electronics

Article

Implicitly Aligning Joint Distributions for Cross-Corpus Speech Emotion Recognition

Cheng Lu [1,2,†], Yuan Zong [1,3,*,†], Chuangao Tang [1,3], Hailun Lian [1,2], Hongli Chang [1,2], Jie Zhu [1,2], Sunan Li [1,2] and Yan Zhao [1,2]

[1] Key Laboratory of Child Development and Learning Science, Ministry of Education, Southeast University, Nanjing 210096, China
[2] School of Information Science and Engineering, Southeast University, Nanjing 210096, China
[3] School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China
* Correspondence: xhzongyuan@seu.edu.cn
† These authors contributed equally to this work.

Abstract: In this paper, we investigate the problem of cross-corpus speech emotion recognition (SER), in which the training (source) and testing (target) speech samples belong to different corpora. This case thus leads to a feature distribution mismatch between the source and target speech samples. Hence, the performance of most existing SER methods drops sharply. To solve this problem, we propose a simple yet effective transfer subspace learning method called joint distribution implicitly aligned subspace learning (JIASL). The basic idea of JIASL is very straightforward, i.e., building an emotion discriminative and corpus invariant linear regression model under an implicit distribution alignment strategy. Following this idea, we first make use of the source speech features and emotion labels to endow such a regression model with emotion-discriminative ability. Then, a well-designed reconstruction regularization term, jointly considering the marginal and conditional distribution alignments between the speech samples in both corpora, is adopted to implicitly enable the regression model to predict the emotion labels of target speech samples. To evaluate the performance of our proposed JIASL, extensive cross-corpus SER experiments are carried out, and the results demonstrate the promising performance of the proposed JIASL in coping with the tasks of cross-corpus SER.

Keywords: cross-corpus speech emotion recognition; domain adaptation; transfer subspace learning; marginal distribution; conditional distribution

Citation: Lu, C.; Zong, Y.; Tang, C.; Lian, H.; Chang, H.; Zhu, J.; Li, S.; Zhao, Y. Implicitly Aligning Joint Distributions for Cross-Corpus Speech Emotion Recognition. *Electronics* **2022**, *11*, 2745. https://doi.org/10.3390/electronics11172745

Academic Editors: Andrea Prati, Luis Javier García Villalba and Vincent A. Cicirello

Received: 28 July 2022
Accepted: 29 August 2022
Published: 31 August 2022

1. Introduction

The research of speech emotion recognition (SER) aims at enabling the computer to automatically understand the emotional states of speech signals [1–3], which provides a more natural interaction between the human and computer. Due to this reason, SER research has attracted the wide attention of researchers from the communities of speech signal processing, affective computing, pattern recognition, and human-computer interaction [1–4]. Over the past several decades, many well-performing SER methods have been proposed and achieved promising performance on widely used speech emotion corpora [3,5,6]. However, it is noted that these SER methods do not consider real-world scenarios [7–10], e.g., the training and testing speech samples may be recorded by different microphones, under different environments, expressed by diverse speakers, or belong to different languages. In this case, the training and testing speech samples possibly have inconsistent feature distributions, which remarkably degrades the performance of the SER methods when the model trained on training data cope with the new testing data. Hence, it is meaningful to investigate a more challenging but interesting SER task, in which the SER model is trained on one (or several) dataset(s) and tested on other datasets, i.e., cross-corpus SER.

Different from the ordinary SER, the labeled training data (source domain) and unlabeled testing data (target domain) in cross-corpus SER come from different speech corpora,

following the definitions of transfer learning (TL) and domain adaptation (DA) [9,11–14]. Consequently, corpus bias [9] has widely existed between the source and target domains, caused by the inconsistent feature distribution of speech samples. To deal with the bias issue in cross-corpus SER tasks, researchers have made great efforts in recent years. For example, in [10], which may be the earliest work to formalize the cross-corpus SER standardly, Schuller et al. systematically defined the setting of cross-corpus SER tasks and investigated how to solve this problem from the elimination of the distribution gap between the source and target speech samples. Additionally, they proposed to use a set of normalization schemes including speaker normalization (SC), corpus normalization (CN), and speaker-corpus normalization (SCN) to investigate the cross-corpus SER.

Subsequently, the methods based on TL and DA gradually began to be applied to cope with cross-corpus SER [9,12,15]. In the work of [12], Zong et al. proposed a domain-adaptive least square regression (DaLSR) model guided by a regularization term consisting of one- and second-order moments to learn a corpus-independent regression matrix. Furthermore, Liu et al. [15] presented a simple yet effective transfer subspace learning method called domain-adaptive subspace learning (DoSL) by only considering the one-order moment (i.e., mean value) to measure the distribution gap between the source and target speech samples. More recently, Song et al. [9] investigated a straightforward transfer subspace learning (TSL) model to bridge the feature distribution gap across corpora by resorting to the maximum mean discrepancy (MMD) [16,17].

In addition to traditional subspace learning-based methods, deep learning-based methods [18–22], e.g., convolution neural networks (CNN) and recurrent Neural networks (RNN), have achieved promising performance in cross-corpus SER, taking advantage of their powerful representation capability. Parray et al. [22] investigated the generalization of several deep learning architectures, e.g., CNN, long short-term Memory(LSTM), and CNN-LSTM, on different cross-corpus SER tasks. In the work of [18,19], deep neural networks were embedded into DA to learn the corpus-invariant features for emotional speech. Moreover, the works of [20,21] utilized domain adversarial learning to reduce the domain shift between training and testing data.

Basically, these methods mainly aim to learn a common emotion feature subspace in which the marginal feature distributions of source and target domains are as close as possible. In the cross-database SER, however, this common subspace is incomplete as the emotion features of speech samples are susceptible to background noise, speaker identity information, and language information, leading to feature confusion [8–10]. To maintain the discriminativeness of emotion features, adapting both marginal and conditional distribution (i.e., joint distribution adaptation (JDA), which has achieved success in image classification [14,23,24]) provides a promising method to deal with cross-corpus SER. Zhang et al. [25] proposed a joint distribution adaptive regression (JDAR) to integrate the conditional distribution into MMD for the fine-gained domain shift alignment. Even so, the JDAR is weak in dealing with outlier samples, leading to large domain discrepancies between the training and testing data.

Inspired by the success of the above TL- and DA-based methods, in this paper, we propose a novel method called joint distribution implicitly aligned subspace learning (JIASL) for the cross-corpus SER problem. Unlike these aforementioned methods, the proposed JIASL has three advantages as follows. (1) It absorbs the idea of recent widely-used approach (i.e., JDA) for the distribution gap alignment [14,23,25], which jointly considers the marginal feature distribution and class-aware conditional distribution. (2) More importantly, it adopts a strategy of reconstructing target speech features by the source features to implicitly remove the feature distribution match between the original source and target speech feature sets instead of directly minimizing statistical moments. (3) Meanwhile, it can also restrict the influence of outlier source features for the reconstruction of the target domain by sparse constraints in joint distribution alignment. Guided by the above advantages, JIASL can learn a corpus invariant projection matrix to predict the emotion labels of target speech samples, although it is merely given the source emotion label information. To

evaluate the proposed JIASL, we design the cross-corpus SER tasks based on three publicly available speech emotion corpora, including EmoDB (Berlin) [26], eNTERFACE [27], and CASIA [28], and conduct extensive experiments. Experimental results showed that compared with current state-of-the-art transfer subspace learning methods, the proposed JIASL achieved more promising performance in coping with the cross-corpus SER tasks.

2. The Proposed Method

In this section, we describe the proposed JIASL in detail and provide its optimization algorithm. Then, we also illustrate its application for cross-corpus SER.

2.1. Notations

Herein, we give some important notations that are needed in formulating JIASL for convenient illustration. According to the task setting of cross-corpus SER [9,12], the speech samples and their emotion labels of source domain are provided, while the ones of the target domain have no labels. Therefore, we denoted the feature matrix of source data as $\mathbf{X}^s = [\mathbf{x}_1^s, \cdots, \mathbf{x}_i^s, \cdots, \mathbf{x}_{n_s}^s] \in \mathbb{R}^{d \times n_s}$ and its label matrix as $\mathbf{L}^s = [\mathbf{l}_1, \cdots, \mathbf{l}_i, \cdots, \mathbf{l}_{n_s}] \in \mathbb{R}^{c \times n_s}$, where $\mathbf{x}_i^s \in \mathbb{R}^{d \times 1}$ is the acoustic feature vector of the i^{th} speech sample. Note that we adopt one-hot labels to represent $\mathbf{l}_i$ (i.e., the i^{th} column of $\mathbf{L}^s$), and $\mathbf{l}_i = [l_{i,1}, \cdots, l_{i,k}, \cdots, l_{i,c}]^T$ is a one-hot vector. The value of its k^{th} entry $l_{i,k}$ is set as 1 if its corresponding speech sample belongs to the k^{th} emotion, while the resting entries are all set as 0. Similarly, the feature matrix of target speech samples can be denoted as $\mathbf{X}^t = [\mathbf{x}_1^t, \cdots, \mathbf{x}_j^t, \cdots, \mathbf{x}_{n_t}^t] \in \mathbb{R}^{d \times n_t}$, where $\mathbf{x}_j^t \in \mathbb{R}^{d \times 1}$ is the feature vector of the j^{th} speech sample in the target domain.

2.2. Formulation of JIASL

The basic idea of the proposed JIASL is very straightforward, i.e., building a subspace learning model to learn an **emotion-discriminative** and **corpus-invariant** projection matrix for cross-corpus SER. Following this idea, we design the optimization problem for JIASL as follows:

$$\min_{\mathbf{U},\mathbf{W},\mathbf{W}^{(i)}} (f_1(\mathbf{U}) + \mu f_2(\mathbf{U}, \mathbf{W}, \mathbf{W}^{(k)})), \tag{1}$$

where $\mathbf{U}$ is a such projection matrix that the proposed JIASL model aims to learn, $\mathbf{W}$ and $\mathbf{W}^{(k)}$ are the reconstruction coefficients, whose detail will be given as follows, and $i \in [1, n_s]$ and $k \in [1, c]$ represent the indexes of speech sample and emotion, respectively. μ is the trade-off parameter to control the balance between two terms.

From Equation (1), it can be found that the objective function of our JIASL has two major terms. Both terms actually correspond to the expectative abilities described in the basic idea of JIASL, i.e., **emotion discriminative** and **corpus invariant**. To this end, $f_1(\mathbf{U})$ is designed as a simple group sparse linear regression loss associated with **emotion-discriminative** ability according to [12,15], which can be formulated as follows:

$$f_1(\mathbf{U}) = \|\mathbf{L}^s - \mathbf{U}^T\mathbf{X}^s\|_F^2 + \lambda\|\mathbf{U}^T\|_{2,1}, \tag{2}$$

where λ is the trade-off parameter for the balance across terms. The projection matrix $\mathbf{U}$ aims to regress the source features $\mathbf{X}^s$ to the label space $\mathbf{L}^s$ to preserve the emotion discrimination in feature learning. Meanwhile, the $\ell_{2,1}$ norm on $\mathbf{U}^T$ seeks the features on some specific dimensions contributed to emotion feature learning through group sparse of the whole row elements.

As for the second term, $f_2(\mathbf{U}, \mathbf{W}, \mathbf{W}^{(k)})$, it corresponds to the **corpus-invariant** ability in emotion feature learning. To achieve this goal, we make efforts from two aspects:

First, instead of directly minimizing their statistical moments such as mean value, covariance, and MMD, we raise the idea of implicitly alleviating the feature distribution mismatch between the source and target speech corpora. The advantage of this method is that the estimation of domain shift is not restricted by the discrepancy measurement function but can be obtained gradually through parameter optimization.

Specifically, inspired by the common subspace learning works [12,29], we adopt a strategy of reconstructing the target speech samples by a part of the source samples to enforce the projection matrix learning in JIASL from the sample-level view, which can be formulated as the following sparse optimization problem:

$$\min_{\mathbf{U},\mathbf{w}_j} (\|\mathbf{U}^T\mathbf{x}_j^t - \mathbf{U}^T\mathbf{X}^s\mathbf{w}_j\|^2 + \tau\|\mathbf{w}_j\|_1), \tag{3}$$

where τ is the trade-off parameter, and $\mathbf{w}_j \in \mathbb{R}^{n_s \times 1}$ denotes the j^{th} column of $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_{n_t}] \in \mathbb{R}^{n_s \times n_t}$ and is the reconstruction coefficient vector corresponding to the j^{th} target speech sample $\mathbf{x}_j^t$. The reconstruction strategy aims to narrow the bias between the source and target domains in the common emotion feature subspace. Meanwhile, in order to avoid the interference of outlier samples in the source domain for the reconstruction, the ℓ_1 norm is embedded into $\mathbf{w}_j$ to achieve the sparseness of redundant samples in the source domain. Furthermore, extending the reconstruction to all the target speech samples, we arrive at the total reconstruction optimization problem, as follows:

$$\min_{\mathbf{U},\mathbf{W}} (\|\mathbf{U}^T\mathbf{X}^t - \mathbf{U}^T\mathbf{X}^s\mathbf{W}\|_F^2 + \tau\|\mathbf{W}\|_1), \tag{4}$$

where $\|\mathbf{W}\|_1 = \sum_{i=1}^{n_t} \|\mathbf{w}_j\|_1$.

Second, our JIASL also absorbs the idea of jointly aligning the marginal and class-aware conditional feature distributions (i.e., JDA) to pursue the fine-gained domain alignment, in which the effectiveness of JDA has been demonstrated in dealing with other domain adaptation tasks [23,25]. By incorporating the JDA idea into Equation (4), the objective function of the above designed reconstruction can be extended to the following formulation according to [23,25], which is eventually served as the $f_2(\mathbf{U}, \mathbf{W}, \mathbf{W}^{(k)})$ for JIASL:

$$\begin{aligned} f_2(\mathbf{U}, \mathbf{W}, \mathbf{W}^{(k)}) &= \|\mathbf{U}^T\mathbf{X}^t - \mathbf{U}^T\mathbf{X}^s\mathbf{W}\|_F^2 + \tau\|\mathbf{W}\|_1 \\ &+ \sum_{k=1}^{c} \|\mathbf{U}^T\mathbf{X}^{t(k)} - \mathbf{U}^T\mathbf{X}^{s(k)}\mathbf{W}^{(k)}\|_F^2 + \tau \sum_{k=1}^{c} \|\mathbf{W}^{(k)}\|_1, \end{aligned} \tag{5}$$

where $\mathbf{W}^{(k)} \in \mathbb{R}^{n_s^{(k)} \times n_t^{(k)}}$ is the k^{th} emotion class-aware reconstruction coefficient matrix, $\mathbf{X}^{s(k)} \in \mathbb{R}^{d \times n_t^{(k)}}$ and $\mathbf{X}^{t(k)} \in \mathbb{R}^{d \times n_t^{(k)}}$, and $n_s^{(k)}$ and $n_t^{(k)}$ denote the numbers of speech sample from the k^{th} emotion class satisfying $n_s^{(1)} + \cdots + n_s^{(c)} = n_s$ and $n_t^{(1)} + \cdots + n_t^{(c)} = n_t$. Similar to $\mathbf{W}$, $\|\mathbf{W}^{(k)}\|_1$ can also be rewritten as $\|\mathbf{W}^{(k)}\|_1 = \sum_{j=1}^{n_t^{(k)}} \|\mathbf{w}_j^{(k)}\|_1$, where $\mathbf{w}_j^{(k)}$ are the j^{th} column in $\mathbf{W}^{(k)}$.

Finally, by substituting Equations (2) and (5) into Equation (1), the total optimization function of the proposed JIASL can be rewritten as follows:

$$\begin{aligned}\min_{\mathbf{U},\mathbf{W},\mathbf{W}^{(k)}} (&\|\mathbf{L}^s - \mathbf{U}^T\mathbf{X}^s\|_F^2 + \lambda\|\mathbf{U}^T\|_{2,1} \\ &+\mu[(\|\mathbf{U}^T\mathbf{X}^t - \mathbf{U}^T\mathbf{X}^s\mathbf{W}\|_F^2 + \tau\|\mathbf{W}\|_1) \\ &+(\sum_{k=1}^{c}\|\mathbf{U}^T\mathbf{X}^{t_{(k)}} - \mathbf{U}^T\mathbf{X}^{s_{(k)}}\mathbf{W}^{(k)}\|_F^2 + \tau\sum_{k=1}^{c}\|\mathbf{W}^{(k)}\|_1)]),\end{aligned} \tag{6}$$

where λ, τ, and μ are all trade-off coefficients to balance the different regularization terms, and the details are illustrated in the experiment section.

2.3. JIASL for Cross-Corpus SER

After constructing the JIASL model, we implement it to the cross-corpus SER task; the details are described as follows. Given the speech features of labeled source and unlabeled target domains, we firstly optimize Equation (6) to achieve the optimal projection matrix $\tilde{\mathbf{U}}$ corresponding to $\mathbf{U}$. Once the $\tilde{\mathbf{U}}$ is learned, we can conveniently predict the target emotion labels. Specifically, suppose we have a speech feature vector denoted by $\mathbf{x}_t^{te}$ from target corpora. Then, its emotion label is determined according to the following criterion:

$$emotion\_label = \arg\min_k \{y_t^{(k)}\}\ (k = 1, \cdots, c), \tag{7}$$

where $y_t^{(k)}$ is the k^{th} entry of label vector $\mathbf{y}_t^{te} = \tilde{\mathbf{U}}^T\mathbf{x}_t^{te}$.

2.4. Optimization of JIASL

Since the optimization function, Equation (6), contains several complex regularization terms, e.g., the ℓ_1 norm and $\ell_{2,1}$ norm, its closed-form solution can not be solved directly. Hence, we adopt the alternating direction method (ADM) [30] to optimize the proposed JIASL according to [12,31]. In detail, as the target label information is unknown, we firstly need to initialize the projection matrix $\mathbf{U}$ to help compute the reconstruction term corresponding to emotion class aware conditional distribution alignment. Then, we repeat the following two major steps until convergence:

(1) Predict the target emotion labels using Equation (7) based on $\mathbf{U}$, and then confirm $\mathbf{X}_t^{(i)}$ and $\mathbf{X}_s^{(i)}$ according to the predicted target emotion labels;

(2) Solve the optimization problem in Equation (6), whose detailed solving procedures are summarized in Algorithm 1.

Algorithm 1 Detailed procedures for solving the optimization problem in Equation (6).

Repeat the following steps until convergence:

1. Fix $\mathbf{W}$ and $\mathbf{W}^{(i)}$, and update $\mathbf{U}$: the optimization problem reduces to the following:

$$\min_{\mathbf{U}} (\|\mathbf{L} - \mathbf{U}^T\mathbf{X}\|_F^2 + \lambda\|\mathbf{U}^T\|_{2,1}),$$

where $\mathbf{X} = [\mathbf{X}^s, \sqrt{\mu}\mathbf{\Delta}^{(0)}, \sqrt{\mu}\mathbf{\Delta}^{(1)}, \cdots, \sqrt{\mu}\mathbf{\Delta}^{(c)}]$ ($\mathbf{\Delta}^{(0)} = \mathbf{X}^{t_{(0)}} - \mathbf{X}^{s_{(0)}}\mathbf{W}$ and $\mathbf{\Delta}^{(i)} = \mathbf{X}^{t_{(i)}} - \mathbf{X}^{s_{(i)}}\mathbf{W}^{(i)}$), and $\mathbf{L} = [\mathbf{L}^s, \mathbf{0}^{(0)}, \mathbf{0}^{(1)}, \cdots, \mathbf{0}^{(c)}]$ ($\mathbf{0}^{(0)}$ and $\mathbf{0}^{(i)}$ are the matrices whose entries are all 0 and have the same size as $\mathbf{\Delta}^{(0)}$ and $\mathbf{\Delta}^{(i)}$).
The above optimization problem is a group sparse linear regression problem proposed by Zheng et al. [31]. We can utilize the ALM method to solve it by rewriting the above objective function as:

$$\|\mathbf{L} - \mathbf{P}\mathbf{X}\|_F^2 + \lambda\|\mathbf{Q}\|_{2,1},\ s.t.\ \mathbf{P} = \mathbf{Q}.$$

Thus, its corresponding augmented Lagrangian function can be represented as:

$$\|\mathbf{L} - \mathbf{P}\mathbf{X}\|_F^2 + tr(\mathbf{M}^T(\mathbf{P} - \mathbf{Q})) + \frac{\eta}{2}\|\mathbf{P} - \mathbf{Q}\|_F^2 + \lambda\|\mathbf{Q}\|_{2,1},$$

where $\mathbf{M}$ and η are the Lagrangian multiplier matrix and a regularization coefficient, respectively. The following optimization of this step can be referred to the work of [31] for more details.

2. Fix $\mathbf{U}$, and update $\mathbf{W}$ and $\mathbf{W}^{(i)}$: in this step, the optimization problem is divided to the following two types of independent problemss:

$$\min_{\mathbf{W}} (\|\mathbf{U}^T\mathbf{X}^t - \mathbf{U}^T\mathbf{X}^s\mathbf{W}\|_F^2 + \tau\|\mathbf{W}\|_1),$$
$$\min_{\mathbf{W}^{(i)}} (\|\mathbf{U}^T\mathbf{X}^{t_{(i)}} - \mathbf{U}^T\mathbf{X}^{s_{(i)}}\mathbf{W}^{(i)}\|_F^2 + \tau\|\mathbf{W}^{(i)}\|_1).$$

For the jth column of $\mathbf{W}$ and $\mathbf{W}^{(i)}$ denoted by $\mathbf{w}_j$ and $\mathbf{w}_j^{(i)}$, their optimal solutions are obtained by solving the following two typical LASSO problem with the SLEP package [32]:

$$\min_{\mathbf{w}_j} (\|\mathbf{z}_j^t - \mathbf{Z}^s\mathbf{w}_j\|_F^2 + \tau\|\mathbf{w}_j\|_1),$$
$$\min_{\mathbf{w}_j^{(i)}} (\|\mathbf{z}_j^{t_{(i)}} - \mathbf{Z}^{s_{(i)}}\mathbf{w}_j^{(i)}\|_F^2 + \tau\|\mathbf{w}_j^{(i)}\|_1),$$

where $\mathbf{z}_j^t = \mathbf{U}^T\mathbf{x}_j^t$, $\mathbf{Z}^s = \mathbf{U}^T\mathbf{X}^s$, $\mathbf{z}_j^{t_{(i)}} = \mathbf{U}^T\mathbf{x}_j^{t_{(i)}}$ and $\mathbf{Z}^{s_{(i)}} = \mathbf{U}^T\mathbf{X}^{s_{(i)}}$.

3. Check convergence: Reaching maximal iterations.

3. Experiments

In this section, we conduct extensive experiments to evaluate the proposed JIASL method and discuss its results compared with the state-of-the-art methods under the cross-corpus SER tasks.

3.1. Speech Emotion Database

The experiments were designed for extensive cross-corpus SER tasks on three widely used speech emotion corpora, i.e., EmoDB (Berlin) [26], eNTERFACE [27], and CASIA [28].

EmoDB is a German emotional speech database consisting of 535 speech samples from seven emotions, i.e., *happiness* (HA), *sadness* (SA), *disgust* (DI), *anger* (AN), *boredom* (BO), *fear* (FE), and *neutral* (NE). Ten German volunteers were induced to express their

emotions with some prepared texts. Each speech sample is recorded with a sample rate of 16 kHz.

eNTERFACE is an English bi-modal emotion database, and we extract its audio in the experiments. There are 1257 samples (without the data of the 6th speaker) in eNTERFACE, and each sample is labeled as one of six types of emotions, i.e., *happiness* (HA), *sadness* (SA), *disgust* (DI), *fear* (FE), *anger* (AN), and *surprise* (SU). The speech sentences are recorded by 43 English speakers with a sample rate of 44 kHz.

CASIA is a Chinese speech emotion corpus and includes 1200 samples with six emotions, e.g., *happiness* (HA), *sadness* (SA), *neutral* (NE), *fear* (FE), *anger* (AN), and *surprise* (SU). The speech utterances are generated by four Chinese volunteers expressing emotions under specific scripts with a sample rate of 16 kHz.

3.2. Experimental Setup

Task Setup and Protocol: The task setting of a cross-corpus SER is that one dataset (or several datasets) is regarded as training data and another dataset is set as target data. Note that we only obtain the data of the target domain without labels in practical cases; thus, the current practice is to use the training samples and their labels in the source domain and the testing samples in the target domain for domain adaptation. By alternatively using either two of the above speech corpora, we were able to design six cross-corpus SER tasks denoted by B→E, E→B, B→C, C→B, E→C, and C→E. B, E, and C are the abbreviations of EmoDB, eNTERFACE, and CASIA, respectively, and the left side of the arrow is the source speech corpus, while the other corresponds to the target one. Noting that, since these three speech corpora have inconsistent label information, we selected the speech samples sharing the same emotion labels in each task. We summarize the sample statistical information in all six cross-corpus SER tasks in Table 1.

Table 1. The sample statistical information in all six cross-corpus SER tasks.

Tasks	Speech Corpus (# Samples from Each Emotion)	# Total
B→E E→B	E (AN: 211, SA: 211, FE: 211, HA: 208, DI: 211)	1052
	B (AN: 127, SA: 62, FE: 69, HA: 71, DI: 46)	375
B→C C→B	C (AN: 200, SA: 200, FE: 200, HA: 200, NE: 200)	1000
	B (AN: 127, SA: 62, FE: 69, HA: 71, NE: 79)	408
C→E E→C	E (AN: 211, SA: 211, FE: 211, HA: 208, SU: 211)	1052
	C (AN: 200, SA: 200, FE: 200, HA: 200, SU: 200)	1000

Input Feature: In the experiments, we chose the INTERSPEECH 2009 Emotion Challenge (IS09) official feature set [33] and INTERSPEECH 2010 Paralinguistic Challenge feature set (IS10) [34] to describe the speech signals. The IS09 feature set consists of 384 elements, including 32 acoustic low-level descriptors (LLDs) and their 12 corresponding functions, which can be extracted by the openSMILE toolkit [35]. The IS10 feature set consists of 1582 elements with 34 LLDs and their 21 corresponding functions in the openSMILE toolkit.

Parameter Setup: We set the trade-off parameters for all the comparison methods by searching from a preset parameter interval and then reported the best evaluation metrics (i.e., WAR and UAR), which correspond to the best parameters. As for our JIASL, we searched for λ, μ, and τ from [0.001:0.001:0.009, 0.01:0.01:0.09, 0.1:0.1:0.9, 1:1:9], where START:STEP:END represents the loop from the start value to the end value with a step.

Evaluation Metric: Two widely used evaluation metrics, i.e., weighted average recall (WAR) and unweighted average recall (UAR) [10], were adopted to serve as the per-

formance measurement for cross-corpus SER. WAR is known as the standard accuracy, denoted as

$$\text{WAR} = \frac{n_{correct}}{n_{all}},$$

where $n_{correct}$ and n_{all} are the numbers of correctly predicted samples and all testing samples, respectively. UAR is defined as the class-wise accuracy, which can be represented as

$$\text{UAR} = \sum_{i=1}^{c} \left(\frac{n^i_{correct} / n^i_{all}}{c} \right),$$

where $n^i_{correct}$ and n^i_{all} represent the numbers of correctly predicted samples and total samples for the ith emotion class, respectively.

3.3. Comparison Methods

We compare our JIASL with recent well-performing state-of-the-art methods for cross-corpus SER tasks. These comparison methods are illustrated as follows.

- ***Baseline method:***
  IS09 or IS10 feature sets with the classifier of SVM [5];
- ***Transfer subspace learning-based methods:***
  Transfer component analysis (TCA) [36];
  Geodesic flow kernel (GFK) [37];
  Subspace alignment (SA) [38];
  Transfer kernel learning (TKL) [13];
  Domain-adaptive subspace learning method (DoSL) [15];
  Joint distribution adaptive regression (JDAR) [25].

3.4. Results and Discussions

Experimental results in terms of WAR are depicted in Table 2. As Table 2 shows, several interesting observations can be found. (1) Based on the IS09 feature set, it is clear to see that the proposed JIASL achieved the best average accuracy of all six cross-corpus SER tasks, reaching 40.19%, which has a remarkable increase of 3.98% compared with the second-highest WAR obtained by GFK [37]. In detail, among all the six cross-corpus SER tasks, it can also be observed that our JIASL achieved the highest WAR in five tasks, i.e., $B \rightarrow C$, $E \rightarrow B$, $B \rightarrow C$, $C \rightarrow B$, and $C \rightarrow E$. (2) The results based on the IS10 feature set reveal that our proposed JIASL obtained more competitive performance than other comparison methods. Furthermore, the JIASL performed best in five cross-corpus SER tasks, i.e., $B \rightarrow C$, $E \rightarrow B$, $C \rightarrow B$, $E \rightarrow C$, and $C \rightarrow E$. Both observations demonstrate the superior performance of the proposed JIASL over recent state-of-the-art transfer subspace learning methods in coping with cross-corpus SER tasks. In addition, we also find that the WAR results based on the IS10 feature set are better than the those based on the IS09 feature set, both in average accuracy and in most subtasks. This is due to the fact that the feature dimension of IS10 is higher than the feature of IS09, containing more emotional information.

Since both the eNTERFACE and Emo-DB datasets used in the experiments are class-imbalanced, we also report the results in terms of UAR for cross-corpus SER tasks to evaluate the performance of our proposed JIASL and the state-of-the-art methods more comprehensively, as shown in Table 3. For Table 3, it is clear that: (1) based on two feature sets, our proposed JIASL obtained the highest average accuracies, in which it reached the 38.42% UAR based on the IS09 feature set (improving by 2.09% compared with the second-highest WAR obtained by DoSL [15]) and 41.98% UAR based on IS10 feature set (increasing by 0.21% compared with the second-highest WAR obtained by JDAR [25])a (2) Among all the six cross-corpus SER tasks, the proposed JIASL based on the IS09 feature set obtained the highest UAR in five tasks, i.e., $B \rightarrow C$, $E \rightarrow B$, $B \rightarrow C$, $C \rightarrow B$, and $C \rightarrow E$.

It also performed the best in the other five tasks, i.e., $B \rightarrow C$, $E \rightarrow B$, $C \rightarrow B$, $E \rightarrow C$, and $C \rightarrow E$. These findings indicate that our proposed method can also achieve the best results in the class-imbalance case. This demonstrates that the joint regularization terms of marginal distribution and class-aware conditional distribution in our JIASL can effectively deal with class-imbalanced datasets.

Table 2. The experimental results in terms of WAR (%) for six designed cross-corpus SER tasks. The best result in each task is highlighted in bold.

Feature	Method	B → E	E → B	B → C	C → B	C → E	E → C	Average
IS09 Feature Set	SVM	28.90	18.93	29.60	34.07	25.10	26.10	27.12
	TCA	30.51	45.07	33.40	42.65	32.32	31.10	35.84
	GFK	32.13	44.53	33.10	46.57	28.14	**32.80**	36.21
	SA	36.06	38.93	34.40	42.16	31.65	30.40	35.60
	DoSL	33.56	40.53	35.80	45.10	28.04	32.60	35.94
	JDAR	36.41	40.27	31.10	43.63	31.56	32.40	35.90
	JIASL (Ours)	**36.88**	**50.40**	**36.50**	**53.68**	**33.17**	30.50	**40.19**
IS10 Feature Set	SVM	34.54	24.53	35.30	35.29	26.79	24.30	30.13
	TCA	32.64	46.78	40.50	54.56	29.75	33.20	39.57
	GFK	36.03	38.67	40.00	47.55	29.09	33.00	37.39
	SA	36.88	41.87	36.80	49.75	33.94	35.60	39.14
	DoSL	35.63	45.00	37.50	48.31	30.52	32.10	38.18
	JDAR	38.02	48.80	**42.70**	52.21	37.64	35.60	**42.50**
	JIASL (Ours)	**38.12**	**49.60**	38.10	**54.66**	**37.83**	**36.00**	42.39

Table 3. The experimental results in terms of UAR (%) for six designed cross-corpus SER tasks. The best result in each task is highlighted in bold.

Feature	Method	B → E	E → B	B → C	C → B	C → E	E → C	Average
IS09 Feature Set	SVM	28.93	23.58	29.60	35.01	25.14	26.10	28.06
	TCA	30.52	44.03	33.40	45.07	32.32	31.10	36.07
	GFK	32.11	42.48	33.10	48.08	28.13	**32.80**	36.17
	SA	36.12	38.95	34.40	45.75	31.59	30.40	36.20
	DoSL	33.50	43.89	35.80	49.03	28.17	32.60	36.33
	JDAR	36.33	39.97	31.10	46.29	31.50	32.40	36.27
	JIASL (Ours)	**36.87**	**44.11**	**36.50**	**49.30**	**33.19**	30.50	**38.42**
IS10 Feature Set	SVM	34.50	28.13	35.30	35.29	26.81	24.30	30.73
	TCA	32.60	44.53	40.50	51.47	29.77	33.20	38.68
	GFK	36.01	40.11	40.00	45.93	29.09	33.00	37.35
	SA	36.82	43.33	36.80	48.45	33.91	35.60	39.15
	DoSL	35.65	43.92	37.50	47.06	30.61	32.10	37.80
	JDAR	37.95	47.80	**42.70**	48.97	37.58	35.60	41.77
	JIASL (Ours)	**38.05**	**48.35**	38.10	**53.64**	**37.76**	**36.00**	**41.98**

Furthermore, comparing the results in Tables 2 and 3, we also have some interesting findings. For instance, in the results of JIASL and JDAR, it is clear to see that both of these two methods promisingly outperformed other comparison methods. This provides evident support to show the better effectiveness of jointly considering marginal and class-aware conditional distribution alignments adopted by JIASL and JDAR than simply considering the marginal one in bridging the distribution gap between two different feature sets. It is also worth mentioning that our JIASL performed better than JDAR, which shows that

for a joint marginal and conditional distribution alignment strategy, our designed implicit method (reconstruction) is more advantageous than the statistical moment-based strategy.

In addition, to evaluate the influence of the proposed optimization terms f_1 and w/o f_2 for cross-corpus SER tasks, we also calculated the cumulative distribution function (CDF) [39] with respect to WAR on the B $\rightarrow$ E task, shown in Figure 1, in which each point represents the WAR results of the specific iteration step, and w/o f_2 and w/o f_1 represent the proposed JIASL without the terms f_2 and f_1, respectively. Figure 1 reveals two main advantages of our proposed JIASL. The first point is that our JIASL with the optimization terms of f_1 and f_2 can achieve the highest WAR result compared to the models of JIASL without f_2 and f_1. The second point is that when the WAR is in the range of [20–26%], the CDF of JIASL is higher than JIASL without f_2 and f_1. This case demonstrates that the proposed JIASL converges faster and has more rounds to reach a high WAR in all iterations.

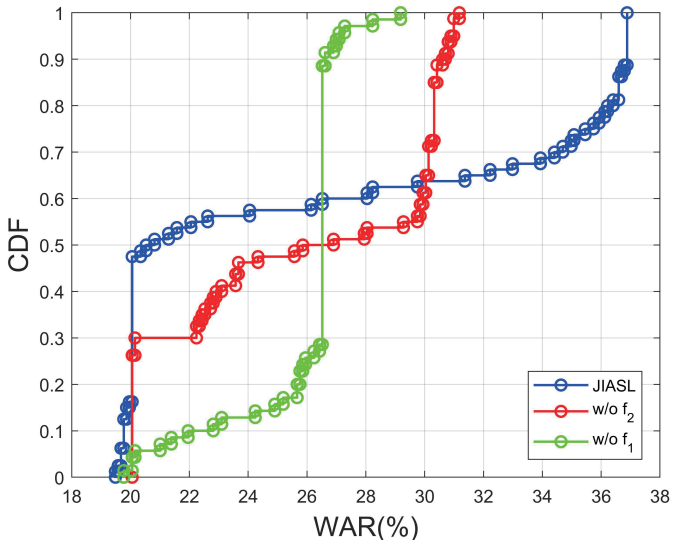
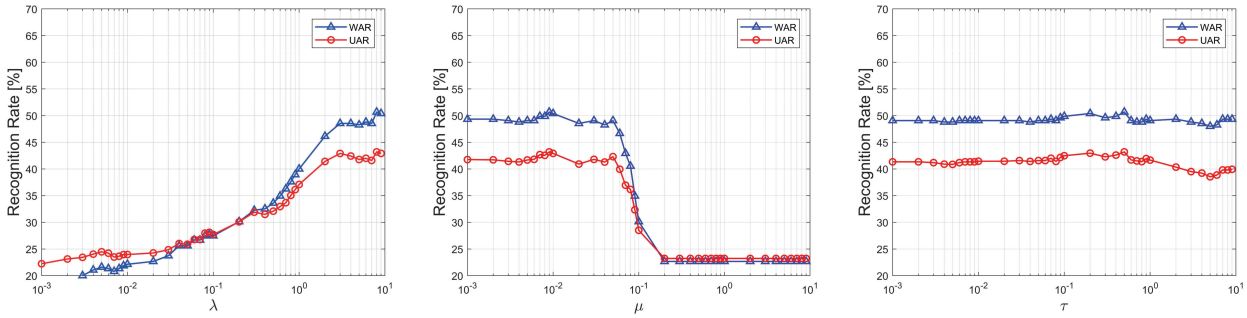


Figure 1. The cumulative distribution function (CDF) with respect to WAR for the cross-corpus SER task of B $\rightarrow$ E, in which w/o f_2 and w/o f_1 represent the proposed JIASL without the terms f_2 and f_1, respectively.

3.5. Parameter Sensitivity Analysis

In our proposed JIASL, three trade-off coefficients (i.e., λ, μ, and τ) are adopted to balance the different loss terms. Hence, we conducted additional experiments for the sensitivity analysis of these parameters to demonstrate the adaptation of JIASL. Figure 2 describes the results of the parameter sensitivity analysis on the cross-corpus SER task of E $\rightarrow$ B, in which the optimal parameters for λ, μ, and τ under this task are 8.0, 0.009, and 0.5, respectively, and the search set of parameters was set as [0.001:0.001:0.009, 0.01:0.01:0.09, 0.1:0.1:0.9, 1:1:9]. From Figure 2a, it is obvious that the recognition accuracies of WAR and UAR vary gently with the λ in [2, 9], demonstrating that JIASL is insensitive to λ in the optimal parameter interval [2, 9]. Similarly, Figure 2b reveals that the proposed method is susceptible to μ in the parameter interval [0.001, 0.4]. Moreover, from Figure 2c, we can observe that the recognition rate is flat enough in the whole parameter interval of τ. These results all indicate that our proposed JIASL is insensitive to three trade-off coefficients.

In addition, it is also interesting to find that the three parameters have specific sensitivities in different parameter intervals. For instance, λ is insensitive in the range of large parameter values, and μ is insensitive in the range of small parameter values. τ is insensitive in the entire parameter search range. This situation also reflects the different contributions of regularization terms to the performance of our JIASL model.



(**a**) trade−off coefficient λ (**b**) trade−off coefficient μ (**c**) trade−off coefficient τ

Figure 2. The parameter sensitivity analysis of three trade−off parameters (i.e., λ, μ, τ) for our proposed JIASL method on the cross-corpus SER task of E $\rightarrow$ B.

4. Conclusions

We propose a novel transfer subspace learning method called joint distribution implicitly aligned subspace learning (JIASL) for cross-corpus SER. The aim of JIASL is to learn an emotion discriminative and corpus-invariant projection matrix to predict the emotion labels of target speech samples. To this end, we first build a sparse linear regression model guided by the labeled source speech samples to endow the projection matrix with the emotion-discriminative ability. Then, a reconstruction regularization term, which implicitly bridges both marginal and emotion class aware conditional feature distribution gaps between two speech corpora, is further designed to enhance the corpus-invariant ability of the projection matrix. Finally, extensive cross-corpus SER experiments on EmoDB, eNTERFACE, and CASIA are conducted to evaluate the proposed JIASL. Experimental results demonstrated the effectiveness and superiority of the proposed JIASL in coping with cross-corpus SER tasks.

Author Contributions: Conceptualization and methodology, C.L., Y.Z. (Yuan Zong), C.T. and H.L.; validation, H.C. and J.Z.; writing—original draft preparation, C.L. and Y.Z. (Yuan Zong); writing—review and editing, S.L. and Y.Z. (Yan Zhao); funding acquisition, C.L. and Y.Z. (Yuan Zong). All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by in part the NSFC under the Grants U2003207 and 61902064, in part by the Jiangsu Frontier Technology Basic Research Project under the Grant BK20192004, in part by the Zhishan Young Scholarship of Southeast University, and in part by the Scientific Research Foundation of Graduate School of Southeast University YBPY1955.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

Abbreviations

The following abbreviations are used in this manuscript:

$\mathbf{X}^s$	The features of labeled speech samples in the source domain
$\mathbf{x}_i^s$	The feature vector of ith speech sample in the source domain
$\mathbf{X}^t$	The features of unlabeled speech samples in the target domain
$\mathbf{x}_j^t$	The feature vector of jth speech sample in the target domain
$\mathbf{L}^s$	The emotion labels of speech samples in the source domain
$\mathbf{l}_i$	The emotion label of the ith speech sample
$l_{i,k}$	The kth entry of one-hot vector $\mathbf{l}_i$
$\mathbf{U}$	The projection matrix
$\mathbf{W}$	The reconstruction coefficient matrix
$\|\cdot\|_F$	The Frobenius norm
$\|\cdot\|_{2,1}$	The $\ell_{2,1}$ norm
$\|\cdot\|_1$	The ℓ_1 norm
n^s	The number of source speech samples
n^t	The number of target speech samples
d	The dimension of the speech feature vector
c	The number of emotions involved in cross-corpus SER tasks

References

1. Schuller, B.; Batliner, A. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*; John Wiley & Sons: Hoboken, NJ, USA, 2013.
2. Busso, C.; Bulut, M.; Narayanan, S.; Gratch, J.; Marsella, S. Toward effective automatic recognition systems of emotion in speech. In *Social Emotions in Nature and Artifact: Emotions in Human and Human-Computer Interaction*; Gratch, J., Marsella, S., Eds.; Oxford University Press: New York, NY, USA, 2013; pp. 110–127.
3. Schuller, B.W. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM* **2018**, *61*, 90–99. [CrossRef]
4. Schuller, B.; Arsic, D.; Rigoll, G.; Wimmer, M.; Radig, B. Audiovisual behavior modeling by combined feature spaces. In Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, Honolulu, HI, USA, 15–20 April 2007; Volume 2, pp. 733–736.
5. Schuller, B.; Vlasenko, B.; Eyben, F.; Rigoll, G.; Wendemuth, A. Acoustic emotion recognition: A benchmark comparison of performances. In Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition & Understanding, Moreno, Italy, 13 November–17 December 2009; pp. 552–557.
6. Akçay, M.B.; Oğuz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun.* **2020**, *116*, 56–76. [CrossRef]
7. Zhang, S.; Zhang, S.; Huang, T.; Gao, W. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Trans. Multimed.* **2017**, *20*, 1576–1590. [CrossRef]
8. Lu, C.; Zong, Y.; Zheng, W.; Li, Y.; Tang, C.; Schuller, B.W. Domain Invariant Feature Learning for Speaker-Independent Speech Emotion Recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 2217–2230. [CrossRef]
9. Song, P. Transfer Linear Subspace Learning for Cross-Corpus Speech Emotion Recognition. *IEEE Trans. Affect. Comput.* **2019**, *10*, 265–275. [CrossRef]
10. Schuller, B.; Vlasenko, B.; Eyben, F.; Wöllmer, M.; Stuhlsatz, A.; Wendemuth, A.; Rigoll, G. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Trans. Affect. Comput.* **2010**, *1*, 119–131. [CrossRef]
11. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1345–1359. [CrossRef]
12. Zong, Y.; Zheng, W.; Zhang, T.; Huang, X. Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression. *IEEE Signal Process. Lett.* **2016**, *23*, 585–589. [CrossRef]
13. Long, M.; Wang, J.; Sun, J.; Philip, S.Y. Domain invariant transfer kernel learning. *IEEE Trans. Knowl. Data Eng.* **2014**, *27*, 1519–1532. [CrossRef]
14. Long, M.; Cao, Y.; Cao, Z.; Wang, J.; Jordan, M.I. Transferable representation learning with deep adaptation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 3071–3085. [CrossRef]
15. Liu, N.; Zong, Y.; Zhang, B.; Liu, L.; Chen, J.; Zhao, G.; Zhu, J. Unsupervised cross-corpus speech emotion recognition using domain-adaptive subspace learning. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5144–5148.
16. Gretton, A.; Borgwardt, K.M.; Rasch, M.J.; Schölkopf, B.; Smola, A. A kernel two-sample test. *J. Mach. Learn. Res.* **2012**, *13*, 723–773.
17. Borgwardt, K.M.; Gretton, A.; Rasch, M.J.; Kriegel, H.P.; Schölkopf, B.; Smola, A.J. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* **2006**, *22*, e49–e57. [CrossRef]

18. Mao, Q.; Xue, W.; Rao, Q.; Zhang, F.; Zhan, Y. Domain adaptation for speech emotion recognition by sharing priors between related source and target classes. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 2608–2612.
19. Deng, J.; Xu, X.; Zhang, Z.; Frühholz, S.; Schuller, B. Universum autoencoder-based domain adaptation for speech emotion recognition. *IEEE Signal Process Lett.* **2017**, *24*, 500–504. [CrossRef]
20. Abdelwahab, M.; Busso, C. Domain adversarial for acoustic emotion recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 2423–2435. [CrossRef]
21. Gideon, J.; McInnis, M.G.; Provost, E.M. Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG). *IEEE Trans. Affect. Comput.* **2019**, *12*, 1055–1068. [CrossRef] [PubMed]
22. Parry, J.; Palaz, D.; Clarke, G.; Lecomte, P.; Mead, R.; Berger, M.; Hofer, G. Analysis of Deep Learning Architectures for Cross-Corpus Speech Emotion Recognition. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 1656–1660.
23. Long, M.; Wang, J.; Ding, G.; Sun, J.; Yu, P.S. Transfer feature learning with joint distribution adaptation. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 2200–2207.
24. Zhu, Y.; Zhuang, F.; Wang, J.; Ke, G.; Chen, J.; Bian, J.; Xiong, H.; He, Q. Deep subdomain adaptation network for image classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 1713–1722. [CrossRef]
25. Zhang, J.; Jiang, L.; Zong, Y.; Zheng, W.; Zhao, L. Cross-Corpus Speech Emotion Recognition Using Joint Distribution Adaptive Regression. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 3790–3794.
26. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. In Proceedings of the Interspeech, Lisboa, Portugal, 4–8 September 2005; Volume 5, pp. 1517–1520.
27. Martin, O.; Kotsia, I.; Macq, B.; Pitas, I. The eNTERFACE'05 audio-visual emotion database. In Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06), Atlanta, GA, USA, 3–7 April 2006; p. 8.
28. Tao, J.; Liu, F.; Zhang, M.; Jia, H. Design of speech corpus for mandarin text to speech. In Proceedings of the the Blizzard Challenge 2008 Workshop, Brisbane, Australia, 1 February 2008; pp. 1–4.
29. Kan, M.; Wu, J.; Shan, S.; Chen, X. Domain adaptation for face recognition: Targetize source domain bridged by common subspace. *Int. J. Comput. Vis.* **2014**, *109*, 94–109. [CrossRef]
30. Lin, Z.; Liu, R.; Su, Z. Linearized alternating direction method with adaptive penalty for low-rank representation. *Adv. Neural Inf. Process. Syst.* **2011**, *24*, 612–620.
31. Zheng, W.; Xin, M.; Wang, X.; Wang, B. A novel speech emotion recognition method via incomplete sparse least square regression. *IEEE Signal Process. Lett.* **2014**, *21*, 569–572. [CrossRef]
32. Liu, J.; Ji, S.; Ye, J. SLEP: Sparse learning with efficient projections. *Ariz. State Univ.* **2009**, *6*, 7.
33. Schuller, B.; Steidl, S.; Batliner, A. The interspeech 2009 emotion challenge. In Proceedings of the Interspeech 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, UK, 6–10 September 2009; pp. 312–315.
34. Schuller, B.; Steidl, S.; Batliner, A.; Burkhardt, F.; Devillers, L.; Müller, C.; Narayanan, S. The INTERSPEECH 2010 paralinguistic challenge. In Proceedings of the INTERSPEECH 2010, Makuhari, Japan, 26–30 September 2010; pp. 2794–2797.
35. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: The munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 1459–1462.
36. Pan, S.J.; Tsang, I.W.; Kwok, J.T.; Yang, Q. Domain adaptation via transfer component analysis. *IEEE Trans. Neural Netw.* **2010**, *22*, 199–210. [CrossRef] [PubMed]
37. Gong, B.; Shi, Y.; Sha, F.; Grauman, K. Geodesic flow kernel for unsupervised domain adaptation. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2066–2073.
38. Fernando, B.; Habrard, A.; Sebban, M.; Tuytelaars, T. Unsupervised visual domain adaptation using subspace alignment. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 2960–2967.
39. Liu, S.; Sinha, R.S.; Hwang, S.H. Clustering-based noise elimination scheme for data pre-processing for deep learning classifier in fingerprint indoor positioning system. *Sensors* **2021**, *21*, 4349. [CrossRef] [PubMed]