the information latent in the modulation of the sound track, we adopt cross-modal emotion distillation (CMED), which transfers robust speech emotion representations from the log-Mel-spectrogram-based model to the EGG-based model.

This paper is organized as follows: Section 2 introduces our materials and methods and presents our proposed model in detail. In Section 3, we discuss the results of our model and the comparison experiments we have conducted. In Section 4, we discuss our work. Finally, Section 5 provides the conclusions of the present work and highlights the expected future works.

## 2. Materials and Methods

### *2.1. Methods*

This paper proposes an efficient framework of relying on only the fundamental frequency extracted from EGG signals to recognize the emotion of speech. We adopted the strategy of cross-modal emotion distillation (CMED) to transfer the latent representations of the acoustic features and enhance the performance of the model. The overall architecture of our framework is shown in Figure 2, which consists of a teacher model and a student model. Each model is composed of three parts: feature extraction, an emotion encoder and a classifier. The output vector of the emotion encoder is regarded as the deep emotion feature, which indicates the distribution of the given utterance's emotion in the deep emotion feature space.

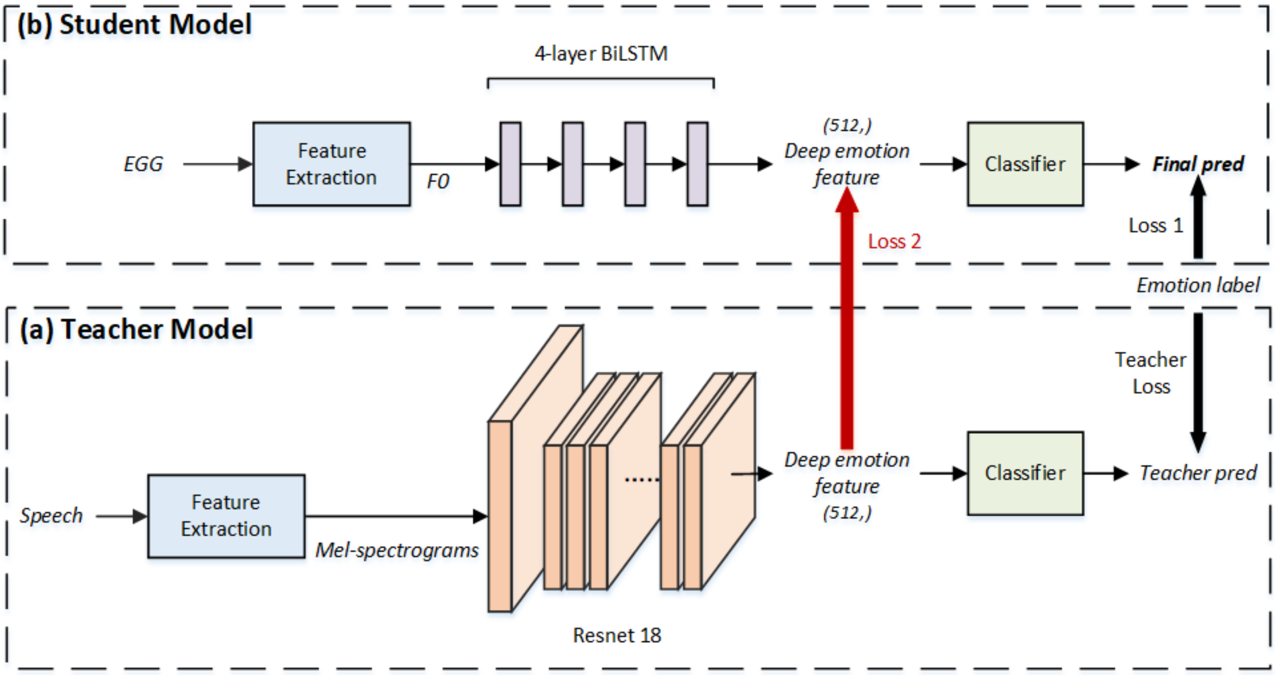

**Figure 2.** Illustration of the proposed framework, which consists of two phases: (**a**) training a strong teacher model based on the acoustic features from speech signals and (**b**) training the student model via cross-modal emotion distillation (CMED) with the pre-trained teacher model.

Based on the teacher-student learning mechanism, our framework consisted of two models: a teacher model and a student model. The adoption of cross-modal emotion distillation was comprised of two stages. Firstly, a strong teacher model based on the acoustic features of speech signals was pre-trained. Secondly, an efficient student model based on EGG signals was trained with the aid of knowledge distilled from the teacher model.

#### 2.1.1. Teacher Model

Considering that EGG signals contain only information regarding the movement of the vocal cord, it is a double-edged sword for the speech emotion recognition task. On the one hand, we can extract the source feature much more easily and exactly, regardless of any noise. However, on the other hand, lacking the information of the sound track may cause a decrease in the recognition accuracy. Thus, we adopted a strong teacher model to lead the student model and meliorate the representations in the deep emotion feature map.

The teacher model is responsible for providing supervision to the student speech emotion model at the level of deep emotion features. The structure of the teacher model consisted of three stages:

1. Selecting and extracting acoustic features from the raw speech signals;

2. Going through a deep learning network to obtain deep emotion features;
3. Using a simple classifier to convert the deep features to the predicted emotion label.

As the primary task of the teacher model is to provide the student model with emotion information associated with the sound track, we chose 80-channel log-Mel-spectrograms as the input of the teacher speech emotion recognition model, which is a typical type of feature containing the information of the sound track [39].

For better modeling of deep emotional feature spaces, we chose the classic ResNet18 (deep residual network [40]) as our encoder, which has been proven to be efficient in many classification tasks. The output of ResNet18 ($dim = 512$), regarded as the deep emotion feature, was then fed in a fully connected layer to generate the predicted emotion label. The output of ResNet18 also acted as a supervision to guide the student model to capture and imitate its embedded emotion representations.

### 2.1.2. Student Model

The student model was based on only EGG signals and was composed of three stages as well. The fundamental frequency ($F_0$) is a commonly-used excitation feature to characterize vocal cord vibration. The fundamental frequency intensity and duration help in the analysis of prosody factors according to Rao et al. [41]. Thus, we chose the fundamental frequency ($F_0$) as the input of our student speech emotion recognition model. The extraction of $F_0$ from EGG signals was based on [42], as shown in Figure 3.

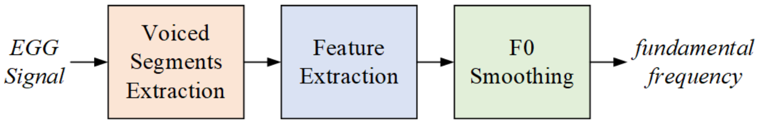

**Figure 3.** The structure of the EGG feature exaction module.

Firstly, we extracted the voiced segments from the raw *EGG* signals according to the short-time logarithmic energy to avoid the effect of unvoiced segments on the estimation of the $F_0$. Then, the fundamental frequency was estimated by the auto-correlation method, which is based on the periodical change in the amplitude of EGG signals, formulated as Equation (1).

$$F_0 = \frac{f_s}{\underset{\frac{f_s}{f_{max}} \leq k \leq \frac{f_s}{f_{min}}}{argmax} \ \sum_{m=0}^{N-1-k} x_{EGG}(m) x_{EGG}(m+k)} \tag{1}$$

where $x_{EGG}(m)$ and $x_{EGG}(m+k)$ are two different sample points of the input EGG signal. $f_s$ is the sampling rate. $f_{max}$ and $f_{min}$ are the maximum and minimum of the $F_0$, respectively.

As the method of $F_0$ extraction will cause erroneous values [43], we adopted the same smoothing method as [44] with bidirectional searching as proposed by Jun et al [45] after the extraction of $F_0$.

When considering the emotion feature encoder, as the fundamental frequency is a 1D acoustic feature with strong temporal correlation, we selected Bi-LSTM (bidirectionally long short-term memory) as the feature encoder for its superior performance in sequence processing and previous speech emotion recognition tasks. After the comparison of different depths of the layer, we finally adopted a 4-layer Bi-LSTM with a fully connected layer as the student speech emotion recognition model, illustrated in Figure 4. Each bidirectionally LSTM contains 512 bidirectional LSTM cells (256 forward nodes and 256 backward nodes) and finally generates a deep emotion feature vector with a length of 512. The dimension of the student feature was restricted the same as the teacher's feature output to learn the distribution of the teacher's features via cross-modal emotion distillation.

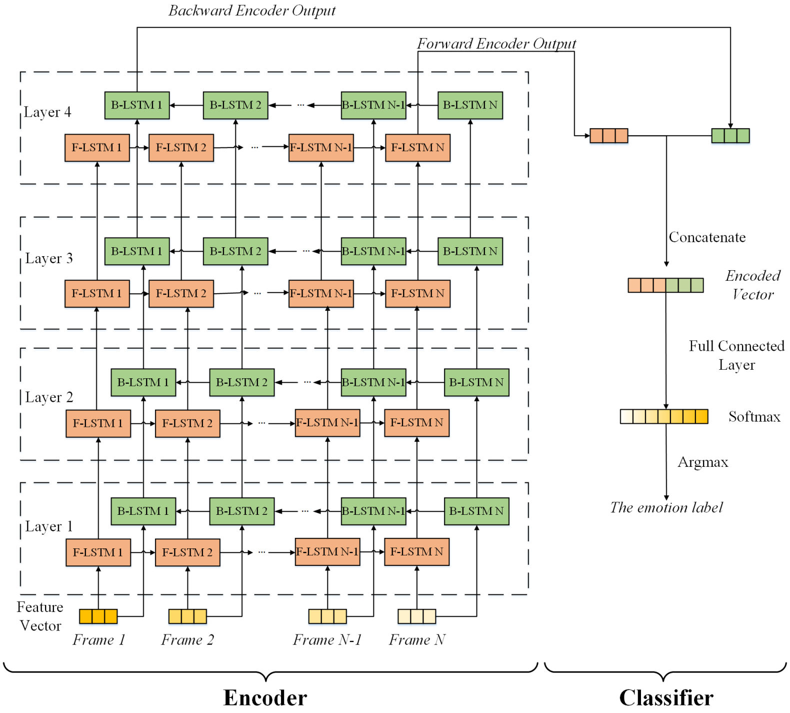

**Figure 4.** The structure of the student speech emotion recognition model based on 4-layer Bi-LSTM.

### 2.1.3. Cross-Modal Emotion Distillation

In order to transfer the emotion information contained in the teacher model's deep emotion features, especially the information corresponding to the soundtrack, we introduced cross-modal emotion distillation to our framework.

Between the response-level and feature-level, we adopted the latter to directly imitate the hidden representation of the teacher's features. Following the protocol of knowledge distillation [32,46–49], we used Kullback–Leibler divergence [50] to minimize the distance between the deep emotion features produced by the teacher and the student models. The procedure of cross-modal emotion distillation is illustrated in Figure 5. The blue parts indicate the procedure of the calculation of the Kullback–Leibler divergence. The teacher's deep emotion features and the student's firstly go through a softmax (for the student's is log softmax) function to normalize their distribution, as defined in Equations (2) and (3), respectively. Then. the Kullback–Leibler divergence $L_{KLdiv}(p^S(\tau)||p^T(\tau))$ is calculated as Equation (4), which is regarded as the KL divergence loss in the total loss function.

The calculation of the Kullback–Leibler divergence can be formulated as:

$$p_m^S(\tau) = log(\frac{\exp(S_m/\tau)}{\sum_{i=1}^{M}\exp(S_i/\tau)}) \tag{2}$$

$$p_m^T(\tau) = \frac{\exp(T_m/\tau)}{\sum_{i=1}^{M}\exp(T_i/\tau)} \tag{3}$$

$$L_{KLdiv}(p^S(\tau)||p^T(\tau)) = \sum_{m=1}^{M} p_m^S(\tau)\log(\frac{p_m^S(\tau)}{p_m^T(\tau)}) \tag{4}$$

where $S$ is the vector of thhe student's deep emotion features, and $T$ represents the teacher's. $\tau$ is a hyper-parameter; in our work, we set $\tau = 2$ according to the results of the comparative experiment. $p_m^S(\tau)$ is the vector of the student's emotion features after log softmax with the hyper-parameter $\tau$ at the point $m$. Finally, $p_m^T(\tau)$ is the teacher's vector after softmax with the hyper-parameter $\tau$ at the point $m$. Equation (4) gives the formulation of Kullback–Leibler divergence.

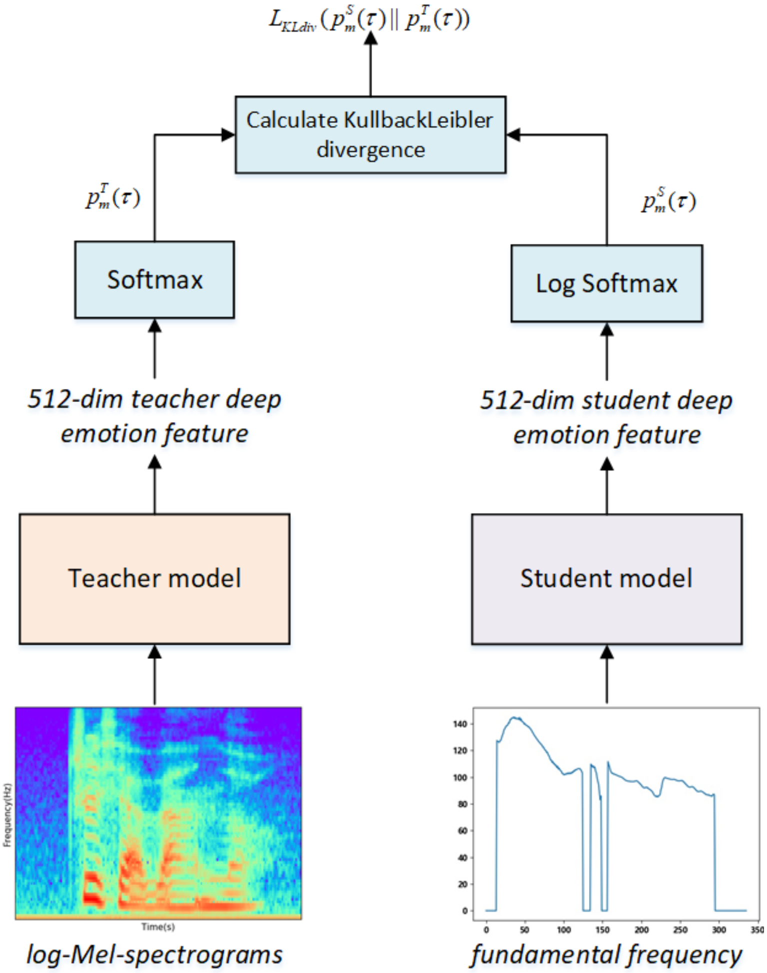

**Figure 5.** The procedure of cross-modal emotion distillation.

Thus, as Equation (5) shows, the total loss function of the student model can be separated in two parts: the KL divergence between two deep emotion features and the cross entropy between the student-predicted label and the ground truth, respectively.

$$Loss = \alpha\tau^2 \cdot L_{KLdiv}(p^S(\tau)||p^T(\tau)) + (1-\alpha) \cdot L_{CrossEntropy}(Q_S, y_{\text{true}}) \tag{5}$$

where $\alpha$ is a hyper-parameter which represents the proportion of the *KL* divergence loss in the loss function. We set $\alpha = 0.6$ according to the results of the comparative experiment. $Q_S$ is the predicted label of the student model, and $y_{true}$ is the ground truth.

## 2.2. Materials

The dataset for our work was the S70 subset of the Chinese Dual-mode Emotional Speech Database (CDESD [51]). This dataset was built by the pattern recognition and human intelligence laboratory affiliated with the Department of Electronics and Information Engineering at Beihang University and collected from 20 speakers aged 21 to 23 (13 men, 7 women) with 7 classes of emotions: happiness, sadness, neutrality, anger, fear, surprise and disgust.

Considering the dataset is acted, not all utterances have been performed well and truly. Thus, we invited 30 volunteers aged 20 to 30 (18 men, 12 women) to evaluate every utterance and select the S70 subset as our dataset, which contains 1323 utterances. The S70 subset indicates that at least 70% of the evaluators can recognize the emotion of the utterance accurately. In the experiment, 80% of the total dataset was chosen as the training set, and the rest was used as the validation test. Table 1 illustrates the data statistics of the

S70 subset of the CDESD Dataset, which shows the number of utterances with different emotions in the training set and the validation set.

**Table 1.** Data statistics of the S70 subset of the CDESD Dataset.

| Emotion | Sadness | Anger | Surprise | Fear | Happiness | Disgust | Neutrality | Total |
|---|---|---|---|---|---|---|---|---|
| Train. Set | 156 | 84 | 294 | 28 | 152 | 67 | 280 | 1061 |
| Val. Set | 38 | 21 | 73 | 7 | 37 | 16 | 70 | 262 |

## 3. Experiments and Results

### *3.1. Implementation Details*

Our proposed SER framework was implemented in PyTorch. For the optimiser, the Adam [52] optimizer was used; meanwhile, the training accuracy was monitored with early stopping set as 8 epochs. The system was trained with a batch size of 16. The initial learning rate was set at $1 \times 10^{-3}$ and decayed to 0.8 every 20 epochs.

### *3.2. Experiments on the Teacher Model*

For the teacher model, we explored our architecture under the following conditions:

1. The choice of acoustic feature input, log-Mel-spectrograms or traditional Mel-spectrograms utilized in [39];
2. Comparing the model structure of ResNet18 to CRNN [39]. The results of these comparative experiments of the teacher model are listed in Table 2.

**Table 2.** Comparison of the teacher model using different methods.

| Model | Unweighted Validation Accuracy (%) |
|---|---|
| CRNN [39] + Mel-spectrograms | 69.53 |
| CRNN + log-Mel-spectrograms | 71.45 |
| ResNet18 + Mel-spectrograms | 80.86 |
| ResNet18 + log-Mel-spectrograms | **81.25** |

Through the results of the comparative experiments, we found that for the selection of the input acoustic feature, under the condition of the same model structure, log-Mel-spectrograms outperform Mel-spectrograms by 1.92% and 0.36%, respectively. This indicates that log-Mel-spectrograms can better characterize the information corresponding to emotion.

As for the architecture of the teacher model, ResNet18 obtained 9.8% (log-Mel-spectrograms input) and 11.33% (Mel-spectrograms input) higher validation accuracy than CRNN while significantly reducing the time it took to converge, which indicates the superiority of ResNet18.

### *3.3. Experiments on the Student Model*

We also discovered how the number of Bi-LSTM layers influenced the results. We conducted experiments with the EGG-based speech emotion recognition model under the conditions of 3, 4 and 5 layers of Bi-LSTM. As shown in Table 3, the 4-layer Bi-LSTM was the best compared to the others. Thus, we set the layer number to 4 to conduct the following cross-modal emotion distillation experiments.

**Table 3.** Comparison of the student model with different layer depths.

| Model | Unweighted Validation Accuracy (%) |
|---|---|
| Bi-LSTM (3 layer) | 53.52 |
| Bi-LSTM (4 layer) | 58.98 |
| Bi-LSTM (>4 layer) | Not converged |

### *3.4. Experiments on Cross-Modal Emotion Distillation*

Based on the aforementioned comparative experiments on the teacher model and the student model, we conducted cross-modal emotion distillation with the following conditions:

1. A pre-trained techer model based on ResNet18 with log-Mel-spectrograms as input;
2. A 4-layer Bi-LSTM student model with $F_0$ as input;
3. The cross-modal emotion distillation was conducted on the feature level.

To find suitable settings for the hyper-parameters, we conducted a series of comparative experiments. Setting $\tau$ from 1 to 4, we found that the most suitable setting for $\tau$ was 2, according to Figure 6.

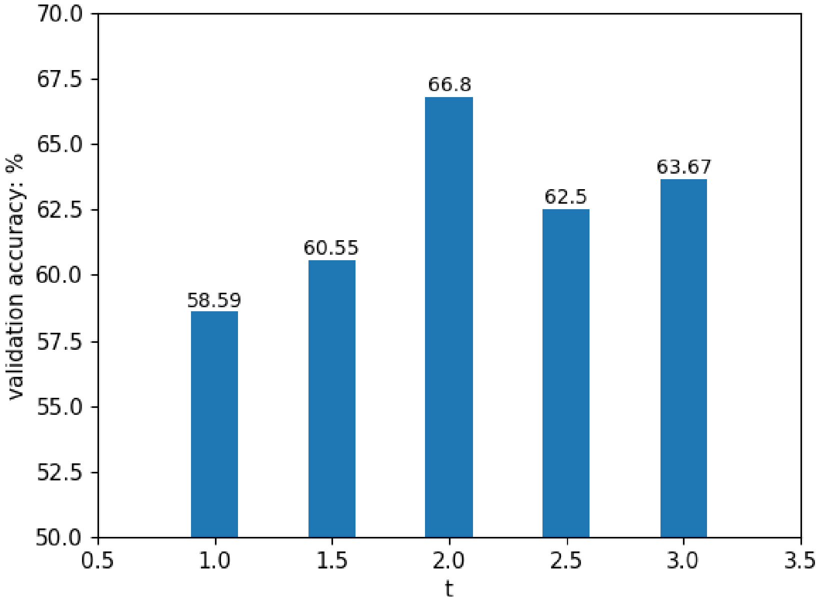

**Figure 6.** Experiments on different settings of the hyper-parameter $\tau$.

To further discover the influence of cross-modal emotion distillation on the results on the condition of $\tau = 2$, we modified the hyper-parameter of $\alpha$, which indicates the proportion of the distillation loss in total loss. Figure 7 indicates that the implementation of cross-modal emotion distillation obviously improves the performance of the model. From $\alpha = 0.2$ to 0.9, the accuracy of the validation dataset increased to varying degrees. The highest validation accuracy was reached when $\alpha = 0.6$, then dropped slowly with the proportion of the categorical cross entropy that was declining. According to the experiment results, we set $\alpha = 0.6$.

The results of our framework are listed in Table 4. It can be observed that the training accuracy is obviously improved by 23.2% and the validation accuracy by 7.82% with the aid of cross-modal emotion distillation. In other words, under the condition of some utterances that are not expressive enough, our framework with CMED obtains better performance and realizes a result comparable to the human subjective evaluation.

**Table 4.** Results of the experiments with cross-modal emotion distillation.

| Model | Unweighted Training Accuracy (%) | Unweighted Validation Accuracy (%) |
|---|---|---|
| Teacher Model | 99.22 | 81.25 |
| Student Model | 66.57 | 58.98 |
| Student Model with CMED | 89.77 | 66.80 |
| Ground Truth [1] | 70.00 | 70.00 |

[1] The ground truth is the average accuracy of all the utterances of the S70 subset according to subjective evaluation.

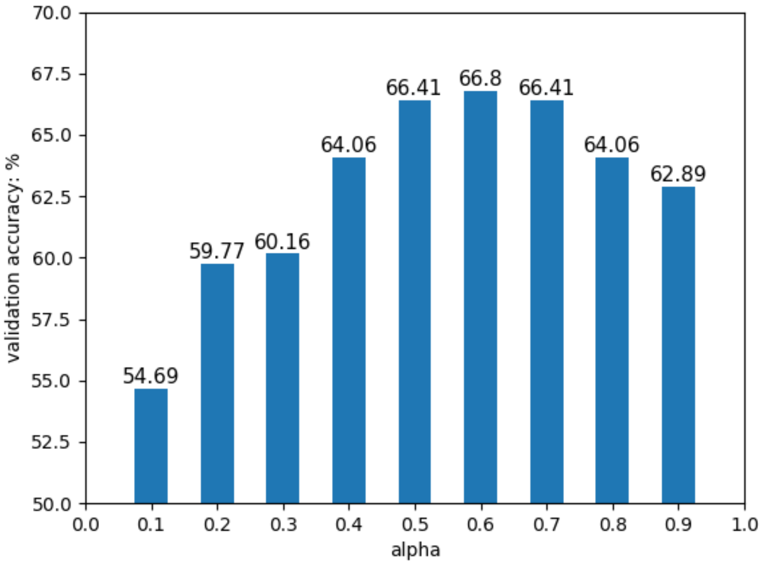

**Figure 7.** Experiments on different proportions of cross-modal emotion distillation in the total loss.

### *3.5. Evaluations and Results*

In this section, we will further evaluate and discuss the results of our framework, especially from the aspect of the different performance between different kinds of emotions. Figures 8 and 9 show the confusion matrix of our student model and the model with CMED, respectively. The horizontal axis is our predicted emotion label and the vertical axis is the true label. Through the contrast of these two confusion matrices, we can find that the validation accuracy of almost every emotion has been consistently improved with the aid of cross-modal emotion distillation. To be specific, for anger, happiness, and neutrality, the validation accuracy has been improved by 15%, 3%, and 17%, respectively, which reflects the superiority of CMED in improving the performance of SER when enough utterances can be acquired. For the emotions that lack utterances, such as fear and disgust, the performance can also be improved by 11% and 7%. Better performance on these types of emotions indicates that with CMED, the ability to distinguish emotions has been improved significantly, regardless of the lack of utterances.

From Figure 9, it can be seen that our framework can classify sadness, surprise, happiness and neutrality better than other emotions. This is because, as can be seen in Table 1, these emotions occupy a larger set of utterances than the others. For example, as there are only 28 utterances of fear in the training set and 7 in the validation set, clearly recognizing fear is a much harder challenge than others, as every mismatch influences the results greatly.

When considering the relationship between emotions, it can be noted that some emotions are more similar to each other and more likely to be mismatched, such as sadness and neutrality, anger, surprise and happiness. This can be explained from the aspect of the characteristics of different emotions. For example, from the dimensional structure theories [53], sadness and neutrality both demonstrate low activation, while anger, surprise and happiness demonstrate high activation. The different levels of activation highly correspond to the average value of the fundamental frequency, which is our model's input.

To further discover the distribution of our 512-dimension deep emotion feature in the emotion feature map, we visualized the feature of the 4-layer Bi-LSTM output using the t-SNE technique [54], which carries out a feature dimension reduction from the original dimension to two-dimensional space. To clearly visualize the distribution, we selected three emotions which are difficult to classify: surprise, happiness and neutrality. Figures 10 and 11 illustrate the distributions of the deep emotion features of the student model and the proposed model with CMED, respectively. The horizontal and vertical axes indicate the x and y axes in two-dimensional space.

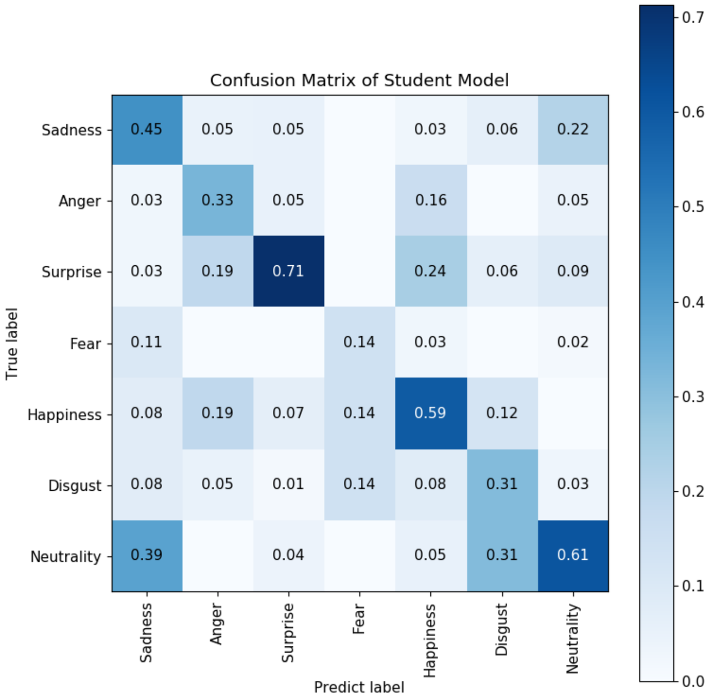

**Figure 8.** Confusion matrix of our student model.

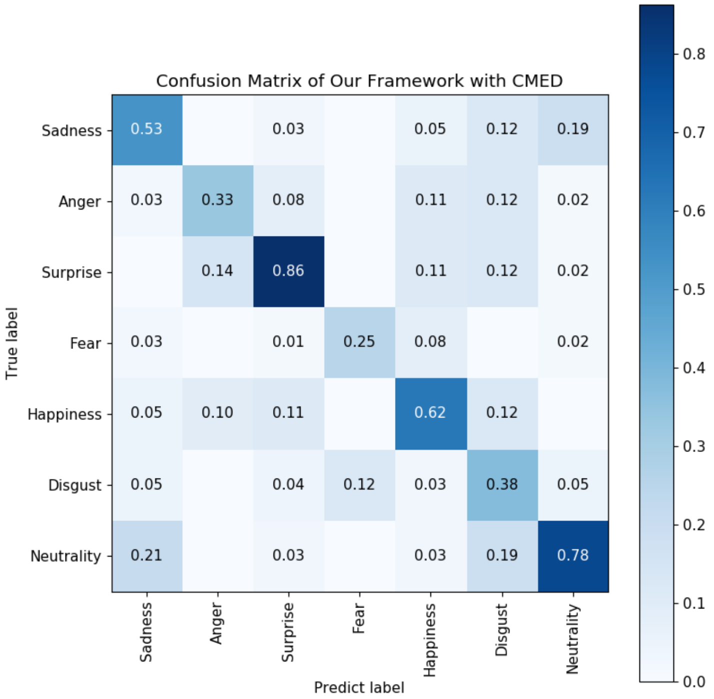

**Figure 9.** Confusion matrix of our framework with CMED.

From Figures 10 and 11, we can intuitively observe that with the aid of cross-modal emotion distillation, the deep emotion features are projected to different clusters clearly and separated from each other. Visualization experiments proved that the cross-modal emotion distillation helps to capture more emotion characteristics and results in a better distribution on the deep emotion feature map.

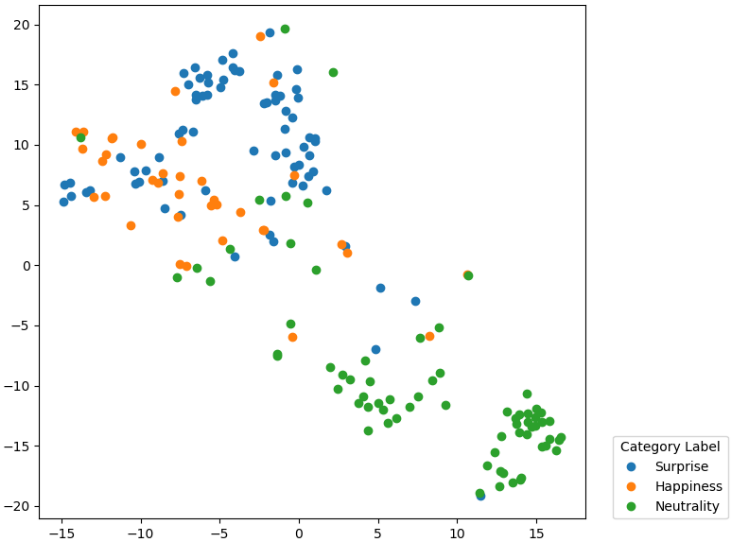

**Figure 10.** Visualization of distributions of deep emotion features of the student model.

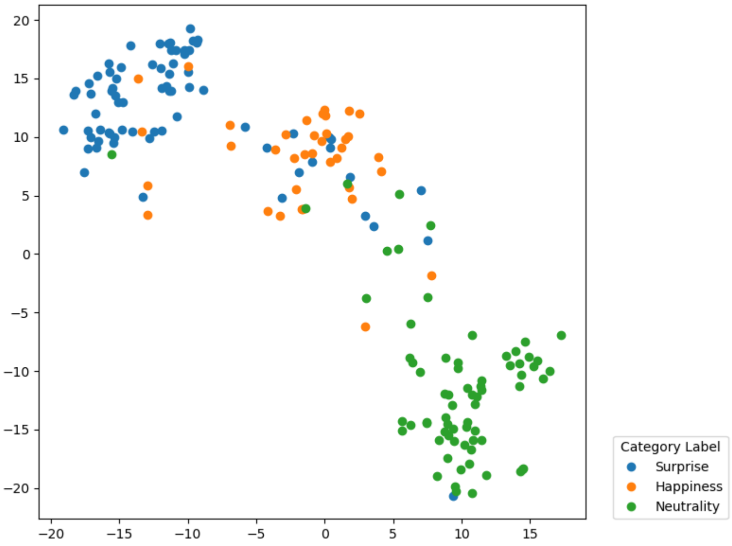

**Figure 11.** Visualization of distributions of deep emotion features of the model with CMED.

### *3.6. Experiments on EMO-DB*

To find out the effect of cross-modal emotion distillation on other languages, we conducted experiments on the Berlin Database of Emotional Speech Berlin EMO-DB [55], which is a classical German dataset with EGG. There are also 7 emotions in this dataset, which are anger, boredom, disgust, fear, happiness, sadness and neutrality. Table 5 illustrates the data Statistics of the EMO-DB, which shows the number of utterances with different emotions in the training set and in the validation set.

**Table 5.** Data statistics of the EMO-DB.

| Emotion | Sadness | Anger | Boredom | Fear | Happiness | Disgust | Neutrality | Total |
|---|---|---|---|---|---|---|---|---|
| Train Set | 96 | 109 | 89 | 97 | 92 | 84 | 83 | 650 |
| Val Set | 24 | 27 | 22 | 24 | 22 | 20 | 20 | 159 |

To test the performance of our framework on German, we retained the hyper-parameters without any adjustment for the new language, i.e., we conducted experiments under the following conditions:

1. Choosing log-Mel-spectrograms as the teacher model input and ResNet18 as the teacher model;
2. Choosing 4-Layer Bi-LSTM as the student model;
3. Setting hyper-parameters $\tau = 2$ and $\alpha = 0.6$ in the model with CMED.

The results of our models on EMO-DB are given in Table 6.

**Table 6.** Results of the experiments on EMO-DB.

| Model | Unweighted Train Accuracy (%) | Unweighted Validation Accuracy (%) |
|---|---|---|
| Teacher Model | 98.12 | 75.00 |
| Student Model | 33.89 | 32.29 |
| Student Model with CMED | 84.13 | **42.71** |

Shown in Table 6, with CMED, the validation accuracy increased from 32.29% to 42.71%, which indicates that cross-modal emotion distillation still works on other languages. However, compared with other research on EMO-DB, the validation accuracy of our framework is not so comparable, which is mainly because of the poor performance of the student model independently. We speculate that this is due to the insufficiency of the training data and the different characteristics of different languages. German may contain more emotional information in the procedure of the movement of the sound track, not the vocal cord. Nevertheless, the improvement of our proposed model with CMED compared to the origin student model still proves the efficiency of CMED on transferring the emotional information.

Figures 12 and 13 show the confusion matrix of our student model and model with CMED. It clearly can be seen that depending on only the student model with a lack of training data and the information of the sound track, we can hardly obtain the emotion information. The output of the student model tends to focus on just part of labels and loses the attention of the others. In contrast, adopting CMED greatly ameliorates this phenomenon. With the aid of CMED, the model can be aware of all the emotions and classify them. Although the result still has room to be improved, it has achieved a great improvement compared with the student model.

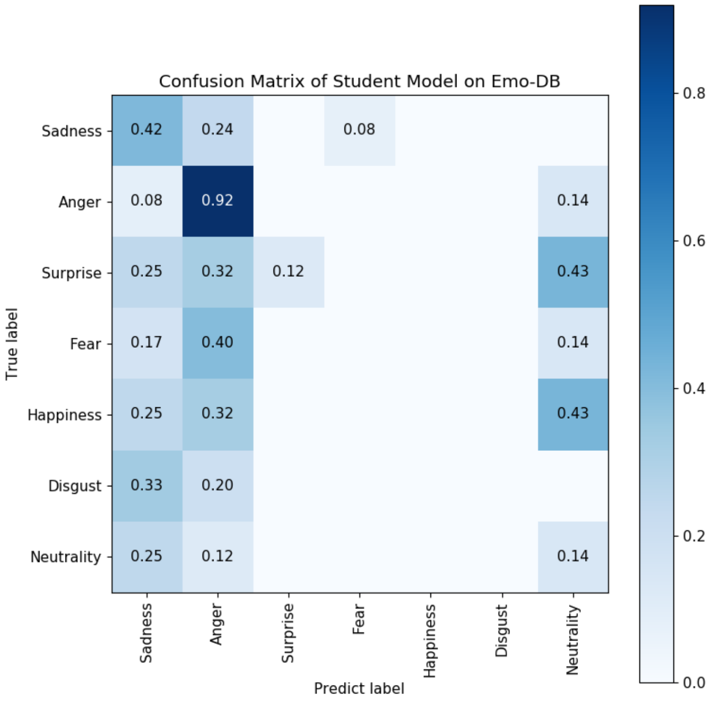

**Figure 12.** Confusion matrix of our student model on EMO-DB.

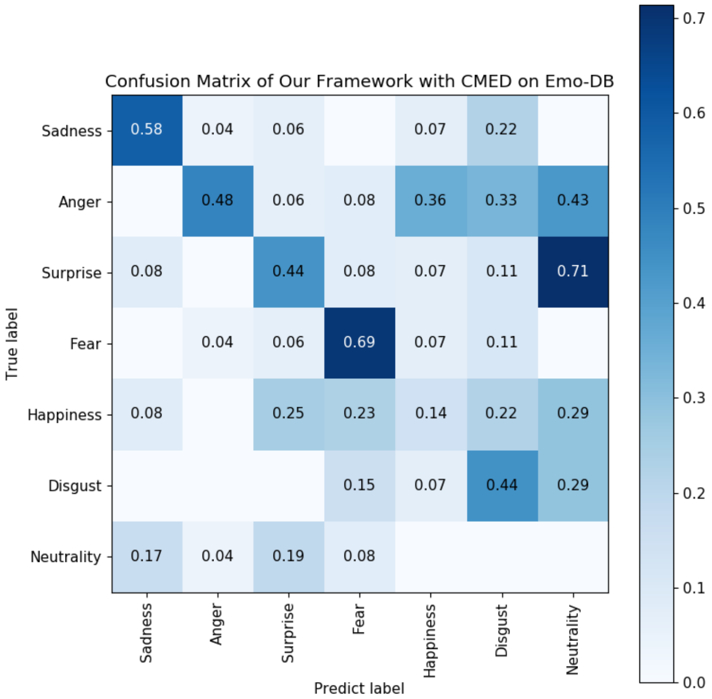

**Figure 13.** Confusion matrix of our framework with CMED on EMO-DB.

Selecting 3 emotions which are difficult to classify, we visualized our output of the student model on EMO-DB by t-SNE. Figures 14 and 15 visualize the distribution of the deep emotion feature in the emotion map. It can be obviously observed that in the student model, the deep emotion features can hardly project to different clusters, while with the aid of CMED, the clusters are much more recognizable.

Through all the experiments above, from the confusion matrix as well as the visualization experiment, there is no doubt of the efficiency of the cross-modal emotion distillation on the EGG-based Chinese speech emotion recognition task. Experiments on EMO-DB illustrate that our cross-modal emotion distillation still works on other languages, successfully transferring information from the teacher model to the student model.

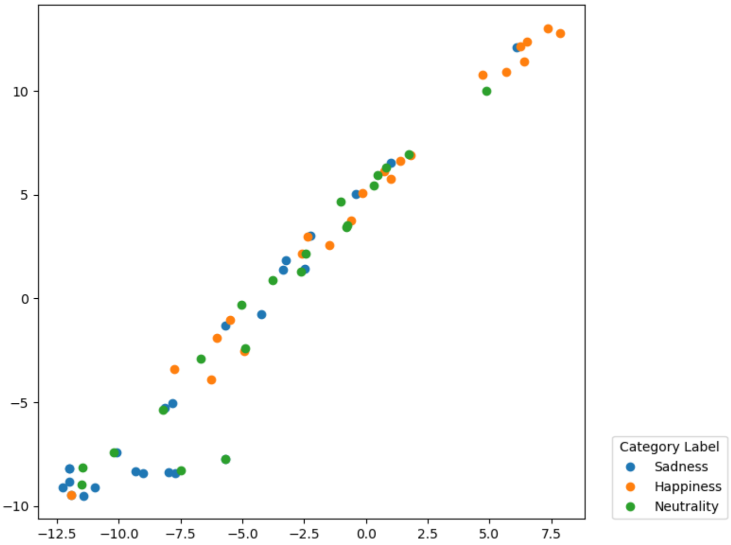

**Figure 14.** Visualization for distributions of deep emotion features of the student model on EMO-DB.

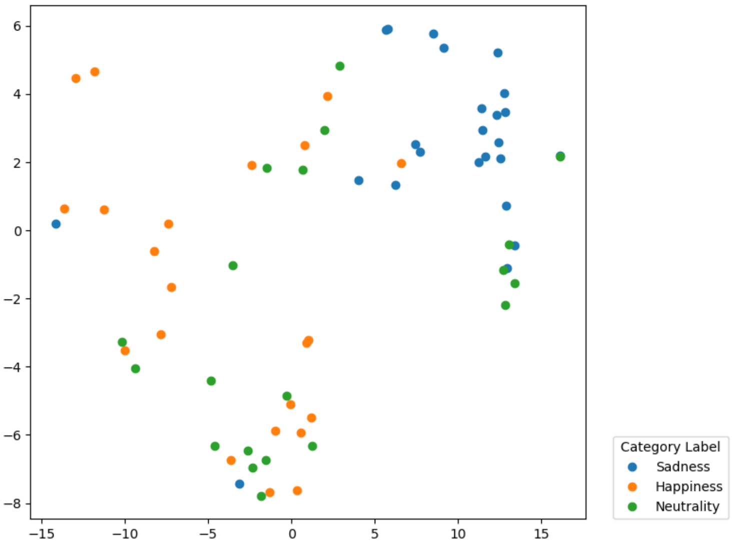

**Figure 15.** Visualization for distributions of deep emotion features of the model with CMED on EMO-DB.

## 4. Discussion

In Section 3, we conducted a series of experiments to explore the best framework.

From the aspect of feature selection, we compared two different classic acoustic features in the teacher model, log-Mel-spectrograms and Mel-spectrograms, and concluded that log-Mel-spectrograms perform better than the latter. As for the choice of the architecture, for the teacher model, by contrasting ResNet18 with CRNN, the result indicated that ResNet18 achieves better performance on classification as well as faster convergence. For the student model, different numbers of layers were explored, and it was concluded that 4 is the best depth for Bi-LSTM. Two experiments regarding the settings of the hyperparameters were conducted to verify the best experimental conditions.

As for the evaluation of the results, we adopted unweighted validation accuracy and the visualization result utilizing t-SNE. Both of the two results on CDESD have proved that CMED works efficiently and can obtain a comparable result in the Chinese speech emotion recognition task. We realized an improvement from 58.98% to 66.80% via cross-modal emotion distillation on the S70 subset of the CDESD, which achieves a comparable accuracy to human subject evaluation.

To explore the performance of our model on other languages, we conducted experiments on EMO-DB with all the experiment conditions remaining the same. With the aid of CMED, we obtained a higher validation accuracy, which improved from 32.29% to 42.71%. For the phenomenon in which the final result cannot reach a similar level with the teacher model, we speculate that it is due to the insufficiency of the training data and the different characteristics of different languages. Nevertheless, the huge improvement observed can still prove that cross-modal emotion distillation can still help to improving the result regardless of the performance of the student model independently.

## 5. Conclusions

In this paper, we provide a new strategy to design an efficient SER model to realize a comparable performance with less input (just one) and simpler model architecture and propose a framework of electroglottograph-based speech emotion recognition via cross-modal emotion distillation. By utilizing an electroglottograph as the input, we can face more complex real circumstances with significant noise and extract the acoustic features we need easily and exactly. To involve the emotion information related to the sound track, we propose cross-modal emotion distillation to transfer emotion information from the teacher model to the student model.

In the future, other teacher and student models will be designed which utilize different features as input to identify the best performance of the architecture and achieve better results in other languages.

**Author Contributions:** Conceptualization, L.C.; Data curation, X.M.; Funding acquisition, L.C. and Q.Z.; Methodology, J.R.; Project administration, X.M. and Q.Z.; Writing—original draft, J.R. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Science Foundation for Young Scientists of China under grant 61603013, National Natural Science Foundation of China under grant 62072021, and the Fundamental Research Funds for the Central Universities (No. YWF-21-BJ-J-534).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Acknowledgments:** The authors of this paper would like to appreciate all subjects involved in the building the dataset used in this study.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J.G. Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.* **2002**, *18*, 32–80. [CrossRef]
2. Ringeval, F.; Michaud, A.; Cifti, E.; Güle, H.; Lalanne, D. AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition. In Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop, Seoul, Korea, 22 October 2018.
3. Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M.A.; Schuller, B.; Zafeiriou, S. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5200–5204. [CrossRef]
4. Neumann, M.; Vu, N.T. Attentive Convolutional Neural Network Based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech. In Proceedings of the Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, 20–24 August 2017; pp. 1263–1267.
5. Kim, J.; Englebienne, G.; Truong, K.P.; Evers, V. Deep Temporal Models using Identity Skip-Connections for Speech Emotion Recognition. In Proceedings of the 2017 ACM on Multimedia Conference, Mountain View, CA, USA, 23–27 October 2017; pp. 1006–1013. [CrossRef]
6. Han, W.; Ruan, H.; Chen, X.; Wang, Z.; Li, H.; Schuller, B. Towards Temporal Modelling of Categorical Speech Emotion Recognition. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018. [CrossRef]
7. Atmaja, B.T.; Akagi, M. Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model. In Proceedings of the 2019 IEEE International Conference on Signals and Systems (ICSigSys), Bandung, Indonesia, 16–18 July 2019; pp. 40–44. [CrossRef]
8. Rajamani, S.T.; Rajamani, K.T.; Mallol-Ragolta, A.; Liu, S.; Schuller, B. A Novel Attention-Based Gated Recurrent Unit and its Efficacy in Speech Emotion Recognition. In Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 6294–6298. [CrossRef]
9. Peng, Z.; Lu, Y.; Pan, S.; Liu, Y. Efficient Speech Emotion Recognition Using Multi-Scale CNN and Attention. In Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 3020–3024. [CrossRef]
10. Helmiyah, S.; Riadi, I.; Umar, R.; Hanif, A. Speech Classification to Recognize Emotion Using Artificial Neural Network. *Khazanah Inform. J. Ilmu Komput. Dan Inform.* **2021**, *7*, 11913. [CrossRef]
11. Tzirakis, P.; Zhang, J.; Schuller, B.W. End-to-End Speech Emotion Recognition Using Deep Neural Networks. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5089–5093. [CrossRef]
12. Sarma, M.; Ghahremani, P.; Povey, D.; Goel, N.K.; Dehak, N. Emotion Identification from Raw Speech Signals Using DNNs. In Proceedings of the Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2–6 September 2018; pp. 3097–3101. [CrossRef]
13. Yu, Y.; Kim, Y.J. Attention-LSTM-Attention Model for Speech Emotion Recognition and Analysis of IEMOCAP Database. *Electronics* **2020**, *9*, 713. [CrossRef]
14. Issa, D.; Demirci, M.F.; Yazici, A. Speech emotion recognition with deep convolutional neural networks. *Biomed. Signal Process. Control* **2020**, *59*, 101894. [CrossRef]

15. Muppidi, A.; Radfar, M. Speech Emotion Recognition Using Quaternion Convolutional Neural Networks. In Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 6309–6313. [CrossRef]
16. Bandela, S.R.; Kumar, T.K. Unsupervised feature selection and NMF de-noising for robust Speech Emotion Recognition. *Appl. Acoust.* **2021**, *172*, 107645. [CrossRef]
17. Tronchin, L.; Kob, M.; Guarnaccia, C. Spatial Information on Voice Generation from a Multi-Channel Electroglottograph. *Appl. Sci.* **2018**, *8*, 1560. [CrossRef]
18. Fant, G. *Acoustic Theory of Speech Production*; De Gruyter Mouton: Berlin, Germany, 1971. [CrossRef]
19. Kumar, S.S.; Mandal, T.; Rao, K.S. Robust glottal activity detection using the phase of an electroglottographic signal. *Biomed. Signal Process. Control* **2017**, *36*, 27–38. [CrossRef]
20. Chen, L.; Mao, X.; Yan, H. Text-Independent Phoneme Segmentation Combining EGG and Speech Data. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 1029–1037. [CrossRef]
21. Paul, N.; Kumar, S.; Chatterjee, I.; Mukherjee, B. Electroglottographic Parameterization of the Effects of Gender, Vowel and Phonatory Registers on Vocal Fold Vibratory Patterns: An Indian Perspective. *Indian J. Otolaryngol. Head Neck Surg.* **2011**, *63*, 27–31. [CrossRef]
22. Macerata, A.; Nacci, A.; Manti, M.; Cianchetti, M.; Matteucci, J.; Romeo, S.O.; Fattori, B.; Berrettini, S.; Laschi, C.; Ursino, F. Evaluation of the Electroglottographic signal variability by amplitude-speed combined analysis. *Biomed. Signal Process. Control* **2017**, *37*, 61–68. [CrossRef]
23. Borsky, M.; Mehta, D.D.; Stan, J.H.V.; Gudnason, J. Modal and Nonmodal Voice Quality Classification Using Acoustic and Electroglottographic Features. *IEEE/Acm Trans. Audio Speech Lang. Process.* **2017**, *25*, 2281–2291. [CrossRef] [PubMed]
24. Liu, D.; Kankare, E.; Laukkanen, A.M.; Alku, P. Comparison of parametrization methods of electroglottographic and inverse filtered acoustic speech pressure signals in distinguishing between phonation types. *Biomed. Signal Process. Control* **2017**, *36*, 183–193. [CrossRef]
25. Lebacq, J.; Dejonckere, P.H. The dynamics of vocal onset. *Biomed. Signal Process. Control* **2019**, *49*, 528–539. [CrossRef]
26. Filipa, M.; Ternstrm, S. Flow ball-assisted voice training: Immediate effects on vocal fold contacting. *Biomed. Signal Process. Control* **2020**, *62*. [CrossRef]
27. Chen, L.; Ren, J.; Chen, P.; Mao, X.; Zhao, Q. Limited text speech synthesis with electroglottograph based on Bi-LSTM and modified Tacotron-2. *Appl. Intell.* **2022**. [CrossRef]
28. Hui, L.; Ting, L.; See, S.; Chan, P. Use of Electroglottograph (EGG) to Find a Relationship between Pitch, Emotion and Personality. *Procedia Manuf.* **2015**, *3*, 1926–1931. [CrossRef]
29. Chen, L.; Mao, X.; Wei, P.; Compare, A. Speech emotional features extraction based on electroglottograph. *Neural Comput.* **2013**, *25*, 3294–3317. [CrossRef] [PubMed]
30. Prasanna, S.R.M.; Govind, D. Analysis of excitation source information in emotional speech. In Proceedings of the 11th Annual Conference of the International Speech Communication Association, Chiba, Japan, 26–30 September 2010; pp. 781–784. [CrossRef]
31. Pravena, D.; Govind, D. Significance of incorporating excitation source parameters for improved emotion recognition from speech and electroglottographic signals. *Int. J. Speech Technol.* **2017**, *20*, 787–797. [CrossRef]
32. Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *Comput. Sci.* **2015**, *14*, 38–39.
33. Afouras, T.; Chung, J.S.; Zisserman, A. ASR is All You Need: Cross-Modal Distillation for Lip Reading. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 2143–2147. [CrossRef]
34. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arxiv:1910.01108.
35. Romero, A.; Ballas, N.; Kahou, S.E.; Chassang, A.; Gatta, C.; Bengio, Y. FitNets: Hints for Thin Deep Nets. In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, 7–9 May 2015.
36. Heo, B.; Kim, J.; Yun, S.; Park, H.; Kwak, N.; Choi, J.Y. A Comprehensive Overhaul of Feature Distillation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 1921–1930. [CrossRef]
37. Albanie, S.; Nagrani, A.; Vedaldi, A.; Zisserman, A. Emotion Recognition in Speech Using Cross-Modal Transfer in the Wild. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Korea, 22–26 October 2018; pp. 292–301. [CrossRef]
38. Li, R.; Zhao, J.; Jin, Q. Speech Emotion Recognition via Multi-Level Cross-Modal Distillation. In Proceedings of the Interspeech 2021, Brno, Czech Republic, 30 August–3 September 2021; pp. 4488–4492. [CrossRef]
39. Zhao, J.; Mao, X.; Chen, L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomed. Signal Process. Control* **2019**, *47*, 312–323.[CrossRef]
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
41. Rao, K.S.; Yegnanarayana, B. Prosody modification using instants of significant excitation. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *14*, 972–980. [CrossRef]
42. Chen, L.; Mao, X.; Compare, A. A new method for speech synthesis combined with EGG. In Proceedings of the National Conference on Man-Machine Speech Communication 2013, Lianyungang, China, 11–13 October 2013.

43. Prukkanon, N.; Chamnongthai, K.; Miyanaga, Y. F0 contour approximation model for a one-stream tonal word recognition system. *AEUE Int. J. Electron. Commun.* **2016**, *70*, 681–688. [CrossRef]
44. Chen, P.; Chen, L.; Mao, X. Content Classification With Electroglottograph. *J. Phys. Conf. Ser.* **2020**, *1544*, 012191. [CrossRef]
45. Xiao, Z. An Approach of Fundamental Frequencies Smoothing for Chinese Tone Recognition. *J. Chin. Inf. Process.* **2001**, *15*, 45–50. [CrossRef]
46. Ma, T.; Tian, W.; Xie, Y. Multi-level knowledge distillation for low-resolution object detection and facial expression recognition. *Knowl.-Based Syst.* **2022**, *240*, 108136. [CrossRef]
47. Wu, J.; Hua, Y.; Yang, S.; Qin, H.; Qin, H. Speech Enhancement Using Generative Adversarial Network by Distilling Knowledge from Statistical Method. *Appl. Sci.* **2019**, *9*, 3396. [CrossRef]
48. Chen, H.; Pei, Y.; Zhao, H.; Huang, Y. Super-resolution guided knowledge distillation for low-resolution image classification. *Pattern Recognit. Lett.* **2022**, *155*, 62–68. [CrossRef]
49. Wang, J.; Zhang, P.; He, Q.; Li, Y.; Hu, Y. Revisiting Label Smoothing Regularization with Knowledge Distillation. *Appl. Sci.* **2021**, *11*, 4699. [CrossRef]
50. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
51. Jing, S.; Mao, X.; Chen, L.; Zhang, N. Annotations and consistency detection for Chinese dual-mode emotional speech database. *J. Beijing Univ. Aeronaut. A* **2015**, *41*, 1925–1934. [CrossRef]
52. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, 7–9 May 2015.
53. Russell, J.A.; Barrett, F.L. Core Affect, Prototypical Emotional Episodes, and Other Things Called Emotion: Dissecting the Elephant. *J. Personal. Soc. Psychol.* **1999**, *76*, 805–819. [CrossRef]
54. Van der Maaten, L.; Hinton, G. Viualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
55. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of german emotional speech. In Proceedings of the Interspeech 2005—Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, 4–8 September 2005.