electronics

MDPI

Article

Single- and Cross-Lingual Speech Emotion Recognition Based on WavLM Domain Emotion Embedding

Jichen Yang [1], Jiahao Liu [1], Kai Huang [2], Jiaqi Xia [1], Zhengyu Zhu [1,3,*] and Han Zhang [4,*]

[1] School of Cyber Security, Guangdong Polytechnic Normal University, Guangzhou 510665, China; nisonyoung@163.com (J.Y.); liujiahao919@gmail.com (J.L.); xswnew66@163.com (J.X.)
[2] Aberdeen Institute of Data Science and Artificial Intelligence, South China Normal University, Foshan 528225, China; 20223801065@m.scnu.edu.cn
[3] Guangzhou Quwan Network Technology Co., Ltd., Guangzhou 510665, China
[4] School of Electronics and Information Engineering, South China Normal University, Foshan 528225, China
* Correspondence: zhuzhengyu0701@163.com (Z.Z.); zhanghan@scnu.edu.cn (H.Z.)

Abstract: Unlike previous approaches in speech emotion recognition (SER), which typically extract emotion embeddings from a trained classifier consisting of fully connected layers and training data without considering contextual information, this research introduces a novel approach. It integrates contextual information into the feature extraction process. The proposed approach is based on the WavLM representation and incorporates a contextual transform, along with fully connected layers, training data, and corresponding label information, to extract single-lingual WavLM domain emotion embeddings (SL-WDEEs) and cross-lingual WavLM domain emotion embeddings (CL-WDEEs) for single-lingual and cross-lingual SER, respectively. To extract CL-WDEEs, multi-task learning is employed to remove language information, marking it as the first work to extract emotion embeddings for cross-lingual SER. Experimental results on the IEMOCAP database demonstrate that the proposed SL-WDEE outperforms some commonly used features and known systems, while results on the ESD database indicate that the proposed CL-WDEE effectively recognizes cross-lingual emotions and outperforms many commonly used features.

Keywords: speech emotion recognition; single lingual; cross lingual

Citation: Yang, J.; Liu, J.; Huang, K.; Xia, J.; Zhu, Z.; Zhang, H. Single- and Cross-Lingual Speech Emotion Recognition Based on WavLM Domain Emotion Embedding. *Electronics* **2024**, *13*, 1380. https://doi.org/10.3390/electronics13071380

Academic Editor: Chang Wook Ahn

Received: 21 January 2024
Revised: 30 March 2024
Accepted: 3 April 2024
Published: 5 April 2024

1. Introduction

Speech emotion recognition (SER) is a technology designed to identify and classify the emotional content conveyed through speech. Its primary objective is to accurately discern the emotional state of the speaker, distinguishing between emotions such as happiness, sadness, anger, or neutrality. This technology finds widespread application across various real-world scenarios, including emotion voice conversion [1–4], emotional text-to-speech [5], and speech emotion applications in movie dubbing [6].

Similar to speaker recognition and speech recognition tasks, an SER system typically comprises a front-end feature extractor and a back-end classifier. In the context of SER, feature extraction and classification are two pivotal components that collaborate to accurately recognize and classify emotional content in speech. Feature extraction involves identifying relevant attributes of the speech signal that are most effective in representing the emotional state of the speaker. On the other hand, the classifier refers to the algorithm used to categorize the extracted features into specific emotional categories.

Previous studies have identified several popular feature extraction techniques for SER, including low-level descriptors (LLDs), the mel spectrogram, the wav2vec representation [7], and feature selection based on genetic algorithms [8–11]. LLDs have been used in studies [12–14], while the mel spectrogram has been used in studies such as [15–19]. Wav2vec, on the other hand, has been used in studies [20–25]. A LLD is a combination of features extracted by the openSMILE toolkit [26], which typically includes the zero-crossing

rate, the root-mean-square of the frame energy, the pitch frequency, the harmonics-to-noise ratio, and mel-frequency cepstral coefficients (MFCC). The LLD aims to capture various acoustic characteristics of speech that are relevant to emotion recognition. The mel spectrogram, on the other hand, is a type of spectrogram that is computed using a mel-scale filter bank. This technique is commonly used in speech processing and music analysis, as it is designed to mimic the human auditory system by emphasizing frequencies that are more perceptually relevant. Mel spectrograms have been shown to be effective in capturing both spectral and temporal information in speech, making them a popular choice for feature extraction in SER. Wav2vec is a self-supervised speech representation (S3R) technique that uses waveform data as input under a pre-trained model. This technique is designed to learn representations of speech that are useful for a variety of downstream tasks, including emotion recognition. Wav2vec has shown promising results in recent studies, as it is able to capture both phonetic and acoustic properties of speech.

Unlike LLDs and the mel spectrogram, which fall into the category of handcrafted features and require significant prior knowledge to design effective extraction methods, wav2vec, like other self-supervised speech representation learning (S3R) approaches, only requires ample unlabeled training data and a Transformer encoder [27] to extract representations. In recent years, S3R has gained significant research attention in the speech and audio signal processing field, with applications including automatic speech recognition (ASR) [7,28,29], phoneme classification [30,31], speaker recognition [28,30,31], voice conversion, and SER [20–22], phoneme segmentation [32], and audio classification [33]. Generally, S3R tends to outperform handcrafted features under the same classifier because it can reveal more comprehensive information within speech, which is often not possible with handcrafted features [30]. This is why S3R has become increasingly popular in the speech and audio signal processing community, including for SER applications.

Previous studies have motivated us to investigate the use of S3R features for SER. The first observation is that only wav2vec has been used for SER, despite being initially proposed for ASR and primarily used for downstream tasks related to preserving source speaker content information, such as in the field of voice conversion [34,35]. However, SER not only involves content information but also speaker-related information [18]. Therefore, wav2vec may not necessarily be the best S3R feature for SER. The second observation is that emotion embedding is typically extracted from a trained classifier based on fully connected (FC) layers, emotion training data, and corresponding label information, with S3R as the input, as seen in [20]. Emotion embedding can be extracted from the trained classifier because different emotions can be well classified and discriminated during classifier training. However, contextual information related to emotion is often neglected in previous studies of emotion embedding extraction. Therefore, there is potential to extract better emotion embedding with contextual information from S3R features for SER. The third observation is that no studies have been conducted on cross-lingual SER using S3R features to date. The features commonly used in the community, such as the mel spectrogram [16], usually contain some unhelpful information for SER, such as language information. In contrast, language information may even degrade performance. Therefore, it is expected that emotion embedding without language information extracted from S3R features will yield better performance for cross-lingual SER.

Given that WavLM [28] was initially developed as a large-scale, self-supervised pre-training model for full-stack speech processing, encompassing both ASR and speaker-related tasks such as speaker verification and speaker diarization, it is reasonable to posit that improved emotion embedding can be derived from the WavLM representation for SER. This can be achieved through the incorporation of contextual information, FC, training data, and corresponding label information. To this end, contextual transformation is employed in this study to extract emotion embedding from the WavLM representation. Moreover, single- and cross-lingual emotion embeddings are extracted to facilitate single- and cross-lingual emotion recognition. Multi-task learning is utilized to extract cross-lingual emotion

embedding by eliminating language information, as it is irrelevant for cross-lingual SER and can be expected to yield promising performance outcomes.

The contribution of the work can be summarized as:

- Firstly, contextual transformation has been applied for the first time in the field of emotion embedding extraction for SER.
- A novel single-lingual WavLM domain emotion embedding (SL-WDEE) is proposed for single-lingual speech emotion recognition. This is achieved by combining an emotional encoder and an emotion classifier at the base of the WavLM representation. The emotional encoder is used to encode the input WavLM representation, while the emotion classifier is employed in the training stage to classify the emotion. The emotion encoder comprises a contextual transformation module, two FCs, and corresponding sigmoid modules.
- A novel cross-lingual WavLM domain emotion embedding (CL-WDEE) is proposed for cross-lingual speech emotion recognition. This is achieved by utilizing multi-task learning from the WavLM representation to extract emotion embedding and simultaneously remove the language information. The CL-WDEE extractor is realized by combining a shared encoder, an emotion encoder, a language encoder, an emotion classifier, and a language classifier. The shared encoder is used to encode the input WavLM representation, while the emotion encoder and the language encoder are employed to encode the shared feature obtained from the shared encoder to extract the CL-WDEE and WavLM domain language embedding (WDLE), respectively. Both the emotion encoder and the language encoder consist of contextual transformation modules, FCs, and sigmoid modules. The emotion classifier and the language classifier are used to classify emotion and language in the training stage, respectively.

The rest of the paper is organized as follows: Section 2 introduces WavLM, and Section 3 introduces the WavLM domain emotion embedding extraction. The experimental result and analysis are given in Section 4, and the conclusion is given in Section 5.

2. WavLM

In this section, we provide an overview of WavLM, including its structure and denoising masked speech modeling.

WavLM is a model that learns universal speech representations from a vast quantity of unlabeled speech data. It has been shown to be effective across multiple speech processing tasks, including both ASR and non-ASR tasks. The framework of WavLM is based on denoising masked speech modeling, where some inputs are simulated to be noisy or overlapped with masks, and the target is to predict the pseudo-label of the original speech masked region. This approach enables the WavLM model to learn not only ASR-related information but also non-ASR knowledge during the pre-training stage [28].

The model architecture of WavLM is depicted in Figure 1, consisting of two key components for encoding the input data. The first component is a CNN encoder, and the second component is a Transformer encoder, which serves as the backbone of WavLM. The output of the first component serves as the input to the second component. The first component comprises seven blocks of temporal convolutional layers with layer normalization and a GELU activation layer. The temporal convolutions utilize 512 channels with strides (5,2,2,2,2,2,2) and kernel widths (10,3,3,2,2,2,2) [28]. The second component is equipped with a convolution-based relative-position embedding layer with a kernel size of 128 and 16 groups at the bottom. Additionally, a gated relative-position bias is employed to enhance the performance of WavLM [28].

To enhance the robustness of the model to complex acoustic environments and to preserve speaker identity, denoising masked speech modeling has been proposed for WavLM [28]. To achieve this, the utterance mixing strategy is utilized to simulate noisy speech with multiple speakers and various background noises during self-supervised pre-training, particularly when only single-speaker pre-training data are available. Moreover, some utterances from each training batch are chosen at random to generate noisy speech.

These utterances are then mixed with either a randomly selected noise audio or a secondary utterance at a randomly chosen region.

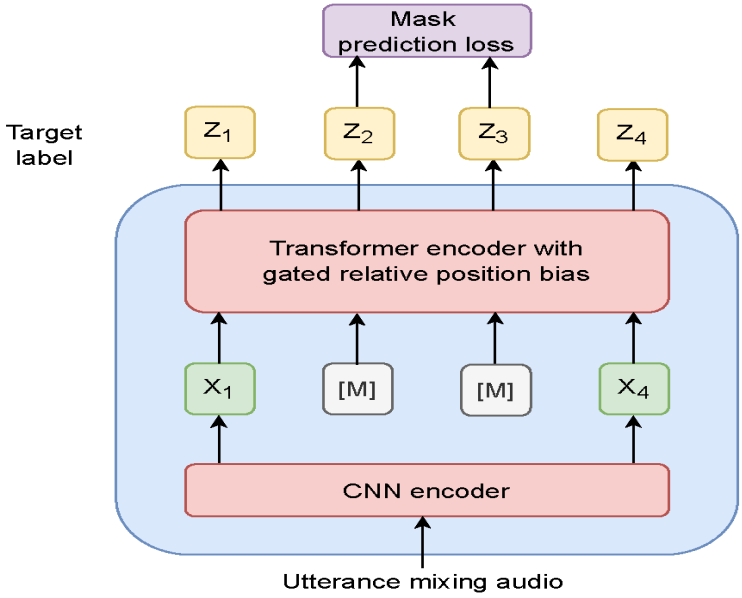

Figure 1. Model architecture of WavLM [28].

3. WavLM Domain Emotion Embedding Extraction

In this section, we introduce the detailed process of extracting the SL-WDEE and CL-WDEE from the WavLM representation, respectively.

3.1. SL-WDEE

The framework of the proposed SL-WDEE extraction method in the training stage is depicted in Figure 2. Here, SL-waveform and SL-WDEE refer to a single-lingual waveform and a single-lingual WavLM domain emotion embedding, respectively. The framework comprises one WavLM pre-trained model, one emotional encoder, and one emotion classifier for the extraction of the SL-WDEE. The emotional encoder comprises the modules of normalization, contextual transformation, two FCs, and one sigmoid module. The emotion classifier only contains one sigmoid, one FC, and one softmax module.

The modules utilized in the proposed SL-WDEE extraction framework play different roles.

- The WavLM pre-trained model is responsible for converting the input SL-waveform into a WavLM representation, which serves as the input for the emotional encoder. The normalization module is utilized to normalize the WavLM representation.
- The contextual transformation module is used to transform the input frame-by-frame information into contextual frame information. Specifically, for each frame, the current frame, its left five frames, and its right five frames are used to form contextual frames. Thus, every input frame information is transformed into 11-frame information by using the contextual transformation.
- The FC module is employed to apply a linear transformation to the input data.
- The sigmoid module is utilized to prevent the generation of values that are too large due to the FC module and transform the input into a range between 0 and 1. For example, given an input x, its sigmoid is as follows:

$$f(x) = Sigmoid(x) = \frac{exp(x)}{1 + exp(x)}, \tag{1}$$

where $f(x)$ is the sigmoid of x.

- The softmax module is used to convert the input into a probability, for instance, given an input $Y = \{y_1, ..., y_N\}$, the softmax of y_i ($i = 1, 2, ..., N$) is as follows:

$$softmax(y_i) = \frac{exp(y_i)}{\sum_{k=1}^{N} exp(y_k)}. \quad (2)$$

In the inference stage, the SL-WDEE can be extracted from the emotional encoder by feeding the input SL-waveform into the WavLM pre-trained model and then into the emotional encoder. In this stage, the output of the emotion classifier is not considered, as it is only used for training.

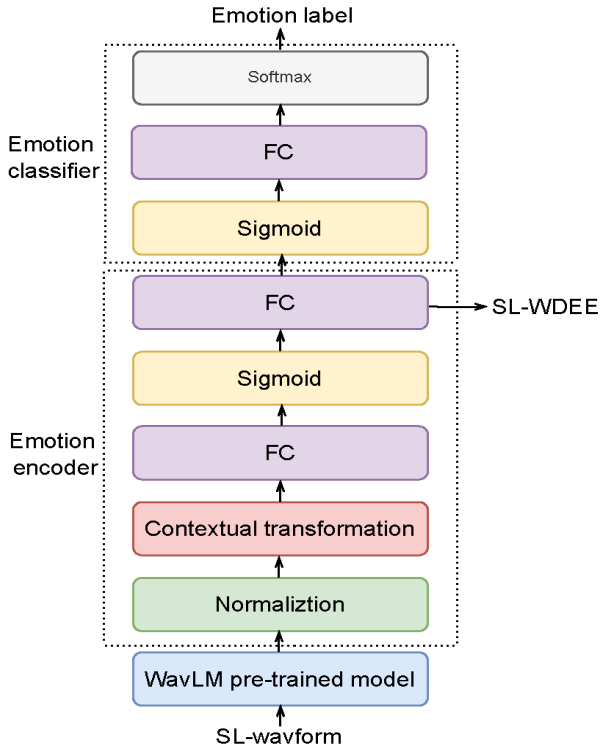

Figure 2. The architecture of the single-lingual WavLM domain emotion embedding (SL-WDEE) extraction in the training stage.

3.2. CL-WDEE

The proposed framework for CL-WDEE extraction based on multi-task learning is illustrated in Figure 3. As depicted in the figure, the framework comprises three encoders and two classifiers.

The three encoders are **the shared encoder**, **the emotion encoder**, and **the language encoder**. Each encoder serves a different purpose, with the **shared encoder** being utilized for all tasks, and the **emotion and language encoders** being specifically designed for emotion classification and language identification, respectively. The differences among the three encoders are as follows:

- **In terms of modules, the shared encoder consists of four modules**, whereas both the emotion encoder and the language encoder consist of three modules. Specifically, the shared encoder contains the *normalization module*, *the contextual transformation module*, *the fully connected (FC)* module, and *the sigmoid module*. Conversely, **the emotion encoder and the language encoder** comprise *two FC modules* and *one sigmoid* module. It should be noted that each module in Figure 3 serves the same function as that in Figure 2.
- **From a functional perspective**, the shared encoder is responsible for extracting shared features that are utilized by both the emotion encoder and the language encoder. The

emotion encoder and the language encoder, on the other hand, are used to encode the shared features and extract the CL-WDEE and WDLE, respectively.

- **The emotion classifier and the language classifier share the same architecture**, which consists of one sigmoid module, one FC module, and one softmax module. Nevertheless, their roles differ, with the emotion classifier being utilized to classify emotions, and the language classifier being employed to classify languages.

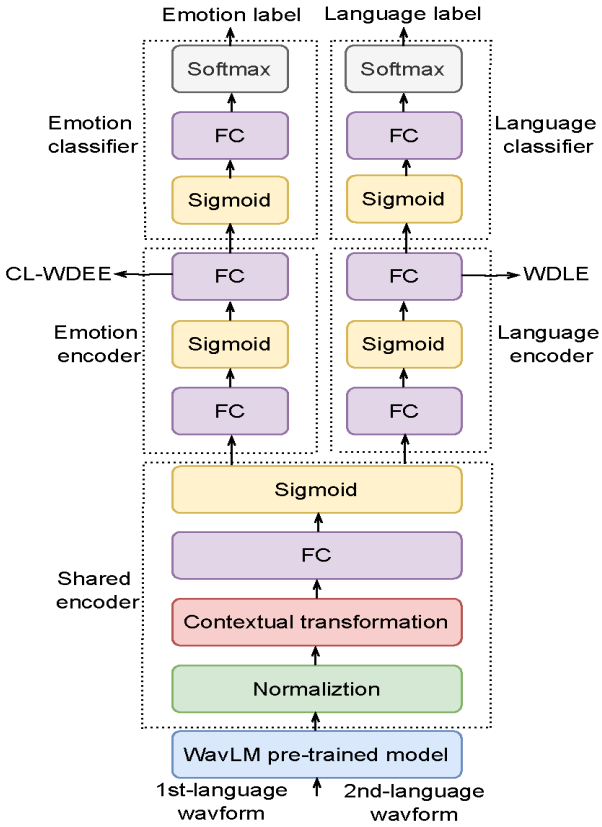

Figure 3. The framework of the proposed cross-lingual WavLM domain emotion embedding (CL-WDEE) extraction based on multi-task learning in the training stage.

During the inference stage, the input cross-lingual waveform is processed through the emotion encoder to extract the CL-WDEE, with the outputs of the emotion classifier, language classifier, and language encoder being disregarded. This is because the outputs of the emotion classifier, language classifier, and language encoder are not relevant for SER, whereas the output of the emotion encoder, i.e., the CL-WDEE, is crucial for SER.

When comparing the extraction of the SL-WDEE in Figure 2 and that of the CL-WDEE in Figure 3, several conclusions can be drawn,

- **The common ground** between them is that both SL-WDEE and CL-WDEE are extracted from the WavLM domain, and that the contextual transformation, FC, and sigmoid modules are utilized in their extraction.
- **The main difference** between them lies in the fact that multi-task learning is employed for the CL-WDEE to eliminate language information with the aid of the language encoder, as depicted in Figure 3. Conversely, there is no need to eliminate language information in the extraction of the SL-WDEE, as shown in Figure 2.
- **The structure of the two extraction methods differs,** with the SL-WDEE extraction consisting of two parts, namely the emotion encoder and the emotion classifier, while the CL-WDEE extraction comprises five parts, which are the emotion encoder, the emotion classifier, the shared encoder, the language encoder, and the language classifier, respectively.

4. Evaluations and Analysis

This section presents the evaluation of the SL-WDEE and CL-WDEE on various databases, with ResNet serving as the classifier. The following subsections provide details regarding the databases used, the experimental setup, the obtained results, and the corresponding analysis.

4.1. Dataset

The proposed SL-WDEE and CL-WDEE were evaluated using the Interactive Emotional Dyadic Motion Capture Dataset (IEMOCAP) [36] and the Emotion Speech Dataset (ESD) [37]. The selection of these datasets was based on the fact that IEMOCAP is the most commonly utilized dataset in the domain of single-lingual SER, while ESD is a parallel English and Chinese emotion dataset that can be utilized for cross-lingual SER.

The IEMOCAP dataset consists of five parts, each of which comprises scripted and impromptu dialogues between two professional male and female actors. The corpus includes a total of nine emotions, namely happy, neutral, angry, sad, excited, fearful, surprised, disgusted, and frustrated. For this study, we follow previous works [13,38] and select only four emotions, namely happy, neutral, angry, and sad. This is because the emotion of being happy is similar to that of being excited, and prior studies such as [13,38] often combine them to increase the number of happy utterances. As a result, the IEMOCAP database comprises 5531 utterances, with the number of neutral, happy, angry, and sad utterances being 1708, 1636, 1103, and 1084, respectively.

The Emotion Speech Dataset (ESD) comprises two parts, namely ESD-Eng for English emotional data and ESD-Chi for Chinese emotional data. Each part consists of 10 speakers, with each speaker having 1750 parallel utterances in five different emotions, namely, neutral, happy, angry, sad, and surprised. This results in a total of 17,500 utterances per part. Additionally, for ESD-Eng, the 1750 utterances have a total word count of 11,015 words and 997 unique lexical words, while for ESD-Chi, the total character count is 20,025 Chinese characters with 939 unique Chinese characters [37].

The summary of IEMOCAP and ESD is given in Table 1.

Table 1. The summary of IEMOCAP and ESD, where ESD has an English part (ESD-Eng) and a Chinese part (ESD-Chi).

Language	Emotion	#IEMOCAP	#ESD		
			Training	Test	Eva
English	Angry	1103	3000	300	200
	Happy	1636	3000	300	200
	Neutral	1708	3000	300	200
	Sad	1084	3000	300	200
	Surprise		3000	300	200
Chinese	Angry		3000	300	200
	Happy		3000	300	200
	Neutral		3000	300	200
	Sad		3000	300	200
	Surprise		3000	300	200

From Table 1, it can be found that both ESD-Eng and ESD-Chi have a training set (Tra), a development set (Dev), and an evaluation set (Eva), respectively. Furthermore, the utterance numbers of Tra, Dev, and Eva are 15,000, 1500, and 1000 in ESD-Eng and ESD-Chi, respectively.

4.2. Experimental Setup

4.2.1. Pre-Trained WavLM Model

The WavLM network was trained on the train-clean-360 subset of the LibriTTS corpus [39] using the same settings as in [30]. The model consisted of a six-layer Transformer encoder with 768 hidden units in each layer, a feed-forward layer comprising 3072 neurons, and 12 attention heads. For further details, please refer to [30].

4.2.2. The Structure of ResNet

In our experiments, all ResNet-based classifiers followed the classic structure of ResNet as constructed in [40,41]. The input features were initially processed, and their shapes were adjusted via a convolutional layer. Subsequently, seven residual blocks followed the first convolutional layer, with each residual block comprising two convolutional layers with a kernel size of 3×7. The input feature of each block was added to the output feature of the block to mitigate the vanishing gradient problem during the training phase. It is worth noting that except for the first two residual blocks, all the other blocks downsampled the feature maps with convolutional strides of $(2, 4)$. The output feature maps from the last block were converted to 128-dimensional features via the adaptive max-pooling layer. The resulting 128-dimensional feature was fed into two fully connected layers, and the final result was obtained from the output of the softmax function. Additionally, the activation function utilized in all residual blocks was the Leaky ReLU, and bottleneck layers were set behind all the activation functions. Moreover, the cross-entropy loss was selected as the loss criterion, and Adam was applied as the optimizer with a momentum of 0.9 and a learning rate of 0.0001.

4.2.3. Evaluation Metric

As with previous works [13,38], the unweighted accuracy (UA), denoting the average accuracy of all emotions, and the weighted accuracy (WA), denoting the overall accuracy, were selected as the evaluation metrics for single-lingual SER on the IEMOCAP corpus and cross-lingual SER on ESD, respectively. The definitions of UA and WA are as follows:

$$UA = \frac{\sum_{k=1}^{K} \frac{x_k}{S_k}}{N}, \tag{3}$$

$$WA = \frac{\sum_{k=1}^{K} x_k}{\sum_{k=1}^{K} S_k}, \tag{4}$$

where x_k represents the number of correctly recognized utterances in the k_{th} emotion category, S_k stands for the total number of utterances in the k_{th} emotion category, and N is the total number of emotion categories.

4.2.4. Experimental Method

The IEMOCAP dataset does not have a separate training, development, and evaluation set. Therefore, following previous work [13,38], a 10-fold cross-validation was performed, and the final performance score was obtained by taking the average of the results. On the other hand, for the ESD dataset, the training set was used to train the model, and the test set and evaluation set were used to evaluate the performance at the test and evaluation stages, respectively.

4.3. Studies on IEMOCAP

4.3.1. The Role of Contextual Transformation

As mentioned earlier, the contextual transformation module in the SL-WDEE plays a crucial role in extracting contextual information. To investigate its role, we removed the contextual transformation module from Figure 2 and obtained a modified feature named SL-WDEE-w/o-CT, where w/o and CT represent without and contextual transformation, respectively. We then compared the performance of SL-WDEE-w/o-CT and SL-WDEE on IEMOCAP using ResNet as the classifier. The experimental results in terms of UA and WA are presented in Table 2.

As shown in Table 2, the SL-WDEE outperformed SL-WDEE-w/o-CT under the ResNet classifier. Specifically, the UA of the SL-WDEE was increased by 1.82%, and the WA of the SL-WDEE was increased by 2.35% compared to SL-WDEE-w/o-CT. This indicates that the CT module is crucial in extracting contextual information. The CT module concatenates the current frame with its left five frames and right five frames, effectively transforming the

short-time window into a long-range window. This allows for the extraction of long-range SER features from short-time windows. These findings confirm our hypothesis that the CT module plays an essential role in the extraction of the SL-WDEE.

Table 2. Comparison of experimental results between SL-WDEE-w/o-CT and SL-WDEE on IEMOCAP under the ResNet classifier in terms of UA and WA.

Feature	Model	UA (%)	WA (%)
SL-WDEE-w/o-CT	ResNet	69.50	68.44
SL-WDEE		71.32	**70.79**

4.3.2. Confusion Matrix

Table 3 presents the normalized confusion matrix on IEMOCAP using the SL-WDEE and ResNet. From this table, several conclusions can be drawn:

- The emotion types "Angry" and "Sad" have a higher recognition rate than the other two types.
- The emotion types "Angry", "Happy", and "Sad" are mainly confused with "Neutral", with error rates of 10.52%, 15.77%, and 15.04%, respectively. This may be since "Neutral" is the closest emotion type to "Angry", "Happy", and "Sad". Moreover, we observe that "Happy", "Sad", and "Angry" are ranked as the first, second, and third nearest distances to "Neutral", respectively.
- The emotion type "Neutral" is mostly confused with "Happy" and "Sad", with error rates of 15.05% and 11.77%, respectively. This is because "Happy" and "Sad" are the first and second nearest distances to "Neutral", respectively.

Table 3. Normalized confusion matrix on IEMOCAP using SL-WDEE and ResNet.

	Angry	Happy	Neutral	Sad
Angry	**0.7697**	0.0952	0.1052	0.0299
Happy	0.0819	**0.6907**	0.1577	0.0697
Neutral	0.0656	0.1505	**0.6663**	0.1177
Sad	0.0371	0.0864	0.1504	**0.7261**

4.3.3. Comparison with Other Domains' Emotion Embedding

As mentioned earlier, the proposed SL-WDEE was obtained from the WavLM domain with the help of the emotion encoder and the training data. However, the mel spectrogram is the most widely used feature in the field of SER, and wav2vec 2.0 has also been used for emotion recognition. Therefore, we compared the performance of the proposed WavLM domain with mel and wav2vec 2.0 domain features on IEMOCAP. To do so, the WavLM pre-trained model module in Figure 2 was first replaced by mel-spectrogram extractors and the wav2vec 2.0 pre-trained model, respectively. Then, the same training data used for training the SL-WDEE extractor were used to train them. Finally, the features obtained from the emotion encoder were then named as single-lingual mel-domain emotion embedding (SL-MDEE) and single-lingual wav2vec 2.0 domain emotion embedding (SL-W2DEE), respectively. The experimental results comparison between the SL-MDEE and SL-WDEE (SL-W2DEE) on IEMOCAP using the ResNet classifier in terms of UA and WA is shown in Table 4.

Table 4 demonstrates that the SL-WDEE achieved better performance than the SL-W2DEE under the ResNet classifier in terms of both UA and WA. This suggests that the WavLM representation can extract more emotion-related information from speech signals compared to W2V2. This can be attributed to the fact that emotions are not only dependent on the content of speech but also on the speaker's characteristics. Additionally, the WavLM representation performs well in both speech recognition and speaker recognition tasks, while wav2vec 2.0 focuses mainly on speech recognition.

Table 4. Comparison of experimental results between SL-MDEE, SL-WDEE, and SL-W2DEE on IEMOCAP using the ResNet classifier in terms of UA and WA.

Domain	Feature	UA (%)	WA (%)
Mel	SL-MDEE	52.37	53.60
W2V2	SL-W2DEE	62.85	62.03
WavLM	SL-WDEE	71.32	70.79

Furthermore, we can observe that both SL-WDEE and SL-W2DEE significantly outperformed the SL-MDEE in terms of UA and WA. This may be because they use different inputs to extract features, and both WavLM representation and wav2vec 2.0 are self-supervised features that are learned from large quantities of unlabeled data, while the mel spectrogram is a handcrafted feature. This finding confirms that self-supervised features can provide more emotional information compared to handcrafted features.

4.3.4. Comparison with Some Known Systems

Table 5 presents the experimental results of the proposed method compared with some known systems on IEMOCAP in terms of UA and WA. In this table, GCN [13] represents a graph convolutional network, and GCN-line and GCN-cycle represent the frame-to-node transformation of the graph construction strategy. DRN stands for a dilated residual network [38], while STL-W2V2-FC and MTL-W2V2-FC denote single-task-learning and multi-task-learning for SER using fully connected (FC) layers, respectively.

Table 5. Experimental results of the proposed method compared with some known systems on IEMOCAP in terms of UA and WA.

System	Feature	Model	UA (%)	WA (%)
LLD-GCN-line [13]	LLD	GCN-line	64.69	61.14
LLD-GCN-cycle [13]	LLD	GCN-cycle	65.29	62.27
LLD-DRN [38]	LLD	DRN	67.40	67.10
STL-W2V2-FC [38]	W2V2	FC	65.11	62.68
MTL-W2V2-FC [38]	W2V2	FC	70.82	68.29
SL-WDEE-ResNet	SL-WDEE	ResNet	71.32	70.79

It is evident from Table 5 that the proposed **SL-WDEE**-ResNet outperforms the other systems in terms of UA and WA on IEMOCAP. This suggests that the proposed system has superior SER capabilities, which can be attributed to the use of the SL-WDEE as input to our system. Furthermore, this finding confirms the effectiveness of the proposed SL-WDEE representation for SER.

4.4. Studies on ESD

4.4.1. Experimental Results and Analysis

Table 6 presents the experimental results on ESD in terms of UA and WA. For this experiment, the CL-WDEE and ResNet were used as the feature representation and classifier, respectively.

Table 6. Experimental results on the development (Dev) and evaluation (Eva) sets of ESD using CL-WDEE and ResNet in terms of UA (%) and WA (%).

Feature	Model	ESD	Dev		Eva	
			UA	WA	UA	WA
CL-WDEE	ResNet	ESD-Eng	91.60	91.60	88.50	88.50
		ESD-Chi	95.60	95.60	91.00	91.00

Table 6 reveals that the UA and WA were the same on the development (evaluation) set of ESD-Eng (ESD-Chi), whereas they differed in the experimental results on the IEMOCAP dataset. This is because each emotion type has the same number of utterances in the development (evaluation) set of ESD. Furthermore, we observe that the performance of the ESD-Eng (ESD-Chi) development set was slightly better than that of the evaluation set. This may be because some similar emotion types in the development set have appeared in the training set, thereby facilitating better recognition. Finally, we note that the performance of ESD-Chi was slightly better than that of ESD-Eng, suggesting that the recognition of emotions in ESD-Chi is relatively easier than that in ESD-Eng.

4.4.2. Confusion Matrix

Table 7 displays the confusion matrix on the ESD evaluation sets using the CL-WDEE and ResNet. The table reveals that the "Sad" emotion category had the highest recognition rate, while the "Happy" and "Surprise" categories had the lowest recognition rates in the ESD-Eng evaluation set. Moreover, we observe the following misclassifications in the evaluation set:

- For "Sad" recognition, there were six, three, and two utterances that were wrongly recognized as "Neutral", "Happy", and "Angry", respectively.
- For "Surprise" recognition, there were 19, 11, and 2 utterances that were wrongly recognized as "Happy", "Angry", and "Sad", respectively.
- For "Happy" recognition, there were 15, 7, and 3 utterances that were wrongly recognized as "Angry", "Neutral", and "Sad", respectively.
- For "Angry" recognition, there were 15, 4, 2, and 1 utterances that were wrongly recognized as "Neutral", "Happy", "Surprise", and "Sad", respectively.
- For "Neutral" recognition, there were 12, 2, 1, and 1 utterances that were wrongly recognized as "Sad", "Angry", "Happy", and "Surprise", respectively.

Table 7. Confusion matrix on ESD evaluation sets using CL-WDEE and ResNet.

Subsets	Emotion	Angry	Happy	Neutral	Sad	Surprise
ESD-Eng	Angry	178	4	15	1	2
	Happy	15	168	7	3	7
	Neutral	2	1	182	14	1
	Sad	2	3	6	189	0
	Surprise	11	19	0	2	168
ESD-Chi	Angry	186	5	0	1	8
	Happy	6	173	1	0	20
	Neutral	1	1	198	0	0
	Sad	1	0	9	190	0
	Surprise	2	36	1	1	160

In contrast to the confusion matrix of the ESD-Eng evaluation set, the recognition rates of "Neutral", "Sad", and "Angry" emotions were higher and equal to 93% in the ESD-Chi evaluation set. Nearly all "Neutral" utterances were correctly recognized, while the recognition rates of "Sad" and "Angry" emotions were the second and third highest, respectively, despite 10 and 14 wrongly recognized utterances. However, the recognition rate of the "Surprise" emotion was the lowest, with 40 utterances that were wrongly recognized, obtaining the worst performance.

4.4.3. Investigation of the Role of Language Information

As mentioned earlier, multi-task learning (MTL) is crucial in removing language-specific information and obtaining a CL-WDEE. We were interested in investigating the role of language information in cross-lingual SER. To this end, we used the SL-WDEE, as shown in Figure 2, to extract features using the training data from ESD-Eng and ESD-Chi. Since the inputs were in two languages, the obtained feature were named as WDEE. Table 8 presents

the experimental results on the evaluation sets of ESD using ResNet as the classifier, in terms of UA and WA, for both WDEE and CL-WDEE.

Table 8. Experimental results on the evaluation sets of ESD between WDEE and CL-WDEE in terms of UA (%) and WA (%).

Feature	Model	ESD	Eva	
			UA (%)	WA (%)
WDEE	ResNet	ESD-Eng	87.60	87.60
		ESD-Chi	90.60	90.60
CL-WDEE		ESD-Eng	88.50	88.50
		ESD-Chi	91.00	91.00

The results in Table 8 indicate that the CL-WDEE outperformed WDEE in terms of UA and WA. This suggests that removing language-specific information using the MTL approach is beneficial for cross-lingual SER. The performance of WDEE is lower, indicating that language-specific information plays a crucial role in recognizing emotions from speech signals. Overall, the experimental results demonstrate the importance of removing language-specific information for achieving better performance in cross-lingual SER. Note that UA equals WA in Table 8, the reason being that the utterance number of every emotion class is the same in ESD-Chin and ESD-Eng.

4.4.4. Comparison with WavLM Representation

We aimed to compare the performance of our proposed CL-WDEE with the WavLM representation (WLMR) on the evaluation sets of ESD in terms of UA or WA. Since we did not have any prior knowledge of the language of the test utterance, we had to consider all scenarios where the models were trained on different training data from ESD. Since there were two types of training data, namely ESD-Eng and ESD-Chi, we trained three models in total, including (i) ESD English (ESD-Eng) training data, (ii) ESD Chinese (ESD-Chi) training data, and (iii) ESD English combined with Chinese (ESD-EngChi) training data. Table 9 presents the experimental results on the evaluation sets of ESD using different training data from ESD for both WLMR and CL-WDEE, with the ResNet classifier, in terms of UA and WA.

Table 9. Experimental results on the evaluation sets of ESD using different training data from ESD for WLMR and CL-WDEE, with the ResNet classifier in terms of UA (%) and WA (%).

Scenario	Feature	Training Data	Eva Data	UA	WA
1	WLMR	ESD-Eng	ESD-Eng	86.90	86.90
2			ESD-Chi	49.90	49.90
3		ESD-Chi	ESD-Eng	45.90	45.90
4			ESD-Chi	90.00	90.00
5		ESD-EngChi	ESD-Eng	84.20	84.20
6			ESD-Chi	88.00	88.00
7	CL-WDEE	ESD-EngChi	ESD-Eng	88.50	88.50
8			ESD-Chi	91.00	91.00

From Table 9, we can draw several conclusions:

1. Good results can usually be obtained when the training data and the test utterances are in the same language, as seen in scenarios 1 and 4. The model trained on ESD-Eng performed well on English utterances, while the model trained on ESD-Chi performed well on Chinese utterances. However, the performance was poor when there was a language mismatch between the training data and test utterances, as observed in scenarios 2 and 3. This is because there is a significant gap between the trained model

and the test utterance when the language does not match. In other words, language can be regarded as a domain, and language mismatch leads to a domain shift.

2. When the WLMR was used as input, the model trained on ESD-EngChi performed better for evaluating Chinese utterances, while it performed slightly worse for evaluating English utterances, compared to the models trained on ESD-Eng or ESD-Chi. This may be due to the fact that the ESD-EngChi training data contained both English and Chinese utterances, making the model more robust to language variations.
3. The CL-WDEE outperformed the WLMR in all scenarios, as seen in the comparison between scenarios 7 and 1, 3, 5, and between scenarios 8 and 2, 4, 6, in terms of UA and WA. This is because the CL-WDEE removes language-specific information, which is known to negatively impact cross-lingual SER performance. These results confirm the importance of removing language information for achieving better cross-lingual SER performance.

4.4.5. Comparison with Other Domains' Emotion Embedding

The proposed CL-WDEE was derived from the WavLM domain by utilizing a shared encoder, emotion encoder, language encoder, and training data. In this study, we aimed to evaluate the performance of the proposed features obtained from the WavLM domain against mel and wav2vec 2.0 domains features on ESD. To achieve this objective, we replaced the WavLM pre-trained model module in Figure 3 with mel-spectrogram extractors (wav2vec 2.0 pre-trained model) and obtained features from the emotion encoder. We named these features as cross-lingual mel-domain emotion embedding (CL-MDEE) and cross-lingual wav2vec 2.0 domain emotion embedding (CL-W2DEE), respectively. Table 10 presents the experimental results comparison between the CL-MDEE and CL-WDEE (CL-W2DEE) on ESD with the ResNet classifier in terms of UA and WA.

Table 10. Experimental results comparison between CL-WDEE and CL-MDEE (CL-W2DEE) on ESD with the ResNet classifier in terms of UA (%) and WA (%).

Domain	Feature	Eva Data	UA	WA
Mel	CL-MDEE	ESD-Eng	82.50	82.50
		ESD-Chi	84.00	84.00
W2V2	CL-W2DEE	ESD-Eng	86.80	86.80
		ESD-Chi	90.70	90.70
WavLM	CL-WDEE	ESD-Eng	88.50	88.50
		ESD-Chi	91.00	91.00

As evident from the results presented in Table 10, the CL-WDEE surpassed CL-M2DEE (CL-MDEE) in terms of UA or WA for the evaluation sets of ESD-Eng and ESD-Chi with the ResNet classifier. This observation suggests that the WavLM representation can extract more emotional information than the W2V2 (mel-spectrogram) representation. One possible explanation for this could be that emotions are not solely related to content, but also to the speaker's characteristics. Additionally, the WavLM representation has shown superior performance in speech recognition and speaker recognition, whereas wav2vec 2.0 only performs well in speech recognition.

Furthermore, it is worth noting that both CL-WDEE and CL-W2DEE demonstrate significantly better performance than CL-MDEE in terms of UA or WA. This is likely because they have different inputs to extract features, and both WavLM representation and wav2vec 2.0 are self-supervised features learned from a large amount of unlabeled data, while Mel-spectrogram is a handcrafted feature. This observation further confirms that self-supervised features can provide more emotion information than handcrafted features.

4.4.6. Comparison with Known System

It should be noted that to date, there have been no reports on ESD for cross-lingual SER. Therefore, a comparison of the proposed (CL-WDEE)-ResNet with other systems for cross-lingual SER on ESD is not feasible. However, ESD has been utilized in previous studies for English SER and Chinese SER [37]. To compare the proposed system's performance with those studies, we present the cross-lingual SER experimental results on SED and corresponding experimental results comparison in Table 11. In this table, LLD features are extracted using the openSMILE toolkit [26], which includes zero-crossing rate, voicing probability, MFCC, and mel-spectrogram. LSTM-FC refers to a LSTM layer followed by a ReLU-activated fully connected layer with 256 nodes [37].

Table 11. Comparison with known systems on ESD in terms of UA (%) and WA (%).

Feature	Model	Training Data	Eva Data	UA	WA
LLD	LSTM-FC	ESD-Eng	ESD-Eng	89.00	89.00
		ESD-Chi	ESD-Chi	92.00	92.00
CL-WDEE	ResNet	ESD-EngChi	ESD-Eng	88.50	88.50
			ESD-Chi	91.00	91.00

As evident from Table 11, the proposed (CL-WDEE)-ResNet performs comparably to LLD-(LSTM-FC) on the evaluation sets of ESD-Eng and ESD-Chi, respectively. It is worth noting that the previous work [37] employed two LSTM-FC models for English SER and Chinese SER, respectively, while the proposed (CL-WDEE)-ResNet model is trained for both ESD-Eng and ESD-Chi. Furthermore, LLD-(LSTM-FC) is designed for single-lingual SER, which can be viewed as a known SER, while (CL-WDEE)-ResNet is designed for cross-lingual SER, which can be viewed as an unknown SER. Therefore, we can conclude that (CL-WDEE)-ResNet has the potential to address the challenge of cross-lingual SER.

5. Conclusions

In summary, this research introduced a novel approach to enhance self-supervised feature-based speech emotion recognition by integrating contextual information. The proposed method leveraged the WavLM domain and contextual cues to extract single-lingual WavLM domain emotion embeddings for single-lingual speech emotion recognition. To tackle the challenge of cross-lingual speech emotion recognition, multi-task learning was employed to remove language-specific information, resulting in the generation of cross-lingual WavLM domain emotion embeddings. An experimental evaluation on the IEMOCAP dataset demonstrated that the proposed approach achieved outstanding performance in recognizing single-lingual speech emotion, attributed to the incorporation of contextual information during feature extraction. Additionally, experimental results on the ESD dataset indicated that the proposed cross-lingual WavLM domain emotion embedding effectively discerned cross-lingual speech emotion and surpassed existing methods. In the future, the proposed method will be further evaluated on challenging datasets such as V2C-Animation [6] to demonstrate its generalizability.

Author Contributions: Conceptualization J.Y. and H.Z.; methodology J.Y. and Z.Z.; writing—review and editing J.L.; software K.H. and J.X.; supervision Z.Z. and H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Science, Technology Program (Key R&D Program) of Guangzhou (2023B01J0004), special projects in key areas of Guangdong Provincial Department of Education (2023ZDZX1006) and Research project of Guangdong Polytechnic Normal University, China (2023SDKYA019).

Data Availability Statement: The data used in this study are public.

Conflicts of Interest: Author Zhengyu Zhu was part time employed by the Guangzhou Quwan Network Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Zhou, K.; Sisman, B.; Li, H. Transforming Spectrum and Prosody for Emotional Voice Conversion with Non-Parallel Training Data. In Proceedings of the Speaker and Language Recognition Workshop (ODYSSEY), Tokyo, Japan, 2–5 November 2020; pp. 230–237.
2. Zhou, K.; Sisman, B.; Zhang, M.; Li, H. Converting Anyone's Emotion: Towards Speaker-Independent Emotional Voice Conversion. In Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH), Incheon, Republic of Korea, 18–22 September 2020; pp. 3416–3420.
3. Zhou, K.; Sisman, B.; Liu, R.; Li, H. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 920–924.
4. Zhou, K.; Sisman, B.; Rana, R.; Schuller, B.W.; Li, H. Emotion Intensity and its Control for Emotional Voice Conversion. *IEEE Trans. Affect. Comput.* **2022**, *14*, 31–48. [CrossRef]
5. Liu, R.; Sisman, B.; Gao, G.; Li, H. Expressive TTS training with frame and style reconstruction loss. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1806–1818. [CrossRef]
6. Chen, Q.; Li, Y.; Qi, Y.; Zhou, J.; Tan, M.; Wu, Q. V2C: Visual voice cloning. *arXiv* **2021**, arXiv:2111.12890v1.
7. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. Wav2vec 2.0: A framework for self-supervised learning of speech representation. In Proceedings of the Annual Conference on Neural Information Processing System, Vancouver, BC, Canada, 6–12 December 2020.
8. Beritelli, F.; Casale, S.; Russo, A.; Serrano, S. A Genetic Algorithm Feature Selection Approach to Robust Classification between "Positive" and "Negative" Emotional States in Speakers. In Proceedings of the IEEE Conference Record of the Thirty-Ninth Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 30 October–2 November 2005; pp. 550–553.
9. Casale, S.; Russo, A.; Serrano, S. Multistyle classification of speech under stress using feature subset selection based on genetic algorithms. *Speech Commun.* **2007**, *49*, 801–810. [CrossRef]
10. Sidorov, M.; Brester, C.; Minker, W.; Semenkin, E. Speech-Based Emotion Recognition: Feature Selection by Self-Adaptive Multi-Criteria Genetic Algorithm. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC), Reykjavik, Iceland, 26–31 May 2014; pp. 3481–3485.
11. Yildirim, S.; Kaya, Y.; Kılıç, F. A modified feature selection method based on metaheuristic algorithms for speech emotion recognition. *Appl. Acoust.* **2021**, *173*, 107721. [CrossRef]
12. Sagha, H.; Deng, J.; Gavryukova, M.; Han, J.; Schuller, B. Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5800–5804.
13. Shirian, A.; Guha, T. Compact graph architecture for speech emotion recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 6284–6288.
14. Jiang, P.; Xu, X.; Tao, H.; Zhao, L.; Zou, C. Convoluitonal-recurrent neural networks with multi attention mechanisms for speech emotion recognition. *IEEE Trans. Cogn. Dev. Syst.* **2022**, *30*, 1803–1814.
15. Chen, M.; He, X.; Yang, J.; Zhang, H. 3-D convolutional recurrent neural networks with Attention model for speech emotion recognition. *IEEE Signal Process. Lett.* **2008**, *25*, 1440–1444. [CrossRef]
16. Cai, X.; Wu, Z.; Zhong, K.; Su, B.; Dai, D.; Meng, H. Unsupervised cross-lingual speech emotion recognition using domain adversarial neural network. In Proceedings of the International Symposium on Chinese Spoken Language Processing (ISCSLP), Hong Kong, China, 24–26 January 2021; pp. 595–602.
17. Fan, W.; Xu, X.; Xing, X.; Chen, W.; Huang, D. LSSED: A large-scale dataset and benechmark for speech emotion recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 641–644.
18. Fan, W.; Xu, X.; Cai, B.; Xing, X. ISNet: Individual standardization network for speech emotion recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 1803–1814. [CrossRef]
19. Li, T.; Wang, X.; Xie, Q.; Wang, Z.; Xie, L. Cross-speaker emotion disentangling and transfer for end-to-end speech synthesis. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 1448–1460. [CrossRef]
20. Cai, X.; Yuan, J.; Zheng, R.; Huang, L.; Church, K. Speecg emotion recognition with multi-task learning. In Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH), Brno, Czech Republic, 30 August–3 September 2021; pp. 4508–4512.
21. Chen, L.W.; Rudnicky, A. Exploring wav2vec 2.0 fine tuning for improved speech emotions recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.

22. Pepino, L.; Riera, P.; Ferrer, L. Emotion recognition from speech using wav2vec 2.0 embeddings. In Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH), Brno, Czech Republic, 30 August–3 September 2021; pp. 3400–3404.
23. Yue, P.; Qu, L.; Zheng, S.; Li, T. Multi-task learning for speech emotion and emotion intensity recognition. In Proceedings of the APISPA Annual Summit and Conference, Chiang Mai, Thailand, 7–10 November 2022; pp. 1232–1237.
24. Liu, M.; Ke, Y.; Zhang, Y.; Shao, W. Speech emotion recognition based on deep learning. In Proceedings of the IEEE Region 10 Conference (TENCON), Hong Kong, China, 1–4 November 2022.
25. Sharma, M. Mutli-lingual multi-task speech emotion recognition using wav2vec 2.0. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 6907–6911.
26. Eyben, F.; Schuller, B. OpenSMILE: The Munich open-source large-scale multimedia feature extractor. *SIGMultimedia* **2015**, *6*, 4–13. [CrossRef]
27. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all your need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 1–11.
28. Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshiola, T.; Xiao, X.; et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1505–1518. [CrossRef]
29. Ravanelli, M.; Zhong, J.; Pascual, S.; Swietojanski, P.; Monteiro, J.; Trmal, J.; Bengio, Y. Multi-task self-supervised learning for robust speech recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6989–6993.
30. Liu, A.T.; Yang, S.; Chi, P.H.; Hsu, P.C.; Lee, H. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6419–6423.
31. Chung, Y.A.; Hsu, W.N.; Tang, H.; Glass, J. An unsupervised autoregressive model for speech representation learning. In Proceedings of the 20nd Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz, Austria, 15–19 September 2019; pp. 146–149.
32. Yang, S.; Liu, A.T.; Lee, H. Understanding self-attention of self-supervised audio transformers. In Proceedings of the 21th Annual Conference of the International Speech Communication Association (INTERSPEECH), Shanghai, China, 14–18 September 2020; pp. 3785–3789.
33. Gong, Y.; Chung, Y.-A.; Glass, J. AST: Audio Spectrogram Transformer. *arXiv* **2021**, arXiv:2014.01778v3.
34. Lin, J.; Lin, Y.Y.; Chien, C.H.; Lee, H. S2VC: A framework for any-to-any voice conversion with self-supervised pretrained representations. In Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH), Brno, Czech Republic, 30 August–3 September 2021; pp. 836–840.
35. Huang, W.C.; Yang, S.W.; Hayashi, T.; Lee, H.Y.; Watanabe, S.; Toda, T. S3PRL-VC: Open-source voice conversion framework with self-supervised speech representations. *arXiv* **2021**, arXiv:2110.06280.
36. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, E.; Povost, E.; King, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
37. Zhou, K.; Sisman, B.; Liu, R.; Li, H. Emotional voice conversion: Theory, database and ESD. *Speech Commun.* **2022**, *137*, 1–18. [CrossRef]
38. Li, R.; Wu, Z.; Jia, J.; Zhao, S.; Meng, H. Dilated residual network with multi-head self-attention for speech emotion recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6675–6679.
39. Zen, H.; Dang, V.; Clark, R.; Zhang, Y.; Weiss, R.J.; Jia, Y.; Chen, Z.; Wu, Y. LibriTTS: A corpus derived from librispeech for text-to-speech. *arXiv* **2019**, arXiv:1904.02882.
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
41. Wu, Q.; Xiong, S.; Zhu, Z. Replay speech answer-sheet on intelligent language learning system based on power spectrum decomposition. *IEEE Access* **2021**, *9*, 104197–104204. [CrossRef]