applied sciences

MDPI

Article

A Simple Unsupervised Knowledge-Free Domain Adaptation for Speaker Recognition

Wan Lin [1,2], Lantian Li [3,*] and Dong Wang [1,*]

[1] Center for Speech and Language Technologies, BNRist, Tsinghua University, Beijing 100084, China; linwan@cslt.org
[2] College of Management, Shenzhen University, Shenzhen 518055, China
[3] School of Artificial Intelligence, Beijing University of Post Telecommunications, Beijing 100876, China
* Correspondence: lilt@bupt.edu.cn (L.L.); wangdong99@mails.tsinghua.edu.cn (D.W.)

Abstract: Despite the great success of speaker recognition models based on deep neural networks, deploying a pre-trained model in real-world scenarios often leads to significant performance degradation due to the domain mismatch between training and testing conditions. Various adaptation methods have been developed to address this issue by modifying either the front-end embedding network or the back-end scoring model. However, existing methods typically rely on knowledge of the network, scoring model, or even the source data. In this study, we introduce a knowledge-free adaptation approach that only necessitates the unlabeled target data. Our core concept is based on the assumption that domain mismatch primarily stems from distributional distortion in the embedding space, such as shifting, rotation, and scaling while maintaining inter-speaker discrimination for data from unknown domains. Building on this assumption, we propose clustering LDA (C-LDA), a full-rank linear discriminant analysis (LDA) based on agglomerative hierarchical clustering (AHC) to compensate for this distortion. This approach does not need any human labels and does not rely on any knowledge of the model in the source domain, making it suitable for real-world applications. Theoretical analysis indicates that with cosine scoring, C-LDA is capable of eliminating distributional distortion related to global shift and within-speaker covariance rotation and scaling. Surprisingly, our experiments demonstrated that this simple approach can outperform more complex methods that require full or partial knowledge, including front-end approaches such as fine-tuning and distribution alignment, and back-end approaches such as unsupervised probabilistic linear discriminant analysis (PLDA) adaptation. Additional experiments demonstrated that C-LDA is insensitive to hyperparameters and works well in both multi-domain and single-domain adaptation scenarios.

Keywords: domain mismatch; unsupervised adaptation; speaker recognition

Citation: Lin, W.; Li, L.; Wang, D. A Simple Unsupervised Knowledge-Free Domain Adaptation for Speaker Recognition. *Appl. Sci.* **2024**, *14*, 1064. https://doi.org/10.3390/app14031064

Academic Editor: Douglas O'Shaughnessy

Received: 31 December 2023
Revised: 20 January 2024
Accepted: 22 January 2024
Published: 26 January 2024

1. Introduction

Speaker verification involves determining whether a given speech corresponds to a claimed speaker. In recent years, deep embedding-based speaker verification systems have achieved remarkable performance [1–3]. A typical verification process with such a system entails a deep embedding model that converts variable-length speech segments into fixed-dimensional dense vectors known as 'speaker embeddings' or 'speaker vectors', followed by a simple back-end scoring using cosine distance to assess the similarity of the vectors from the enrollment and test segments. The x-vector model [3] is among the most popular deep embedding models, and with the advancement of more comprehensive architectures, improved pooling methods, and training strategies, the x-vector model has demonstrated state-of-the-art performance in complex acoustic environments [4–15].

However, due to the complexity and variability of acoustic environments, and the limited data coverage during training, well-developed models often experience significant performance degradation when deployed in real-world domains that differ from the training domain, which is known as the domain mismatch issue [16–18]. To adapt a pre-trained

model to an unknown working condition, existing practices can be broadly categorized into two types of approaches: (1) **Front-end network adaptation** [19–21], which involves fine-tuning the speaker embedding network to align the embedding vectors with the target environment. (2) **Back-end scoring model adaptation** [22–24], which modifies the parameters of the scoring model while keeping the front-end network unchanged.

Recent adaptation approaches mostly fall into the front-end category, partly due to the diminishing use of complex back-end models like probabilistic linear discriminant analysis (PLDA) [25]. Moreover, most research efforts have focused on unsupervised adaptation [26–28], addressing the challenge of speaker labeling for target-domain data collected from a deployed system, owing to labeling cost and privacy concerns.

A drawback of existing methods is that they require knowledge of the system and/or the source data used to train the model. For example, front-end adaptation methods require the model to be accessible, allowing modification of its parameters. Similarly, PLDA adaptation methods require access to the between-speaker and within-speaker covariances of the original PLDA model [22]. Additionally, most distribution alignment approaches even mandate the replay of the source data [29–31]. Requiring knowledge of the system is often impractical in many situations, not to mention the challenges associated with accessing the source data.

In this paper, we introduce a knowledge-free adaptation approach. This concept was inspired by an intriguing observation that speaker embeddings of unknown domains are still separable. As depicted in Figure 1, one can find that while the performance in equal error rate (EER) significantly decreases when the data is from an unknown domain, the discriminability of the embedding vectors remains relatively stable. This suggests that the primary issue of domain mismatch may not be the reduction in discriminability, but rather a systematic distortion of the distribution of the embedding vectors, including shifting, rotation, and scaling. If this hypothesis holds, it may be feasible to compensate for this distortion through a simple linear transform in the speaker embedding space.

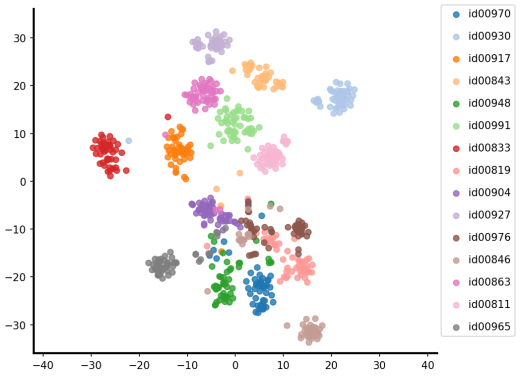

Figure 1. The t-SNE plot visualizes the speaker vectors of CN-Celeb1 extracted by a model pre-trained on VoxCeleb. Each color represents a speaker.

This leads to the development of our clustering-based unsupervised linear adaptation approach, which involves four steps: (1) Utilizing a pre-trained model from the source domain to extract speaker embedding vectors from unlabeled data in the target domain. (2) Generating pseudo labels for these vectors using a clustering algorithm. (3) Training a full-rank linear discriminant analysis (LDA) model [32] based on the pseudo-labels. (4) Using the resulting LDA matrix to transform speaker embedding vectors of speech from the target domain, compensating for potential distributional distortion. We refer to this clustering-based full-rank LDA model as *C-LDA* .

In theory, the full-rank LDA does not alter the discrimination capacity of the embedding vectors; instead, it aims to discover a new set of coordinates where both within-class variance and between-class variance align along the coordinate axes. This enables speaker vectors, after the LDA transform, to be suitable for cosine scoring. Therefore, C-LDA should

be viewed as an embedding normalization approach. It is important to note that C-LDA does not require any knowledge of the system and the source data, making it *knowledge-free*. Interestingly, this simple linear transform, derived without reference to any prior knowledge, yields notably good performance, even outperforming approaches that require full or partial knowledge and significant computation. In summary, the contribution of this paper is three-fold:

(1) We introduced a novel speaker recognition adaptation approach called C-LDA. This method is simple (involving a linear transform), unsupervised (utilizing clustering without human labels), and knowledge-free (requiring no knowledge of models and source data). These characteristics make it highly suitable for real-world applications.
(2) It was demonstrated that the C-LDA adaptation is highly effective and even outperforms the more complex front-end and back-end approaches in both multi-domain and single-domain adaptation scenarios. Additionally, the performance of C-LDA is robust against the setting of hyperparameters.
(3) The success of C-LDA further demonstrated our hypothesis that domain mismatch faced by speaker recognition systems based on deep models is primarily attributed to distributional distortion in the embedding space, rather than a decrease in discriminability.

2. Related Work

Researchers have long been concerned with the issue of distributional distortion in speaker vectors when a pre-trained model is tested in a mismatched acoustic condition and have proposed many approaches.

2.1. Back-End Adaptation

Early approaches primarily focused on back-end PLDA domain adaptation methods, which involved adjusting PLDA parameters to adapt to the target domain data. For example, these methods utilize the target domain data to adjust the between-class and within-class covariance matrices [33,34]. The correlation alignment (CORAL) method aligned the covariance matrices of the source and target domain data through linear mapping, then mapped the source domain data into pseudo-target domain data, which were used to retrain PLDA [30]. Lee et al. [23] proposed CORAL+, a variation of CORAL that avoided using source domain data. Instead, it directly updated the within-class and between-class covariance matrices in PLDA based on the relationship between the full covariance matrices of the source and target domains. Li et al. [35] combined CORAL and CORAL+, simultaneously performing feature mapping and model alignment.

2.2. Front-End Adaptation

With the increasing adoption of the maximum margin criterion, more research relies on a straightforward cosine distance for scoring, and domain mismatch problems are primarily addressed by modifying the front-end embedding network. One approach is rooted in adversarial learning [36–38], where an additional adversarial domain discriminator is introduced to compel the network to learn domain-invariant features. Another approach is based on distribution alignment [28,39,40], leveraging first-order or second-order statistics to estimate the data distributions of the source and target domains and minimize their difference. For instance, deep CORAL [29] optimizes the network to bring the covariance matrices of the source and target domains closer, while maximum mean discrepancy (MMD) [39] minimizes the difference between the means of the two domains in the Hilbert space.

In recent years, self-supervised learning (SSL) has made remarkable advances in speaker recognition, leading researchers to develop self-supervised strategies to tackle domain mismatch challenges. Chen et al. [41] jointly trained on supervised loss from the source domain and contrastive self-supervised loss from the target domain. Building on this, Mao et al. [42] proposed a clustering-based unsupervised domain adaptation (UDA) framework and conducted supervised training on both the source and target data. The

core idea of these methods is to optimize the target domain data while maintaining strong recognition capabilities on the source domain. In our previous work [43], we introduced a multi-domain adaptation method that combines the concept of SSL adaptation [41] and distribution alignment [39], where the aligned distributions are from several parallel target domains. This represents the first attempt at simultaneous adaptation to multiple target domains.

2.3. Embedding Adaptation

Some previous studies have proposed adaptation methods in the embedding space. Alam et al. [30] employed a linear transform to align the covariance of the target domain with the source domain and then used the transform to convert source-domain embeddings to the target domain to facilitate PLDA training. Lin et al. [31] proposed to align speaker embeddings from different domains in the latent space derived from auto-encoders, and the alignment is based on the MMD criterion. Cai et al. [44] presented a deep normalization approach that regulates speaker embeddings to linear Gaussians using a normalization flow model. The regulated embeddings exhibit better robustness, but not specifically for solving the domain mismatch problem. Li et al. [45] also designed a linear transform to mitigate domain mismatch, although the mismatch in their study is between enrollment and test, rather than training and deployment.

The proposed C-LDA approach belongs to the 'embedding adaptation' category. Unlike previous studies that try to transform to [30] or align with [31] the source domain, C-LDA is self-contained: it does not use any information from the source domain. Instead, it only utilizes the clustering structure of the embeddings in the target domain and normalizes the embeddings according to this structure.

3. Methodology

The models trained in the source domain often suffer from significant performance degradation when applied to the target domain. This degradation can be attributed to two possible reasons. First, the discriminability of the embedding vectors deteriorates, making it challenging to distinguish between different speakers. Second, the embedding vectors undergo systematic distortion, rendering them incompatible with the back-end scoring model. For the former issue, optimization of the embedding network is required, while the latter can be resolved through a straightforward linear transform.

This paper leans more toward the latter possibility and introduces a method known as clustering-based linear discriminant analysis (C-LDA). We employ linear discriminant analysis (LDA) to perform this linear transform and use pseudo-labels produced by agglomerative hierarchical clustering (AHC) [46] to achieve *unsupervised* domain adaptation.

3.1. Normalization by Full-Rank LDA

LDA has been widely adopted in the back-end for preprocessing embedding vectors before scoring. The primary objective of LDA is to identify directions that offer the most discrimination for speakers and to eliminate those with weaker correlations to speakers, thereby enhancing the discriminability among speaker vectors [47].

In recent years, deep embedding models based on the maximum margin criterion have simplified the paradigm of speaker recognition systems. By using these 'advanced' models, all the dimensions in the embedding vectors contain speaker information, so a simple cosine scoring on raw speaker embeddings can achieve good performance. As a result, the utilization of LDA has declined. In this paper, we re-discover the merit of LDA from another perspective: instead of using it to identify discriminative dimensions, we use it to 'normalize' the distribution of embeddings from unknown domains. The theory is established as follows.

Intuitively, LDA aims to identify directions W that maximize the discrimination between embedding vectors from different speakers. The pursuit of the directions entails taking into account both the between-class variance and the within-class variance. Therefore,

the criterion for determining the optimal direction is to maximize the ratio of between-class variance to within-class variance. This objective can be formulated as follows:

$$S_B W = \lambda S_W W \tag{1}$$

where S_B and S_W are the between-class covariance matrix and the within-class covariance matrix, respectively.

The solution to this objective is to find the eigenvectors of ${S_W}^{-1}S_B$, with λ representing the corresponding eigenvalues. By arranging all the eigenvectors of ${S_W}^{-1}S_B$ in descending order according to their eigenvalues, we obtain the principal directions of discriminability, ranked from strongest to weakest. In speaker recognition systems, it is customary to select the top 125 or 150 eigenvectors to construct the eigenspace of LDA. This helps mitigate the influence of less discriminative directions, thus enhancing speaker discriminability.

In this work, we utilize a full-rank LDA, and there is no eigenvector selection process as previously described. Thus, it does not impact the discriminability of speaker vectors; instead, it serves as a linear transform to normalize the distribution of speaker vectors. Previous studies have demonstrated that a full-rank LDA can be decomposed into the following steps [32]:

(1) **Global shift**: Perform centering (mean subtraction) to normalize the data to a zero-mean distribution.
(2) **Rotation (first)**: Rotate the coordinate to align the axes to the principal directions of the within-class covariance.
(3) **Scaling**: Scale the coordinate axes to unify the within-class variance along each axis, known as whitening.
(4) **Rotation (second)**: Rotate the coordinate again to align the axes to the principal directions of between-class variance (note that since within-class variance has been whitened, the subsequent rotation of the coordinate will not affect the within-class variance).

From these steps, it is evident that the effect of a full-rank LDA is to align the principal directions of within-class variance and between-class variance with the coordinate axes. In other words, after the transformation, the within-speaker covariance and between-class covariance matrices are all diagonalized, which means the dimensions of the transformed embeddings are mutually uncorrelated. In summary, this full-rank LDA results in a normalized distribution with elegant coordinate alignment.

In theory, step (2) and step (4) do not change the cosine score as they are both orthogonal transformations. Step (1) is important, as the location of the origin severely impacts cosine scores. Step (3) is important because, without such scaling, the between-class distance will be affected by the within-class variation that varies across dimensions. For instance, when computing cosine distance, the between-class distance in a particular dimension should be more de-emphasized if this dimension exhibits a large within-class variation, as the large distance mainly reflects a large within-class variation. Importantly, if the within-class variation is not whitened, the cosine distance will be biased towards directions with large within-class variance.

According to this analysis, if the main impact of domain mismatch in the embedding space is global shift and within-class distributional distortion, then the full-rank LDA can provide reasonable compensation. Another key insight is that LDA normalization is not affected by the between-class covariance. The positive side is that it does not require an accurate between-class covariance, which is often difficult to estimate in particular with limited data. The negative side, however, is that the LDA normalization cannot compensate for between-class distortion. It should be careful when interpreting the statement "LDA does not require between-class covariance": it does not mean estimating the LDA transform does not require a between-class covariance, it just states that any estimation of the between-class covariance does not change the results with cosine scoring.

3.2. AHC Clustering

LDA training requires speaker labels of speech from the target domain. To avoid the cost associated with manual labeling, the clustering approach is used to generate pseudo-labels. We anticipate that the clustering approach can produce reasonable pseudo-labels for two reasons: (1) As shown in Figure 1, the speaker embeddings of the target-domain speech are largely separated, so can be well clustered; (2) From the theoretical analysis, only the within-class variation is important in C-LDA, which substantially reduces the demand for the accuracy of the clustering.

The agglomerative hierarchical clustering (AHC) algorithm [46] was adopted in this study. AHC is a bottom-up hierarchical clustering process that initially treats each sample as one cluster and iteratively merges the two most similar clusters until a predetermined stopping criterion is satisfied. Due to its simplicity and effectiveness, AHC has been widely used in unsupervised speaker adaptation [22,34,48]. We use AHC to cluster the embedding vectors of the target domain data and use the cluster index as the pseudo-speaker label for each embedding.

For simplicity, we manually set the maximum number of clusters and use cosine distance as the distance metric. Specifically, when the number of clusters does not reach the predefined maximum cluster, in each iteration, we enumerate all possible cluster pairs to merge. For each pair, we compute the cosine distance between each embedding vector and the mean vector of the whole embeddings in the pair of clusters, and compute the sum of the distances. Subsequently, we select the pair with the smallest sum of distances and merge the two clusters as a single one. Although this process differs from some typical AHC implementations, it has been demonstrated to be effective in our experiments.

Formally, in the p-th iteration, assuming C^p as the set of existing clusters, the two clusters C_a^p and C_b^p are merged if they satisfy the following criterion:

$$\left\{C_a^p, C_b^p\right\} = \underset{C_i^p, C_j^p \in C^p, i \neq j}{\arg\min} \sum_{x_{ij}^p[k] \in \left\{C_i^p, C_j^p\right\}} \cos\left(x_{ij}^p[k], \bar{x}_{ij}^p\right) \tag{2}$$

where C_i^p represents cluster i at the initial stage of the p-th iteration. $x_{ij}^p[k]$ denotes the k-th embedding of the merging cluster of C_i^p and C_j^p, $\bar{x}_{ij}^p$ is the mean of all the embeddings that belong to C_i^p and C_j^p.

4. Experiments

4.1. Datasets

We used VoxCeleb2 [49] as the source-domain dataset and CN-Celeb1 as the target-domain dataset [50]. VoxCeleb2 is a large-scale corpus containing over 1 million utterances from 5994 speakers. It was collected from YouTube, with most of the speech data being in English. CN-Celeb1 is a multi-genre Chinese corpus comprising over 125 k utterances from 997 speakers, spanning 11 distinct genres. The number of utterances, speakers, and hours in these 11 genres all vary, representing natural and complex domain variations. The entire dataset is divided into two parts: CNC1.dev, which contains 797 speakers for domain adaptation, and CNC1.eval, which consists of 200 speakers for performance evaluation. More detailed information can be found in Table 1. We highlight that CN-Celeb1 is highly complex and their genres can be regarded as different domains, so our task is essentially a multi-domain adaptation, a scenario seldom explored before. Note that the two datasets were well constructed and silence was well purged, so no voice activity detection (VAD) was employed in training and testing.

Table 1. The datasets of CN-Celeb1 used in our experiment. '#' denotes number.

Datasets	# of Spks	# of Utts	# of Hours	# of Avg Dur
CN-Celeb1	997	126,532	271.71	7.73
CNC1.dev	797	107,953	228.01	7.60
CNC1.eval	200	18,579	43.71	8.47
CNC1.dev.interview	637	53,035	120.94	8.21
CNC1.dev.entertainment	351	18,443	27.40	5.35
CNC1.dev.singing	250	10,544	23.83	8.14

4.2. System Configuration

For the front-end embedding model, we adopted the x-vector architecture. The acoustic features are 80-dimensional Filter Banks (Fbanks). The primary architecture comprises three key components: (1) An ECAPA-TDNN [4] backbone, which is currently the most popular TDNN-series model for speaker verification, for learning frame-level features; (2) An attentive statistic pooling (ASP) [6] layer for aggregating frame-level features and a dense layer that projects the pooled features into an utterance-level x-vector; (3) A fully connected layer to classify different speakers in the training data. Table 2 shows the topology of this model.

Table 2. The topology of the ECAPA-TDNN model used in our experiment. '#' denotes number.

Layer	Kernel Size	Stride	Dilation	Output
Input	–	–	–	80 × 200
Conv1D	1 × 5	1 × 1	1 × 1	1024 × 200
SE-Res2Block1	1 × 3	1 × 1	1 × 2	1024 × 200
SE-Res2Block2	1 × 3	1 × 1	1 × 3	1024 × 200
SE-Res2Block3	1 × 3	1 × 1	1 × 4	1024 × 200
SE-Res2Block4	1 × 1	1 × 1	1 × 1	1536 × 200
Pooling		ASP		3072 × 1
Dense		–		192 × 1
Dense		AAM-Softmax		# of Spks

During the pre-training phase, the front-end embedding model was trained for 80 epochs. Each mini-batch involved 256 2-s segments randomly sampled from 5994 speakers in the VoxCeleb2 dataset. The additive angular margin softmax (AAM-Softmax) [51] was employed to discriminate speakers, with a margin value of 0.2 and a scale factor of 30. In addition, extensive data augmentation techniques were used, including SpecAugment [52], additive noise and room impulse response (RIR) simulation. Specifically, for SpecAugment, we randomly masked 0 to 10 frames in the time domain and 0 to 8 channels in the frequency domain. Additive noise was sourced from audio clips in the MUSAN corpus [53], and simulated filters of small, medium, and large rooms released in [54] were used for RIR. The model was trained using the Adam optimizer with an initial learning rate set to 0.001, which decreased by 5% each epoch. Once the learning rate dropped below 0.0001, it remained fixed.

During the domain adaptation phase, for all the considered front-end adaptation methods, the mini-batch size was set to 128 and the initial learning rate was set to 0.0005 decreased by 5% at each epoch. The duration of the input segment is kept at 2 s. If it is less than 2 s, zero-padding will be applied. The entire adaptation process involved 20 epochs. The same data augmentation techniques as in the pre-training phase were used, except that SpecAugment was excluded. For C-LDA, the complete utterance is fed into the front-end

model to obtain speaker embedding, and then agglomerative hierarchical clustering (AHC) was used to generate pseudo-speaker labels with which the full-rank LDA was trained. The number of clusters was set to 800, and the dimensionality of LDA-transformed x-vector was set to 192. Further analysis of these two parameters will be discussed in Section 4.4.2.

For model evaluation, we adopted two performance metrics. The primary metric used is the equal error rate (EER), which indicates the threshold where the false acceptance rate (FAR) equals the false rejection rate (FRR). The alternative metric is the minimum detection cost function (minDCF), which calculates the minimum detection cost by considering the trade-off between acceptance and rejection errors. Note that we used $C_{miss} = C_{fa} = 1$ and $P_{tar} = 0.05$ in the cost function.

4.3. Main Results

4.3.1. Embedding vs. Front-End

We first compare C-LDA with several front-end adaptation methods. All the comparison methods use cosine scoring. The results tested on CNC1.eval, in terms of EER and minDCF, are reported in Table 3.

Table 3. Performance comparison among various adaptation methods. The pre-trained model was trained on VoxCeleb. CNC1.dev was used for adaptation and CNC1.eval was used for test. 'FT' denotes fine-tuning, 'LDA' denotes linear discriminant analysis; 'C-' and 'S-' denote 'clustering' and 'supervised', respectively.

Adaptation	Method	Supervision Label	Knowledge	EER(%)	minDCF ($P_{tar} = 0.05$)
	Pretrain	-	-	14.22	0.5137
Front-End	FT	Speaker	NeuralNet	9.50	0.3991
	C-FT	-	NeuralNet	11.08	0.4820
	SSL-DA	Domain	NeuralNet	11.54	0.4551
Embedding	C-LDA	-	-	10.66	0.4122
	S-LDA	Speaker	-	9.75	0.4047

The 'Pretrain' method represents the pre-trained model with VoxCeleb2, which serves as the source-domain model. The other methods listed, which all use cosine scoring, are domain adaptation techniques built upon the pre-trained model. Although our focus is unsupervised methods, results of supervised methods are also reported to show the upper bound of the performance of the corresponding unsupervised methods. Note that we exclude those methods that require source data, e.g., Deep CORAL [29] and MMD [39], as they are infeasible in real-life deployment systems. The methods involved in Table 3 are summarized below.

(1) **Fine-tuning (FT)**: The standard front-end adaptation approach, utilizing the development data from CNC1.dev. It uses genuine speaker labels of the speech, requires the knowledge of the source-domain model, and modifies its parameters by back-propagation.
(2) **Clustering fine-tuning (C-FT)**: The same as FT, except that the speaker labels of the adaptation data are pseudo labels from the AHC algorithm (the same as C-LDA). This system is mainly used to test the quality of the pseudo labels.
(3) **SSL-DA** [43]: Front-end adaptation by (1) self-supervised training with CNC1.dev; (2) distribution alignment between different genres. It does not require speaker labels but needs genre labels.
(4) **Clustering LDA (C-LDA)**: The proposed method of this paper. It does not require any supervision or any knowledge of the front-end model.

(5) **Supervised LDA (S-LDA)**: The same as C-LDA, except that it uses genuine speaker labels. S-LDA presents the upper bound for the capability of C-LDA in compensating for the distributional distortion, by eliminating label errors.

Several noteworthy observations can be obtained from Table 3. For clearance, only the EER results are referred to in the discussion.

- With all the adaptation methods, the results of the adapted systems are consistently better than that of the pre-trained source-domain model (14.22%). This confirms the impact of domain mismatch and underscores the necessity of domain adaptation techniques.
- FT vs. C-FT (9.50% vs. 11.08%) and S-LDA vs. C-LDA (9.75% vs. 10.66%): It reveals that clustering-based unsupervised learning methods, while not reaching the performance of supervised learning with genuine speaker labels, can still effectively alleviate the domain mismatch problem and achieve performance close to that of the supervised methods. This demonstrated that simple hierarchical clustering can produce high-quality pseudo labels.
- FT vs. S-LDA (9.50% vs. 9.75%): Both of which are supervised, and the difference is that they adapt the embedding network and the embedding space, respectively. It can be observed that the performance of S-LDA is quite close to that of FT. This strongly supports our hypothesis that the domain mismatch issue can largely be attributed to the distributional distortion in the embedding space, and a simple linear mapping can largely eliminate this distortion.
- C-FT vs. C-LDA (11.08% vs. 10.66%): Both of which are unsupervised and rely on pseudo-labels. It can be seen that the embedding adaptation method C-LDA outperforms the front-end adaptation method C-FT. On one hand, it reaffirms that a linear transform such as full-rank LDA can largely eliminate the distributional distortion caused by domain mismatch, even though the LDA is trained with inaccurate speaker labels. On the other hand, more importantly, it indicates that for noisy pseudo-labeled data, compared to fine-tuning the entire front-end network, the simpler embedding adaptation is perhaps more effective. This relative superiority of C-LDA might be attributed to its simple functional form that prevents over-fitting to the errors in the pseudo labels.
- The performance of the SSL-DA approach is relatively weaker (11.54%) compared to other unsupervised methods, though it is still much better than the pre-trained model (14.22%). We guess it is because self-supervised learning is not powerful enough to address the complex multi-domain adaptation problem due to the weak supervision signal from contrastive pairs.

It should be highlighted that among all these adaptation methods, C-LDA is the only one that does not require any supervision and knowledge of the source-domain model, even though its performance (10.66%) is substantially better than the pre-trained model (14.22%) and close to the supervised fine-tuning (9.50%). It also substantially outperforms the more complex SSL-DA approach that not only requires knowledge of the embedding model but also needs genre labels.

4.3.2. Embedding Visualization

To demonstrate the impact of different methods on the speaker embedding vectors, we analyze all these methods by visualization. We randomly selected 15 speakers from CNC1.eval and sampled 50 utterances from each of these speakers. The t-SNE toolkit [55] was applied to project the speaker embeddings of these utterances to a two-dimensional space, as illustrated in Figure 2.

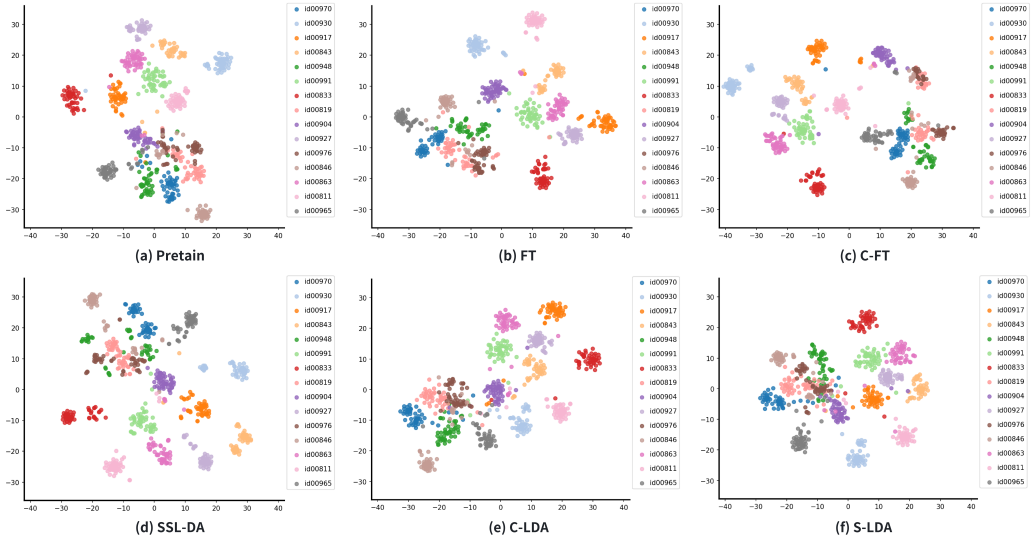

Figure 2. Embeddings produced by the pre-trained model and the models with different adaptation methods. The pictures are plotted by t-SNE. Each point represents an utterance and each color represents a speaker. The utterances are randomly selected from CNC1.eval.

Firstly, despite the pre-trained model experiencing a notable performance decrease in EER and minDCF, the embeddings of distinct speakers remain well separated. This suggests that the pre-trained source-domain model still retains strong discriminative capability even when applied to cross-domain data. Furthermore, none of the adaptation methods, including supervised fine-tuning, demonstrate a significant enhancement in discriminating the agglomerative speakers.

Secondly, S-LDA and C-LDA essentially denote a full-rank linear transform within the embedding space. Theoretically, they should not enhance speaker discrimination in the speaker embeddings, as evidenced in the t-SNE plots Figure 2e,f. However, they do yield significant performance improvements in terms of EER and minDCF. This implies that by merely rotating and scaling the speaker vectors, it is feasible to effectively counteract the performance decline resulting from domain mismatch. This discovery further reinforces our hypothesis that a distributional distortion in the speaker embedding space can predominantly characterize the impact of domain mismatch.

4.3.3. Embedding vs. Back-End

Although complex back-end scoring models such as PLDA have faded in most speaker recognition systems, we still compare some representative back-end adaptation methods with C-LDA, to gain further insight into the merit of the proposal. Table 4 presents the comparative results between C-LDA/S-LDA and several representative PLDA adaptation methods. In all these methods, the pre-trained front-end model is frozen and is used to extract the embedding vectors. The PLDA adaptation methods considered in the comparative study are listed below.

(1) **Supervised PLDA (S-PLDA)**: PLDA re-trained with CNC1.dev, the development data for the target domain, using the genuine speaker labels.
(2) **Adaptive PLDA (A-PLDA)**: An off-the-shelf unsupervised PLDA adaption approach provided by Kaldi [56]. It estimates the change in the between-speaker and within-speaker covariance by the change on the *total covariances* from the source domain to the target domain. Note that no speaker labels are required.
(3) **CORAL+** [23]: An unsupervised PLDA adaption method, with the idea of distribution alignment between the source and target domain. This approach has shown better performance than several competitive approaches [23].

(4) **Clustering PLDA (C-PLDA)**: The same as S-PLDA, except that the speaker labels are pseudo labels produced by clustering. It is the probabilistic version of C-LDA. Note that C-PLDA does not require the parameters of the source-domain PLDA, so it is knowledge-free, just like C-LDA.

Table 4. Performance comparison between C-LDA and other back-end adaptation methods on CNC1.eval. 'C-' and 'S-' signify 'clustering' and 'supervised', respectively.

Adaptation/Scoring	Method	Supervision Label	Knowledge	EER(%)	minDCF (P_{tar} = 0.05)
	Pretrain	-	-	14.22	0.5137
Embedding/Cosine	S-LDA	Speaker	-	9.75	0.4047
	C-LDA	-	-	10.66	0.4122
Back-End/PLDA	S-PLDA	Speaker	-	8.87	0.3684
	A-PLDA	-	Covariance	10.86	0.4198
	CORAL+	-	Covariance	10.54	0.4299
	C-PLDA	-	-	10.11	0.3979

The experimental results are shown in Table 4, which involves several interesting observations as follows. Again, we only refer to the EER results in the discussion.

- All the embedding and back-end adaptation methods outperform the baseline result without any adaptation (14.22%), demonstrating the efficiency of these methods.
- In all the methods, S-PLDA shows the best performance (9.57%). This result is close to the one obtained with front-end fine-tuning (9.50%). It indicates that the PLDA model, if well-trained in the target domain, can effectively solve the domain mismatch problem.
- C-LDA (10.66%) vs. C-PLDA (10.11%) and S-LDA (9.75%) vs. S-PLDA (8.87%): It can be observed that C-PLDA performs better than C-LDA, and S-PLDA outperforms S-LDA. This is expected as PLDA scoring utilizes the between-class information and conducts minimum-risk Bayesian decision [57].
- C-LDA (10.66%) vs. S-LDA (9.75%) and C-PLDA (10.11%) vs. S-PLDA (8.87%): This comparison indicates that using pseudo labels produced by clustering leads to inferior performance compared to using the genuine labels; however, compared to the baseline result, the improvement with the cost-free pseudo labels is highly significant. We also note that the disparity between C-LDA and S-LDA (10.66% vs 9.75%) is smaller compared to that between C-PLDA and S-PLDA (10.11% vs 8.87%). This seems to indicate that *LDA relies less on the accuracy of speaker labels compared to PLDA*; thus, it is more suitable for unsupervised learning. We hypothesize that this is because LDA does not need an accurate between-class covariance while PLDA does need one.
- C-PLDA (10.11%) vs. A-PLDA (10.86%) vs. CORAL+ (10.54%): All these three methods are unsupervised and are based on PLDA back-end. It shows that C-PLDA performs significantly better than the other two PLDA-based back-end adaptation methods, although it does not require any knowledge of the source-domain PLDA parameters. This further strengthens the evidence that *the pseudo labels generated by the clustering algorithm convey reasonable speaker-related information and can be used to train a strong PLDA model*. This, in turn, supports the premise that the embeddings extracted by the pre-trained network maintain sufficient discriminant strength in terms of speakers; otherwise, the simple clustering cannot generate such high-quality pseudo labels.
- C-LDA vs. A-PLDA (10.66% vs. 10.86%) vs. CORAL+ (10.54%): In comparison, C-LDA is simpler than the other two PLDA-based domain adaptation methods, both in the adaptation process and scoring process, but obtained similar or even better performance. This suggests that if the distributional distortion can be well alleviated, a simple cosine scoring is sufficient to gain good performance. Considering the

simplicity and the good acceptance of cosine scoring, *C-LDA is preferable compared to A-PLDA, CORAL+, and even to the more powerful but also more complex C-PLDA*. This argument will be further strengthened in the single-domain adaptation experiments presented shortly.

4.4. Further Study

4.4.1. Single-Domain Adaptation

Previous experiments have successfully validated the superior performance of C-LDA in CNC1.eval, which involves multiple genres so can be regarded as a multi-domain adaptation test. To further assess its potential, we chose three genres (interview, entertainment, and singing) to test its performance, where each genre is treated as a particular domain. The three genres were chosen because there is abundant data in these genres (ref. Table 1), making them suitable for evaluating the performance of C-LDA in single-domain adaptation.

To perform the evaluation, we need to construct the evaluation trials as they are not involved in the standard CNC1.eval set. For each speaker of each genre, we concatenated the utterances to create a 20 s enrollment segment, while the remaining utterances were kept for testing. By cross-pairing these enrollment and test utterances, we conducted single-domain evaluation. All the unsupervised methods were considered, and the experimental results are shown in Table 5. It is important to note that for every single genre, we independently adapt the source-domain model (SSL-DA and C-FT), the back-end model (A-PLDA, CORAL+, C-PLDA) or the embeddings (C-LDA).

Table 5. EER performance on multi/single-domain adaptation with all considered unsupervised methods. Note that the column CNC1.dev shows the multi-domain adaptation.

Scoring	Method	Domain			
		CNC1.dev	CNC1.dev. Interview	CNC1.dev. Entertainment	CNC1.dev. Singing
	Pretrain	14.22	c9.35	10.88	28.08
Cosine	SSL-DA	11.54	7.02	7.89	16.93
	C-FT	11.08	7.62	8.99	21.54
	C-LDA	10.66	6.21	7.12	18.57
PLDA	A-PLDA	10.86	6.53	7.29	18.89
	CORAL+	10.54	6.72	7.36	19.23
	C-PLDA	10.11	5.35	7.18	18.89

Firstly, C-LDA excels in both multi-domain and single-domain scenarios due to the limited data for each single domain compared to the multi-domain scenario. The limited data provide a relative advantage for simple adaptation methods such as LDA, as complex front-end adaptation may suffer from over-fitting. This explains why C-LDA outperforms SSL-DA and C-FT. Additionally, the limited data, especially the limited number of speakers, may prevent an accurate estimation of between-speaker covariance. This does not impact C-LDA, as the between-class covariance does not affect its performance when scoring is based on cosine distance. However, for PLDA, the between-class covariance is essential and involved in the score computation, explaining why C-LDA outperforms A-PLDA, CORAL+, and C-PLDA in most cases.

Secondly, in single-domain scenarios, SSL-DA outperforms C-FT, although it is weak in the multi-domain test. This could be attributed to the fact that there is no need to perform domain alignment for SSL-DA in the single-domain test, and contrastive loss seems mild and causes little over-fitting when used to adapt the front-end neural net. In contrast, C-FT seems to severely over-fit the training data, especially the errors in the pseudo labels.

Thirdly, SSL-DA outperforms all other methods in the singing domain, whereas this is not the case in other domains. This could be because the singing domain differs significantly from the source domain (mostly interview speech). This indicates that if there

is a significant domain shift, adapting the front-end network is still necessary, making SSL-DA better than the embedding adaptation (C-LDA) and back-end adaptation (A-PLDA, CORAL+, C-PLDA) methods. Additionally, preventing over-fitting is important, which is why SSL-DA outperformed C-FT.

Finally, C-PLDA exhibits superior performance in the interview domain compared to C-LDA but does not demonstrate performance advantages in the entertainment and singing domains. This can be explained by the data volume: there are more interview data than entertainment and singing data, which gives C-PLDA more advantage over C-LDA in the interview domain. This is because C-PLDA requires estimating more parameters than C-LDA, i.e., the between-class covariance.

4.4.2. Parameter Sensitivity

This experiment explores the impact of hyper-parameter settings on C-LDA performance, including the dimensionality of the LDA projection space (referred to as LDA dimensionality) and the number of clusters in AHC, on the results of C-LDA. The results are illustrated in Figure 3.

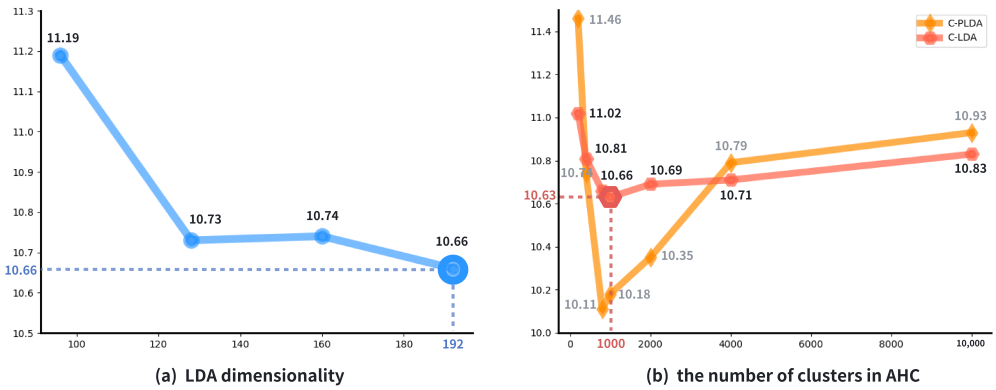

Figure 3. The impact of (**a**) the dimensionality of LDA projection space and (**b**) the number of clusters in AHC on the results of C-LDA and C-PLDA.

Firstly, we set the number of clusters at 800 (the true number of speakers in CNC1.dev is 797) and examined the impact of the LDA dimensionality. From Figure 3a, it is evident that the optimal LDA dimension for C-LDA is 192, which matches the dimension of the original embedding vector. This strongly supports our hypothesis that the embedding vectors, especially those trained with AAM softmax, fully encapsulate speaker information even for cross-domain data. Therefore, no dimensions should be discarded, and simple rotation and scaling are sufficient to make it suitable for cosine scoring.

Secondly, we fixed the LDA dimension at 192 and examined the impact of the number of clusters in the agglomerative hierarchical clustering. As shown in Figure 3b, C-LDA is highly robust against the setting of the number of clusters, with negligible impact on performance. Even when the value is set to 5 times the actual number of speakers (e.g., 4000), the EER only increases by 0.47%. Furthermore, we note that overestimating the number of clusters has a more tolerable impact than underestimating it. In summary, all observations demonstrate that C-LDA is not sensitive to the setting of the clustering algorithm, which is particularly important for its practical usage.

For comparison, we also plot the C-PLDA results in Figure 3b. It is evident that C-PLDA is more impacted by the setting of the number of clusters. When the number is set to be the number of true speakers, C-PLDA achieves better performance than C-LDA; however, if the number is set far from the correct value, C-PLDA shows worse performance than C-LDA. This is because with an incorrect setting, the between-class covariance is poorly estimated, leading to suboptimal performance. Since C-LDA is irrelevant to the between-class covariance, it is less sensitive to this setting.

4.4.3. Shift, Rotation and Scaling

In the final experiment, we analyze which part of the C-LDA transform resolves the distributional distortion problem. As discussed in Section 3.1, there are four operations in the LDA transform: (1) Global shift to centralize the data; (2) Axes rotation to diagonalize the within-class covariance; (3) Scale to whiten the within-class covariance; (4) Axis rotation again to diagonalize the between-class covariance. Table 6 provides the performance with these operations applied successively, with the test setting consistent with the main experiment reported in Section 4.3.

The results show that the global shift offers the most substantial improvement, and scaling also contributes, although not as notably as the global shift. Axis rotation does not impact the performance as cosine similarity is invariant to the rotation operation. This indicates that the main cross-domain distortion is attributed to the global shift in the embedding space, at least with cosine scoring.

Table 6. Performance after each stage of the C-LDA operations.

Method	EER(%)	minDCF (P_{tar} = 0.05)
Pretrain	14.22	0.5137
Global Shift	11.48	0.4351
+Rotation (1st)	11.48	0.4351
++Scaling	10.66	0.4122
+++Rotation (2nd)	10.66	0.4122

5. Discussion

We advocate for C-LDA for the following reasons:

(1) If the target-domain speaker vectors extracted by the source-domain model remain highly separable in the embedding space, relatively accurate pseudo-labels can be obtained through simple clustering, which then allows for a reasonably good C-LDA model. Fortunately, modern speaker embedding models trained with max margin loss seem to maintain this cross-domain discrimination.
(2) Full-rank LDA involves only a linear transform, so it is less prone to overfitting issues compared to front-end adaptation methods, especially with potential errors in the pseudo labels.
(3) It does not depend on accurate between-class covariance, making it superior to back-end adaptation methods including C-PLDA, especially when the number of speakers in the adaptation data is limited.
(4) C-LDA is a knowledge-free approach and does not require knowledge of both the front-end and back-end models, providing an important advantage for practical usage compared to most existing methods.

These advantages of C-LDA mentioned above are also its shortcomings. For example, if the discrimination power of the target-domain embeddings is lost, then C-LDA will completely fail. Moreover, its linear form and the neglect of between-class covariance limit its capacity and make it unable to handle more complex conditions than the test scenarios in our experiment (although CN-Celeb is one of the most complex datasets so far). Finally, its lack of knowledge about the front-end and back-end models prevents it from leveraging the information there. Nevertheless, these limitations do not pose much concern so far, at least in our sufficiently complex experiments.

6. Conclusions

This paper explores the issue of domain mismatch in speaker verification tasks based on deep embedding networks. We hypothesize that the primary problem in domain mismatch lies not in the decrease in speaker discriminability, but instead of the presence of systematic distortions in the embedding space such as shifting, rotation, and scaling.

In light of this assumption, we propose a distribution normalization approach using a full-rank LDA and train it in an unsupervised manner by using pseudo labels produced by hierarchical clustering. We term this method as C-LDA. Theoretical analysis shows that C-LDA when combined with cosine scoring, can compensate for global shift and scaling on within-class covariance.

Our experiments were conducted using VoxCeleb2 as the source-domain dataset and CN-Celeb1 as the target-domain dataset, which contains multiple genres and represents a multi-domain adaptation test scenario. The results highlight that C-LDA achieves an EER of 10.66% on the CN-Celeb1 evaluation set, without any knowledge of the source model and any human labels for the target data. This performance surpasses that of most of the compared front-end and back-end adaptation methods. In more detailed experiments, it was found that C-LDA shows even more performance superiority to single-domain tests where the adaptation data is limited. Importantly, C-LDA is not sensitive to the settings of the clustering algorithm, making it highly suitable for practical usage. All these findings strongly support our distributional distortion assumption and underscore C-LDA as a vital tool to tackle the domain-mismatch problem.

Future work involves investigating scenarios with serious domain mismatch, such as the singing genre in our test. For these scenarios, the distributional distortion assumption may not hold and linear compensation could fail. We have seen this trend in our experiment, where self-supervised learning for the front-end model achieves better performance than C-LDA and other adaptation methods. Dedicated research is required to study what happens in the embedding space before designing the solution. Another area of research is to investigate frame-level compensation rather than embedding-level compensation. Visualization tools are important to understand the behavior of the features, and layer-by-layer compensation is considered.

Author Contributions: Conceptualization, W.L., L.L. and D.W.; Methodology, W.L., L.L. and D.W.; Validation, W.L.; Formal analysis, W.L.; Resources, L.L. and D.W.; Writing—original draft, W.L.; Writing—review & editing, L.L. and D.W.; Visualization, W.L.; Supervision, L.L. and D.W.; Funding acquisition, L.L. and D.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (NSFC) under Grants No.62171250/62301075 and also the Fundamental Research Funds for the Central Universities of China.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in VoxCeleb2 at https://doi.org/10.1016/j.csl.2019.101027, reference number [49] and in CN-Celeb1 at https://doi.org/10.1109/ICASSP40776.2020.9054017, reference number [50].

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

References

1. Li, C.; Ma, X.; Jiang, B.; Li, X.; Zhang, X.; Liu, X.; Cao, Y.; Kannan, A.; Zhu, Z. Deep speaker: An end-to-end neural speaker embedding system. *arXiv* **2017**, arXiv:1705.02304.
2. Snyder, D.; Garcia-Romero, D.; Povey, D.; Khudanpur, S. Deep neural network embeddings for text-independent speaker verification. In Proceedings of the INTERSPEECH, Stockholm, Sweden, 20–24 August 2017; Volume 2017, pp. 999–1003.
3. Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; Khudanpur, S. X-vectors: Robust dnn embeddings for speaker recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5329–5333.
4. Desplanques, B.; Thienpondt, J.; Demuynck, K. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. *arXiv* **2020**, arXiv:2005.07143.
5. Zhou, T.; Zhao, Y.; Wu, J. Resnext and res2net structures for speaker verification. In Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 19–22 January 2021; pp. 301–307.
6. Okabe, K.; Koshinaka, T.; Shinoda, K. Attentive statistics pooling for deep speaker embedding. *arXiv* **2018**, arXiv:1803.10963.

7. Tang, Y.; Ding, G.; Huang, J.; He, X.; Zhou, B. Deep speaker embedding learning with multi-level pooling for text-independent speaker verification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6116–6120.
8. Xie, W.; Nagrani, A.; Chung, J.S.; Zisserman, A. Utterance-level aggregation for speaker recognition in the wild. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 5791–5795.
9. Gao, Z.; Song, Y.; McLoughlin, I.; Li, P.; Jiang, Y.; Dai, L.R. Improving Aggregation and Loss Function for Better Embedding Learning in End-to-End Speaker Verification System. In Proceedings of the INTERSPEECH, Graz, Austria, 15–19 September 2019; pp. 361–365.
10. Wang, S.; Rohdin, J.; Plchot, O.; Burget, L.; Yu, K.; Černockỳ, J. Investigation of specaugment for deep speaker embedding learning. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7139–7143.
11. Liu, T.; Lee, K.A.; Wang, Q.; Li, H. Disentangling Voice and Content with Self-Supervision for Speaker Recognition. *arXiv* **2023**, arXiv:2310.01128.
12. Cai, D.; Cai, W.; Li, M. Within-sample variability-invariant loss for robust speaker recognition under noisy environments. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6469–6473.
13. Zhang, C.; Yu, M.; Weng, C.; Yu, D. Towards robust speaker verification with target speaker enhancement. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 6693–6697.
14. Li, L.; Liu, R.; Kang, J.; Fan, Y.; Cui, H.; Cai, Y.; Vipperla, R.; Zheng, T.F.; Wang, D. CN-Celeb: Multi-genre speaker recognition. *Speech Commun.* **2022**, *137*, 77–91. [CrossRef]
15. Dua, M.; Sadhu, A.; Jindal, A.; Mehta, R. A hybrid noise robust model for multireplay attack detection in automatic speaker verification systems. *Biomed. Signal Process. Control* **2022**, *74*, 103517. [CrossRef]
16. Wang, X.; Li, L.; Wang, D. VAE-based domain adaptation for speaker verification. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 535–539.
17. Villalba, J.; Chen, N.; Snyder, D.; Garcia-Romero, D.; McCree, A.; Sell, G.; Borgstrom, J.; García-Perera, L.P.; Richardson, F.; Dehak, R.; et al. State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and speakers in the wild evaluations. *Comput. Speech Lang.* **2020**, *60*, 101026. [CrossRef]
18. Lin, W.; Mak, M.W.; Li, N.; Su, D.; Yu, D. A framework for adapting DNN speaker embedding across languages. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2810–2822. [CrossRef]
19. Zhang, C.; Ranjan, S.; Hansen, J.H. An Analysis of Transfer Learning for Domain Mismatched Text-independent Speaker Verification. In Proceedings of the Odyssey, Stockholm, Sweden, 26–29 June 2018; pp. 181–186.
20. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1345–1359. [CrossRef]
21. Li, J.; Han, J.; Song, H. CDMA: Cross-Domain Distance Metric Adaptation for Speaker Verification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 7197–7201.
22. Mccree, A.; Shum, S.; Reynolds, D.; Garcia-Romero, D. Unsupervised Clustering Approaches for Domain Adaptation in Speaker Recognition Systems. In Proceedings of the Speaker and Language Recognition Workshop (Odyssey 2014), Joensuu, Finland, 16–19 June 2014; pp. 265–272.
23. Lee, K.A.; Wang, Q.; Koshinaka, T. The CORAL+ algorithm for unsupervised domain adaptation of PLDA. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 5821–5825.
24. Wang, Q.; Okabe, K.; Lee, K.A.; Koshinaka, T. Generalized domain adaptation framework for parametric back-end in speaker recognition. *IEEE Trans. Inf. Forensics Secur.* **2023**, *18*, 3936–3947. [CrossRef]
25. Burget, L.; Plchot, O.; Cumani, S.; Glembek, O.; Matějka, P.; Brümmer, N. Discriminatively trained probabilistic linear discriminant analysis for speaker verification. In Proceedings of the 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP), Prague, Czech Republic, 22–27 May 2011; pp. 4832–4835.
26. Hu, H.R.; Song, Y.; Liu, Y.; Dai, L.R.; McLoughlin, I.; Liu, L. Domain robust deep embedding learning for speaker recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 7182–7186.
27. Li, J.; Liu, W.; Lee, T. EDITnet: A Lightweight Network for Unsupervised Domain Adaptation in Speaker Verification. *arXiv* **2022**, arXiv:2206.07548.
28. Hu, H.R.; Song, Y.; Dai, L.R.; McLoughlin, I.; Liu, L. Class-aware distribution alignment based unsupervised domain adaptation for speaker verification. In Proceedings of the INTERSPEECH, Songdo, Republic of Korea, 18–22 September 2022.
29. Sun, B.; Saenko, K. Deep coral: Correlation alignment for deep domain adaptation. In *Proceedings of the Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, 8–10 and 15–16 October 2016*; Proceedings, Part III 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 443–450.

30. Alam, M.J.; Bhattacharya, G.; Kenny, P. Speaker verification in mismatched conditions with frustratingly easy domain adaptation. *Odyssey* **2018**, *25*, 176–180.
31. Lin, W.W.; Mak, M.W.; Li, L.; Chien, J.T. Reducing domain mismatch by maximum mean discrepancy based autoencoders. *Odyssey* **2018**, *23*, 162–167.
32. Izenman, A.J. Linear discriminant analysis. In *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 237–280.
33. Garcia-Romero, D.; McCree, A. Supervised domain adaptation for i-vector based speaker recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 4047–4051.
34. Garcia-Romero, D.; McCree, A.; Shum, S.; Brummer, N.; Vaquero, C. Unsupervised domain adaptation for i-vector speaker recognition. In Proceedings of the Odyssey: The Speaker and Language Recognition Workshop, Joensuu, Finland, 16–19 June 2014; Volume 8.
35. Li, R.; Zhang, W.; Chen, D. The CORAL++ algorithm for unsupervised domain adaptation of speaker recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 7172–7176.
36. Wang, Q.; Rao, W.; Sun, S.; Xie, L.; Chng, E.S.; Li, H. Unsupervised domain adaptation via domain adversarial training for speaker recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4889–4893.
37. Yin, Y.; Huang, B.; Wu, Y.; Soleymani, M. Speaker-invariant adversarial domain adaptation for emotion recognition. In Proceedings of the 2020 International Conference on Multimodal Interaction, Virtual Event, The Netherlands, 25–29 October 2020; pp. 481–490.
38. Wang, Q.; Rao, W.; Guo, P.; Xie, L. Adversarial training for multi-domain speaker recognition. In Proceedings of the 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP), Hong Kong, China, 24–27 January 2021; pp. 1–5.
39. Lin, W.; Mak, M.M.; Li, N.; Su, D.; Yu, D. Multi-level deep neural network adaptation for speaker verification using MMD and consistency regularization. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6839–6843.
40. Zhou, Z.; Chen, J.; Wang, N.; Li, L.; Wang, D. An Investigation of Distribution Alignment in Multi-Genre Speaker Recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024.
41. Chen, Z.; Wang, S.; Qian, Y. Self-supervised learning based domain adaptation for robust speaker verification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 5834–5838.
42. Mao, H.; Hong, F.; Mak, M.w. Cluster-Guided Unsupervised Domain Adaptation for Deep Speaker Embedding. *IEEE Signal Process. Lett.* **2023**, *30*, 643–647. [CrossRef]
43. Lin, W.; Li, L.; Wang, D. Multi-Domain Adaptation by Self-Supervised Learning for Speaker Verification. *arXiv* **2023**, arXiv:2309.14149.
44. Cai, Y.; Li, L.; Abel, A.; Zhu, X.; Wang, D. Deep normalization for speaker vectors. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *29*, 733–744. [CrossRef]
45. Li, L.; Wang, D.; Kang, J.; Wang, R.; Wu, J.; Gao, Z.; Chen, X. A principle solution for enroll-test mismatch in speaker recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 443–455. [CrossRef]
46. Murtagh, F.; Contreras, P. Algorithms for hierarchical clustering: An overview. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2012**, *2*, 86–97. [CrossRef]
47. Dehak, N.; Dehak, R.; Kenny, P.; Brümmer, N.; Ouellet, P.; Dumouchel, P. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In Proceedings of the Tenth Annual Conference of the International Speech Communication Association, Brighton, UK, 6–10 September 2009.
48. Misra, A.; Hansen, J.H. Maximum-likelihood linear transformation for unsupervised domain adaptation in speaker verification. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 1549–1558. [CrossRef]
49. Nagrani, A.; Chung, J.S.; Xie, W.; Zisserman, A. Voxceleb: Large-scale speaker verification in the wild. *Comput. Speech Lang.* **2020**, *60*, 101027. [CrossRef]
50. Fan, Y.; Kang, J.; Li, L.; Li, K.; Chen, H.; Cheng, S.; Zhang, P.; Zhou, Z.; Cai, Y.; Wang, D. CN-Celeb: A challenging Chinese speaker recognition dataset. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7604–7608.
51. Xiang, X.; Wang, S.; Huang, H.; Qian, Y.; Yu, K. Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition. In Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 1652–1656.
52. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv* **2019**, arXiv:1904.08779.
53. Snyder, D.; Chen, G.; Povey, D. Musan: A music, speech, and noise corpus. *arXiv* **2015**, arXiv:1510.08484.

54. Ko, T.; Peddinti, V.; Povey, D.; Seltzer, M.L.; Khudanpur, S. A study on data augmentation of reverberant speech for robust speech recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 5220–5224.
55. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
56. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; et al. The Kaldi speech recognition toolkit. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, Waikoloa, HI, USA, 11–15 December 2011*; IEEE Signal Processing Society: Piscataway, NJ, USA, 2011; number CONF.
57. Wang, D. A simulation study on optimal scores for speaker recognition. *EURASIP J. Audio Speech Music. Process.* **2020**, *2020*, 18. [CrossRef]