information

MDPI

Article

Whisper40: A Multi-Person Chinese Whisper Speaker Recognition Dataset Containing Same-Text Neutral Speech

Jingwen Yang and Ruohua Zhou *

School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 102616, China; 2108110021031@stu.bucea.edu.cn
* Correspondence: zhouruohua@bucea.edu.cn

Abstract: Whisper speaker recognition (WSR) has received extensive attention from researchers in recent years, and it plays an important role in medical, judicial, and other fields. Among them, the establishment of a whisper dataset is very important for the study of WSR. However, the existing whisper dataset suffers from the problems of a small number of speakers, short speech duration, and lack of neutral speech with the same-text as the whispered speech in the same dataset. To address this issue, we present Whisper40, a multi-person Chinese WSR dataset containing same-text neutral speech spanning around 655.90 min sourced from volunteers. In addition, we use the current state-of-the-art speaker recognition model to build a WSR baseline system and combine the idea of transfer learning for pre-training the speaker recognition model using neutral speech datasets and transfer the empirical knowledge of specific network layers to the WSR system. The Whisper40 and CHAINs datasets are then used to fine-tune the model with transferred specific layers. The experimental results show that the Whisper40 dataset is practical, and the time delay neural network (TDNN) model performs well in both the same/cross-scene experiments. The equal error rate (EER) of Chinese WSR after transfer learning is reduced by 27.62% in comparison.

Keywords: speaker recognition; deep learning; whisper dataset; transfer learning; audio reverse

Citation: Yang, J.; Zhou, R. Whisper40: A Multi-Person Chinese Whisper Speaker Recognition Dataset Containing Same-Text Neutral Speech. *Information* **2024**, *15*, 184. https://doi.org/10.3390/info15040184

Academic Editor: Ralf Krestel

Received: 28 February 2024
Revised: 26 March 2024
Accepted: 26 March 2024
Published: 28 March 2024

1. Introduction

Whisper, as an informal form of communication, has received much attention from researchers in recent years [1,2]. Conveying information by speaking in a low voice can avoid being heard by others, and this form of communication is often used in situations where confidentiality is required or to avoid disturbing others. For example, when people need to communicate credit card information and passwords to a biometric system, they speak in a whisper manner to avoid overhearing the content; in some cases, criminals may use whispered speech in telephone conversations to avoid sharing voiceprints with forensic examiners, and for laryngectomy patients, whispered speech is the only form of vocalization. Compared to neutral speech, whispered speech is characterized by a lack of fundamental frequency [3], a flat spectral slope [4], a low signal-to-noise ratio, and a greater susceptibility to environmental influences. This makes research in the field of whispering even more challenging.

The current research on whispered speech mainly includes the task of normal/whisper speech classification and detection, whisper-normal conversion [5], and WSR. The rapid development of Voice Assistant (VA) systems has put forward requirements for classification tasks of whispered and normal speech. Among them, features based on Spectral Information Entropy (SIE) ratios with Gaussian Mixture Model (GMM) classifiers have been proposed for the classification of whispered speech from five different types of speech signals [6]. Deep Neural Networks (DNNs) and Long Short-Term Memory (LSTM) architectures with logarithmic filter bank energies have been proposed for the frame-level whispered-speech-detection task [7]. However, the decoding of LSTM takes more time

and relies on contextually neighboring frames, which affects real-time decision-making. Furthermore, Convolutional Neural Network (CNN)-based classifiers are able to detect whispered speech at the utterance-level in the presence of unbalanced class learning [8]. The spectral phase-based features proposed in [9] can better capture the excitation source information and are effective for whispered speech detection.

For the whisper-normal conversion task, it can be summarized into two main approaches. The first method is the GMM-based conversion method [10,11]. This method decomposes the speech signal from the vocoder into specific speech feature parameters. These include speech spectral features, speech fundamental frequency features, and speech non-periodic features. The three features are trained on three GMM whisper conversion models, respectively. Finally, the predicted speech feature parameters are reconstructed into normal speech by the vocoder. However, since the method is based on feature inter-frame mapping and the speech signal is a continuous signal, this results in poor spectral smoothness of the converted speech using this method. The second method mainly starts from DNN [12]. The whispered speech conversion method based on a Bidirectional Long Short-Term Memory (BiLSTM) network proposed by Meenakshi et al. [13] is able to handle the feature sequence context information effectively. It has the ability to describe the time–domain relationship of speech signals. The experimental results show that this method generates high-quality speech with better speech listening comfort.

Unlike neutral speech speaker recognition, where enrollment is performed using neutral and/or whispered speech data during WSR, only whispered speech data are used in the test phase. This requires the whisper speaker recognition system to be robust to whispered speech without compromising the performance of the neutral speech speaker recognition system as well.

Current mainstream WSR methods can be broadly classified into three categories. The first class of methods focuses on feature extraction [14–18], such as the Mel-scale Frequency Cepstral Coefficient (MFCC), Weighted Instantaneous Frequency, Auditory-inspired Amplitude Modulation Features (AAMFs), Formant-gaps (FoGs) features, and Weighted Modified Linear frequency cepstrum coefficients. Among them, MFCC has been used with GMM and has been used for the speaker recognition task. The method uses maximum a posteriori estimation. The mean hyper-vector of the generalized background model is applied to the speaker frames. From this, the GMM mean hypervector for each speaker is obtained, and a more compact "identity vector" (i-vector) is extracted from this mean hyper-vector. Sarria et al. combined weighted instantaneous frequency features, AAMF, and FoGs features with mean envelope coefficients for WSR. The results show a relative improvement of 3.79% in the effect of FoGs features under whispered speech test conditions, but further research on large corpora is necessary. The second class of methods focuses on feature transformations, such as frequency warping and Feature Mapping (FM) based on DNN [19,20]. Naini A R et al. [21] trained the FM function by optimizing the mean square error between MFCC features and neutral speech features under Dynamic Time Warping (DTW). This method retains only the speaker information and achieves the mapping from whispered speech to neutral speech by maximizing the cosine similarity between the neutral speech and whispered speech i-vectors, achieving a relative improvement of 24%. However, the Mean Squared Error (MSE) does not distinguish between speaker-specific factors, and therefore, MSE-based objective functions are considered inappropriate for WSR. The third category of approaches focuses on the complementary properties of whispered and neutral speech. Sarria-Paja and Falk, in 2017, proposed three new speech feature types to train three different speaker verification systems. Under neutral and whispered speech conditions, the EER improved by 66% and 63%, respectively, compared to the baseline using the traditional MFCC coefficients. In [22], a new modeling technique was introduced by Vestman et al. This technique involves long-term speech analysis based on the joint utilization of Frequency Domain Linear Prediction and Time-varying Linear Prediction (FDLP-TVLP). In their experiments, they used a corpus of CHAINs to test the speech mismatch condition

and demonstrated that the FDLP-TVLP features improved the speaker identification EER by 7–10% over the standard MFCC features.

Because of the limited research on WSR, few studies have attempted to explore its relationship with neutral speech speaker recognition. Significant progress has been made in the research on neutral speech speaker recognition. Among them, factor analysis and DNN-embedding-based methods [23–25] are considered to be the most advanced in the field of speaker recognition. The mainstream methods are TDNN-based x-vector, ResNet34-based r-vector, and ECAPA-TDNN [26]-based speaker embedding extraction. Also, since TDNN is a highly competitive model in the field of speaker recognition, it does not require the precise localization of tokens during its training and is able to express the relationship of speech features in the time dimension. Therefore, this network is introduced in the methodology of this paper. Meanwhile, this paper, for the first time, uses the ResNet34 and ECAPA-TDNN model to train the WSR system and compares the effect with the TDNN model to find the most suitable recognition model for whispered speech. Despite the displacement of the low-frequency resonance peaks in whispered speech compared to neutral speech, it also contains a large amount of content and speaker information. Therefore, the feature and embedding extraction methods for neutral speech speaker recognition are instructive for WSR. Transfer learning is a method of transferring knowledge structures from related domains to complete or improve the task in the target domain, which has been widely used in the fields of speech synthesis and target recognition [27]. Therefore, in this paper, transfer learning is introduced to use the pre-training specific network layer parameters of TDNN as the initial values of the model for WSR. These parameters are used to initialize and fine-tune the parameters of the model trained on the whisper data in order to improve the accuracy of the WSR.

In addition, it is very difficult to introduce advanced DNN-based neutral speech speaker recognition methods into the field of whispered speech due to the lack of a large whisper dataset, as these methods all require large amounts of speaker audio data to build more robust speaker models. Based on this paper, a new whisper dataset, Whisper40, is built, and a WSR system in transfer learning mode is constructed. This dataset consists of 40 people, and each speaker contains neutral speech and whispered speech of the same-text to study the performance of the WSR system across scenarios. Secondly, this paper introduces advanced models in the field of speaker recognition into the WSR system and, at the same time, introduces a new data-augmentation method of audio reverse, as well as same/cross-scene experiments, to further compare the differences between whispered speech and neutral speech. Finally, the small scale of WSR training data makes WSR ineffective. In this paper, based on the idea of transfer learning, the empirical knowledge of specific network layers in the pre-trained model is transferred to the WSR model. The self-collected Whisper40 data and CHAINs data are also utilized for fine-tuning until fitting. The experimental results show that the self-collected Whisper40 dataset is practical, and the adopted data-augmentation method effectively reduces the EER of the WSR system, and the EER of the Chinese WSR after transfer learning is relatively reduced by 27.62%.

The paper is structured as follows. Firstly, the created Whisper40 dataset is compared with existing whisper datasets. Then, we document the collection and organization of Whisper40 and describe the structure of the dataset in detail. Finally, we constructed a WSR system under a transfer learning pattern, which proves that the self-collected dataset has practicality and also shows the importance of the TDNN model in the WSR, as well as the effectiveness of the transfer learning strategy under a small data volume.

2. Related Work

Speech datasets are the research basis for conducting speech-related technologies and are indispensable resources for speech signal research. However, the research on whispered speech is not yet mature at home and abroad, and there is no large-scale and relatively standardized whispered speech dataset, which severely limits the research on WSR. In [1], a cross-modal speaker verification experiment was conducted by combining the fractional

calibration and embedding compensation methods for vocal modes with an x-vector-based speaker verification system, which shows that further exploitation of advanced knowledge of speaker recognition technology is essential.

Due to the special characteristics of whispered speech pronunciation and its low signal-to-noise ratio, the collection of a whispered dataset is more difficult than the creation of neutral speech datasets. Only a small number of whispered speech corpuses are available, and it has not been possible to evaluate the performance of cross-modal WSR systems. The Institute of Acoustics of Nanjing University, with support from the Natural Science Foundation of China, has established a Chinese whisper dataset containing 1172 characters and 98 near-syllabic words in a single female voice [28]. Each word in this dataset was recorded 20 times in whispered speech and 1 time in neutral speech; each group of words was recorded 10 times in whispered speech and 1 time in neutral speech. Due to the small number of people in this dataset and the short duration of each audio sample, it is not possible to effectively evaluate the WSR system. A small English whispered dataset was recorded in text [29] using simultaneous dual recording with a throat microphone and a normal microphone, using both whispered and neutral speech for each sentence of each speaker. However, its size is so small that in order to improve the recognition rate, only multiple methods of adaptation can be performed to eliminate the effect of insufficient data. The WTIMIT dataset is constructed after the TIMIT dataset [30], which is usually used for the study of the automatic recognition of phonemes. It is a systematically organized collection of paired whispered speech and neutral speech produced by multiple speakers. The first phase of the project collects the speech of 20 Singaporeans, and the second phase collects the speech of 28 North Americans. The dataset is designed to be phonetically balanced and large enough to support the statistical data required for training acoustic models in speech recognition. The CHAINs dataset [31] contains 36 speakers, 8 of which are from the UK and the USA, and the remaining 28 are from the east of Ireland, each with 37 utterances. The CHAINs dataset was recorded in six different modes, including personal speech, repetition, synchronized speech, repetitive synchronized mimicry, rapid speech, and whispered speech, with personal speech and whispered speech facilitating verification experiments on whispered speakers. However, the current wTIMIT and CHAINs datasets are all recorded from English speakers, so it is necessary to build a new corpus in the Chinese language. Focus on the potential impact of linguistic differences on the WSR system.

In this paper, we present a combined dataset of whispered speech and neutral speech, labeled with labels, such as speaker, gender, and mode of speech. Its comparison with existing whispered datasets is shown in Table 1. This paper also demonstrates the usefulness of the Whisper40 dataset by implementing a baseline WSR system. In the following sections, we describe in detail the content structure of the dataset and the data collection process, and experimentally analyze the performance of state-of-the-art speaker recognition methods on this dataset. The dataset is available at https://github.com/Lijingze666/whisper40 (accessed on 1 February 2024) at any time.

Table 1. Comparison of Whisper40 dataset with existing whisper datasets.

Dataset	Number of Male Speakers	Number of Female Speakers	Number of Neutral Speech Samples Per Speaker	Number of Whisper Speech Samples Per Speaker
CHAINs	16	20	37	37
wTIMIT	24	24	450	450
Whisper40	21	19	40	40

3. Data Collection

3.1. Speaker Recruitment

In this paper, a new whisper dataset called Whisper40 has been created. Producing a corpus generally requires the consideration of speaker information, utterance design

specifications, recording specifications, data storage technology specifications, and dataset labeling specifications. 40 students were recruited at Beijing Architecture University to participate in the recording of this dataset, including 21 male students and 19 female students, ranging from 22 to 26 years old. The speakers were all native Chinese speakers with no language barriers and were able to speak in whispers. Considering the universal applicability of whispered speech, there are no special requirements for the origin and accent of the speaker. Moreover, these speakers had different places of origin and most of their words were influenced by dialects to some extent, enhancing the diversity of data for speaker verification. Each speaker recorded 20 speeches at a time, with an interval of 2–5 days between each recording. The gender and age distribution of the speakers is shown in Figure 1.

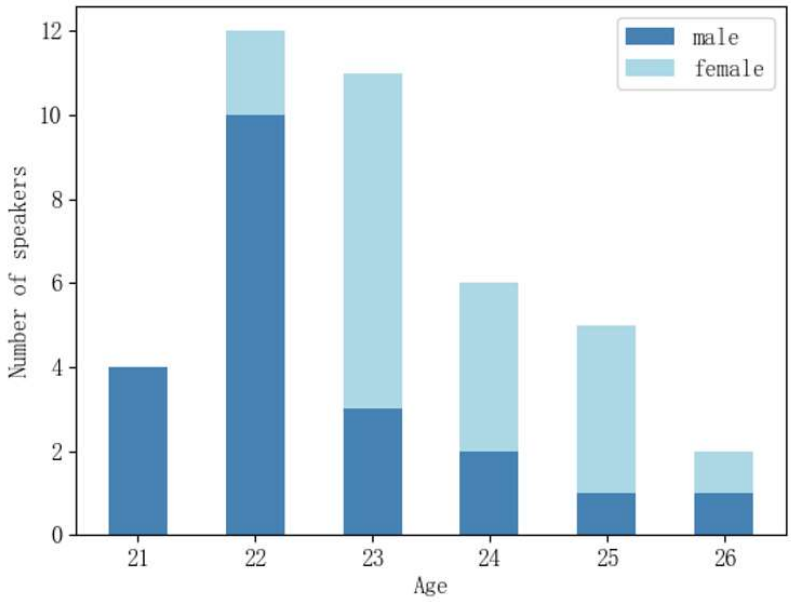


Figure 1. Gender and age distribution of speakers in the Whisper40 dataset.

Before whispered speech collection, each speaker was first familiarized with the corpus content and the characteristics of whispered speech. The speakers read aloud in Mandarin, making their speech clear and fluent. Due to the long sound length of whispered speech, special attention is needed when reading aloud. There should be a certain interval in the middle of reading a sentence. Due to the special nature of whispered speech recording and the high sensitivity of the equipment, the speaker is required to follow the arrangements as much as possible during the recording and try to avoid small sounds (such as moving tables and chairs). In the process of reading aloud, try to keep the voice level consistent and avoid obvious ups and downs.

3.2. Recording Settings

Because whispered speech is pronounced with less energy and a low signal-to-noise ratio, it places demands on the recording environment. Speakers were asked to read prescribed content in a quiet environment in front of a 16-way microphone array or a computer microphone. There was no obstruction between the speaker and the recording device during the recording process to avoid blocking the airflow to be picked up by the microphone. The ambient noise in the recording room was measured using Audio Toolbox before each recording session, at approximately 30 dB each time, and the weather for the day was also assessed before recording. Windy and thunderstorms were avoided to prevent outdoor noise from interfering with whispered speech. The voice quality was checked after each recording. If it did not meet the standard, it was re-recorded to ensure the quality of the whisper dataset.

In order to be used effectively for the study of WSR, a combination of long and short sentences was chosen for the content of the corpus. Each speaker reads uniform textual content consisting of 40 sentences of neutral speech and 40 sentences of whispered speech, of which, the first 30 are long sentences with an average duration of 15.17 s and the last 10 are short sentences with an average duration of 4.01 s. The sampling rate of these recordings was usually 16 kHz or 44 kHz. The audio files were recorded in 2-channel stereo format and then converted to mono 16 kHz sampling rate wav format after preprocessing. The length distribution of whispered speech is shown in Figure 2.

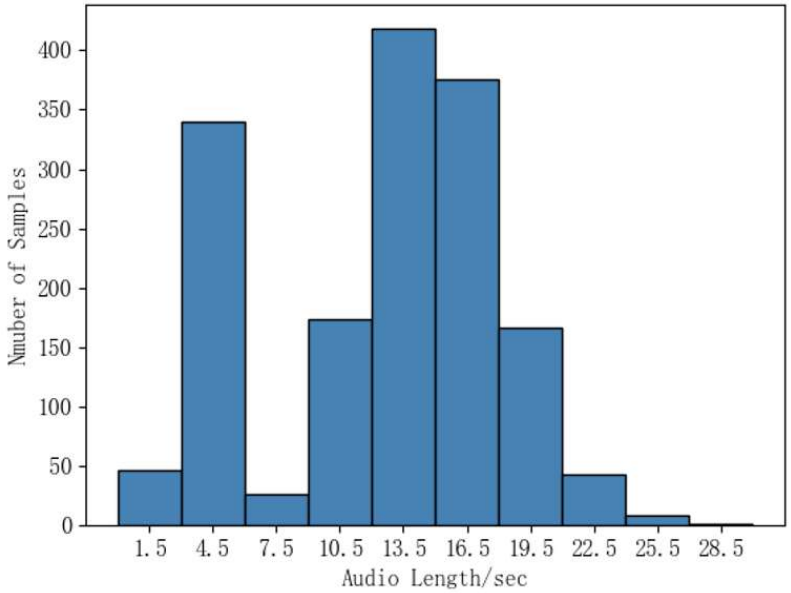


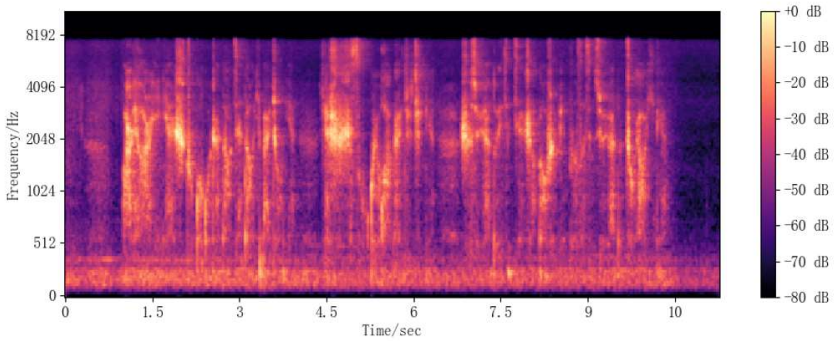
Figure 2. Distribution of audio lengths in the Whisper40 dataset.

3.3. Dataset Organization

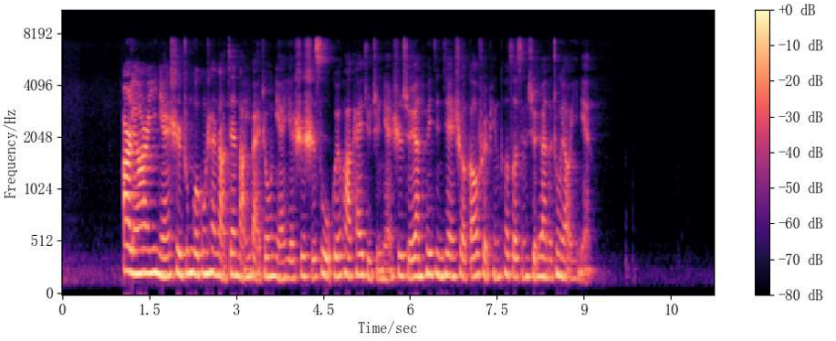
The dataset contains a total of 3200 utterances, including 1600 each of neutral speech and whispered speech, totaling 655.897 min of speech data. It consists of 40 speakers (21 males and 19 females) performing neutral speech and whispered speech in Chinese. Each speaker is assigned a different number and is divided into a training set and an evaluation set in order to verify the reproducibility of the experiment and to conduct follow-up studies. Each speaker title has subfolders containing neutral speech files and whispered speech files. Each speaker audio file was divided into a training subset and a test subset. In addition, the 40 speakers in the dataset were divided into two subsets as shown in Table 2. One speaker's 5 s neutral speech and whispered speech audio were selected and plotted with Mel spectrograms, as shown in Figure 3. It can be seen that when whispered speech is produced, the vocal folds remain open but do not vibrate, so there is no fundamental frequency and the energy is low.

Table 2. Statistics for the Whisper40 dataset.

Dataset	Enroll	Test	All
# of whisper speech/per speaker	3	37	40
# of whisper speech/all speaker	120	1480	1600
# of whisper speech seconds	1654.82	18,149.75	19,804.57
# of neutral speech seconds	1611.66	17,937.56	19,549.22
Max # of seconds/speaker	22.12	30.46	30.46
Min # of seconds/speaker	8.10	1.96	1.96
Avg # of seconds/speaker	13.79	12.26	12.38



(**a**)



(**b**)

Figure 3. Mel spectrograms of whispered speech compared to neutral speech. (**a**) Mel spectrograms of whispered speech. (**b**) Mel spectrograms of neutral speech.

4. Methods

In this section, we construct a TDNN model in transfer learning mode for WSR. Also, for the first time, ResNet34 and ECAPA-TDNN models that have performed well in the field of speaker recognition have been introduced into WSR system, in order to compare and discuss the effect.

4.1. TDNN Model Structure

TDNN is a feedforward neural network with a multilayer structure. Unlike fully connected feedforward neural networks, the output units of TDNN are only connected to the input units within a certain time range. Thus, it captures the temporal relationships between neighboring frames. It aggregates variable-length inputs across time to create fixed-length representations that capture speaker features. The shallow network units are connected over a narrow time range, while the deep network learns information over a wider time range. The network structure of TDNN is shown in Figure 4.

The statistical pooling layer in the network architecture is responsible for mapping frame-level features to segment-level features. Speaker embeddings are extracted from the bottleneck layer before the output layer. The method follows an end-to-end system that uses a delayed DNN to generate embeddings combined with a similarity metric. It performs embedding comparisons by means of independently trained classifiers (e.g., Probabilistic Linear Discriminant Analysis, PLDA). Firstly, short-time, frame-level context is extracted using time delay. The statistical pooling layer aggregates the input segments and calculates the mean and standard deviation. Afterwards, the speaker is classified by the DNN, and the resulting segment-level speaker embedding [32] is called an x-vector. The detailed parameters of this architecture are shown in Table 3, where N denotes the number of speakers in the training set.

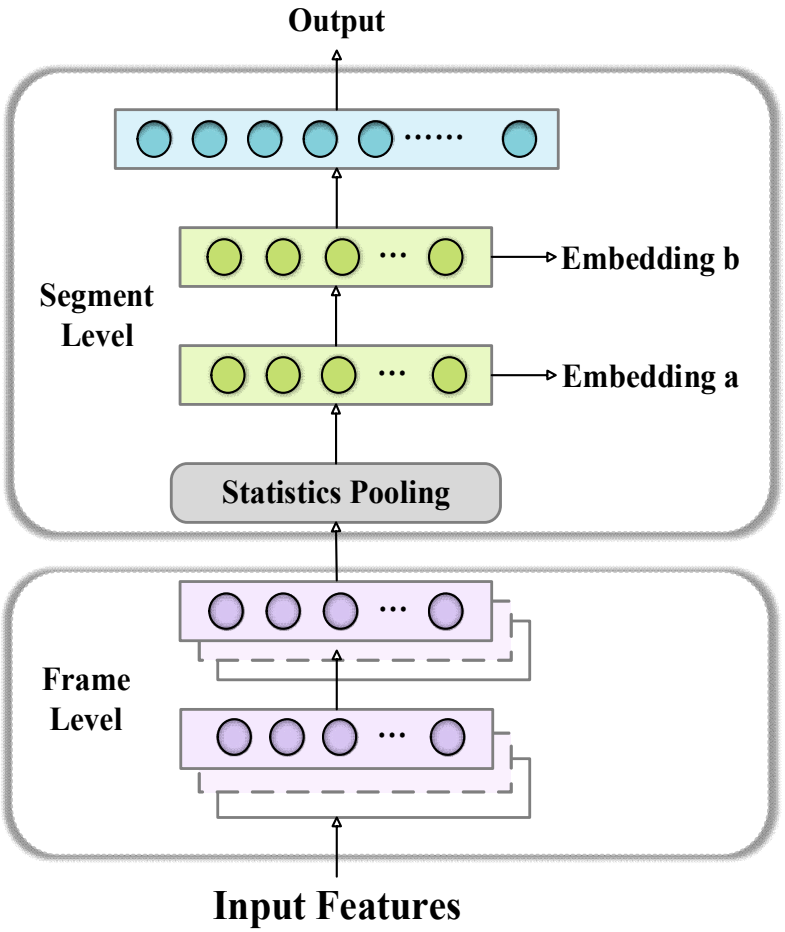


Figure 4. The network structure of TDNN.

Table 3. Detailed structural parameters of the TDNN model.

Layer	Layer Context	Total Context	Output
Frame1	{−2, +2}	5	512
Frame2	{−2, 0, +2}	9	512
Frame3	{−3, 0, +3}	15	512
Frame4	{0}	15	512
Frame5	{0}	15	1500
Statistic Pooling	[0, T]	T	3000
Segment6	{0}	T	512
Segment7	{0}	T	512
Softmax	{0}	T	N

In the method of this paper, the x-vector of the training set is used to train the PLDA model, which is subsequently used for recognition scoring. The parameters ξ, and Σ of the PLDA model are iteratively solved based on the training data in combination with the classical data clustering algorithm (Expectation Maximization, EM). In the testing phase, it is calculated whether two audios are generated in the same speaker space, regardless of intra-class spatial differences.

Use the log-likelihood ratio to calculate the score given in Equation (1).

$$Score = \log \frac{P(\eta_1, \eta_2 | H_s)}{P(\eta_1 | H_d) P(\eta_2 | H_d)} \quad (1)$$

where η_1 and η_2 represent the embedding of the two test sentences. The hypothesis that the two voices are from the same space is H_s, and the hypothesis that they are from different spaces is H_d. $P(\eta_1, \eta_2|H_s)$ is the likelihood function of the two audios being from the same space, and $P(\eta_1|H_d)P(\eta_2|H_d)$ represents the likelihood function of different spaces, respectively. The degree of similarity between the two test speeches is measured by calculating the log-likelihood ratio. The higher the score, the higher the likelihood that the two voices belong to the same speaker.

4.2. WSR in Transfer Learning Mode

Transfer learning is an approach to deep learning that refers to the impact of one type of learning on another or the impact of acquired experience on the completion of other activities. Transfer learning is widespread in the learning of various knowledge, skills, and social norms. With the emergence of more and more deep learning application scenarios, a large amount of labeled data is required to achieve better supervised learning performance. This is a task that requires huge human and material resources. Therefore, transfer learning is receiving more and more attention. The main idea of transfer learning is to transfer labeled data or knowledge structures from related domains to complete or improve the learning effect of the target domain or task.

Extracting speaker features using DNNs is currently a widely interested research area in speaker recognition, e.g., the representation called x-vector is extracted from TDNN. And in this embedded system, artificially increasing the training data using noise and reverberation is a very effective strategy to improve the recognition performance. Most of the current research for WSR is based on i-vector, so it is very necessary to extend the advanced DNN-based normal speaker recognition methods to the field of whisper. Meanwhile, considering the scarcity of the whispered corpus and the large gap between neutral speech and whispered speech, this paper adopts a transfer learning method. The TDNN embedding extraction network is trained using the normal source domain dataset. It is determined that its 5-layer frame-level layer, 2-layer segment-level layer, and pooling layer can be simultaneously applied under the whispered data of the target domain. Transfer the empirical knowledge learnt from neutral speech to the whispered speech domain simultaneously with fine-tuning to help improve the target domain WSR performance.

The flow of the WSR system based on transfer learning proposed in this paper is shown in Figure 5. The whole recognition process consists of a training phase and a testing phase. The training phase requires a number of training whispered speech segments of the speaker, and the feature parameters of these speech segments are extracted to be used as criteria to train the system to learn and thus build a specific speaker module. In the testing phase, PLDA scores are used to measure the similarity between the enroll audio and the test audio in order to calculate the embedding similarity. The feature extraction process converts the input audio into spectrogram features. The acoustic features are extracted by the speaker embedding extractor in order to be able to obtain vectors that contain the features of the speakers. Different neural networks trained with different embedding extractors can obtain different speaker representations. The main purpose of extracting the speaker representation vectors is to map the speaker acoustic information to the specified speaker identity.

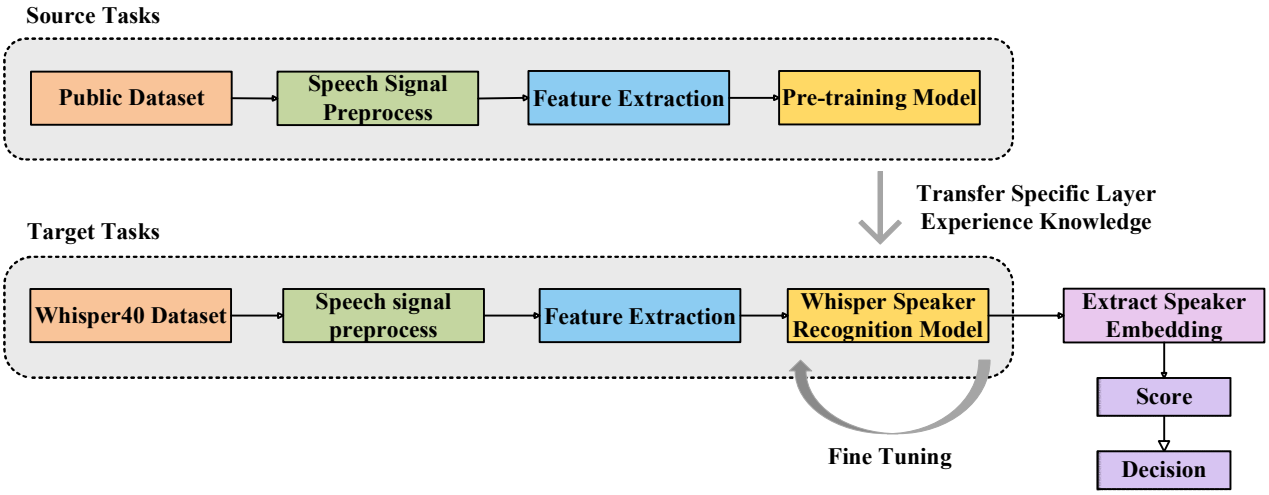


Figure 5. Flowchart of a WSR system based on TDNN and transfer learning.

4.3. Metrics

Uniform metrics are needed for different recognition techniques to measure the strengths and weaknesses of the systems. Unlike the mean average precision (mAP) metric commonly used in classification problems, the commonly used scoring metrics in speaker recognition are EER and the Minimum Detection Cost Function (minDCF). EER is one of the common model evaluation metrics, commonly used in speaker recognition, fingerprint recognition, face recognition and other fields and is closely related to the false rejection rate (FRR) and false acceptance rate (FAR). For the binary classification problem, the combination of its actual classification and predicted classification can be divided into Table 4.

Table 4. Combined division of actual and predicted classifications.

Actual Classification	Predictive Classification	
	Positive	**Negative**
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

FAR means the proportion of acceptance in the sample that should not have been accepted, the formula is as in (2):

$$FAR = \frac{FP}{FP + TN} \tag{2}$$

FRR means the proportion of rejections in samples that should not have been rejected, with the formula as in (3):

$$FRR = \frac{FN}{TP + FN} \tag{3}$$

In speaker recognition systems, *FAR* and *FRR* can be weighed by setting different thresholds. The higher the security required by the system, the higher the threshold should be set, and the stricter the acceptance conditions, i.e., the lower the *FAR*, but the higher the *FRR*. On the contrary, if the system pursues better user experience, the lower the threshold should be, the more lenient the acceptance conditions are, the higher the FAR is, but the lower the *FRR* is. A monotonically decreasing Detection Error Tradeoff (DET) curve is usually plotted with FAR as the horizontal coordinate and *FRR* as the vertical coordinate. This form of DET curve is more conducive to observing the tradeoff between the two types of errors as the discrimination threshold changes. The smaller the area underneath the DET curve in this setting, the better the system's ability to discriminate. On the DET curve, at

the intersection of the first quadrant angular bisector and its intersection, the *FAR* and *FRR* values are equal, and this error rate is EER. At this time, *FAR*, *FRR*, and EER are equal. The smaller the value of EER the better the performance of the system, and it is an important parameter to measure the performance of the system.

5. Experiment and Results

5.1. Experimental Setup

5.1.1. Datasets

The CHAINS corpus contains 36 speakers, 8 from the UK and USA and the remaining 28 from the East of Ireland, each with 37 utterances. It is used as training data for same/cross-scene experiments, as well as training and evaluation data for transfer learning under the English target domain. The CHAINS data were recorded in six different modes with varying sound intensities, in which we consider the neutral speech and whispered speech of 36 speakers recorded at 44.1 kHz. The Whisper40 dataset was used as evaluation data for same/cross-scene experiments, as well as training and evaluation data for transfer learning in the Chinese target domain. In the same/cross-scene experiments the whole dataset is randomly divided into a test set and an enroll set with a ratio of 37:3. Three random audio files for each speaker become the enrollment data. Each piece of enrolled data is kept at full length, and the rest of the files will be tested.

Aishell [33] and the VoxCeleb1 corpus were used as source domain data in English and Chinese under transfer learning, respectively. The Aishell dataset has a total recording length of 178 h. The dataset was recorded with the participation of 400 speakers from different accent regions in China. After being written and annotated by professional speech proofreaders and passing a strict quality check, the text correctness rate of this database is above 95%. The dataset is divided into a training set of 340 speakers, a development set of 20 speakers, and a test set of 40 speakers, each of which contains roughly three hundred utterances. The Voxceleb1 dataset contains more than 150,000 utterances for 1251 celebrities. The data are derived from real-life scenarios in natural environments, most of the speech has some degree of noise, and the speakers cover many different ethnicities. The audio is all mono 16 kHz, 16-bit speech, and the content type is a text-independent dataset. The entire dataset is divided into a training set of 1211 speakers and a test set of 40 speakers.

5.1.2. Training Setup

Before transfer learning, three speaker recognition models, TDNN, ResNet34, and ECAPA-TDNN, are trained on the source domain data for the speaker classification detection task, respectively. The input features consist of a combination of 23-dimensional MFCC and 3-dimensional pitch with a frame length of 25 ms, averaged and normalized over a sliding window of up to 3 s. Where the MFCC features pre-emphasize, frame-split and add Hamming window processing of the signal and then perform a short-time Fourier change to obtain its spectrum, and then the spectrum is squared, the energy in each filter band is superimposed, and the kth filter output power spectrum is as follows:

$$X[k] = E(t, f) \tag{4}$$

Finally, the output of each filter is taken logarithmically to obtain the logarithmic power spectrum of the corresponding frequency band, and the inverse discrete cosine transform is performed to obtain the L MFCC factors:

$$C_n = \sum_{k=1}^{M} logX[k]\cos\left(\frac{\Pi(k-0.5)n}{M}\right) n = 1, 2, \cdots, L \tag{5}$$

It also uses energy-based SAD to filter out non-speech frames during training. To optimize the speaker recognition model, the Adam optimizer [34] is used with a learning rate of 0.0001 and a weight decay of 0.3 If the verification loss does not change in two periods, the learning rate is reduced by a factor of 0.3. The Adam optimizer defines two exponentially

weighted averages: the first one is the exponentially weighted average of the gradients, and the second one is the exponentially weighted average of the gradient squared. These two weighted averages are used to adjust the learning rate for each parameter, and the update rule for this optimizer is as follows:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \tag{6}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \tag{7}$$

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t} \tag{8}$$

$$\hat{v}_t = \frac{v_t}{1-\beta_2^t} \tag{9}$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \varepsilon} \hat{m}_t \tag{10}$$

All three models are trained using the AM-Softmax loss function, which improves on the angular distance loss by adding a hyperparameter λ that drives the difficulty of the classification boundaries, while a hyperparameter s is introduced to improve the convergence speed, as shown in (11).

$$\begin{aligned} L_{AMS} &= -\frac{1}{n}\sum_{i=1}^{n} log \frac{e^{s(cos\theta_{y_i}-m)}}{e^{s(cos\theta_{y_i}-m)} + \sum_{j=1, i\neq y_i}^{c} e^{scos\theta_j}} \\ &= -\frac{1}{n}\sum_{i=1}^{n} log \frac{e^{s(W_{y_i}^T f_i - m)}}{e^{s(W_{y_i}^T f_i - m)} + \sum_{j=1, j\neq y_i}^{c} e^{s} W_j^T f_i} \end{aligned} \tag{11}$$

5.1.3. Data Augmentation

Data augmentation is a commonly used technique to effectively improve the performance and generalization of models in machine learning and deep learning tasks. In this paper, the number and diversity of existing data samples are increased by applying various transformations and expansions to the training data to improve the robustness of the model. The data-augmented neural network can learn new data without overfitting and improve the performance. Currently, a video rewinding technique has been widely used as an enhancement method with positive results in the field of image recognition [35,36]. It stores all the frames of the video first and then encodes them in reverse order. Inspired by video rewinding, we introduce audio rewinding for the first time in order to expand the amount of data for Chinese whispered speech. Since different frames of audio may correspond to different syllables, when rewinding is performed, the syllables are rearranged. This produces audio with different contents, tones, and rhythms from the original speech without any noise and without changing the audio quality. This method expands the target domain whispered data and increases the diversity of the data. The total amount of data is twice as much as the original data. Noise enhancement is performed on the training set of both source and target domain data. Using the simulated RIR described by Ko et al. in [37], the reverberation itself was performed using the multi-conditional training tool in Kaldi. For additive noise, the MUSAN dataset was used, which consists of more than 900 noises, 42 h of music in various genres, and 60 h of speech in 12 languages of three types. After data augmentation, the amount of data becomes five times the original data.

5.1.4. Same/Cross-Scene Experimental Setup

If the audio to be evaluated is in a modality unfamiliar to the evaluator, both human and machine recognition performance will suffer. Most previous speaker recognition systems have been experimented on with homologous data. The effect of cross-scene training and test data on the performance and generalization ability of speaker recognition systems has been investigated experimentally. In this paper, experiments on same-scene and cross-scene data are conducted to verify the effect of WSR in different scenarios. In the case

of same-scene experiments, the training and evaluation sets consist of the same language, either whispered or neutral speech. In the cross-scene experiments, i.e., the training set includes the training set of whispered speech alone, the training set of neutral speech alone, and the overall training set of the combination of normal speech and whispered speech. The evaluation set includes the whispered speech alone enroll set, the neutral speech alone enroll set, and the overall enroll set. In the same-scenario experiments, the training data were the entire whispered speech data from CHAINs. The enrollment set is three randomly selected whispered speeches for each speaker in the Whisper40 dataset, and all the remaining 37 whispered speeches are used for testing to ensure the same speech style throughout the recognition process. In the cross-scene experiments, the neutral speech portion of the CHAINs and the Whisper40 data are introduced in the training phase and the evaluation phases, respectively, to achieve the comparison of different speech styles. The input features for training consisted of a combination of 23-dimensional MFCC and 3-dimensional pitch and non-speech frames were filtered out using energy-based SAD.

5.1.5. Transfer Learning Experimental Setup

In this paper, 340 speakers from Aishell and 1211 speakers from VoxCeleb1 in the speaker recognition domain are used as the Chinese and English source domain datasets, respectively. The speaker classification task is performed by learning different models. The self-recorded Whisper40 dataset is divided in a 32:8 ratio as training and test samples for the target domain task. The publicly available CHAINs dataset was divided in a ratio of 30:6 for the English target domain training set and test set, respectively. For all test speakers, three randomly selected whispered speech audios are used as the enrollment set, and all remaining audios are used for test. The same network layers are transferred for the same network model in different target domains.

Three networks, TDNN, ResNet34, and ECAPA-TDNN, were selected as the pre-training model structure before transfer learning. Among them, ResNet34 is a deep residual model composed of 34 layers of convolutional neural networks. It can effectively solve the problem of gradient vanishing in deep neural networks, which makes the network better learn abstract features. The main characteristic of this network is the introduction of residual connections. The residual connection allows the network to directly skip certain layers in the information transfer process. It is able to better retain the original information, enabling the network to learn more deeply. In the field of speaker recognition, it can effectively extract advanced features in the audio signal, for example, the speaker's voice characteristics and pronunciation habits, etc., thus realizing high-precision speaker recognition. ECAPA-TDNN combines both the powerful performance of TDNN and the advantages of the ECAPA strategy. First, TDNN uses local receptive fields with variable-sized sliding windows of time to capture local patterns in the input sequence. The Extended Channel Attention module adaptively assigns the weights of each channel to each layer of the neural network. This allows the model to better focus on features relevant to speaker recognition, thus improving performance. Information propagation takes place through residual connections and dense connections. Together, these two types of connections enhance the expressive power of the network, allowing information to flow more efficiently between layers. Meanwhile, the network uses a global pooling strategy to aggregate the last layer of the feature map. Combining global average pooling and global maximum pooling generates more discriminative speaker embeddings. To further improve the model performance, the embeddings are refined using a multi-head self-attention mechanism. This improves the accuracy of speaker recognition.

When performing transfer learning, the structure and characteristics of the model to be transferred are determined firstly, and secondly, the network layer parameters to be transferred are determined. For the x-vector model, the 5-layer frame-level layer, 2-layer segment-level layer, and pooling layer are transferred; the ResNet34 model is selected to transfer the residual layer and pooling layer; and the ECAPA-TDNN model transfers the Res2Block module layer, pooling layer, and fully-connected layer. After saving the network

parameters to be transferred, these parameters are used to initialize the parameters of the models trained on the whispering data. The parameters that have been transferred replace the parameters of the whispering model, and the unmigrated parameters undergo normal random initialization. After parameter initialization, fine-tuning is performed on the whispered dataset until the model is fitted. Since the number of target classifications before and after the transfer is not the same, the final output layer and the number of output nodes will be different, so the option of transferring the output layer is not chosen throughout the experiment.

5.2. Results

5.2.1. Same-Scene Experiments

In the same-scene experiments, the training data were the entire CHAINs whisper data, and the MUSAN and RIR datasets were used for data augmentation for each training. When training the model using whispered speech from the CHAINs dataset, three randomly selected audios from the whispered portion of Whisper40 for each speaker constituted the enrolment set, and the remaining 37 whispered audios for each person were the test set. When training the model using neutral speech from the CHAINs dataset, again three randomly selected audios from each speaker in the Whisper40 neutral speech section were used as the enrolment set. The remaining 37 neutral speeches from each person are the test set. At the same time, the audio reverse strategy was combined to verify the effectiveness of this data enhancement tool on the WSR results. The experimental results of training the WSR system using the three models in the same scenario are shown in Table 5.

Table 5. Experimental results in the same-scene (training and evaluation sets are both whispered or normal speech).

Train	Enroll	Test	Audio Reverse	ResNet34 (%)	ECAPA-TDNN (%)	TDNN (%)
Whisper	Whisper	Whisper	No	18.58	23.31	13.72
Whisper	Whisper		Yes	18.18	20.54	11.82
Neutral	Neutral	Neutral	Yes	11.87	15.21	11.73

From the results, it can be observed that the three WSR models show different performances when the condition of using whispered speech for both training and evaluation is considered. For whispered speech, the TDNN model performs the best. The lowest EER of 13.72% is obtained without the use of audio reversion. After the introduction of the audio inversion strategy, the EER of all three models decreases. Among them, the TDNN model still shows the best performance, with a relative reduction of 13.85% in the EER compared to the pre-reverse. For the same-scene experiments with neutral speech, the EER for speaker recognition is lower than that for whispered speech because its audio has a wider frequency range and pitch variation and can contain more information about the identity of the speaker. Although the ResNet34 model has the greatest performance improvement, with a 34.71% relative decrease in its EER, the TDNN model still performs optimally.

5.2.2. Cross-Scene Experiments

The experimental results obtained using cross-speech style data for training and testing are shown in Table 6. For the cross-scene experiments, the training data were both noise-added and reversed. The training set is CHAINs data, and the evaluation set is Whisper40 data. "All" under the training set in the table indicates the combination of all whispered speech and all neutral speech from the CHAINs data set. The "All" under the enroll set indicates the combination of the Whisper40 whispered speech enrollment set and the neutral speech enrollment set with a total of six audios.

Table 6. Experimental results in cross-scene (training and evaluation sets in different speech patterns).

Train	Enroll	Test	ResNet34 (%)	ECAPA-TDNN (%)	TDNN (%)
Neutral	Neutral	Neutral	11.87	15.21	11.73
Neutral	All		14.64	17.77	19.69
Whisper	Neutral		19.12	19.76	17.77
Whisper	All		25.59	27.43	25.37
All	Neutral		11.37	21.25	13.43
All	All		16.13	21.25	20.26
Neutral	Whisper	Whisper	21.42	24.26	15.68
Neutral	Neutral		35.61	42.30	34.05
Neutral	All		26.15	34.26	22.43
Whisper	Whisper		18.18	20.54	11.82
Whisper	Neutral		34.12	36.82	30.07
Whisper	All		20.81	23.51	17.97
All	Whisper		18.04	22.30	13.11
All	Neutral		34.26	33.38	24.46
All	All		25.14	25.61	13.04

The above results show that the ResNet34 model results are competitive when the test speech is neutral speech. However, when the test speech is whispered speech, using the TDNN model can better learn the speaker's identity information from the whispered speech, regardless of the type of speech used for training and enrollment. And it shows more robustness between cross-scene test data. Recognition is worst when the training and enrollment data are neutral speech and the test data are whispered speech. When the model is trained with neutral speech, all speech is registered, and whispered speech is tested, the recognition EER of the TDNN model decreases from 34.05% to 22.43%, which is a relative reduction of 34.13%. This is due to the fact that speaker identity features corresponding to the test audio can be extracted from the whispered speech in the enrollment phase. In the condition of whispered speech training, if whispered speech is also included in the enrollment, the recognition error rate will be reduced again. Considering only whispered speech as the test set, both the ResNet34 and ECAPA-TDNN models achieve the best results in the cross-scene experiments when "All" the audio is used as the training set for the model and the enrollment set consists of only the whispered data. However, it is still the TDNN model that performs best.

5.2.3. Transfer Learning Experiments

The training process of the WSR model based on transfer learning is carried out by pre-training on the source domain dataset and then transferring certain network layers of the pre-trained model to the target domain dataset for fine-tuning. Thus, the generalization ability of the WSR model is substantially improved [38]. In the experiments, Aishell and VoxCeleb1 were used as the Chinese and English source domain data, respectively, to train the pre-trained models. Whisper40 was used as the Chinese target domain data and was divided into training and evaluation sets with a ratio of 32:8. The CHAINs whispered part was used as the English target domain data and was divided into training and evaluation sets with a ratio of 30:6. The experimental results are shown in Tables 7 and 8.

Table 7. Experimental EER under transfer learning in Chinese target domain.

Transferred Model	Pre-Training Model	Result	
	Training Dataset	Before Transferred (%)	After Transferred (%)
ResNet34	aishell	16.03	13.46
	aishell-aug		13.14
ECAPA-TDNN	aishell	18.59	18.91
	aishell-aug		15.71
TDNN	aishell	15.06	12.50
	aishell-aug		10.90

Table 8. Experimental EER under transfer learning in English target domain.

Transferred Model	Pre-Training Model	Result	
	Training Dataset	Before Transferred (%)	After Transferred (%)
ResNet34	Voxceleb1	7.870	6.481
	Voxceleb1-aug		6.019
ECAPA-TDNN	Voxceleb1	8.333	7.406
	Voxceleb1-aug		7.870
TDNN	Voxceleb1	8.796	7.407
	Voxceleb1-aug		6.944

From the results, it can be observed that due to the small amount of whispered data, the introduction of a transfer learning approach is able to achieve better performance on the WSR task. And it can reduce the training time and resource consumption. The transfer learning approach achieves the best results because the source domain task and the target domain task are of the same language variety. Previous same/cross-scene experiments have verified that the TDNN model performs optimally on the WSR task. The performance of the model improvement under transfer learning is also most obvious when both the source and target domains are Chinese audio. When the pre-trained model is trained with the un-noised Chinese source domain data, the EER of the WSR system trained by the TDNN model before and after transfer is 15.06% and 12.50%, respectively, which is a relative reduction of 16.70%. However, the recognition performance of the ECAPA-TDNN model is slightly reduced. When the pre-trained model uses the augmented Aishell corpus, its model effect is further improved. The transfer learning process can achieve the goal of further reducing the error rate of Chinese WSR by learning more advanced experience on the source domain model. The error rates of the three WSR models after transfer are 10.90%, 13.14%, and 15.71%, respectively. Among them, the TDNN model shows the most obvious improvement, with a 27.62% improvement in recognition performance after transfer learning. When using the English source domain data to train the pre-training model, it can be observed from Table 8 that the ResNet34 model shows strong competitiveness and obtains the best recognition effect among the three models with an EER of 6.019%. When the network layer of the TDNN model trained by the data enhancement method is transferred, the EER of English WSR recognition reaches 6.944%. Compared to the pre-transfer, learning recognition improvement is the most obvious, and the EER is reduced by 21.06% compared to the transfer learning. The transfer learning experiments demonstrate the usefulness of the Whisper40 dataset and the good generalizability of the TDNN model for whispered speech.

5.2.4. Experiments with Different Noise Conditions

In real-world environments, noise can lead to significant performance degradation of WSR systems, which is an issue that requires special attention. In order to investigate the change in performance of a WSR system from a specific environment to a common environment, we conducted experiments under different noise conditions. The model is chosen to be the TDNN network that performs well on the whispered dataset. The

experimental results before and after transfer learning are tested for the Chinese WSR task under different Signal-to-Noise Ratios (SNRs) (0 dB, 5 dB, 10 dB and 15 dB). The training data for the pre-training model are the data-enhanced Aishell dataset. The Whisper40 dataset is used as the target domain data and is divided into the training and evaluation sets in a ratio of 30:6. The experimental results are shown in Table 9.

Table 9. Experimental EER under transfer learning at different SNRs.

SNR	Before Transferred (%)	After Transferred (%)
0 dB	25.00	20.83
5 dB	20.51	19.55
10 dB	20.19	18.27
15 dB	19.23	15.71
Baseline	15.06	10.90

Table 9 demonstrates the effect of varying the noise conditions on the performance of the WSR system. The stronger the applied noise, the lower the SNR value. A noise level of 0 dB indicates that sound and noise have the same energy. From the table, it can be observed that as the noise diminishes, i.e., the signal-to-noise ratio decreases from low to high, and the recognition EER gradually decreases. Regardless of whether the transfer learning method is used or not, the best results in the noise condition are achieved systematically when the SNR is 15 dB. Compared to the 0 dB condition, the recognition error rate is relatively reduced by 23.08%. The transfer learning method also demonstrated a large advantage under noisy conditions. The EER was reduced after transfer learning at different signal-to-noise ratios. When the SNR is 15 dB, the result after transfer learning is relatively improved by 18.30% compared with the result before transfer learning. However, it is worth noting that the performance of the WSR system is not as good as the un-noised baseline results regardless of the noise condition. This experimental result also demonstrates the good utility of the self-collected Whisper40 dataset even in noisy environments. It facilitates subsequent WSR studies in complex environments.

6. Conclusions

In this paper, a Chinese whispered speaker recognition dataset, Whisper40, is built. The dataset also contains neutral speech with the same-text as the whispered speech, in order to explore the differences between the whispered speech and the neutral speech. Whisper40 consists of 40 speakers, which fills in the gap of the Chinese whispered speech acoustic pattern recognition dataset. In this paper, ResNet34 and ECAPA-TDNN models, which have good performance in the field of speaker recognition, are introduced into the field of whispering for the first time. A baseline system for WSR in transfer learning mode is constructed, while the data augmentation method of audio reverse is used in the training process. The experimental results show that the self-collected Whisper40 dataset is practical and has strong confidence for WSR, and the TDNN model performs optimally in both same-scene and multi-scene experiments. It also confirms that the transfer learning method can effectively improve the effectiveness of the WSR model under data-poor conditions. The Whisper40 dataset will be a valuable resource for WSR and cross-scene speaker recognition. Future work will investigate how to extract more accurate speaker identity features from whispered speech in order to improve the robustness of the model to whispered speech and hence the recognition accuracy.

Author Contributions: Conceptualization and methodology, J.Y. and R.Z.; software, validation, and formal analysis, J.Y. and R.Z.; writing—original draft preparation, J.Y.; writing—review and editing, J.Y. and R.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The codes used in this manuscript are available from the corresponding author upon request.

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

References

1. Prieto, S.; Ortega, A.; López-Espejo, I.; Lleida, E. Shouted and whispered speech compensation for speaker verification systems. *Digit. Signal Process.* **2022**, *127*, 103536. [CrossRef]
2. Naini, A.R.; Rao, A.; Ghosh, P.K. Whisper to Neutral Map** Using I-Vector Space Likelihood and a Cosine Similarity Based Iterative Optimization for Whispered Speaker Verification. In Proceedings of the 2022 National Conference on Communications (NCC), Mumbai, India, 24–27 May 2022; pp. 130–135.
3. Kim, K.; Kim, J.; Voight, A.; Ji, M. Listening to the screaming whisper: A voice of mother caregivers of children with autistic spectrum disorder (ASD). *Int. J. Qual. Stud. Health Well-Being* **2018**, *13*, 1479585. [CrossRef] [PubMed]
4. Fan, X.; Hansen, J.H.L. Speaker identification for whispered speech based on frequency warping and score competition. In Proceedings of the Ninth Annual Conference of the International Speech Communication Association, Brisbane, Australia, 22–26 September 2008.
5. Patel, M.; Parmar, M.; Doshi, S.; Shah, N.; Patil, H.A. Novel Inception-GAN for Whisper-to-Normal speech conversion. In Proceedings of the 10th ISCA Speech Synthesis Workshop (SSW 10), Vienna, Austria, 20–22 September 2019; pp. 87–92.
6. Zhang, C.; Hansen, J.H.L. Analysis and classification of speech mode: Whispered through shouted. In Proceedings of the Interspeech 2007, Antwerp, Belgium, 27–31 August 2007; Volume 7, pp. 2289–2292.
7. Raeesy, Z.; Gillespie, K.; Ma, C.; Drugman, T.; Gu, J.; Maas, R.; Rastrow, A.; Hoffmeister, B. LSTM-based whisper detection. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018; pp. 139–144.
8. Ashihara, T.; Shinohara, Y.; Sato, H.; Moriya, T.; Matsui, K.; Fukutomi, T.; Yamaguchi, Y.; Aono, Y. Neural Whispered Speech Detection with Imbalanced Learning. In Proceedings of the INTERSPEECH, Graz, Austria, 15–19 September 2019; pp. 3352–3356.
9. Shah, N.J.; Shaik MA, B.; Periyasamy, P.; Patil, H.A.; Vij, V. Exploiting phase-based features for whisper vs. speech classification. In Proceedings of the 2021 29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland, 23–27 August 2021; pp. 21–25.
10. Toda, T.; Shikano, K. NAM-to-speech conversion with Gaussian mixture models. In Proceedings of the INTERSPEECH2005: The 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, 4–8 September 2005.
11. Nakamura, K.; Toda, T.; Saruwatari, H.; Shikano, K. Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech. *Speech Commun.* **2012**, *54*, 134–146. [CrossRef]
12. Lian, H.; Hu, Y.; Yu, W.; Zhou, J.; Zheng, W. Whisper to Normal Speech Conversion Using Sequence-to-Sequence Mapping Model with Auditory Attention. *IEEE Access* **2019**, *7*, 130495–130504. [CrossRef]
13. Sun, L.; Kang, S.; Li, K.; Meng, H. Voice conversion using deep bidirectional long short-term memory based recurrent neural networks. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, South Brisbane, Australia, 19–24 April 2015; pp. 4869–4873.
14. Sarria-Paja, M.; Falk, T.H. Fusion of bottleneck, spectral and modulation spectral features for improved speaker verification of neutral and whispered speech. *Speech Commun.* **2018**, *102*, 78–86. [CrossRef]
15. Naini, A.R.; Rao, A.; Ghosh, P.K. Formant-gaps features for speaker verification using whispered speech. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6231–6235.
16. Zhang, Q.; Zhao, H.; Gong, C. Whispered Speaker Identification Based on Factor Analysis and Feature Mapping. *J. Data Acquis. Process.* **2016**, *31*, 362–369. [CrossRef]
17. Gong, C.; Zhao, H.; Tao, Z.; Zhang, Q. Speaker factor analysis of whispered speech from global spectral features. *Acta Acust.* **2014**, *39*, 281–288. [CrossRef]
18. Huang, W.; Peng, Y.; He, S. Vocal effort in speaker recognition based on MAP+CMLLR. *J. Comput. Appl.* **2017**, *37*, 906–910.
19. Gu, X.; Zhao, H.; Lü, G. An application in whispered speaker identification using feature and model hybrid compensation method. *Acta Acust.* **2012**, *37*, 198–203. [CrossRef]
20. Sarria-Paja, M.; Senoussaoui, M.; O'Shaughnessy, D.; Falk, T.H. Feature mapping, score-, and feature-level fusion for improved normal and whispered speech speaker verification. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5480–5484.
21. Naini, A.R.; Rao, A.; Ghosh, P.K. Whisper to Neutral Mapping Using Cosine Similarity Maximization in i-Vector Space for Speaker Verification. In Proceedings of the INTERSPEECH, Graz, Austria, 15–19 September 2019; pp. 4340–4344.
22. Vestman, V.; Gowda, D.; Sahidullah, M.; Alku, P.; Kinnunen, T. Speaker recognition from whispered speech: A tutorial survey and an application of time-varying linear prediction. *Speech Commun.* **2018**, *99*, 62–79. [CrossRef]
23. Variani, E.; Lei, X.; McDermott, E.; Moreno, I.L.; Gonzalez-Dominguez, J. Deep neural networks for small footprint text-dependent speaker verification. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 4052–4056.

24. Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; Khudanpur, S. X-vectors: Robust dnn embeddings for speaker recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5329–5333.
25. Li, C.; Ma, X.; Jiang, B.; Li, X.; Zhang, X.; Liu, X.; Zhang, X.; Liu, X.; Cao, Y.; Kannan, A.; et al. Deep speaker: An end-to-end neural speaker embedding system. *arXiv* **2017**, arXiv:1705.02304.
26. Desplanques, B.; Thienpondt, J.; Demuynck, K. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv* **2020**, arXiv:2005.07143.
27. Li, S.; Wei, Z.; Zhang, B.; Hong, B. Target recognition using the transfer learning-based deep convolutional neural networks for SAR images. *J. Univ. Chin. Acad. Sci.* **2018**, *35*, 75–83.
28. Yang, L.; Li, Y.; Xu, B. The Establishment of a Chinese Whisper Database and Perceptual Experiment. *J. Nanjing Univ. Nat. Sci.* **2005**, *41*, 311–317.
29. Jou, S.C.; Schultz, T.; Waibel, A. Whispery speech recognition using adapted articulatory features. In Proceedings of the Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, PA, USA, 23 March 2005; Volume 1, pp. I/1009–I/1012.
30. Zue, V.; Seneff, S.; Glass, J. Speech database development at MIT: TIMIT and beyond. *Speech Commun.* **1990**, *9*, 351–356. [CrossRef]
31. Cummins, F.; Grimaldi, M.; Leonard, T.; Simko, J. *The CHAINS Corpus (Characterizing Individual Speakers)*; School of Computer Science and Informatics, University College Dublin: Dublin, Ireland, 2006.
32. Du, Y.; Zhou, R. Multi-model Fusion VoxSRC22 Speaker Diarization System. *Comput. Eng. Appl.* **2024**, 1–9. Available online: https://kns-cnki-net.door.bucea.edu.cn/kcms/detail/11.2127.TP.20230328.1100.020.html (accessed on 1 February 2024).
33. Bu, H.; Du, J.; Na, X.; Wu, B.; Zheng, H. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In Proceedings of the 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA), Seoul, Republic of Korea, 1–3 November 2017; pp. 1–5.
34. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
35. Zhang, H.; Wang, M.; Hong, R.; Chua, T.S. Play and rewind: Optimizing binary representations of videos by self-supervised temporal hashing. In Proceedings of the 24th ACM International Conference on Multimedia, Melbourne, Australia, 15–19 October 2016; pp. 781–790.
36. Rewind, Play, Fast forward: The past, present and future of the music video. In *Rewind, Play, Fast Forward*; Transcript Verlag: Bielefeld, Germany, 2015.
37. 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 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 5220–5224.
38. Ji, W.; Yang, M.; Li, Y.; Zheng, H. Parkinson's Disease Detection Method Based on Masked Self-supervised Speech Feature Extraction. *J. Electron. Inf. Technol.* **2023**, *45*, 3502–3510.