applied sciences

MDPI

Article

Addressing Text-Dependent Speaker Verification Using Singing Speech

Yan Shi [1,†], Juanjuan Zhou [1,†], Yanhua Long [1,*], Yijie Li [2] and Hongwei Mao [1,*]

[1] SHNU-Unisound Joint Laboratory of Natural Human-Computer Interaction, Shanghai Normal University, Shanghai 200234, China
[2] Beijing Unisound Information Technology Co., Ltd., Beijing 100028, China
* Correspondence: yanhua@shnu.edu.cn (Y.L.); maohw2007@shnu.edu.cn (H.M.)
† These authors contributed equally to this work.

Received: 9 May 2019; Accepted: 26 June 2019; Published: 28 June 2019

Abstract: The automatic speaker verification (ASV) has achieved significant progress in recent years. However, it is still very challenging to generalize the ASV technologies to new, unknown and spoofing conditions. Most previous studies focused on extracting the speaker information from natural speech. This paper attempts to address the speaker verification from another perspective. The speaker identity information was exploited from singing speech. We first designed and released a new corpus for speaker verification based on singing and normal reading speech. Then, the speaker discrimination was compared and analyzed between natural and singing speech in different feature spaces. Furthermore, the conventional Gaussian mixture model, the dynamic time warping and the state-of-the-art deep neural network were investigated. They were used to build text-dependent ASV systems with different training-test conditions. Experimental results show that the voiceprint information in the singing speech was more distinguishable than the one in the normal speech. More than relative 20% reduction of equal error rate was obtained on both the gender-dependent and independent 1 s-1 s evaluation tasks.

Keywords: speaker verification; singing speech; corpus; X-vector

1. Introduction

Automatic speaker verification (ASV) is the verification of a speaker's identify based on his/her speech signals [1]. It is an important biometric technology and can be widely used in access control security authentication, personalized services, etc. Performances of ASV systems have been significantly improved by recent advances in speech technology. The state-of-the-art ASV systems are robust to session and channel variations [2–5]. However, they are vulnerable to spoofing attacks, such as spoofed speech produced using either text-to-speech or voice conversion technologies [6,7]. It is still very challenging to use the current technologies for real ASV applications.

To improve the robustness of ASV systems under real applications, most previous works focused on exploring new speaker modeling algorithms, discriminative biometric patterns, or finding countermeasures to eliminate the acoustic mismatches between training and testing or spoofing speech, such as the self-attentive speaker embeddings in [8] for a better speaker identity representation, the end-to-end speaker modeling framework in [9], the light convolution neural network [10] and attentive filtering network modeling architectures in [11] for detecting spoofing utterances, etc. [12,13].

In the literature, most of these previous works extracted the speaker information from natural speech, either the read speech or spontaneous or contextual speech [1]. We only found very few works focus on exploiting robust speaker information from other prospectives, such as speaker recognition in [14–16], where the humming was investigated to extract the speaker identity information.

Specifically, in [14,15], various acoustical features such as the linear prediction cepstral coefficients, the conventional Mel-frequency cepstral coefficients (MFCCs) and perceptually linear prediction coefficients were evaluated for humming speech. In [16], the variable length Teager energy based mel rrequency cepstral coefficients was proposed to identify speakers from their hum. Moreover, the speaker-dependent characteristics were extracted from the nasals for forensic speaker recognition [17]. In [18], the humming, singing and speech were compared and evaluated as biometric signal. They found that the humming sounds are better for capturing speaker-specific characteristics than speech and singing. In [19], GMM mean supervector was used to improve performance of speaker clustering with speech from both reading and singing. The authors of [20,21] proposed the timbre and vibrato-motivated acoustic features for singer identification. All of these previous works provide a good reference for us to study speaker recognition from a new perspective. However, we have not found any previous works that examine and compare the speaker verification performances between using the natural Mandarin reading speech and singing speech. For example, the authors of [14–16,18] only designed text-independent speaker verification tasks; their speech corpus and songs were all of Hindi languages and only a second-order polynomial classifier was used in their speaker classification. In [14,16], only humming speech was examined for human verification and identification, no comparison results was presented, such as the performance comparison between the humming and the natural reading speech.

In this paper, we also focus on the speaker verification using the singing speech. However, the big difference between this work and the above-mentioned previous works is that we focus on examining and comparing the effectiveness of using normal Mandarin reading speech and their corresponding singing speech for short-time text-dependent speaker verification. Different features and three types of speaker modeling approaches were also investigated, including the conventional and state-of-the-art deep neural network based techniques. Our motivation is to explore using the singing speech as an input to ASV system, because we want to see whether the singing speech is more effective for voiceprint information protection in real-world ASV applications, such as personalized accessing of WeChat, QQ accounts, etc.

Firstly, we designed a new corpus for short-time text-dependent ASV experiments. We released it on the Zenodo website (https://zenodo.org/record/3241566) and put our implementation code in the Github repository (https://github.com/Moonmore/Speaker-Verification) for public research. Based on this corpus, we performed the text-dependent (TD) ASV comparison experiments using either the natural speech or the singing speech, or both of them. Then, we focused on the TD ASV experiments to exploit the effectiveness of natural and singing speech for speaker verification. We first examined the speaker discrimination between natural and singing speech in both the Normalized Cross Correlation Function (NCCF) coefficients [22] and 2-D Mel Frequency Cepstral Coefficients (MFCC) feature spaces. Then, the conventional Gaussian Mixture Model (GMM), Dynamic Time Warping (DTW) and the state-of-the-art deep neural network were investigated for speaker modeling. Preliminary results show that the voiceprint information in the singing speech was more distinguishable than the one in the natural reading speech for the short-time gender-dependent as well as independent ASV tasks.

2. Corpus

Since there is no publicly released Mandarin corpus that meets our motivation, we designed a new corpus for our research. It includes both the normal reading and singing speech, thus we named this corpus as “RSS”. A detailed description of RSS is shown in Table 1. It consists of 20 speakers, including 10 male and 10 female undergraduate students. This study was our preliminary work on exploring and comparing the Mandarin singing speech and normal reading speech for speaker verification. Our motivation was to examine the effectiveness of different speaking styles for speaker verification. Therefore, to eliminate the interference from complicate recording setups, we only selected the undergraduate students with age ranges from 22 to 24 as our target speakers.

We selected 10 lyrics segments that are familiar to everyone as the text for audio recording. Around 5–15 utterances were included in each lyrics segments. Three music styles of these lyrics were selected: pop music, classical music and country music. These music genres are popular to most of the students in our university. During speech recording, the speaker used a common laptop built-in microphone to sing and read the given lyrics in a quiet lab environment. We used "Audition" as our speech recording and editing software. To create reasonable comparative experiments between normal and singing speech, for each speaker and each lyrics text, we recorded it twice for reading speech and twice for singing speech in two days. Each lyrics segment covered around 10–30 s speech. Each recording was formatted as 16 kHz, 16 bit WAV file. We chose the specific wav file format to record the audio, because this format is more generally used for human–machine interaction applications. Moreover, as 16 kHz WAV format is also normally used in automatic speech recognition applications, it would be better to keep the same speech input setup to make things compatible in real-world applications. All of the recordings are in Mandarin.

Table 1. RSS corpus description.

Item	Details
Speaker	20 undergraduate students (10 male, 10 female)
Language	Mandarin
Format	16,000 Hz, 16 bit, 1 channel
Text	10 lyrics of pop, classical and country songs
Biometric signal	reading speech and corresponding singing speech
Microphone	common laptop built-in microphone
Recording software	Audition
Acoustic environment	quiet lab environment

Table 2 gives the details of the ASV tasks we designed based on the "RSS" corpus. In most application scenarios, the enrollment and test data per speaker were normally around 1–3 s. Therefore, we should cut our long recordings into 1 s and 3 s short segments. Taking the 1 s-1 s-GD task with reading speech as example, for each of the 20 target speakers, he/she had 10 distinct lyrics that cover 10–30 s reading speech, and each lyric was recorded twice. To construct our ASV tasks, we first cut all these long speech segments into many small segments according to the completeness of the sentence text (lyric). For all of the 10–30 s audio files with the same lyric, we segmented them according to the same lyric integrity. Then, we picked out all of the short segments with length around 1 s. Finally, we chose the hold-out cross-validation method to design our experiments. Because each speaker had two recordings with the same lyrics, we randomly selected one of the two short segments with the same lyric text as the 1 s training (enrollment) segment of target speaker and the other as test. There was no overlap between the speech of speaker training and test. This test segment was used to test all of the other target speaker models. Therefore, from all of these selected 1 s segments, we obtained 600 segments as the target speaker enrollment speech, 300 for males and 300 for females. There were some enrollment segments belong to the same speaker, but they were treated as separate speakers to have their own speaker models in our experiments. This is because, to make the observations in our experiments more general, we should try to increase the number of the target enrollment speakers as much as possible.

To construct the gender-dependent test trials, such as for one speaker of 10 females with one lyric, we then had 30 enrollment segments (1 s per each) to train 30 separate target speaker models, and 30 segments for test. Considering the 10 lyrics, we generated $30 \times 10 \times 10 = 3000$ trials, 300 of which were target trials (test and training segments belong to the same speaker), others were non-target trials (test speech belong to imposter). For the 3 s-3 s tasks, we processed the data in a similar way to construct the speaker model training set and test trials. Audio files of singing were processed in the same way as the ones of reading speech to construct the exactly the same experimental setup, except

for the input segments were singing. More details about RSS data and experimental configuration files, please refer to the Zenodo website and Github repository mentioned in the Introduction.

Table 2. Text-dependent short-time ASV task description.

Task	#Target Speakers	#Test Segments	#Target Trials	#Non-Target Trials
1 s-1 s-GD	600	600	300 male, 300 female	2700 male, 2700 female
1 s-1 s-GI	600	600	600	11,400
3 s-3 s-GD	300	300	150 male, 150 female	1350 male, 1350 female
3 s-3 s-GI	300	300	300	5700

To see whether the behavior of speaker discrimination between two speaking styles is gender-dependent or not, we designed both GD and GI ASV tasks according to the above steps. We hope that the performance difference between male and female may motivate researchers to propose new gender-dependent features or methods to improve the final ASV systems. In Table 2, "GD" and "GI" refer to gender-dependent and gender-independent, respectively. "1 s-1 s" refers to the duration of speaker enrollment and test speech are both 1 s, and "3 s-3 s" means a similar case. In this study, all experiments were performed on these eight ASV tasks. Furthermore, we released our corpus publicly for research purpose only. The free download website can be found in the footnote of the Introduction. For a better comparison, two independent ASV tasks were constructed for experiments. They had the same trials configurations as shown in Table 2, but one for singing speech, the other for normal reading speech.

3. Speaker Identity Discrimination in Different Feature Space

The theoretical principle of speaker verification is that each person's voice has its unique characteristics. The unique property is determined primarily by two factors, the size of the acoustic cavity and the manner in which the vocal organ is manipulated [23]. In speaker verification, these properties are included in the acoustic feature space. In this section, we focus on exploring the speaker identity discrimination in two feature spaces, the pitch and the MFCC feature spaces. In these spaces, we examined the feature discrimination between normal speaking (reading) and singing speech of the same speaker, and the speaker discrimination under the same speech style between different speakers.

3.1. Pitch Discrimination

Pitch is one of the most important features for describing the excitation source property in speech signal processing [24]. We used the Normalized Cross Correlation Function (NCCF) coefficients [22] extracted by using the Kaldi pitch tracker [25] to represent the pitch information for our ASV. The NCCF is not the normal pitch value; it is a pitch estimation method that is very similar to the autocorrelation function, but it better follows the rapid changes in pitch and the amplitude of speech signal. In comparison with the normal autocorrelation function, the peaks corresponding to pitch period in the NCCF are more prominent and less affected by the rapid variations in the signal amplitude. In [22], the effectiveness of NCCFs has been proved on speech recognition tasks. Figure 1 shows the NCCF contour of the same female speaker's reading and singing speech with the same text. Figure 2 shows the NCCF contour of reading and singing speech between two different female speakers, given with the same text.

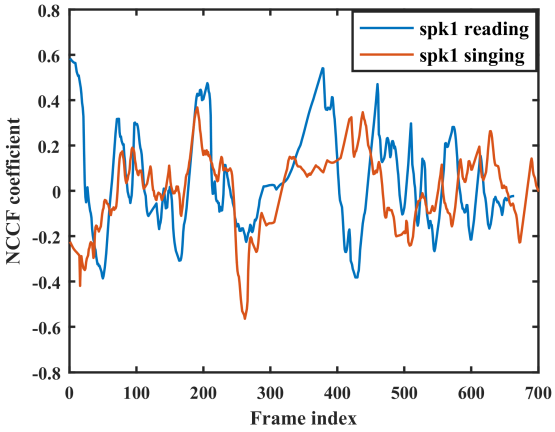

Figure 1. NCCF contour of the same speaker's reading and singing speech with the same text.

In Figure 1, it is clear that the NCCF contour of reading speech deviates far from the one of singing speech, with the same speaker and the same text. By comparing the NCCF trajectories in Figure 2, we can see that the difference of pitch information between two different speakers from singing speech is larger than the one from reading speech, even under the text-dependent task.

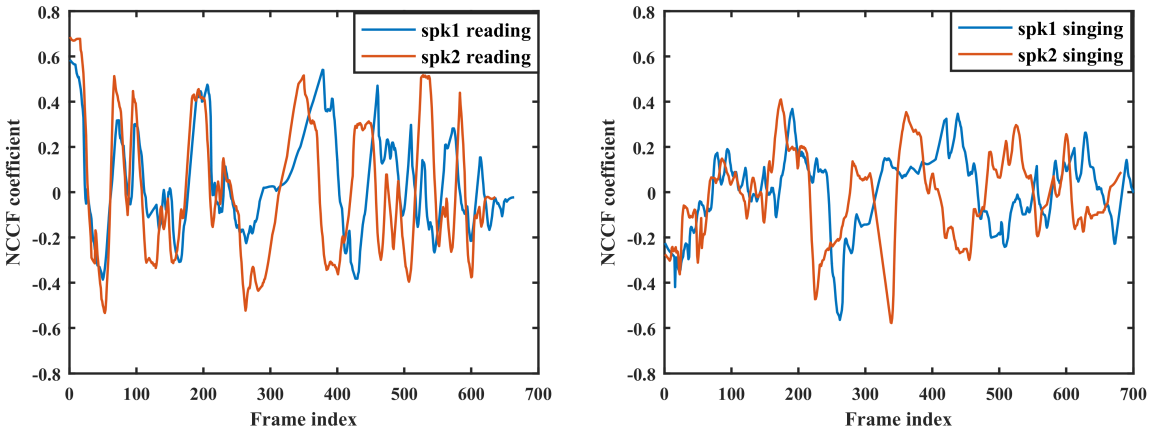

Figure 2. NCCF pitch contour of reading (**left part**) and singing (**right part**) speech between two females with the same text.

3.2. MFCC Discrimination

MFCC is a conventional frequency-domain auditory perception cepstral coefficient method. MFCCs have been widely used in both the speech and speaker recognition applications [1]. We first extracted the 20-dimensional (c0–c19) MFCC features using the Kaldi Toolkit [25]. Then, to visualize the MFCC discrimination between different speakers more clear, we applied the Principal Component Analysis (PCA) algorithm to reduce the 20-dimensional MFCCs to a two-dimensional feature space. Finally, we performed the feature analysis in this low-dimensional space. Similar to the above NCCF information contour, Figure 3 demonstrates the two-dimensional PCA feature distribution of MFCCs of both the reading and singing speech between two different female speakers, given the same text.

In Figure 3, we can easily observe that the overlap of orange dots and blue circles in the left subfigure is less than that in the right subfigure. This indicates that the speaker discrimination in MFCC feature space of singing speech is larger than the discrimination in reading speech feature space. By comparing the same color parts of the left and right subfigures, it can be seen that, even for the same speaker with the same text, there is also large difference between singing and reading speech in the MFCC feature space. Therefore, we guess that it may be better for us to use the singing speech instead of normal reading speech to characterize a speaker's identity. Actually, even if we see discriminate

information in the feature space, we cannot guarantee that the speaker model would be able to exploit the discriminations very well. Therefore, we provide detail validation experiments in next sections to see what would happen.

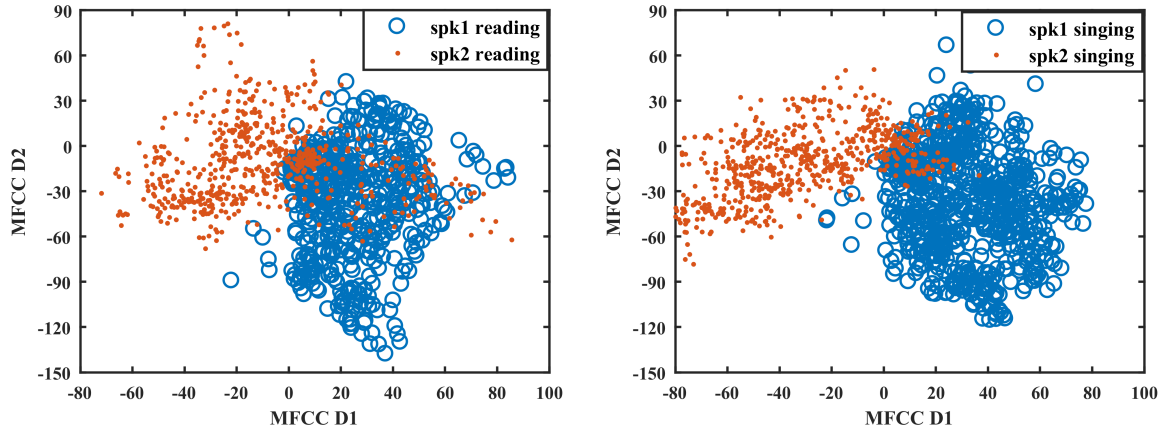

Figure 3. MFCC PCA feature distribution of reading (**left**) and singing (**right**) speech between two females with the same text.

4. Speaker Verification Systems

This section describes the features and three types of speaker verification systems developed for this study: the GMM-based, DTW-based and the state-of-the-art deep neural network based systems.

Features: As shown in Section 3, we extracted 20-dimensional MFCC, then applied PCA to reduce them to two dimensions for a better visualization. However, to preserve more detail information in the features, we did not use the reduced two-dimensional feature to train our speaker verification systems. All of the features used for GMM-based, DTW-based and deep neural network based ASV systems were 61-dimensional: the 20-dimensional MFCCs, their delta and delta-delta dynamic features [26] and the one-dimensional NCCF feature. All of these features were extracted using a 25 ms hamming window with a 10 ms frame shift. An energy-based voice activity detection (VAD) was applied to remove the silence.

GMM-based system: We used the conventional GMM to model the speaker identity for each enrollment speaker. Assuming a d-dimensional input feature for each speech segment, the GMM with M mixtures is:

$$p(x|\lambda) = \sum_{i=1}^{M} w_i g(x|\mu_i, \Sigma_i) \tag{1}$$

where $\lambda = \{w_i, \mu_i, \Sigma_i\}$ is the GMM model, and $g(x|\mu_i, \Sigma_i)$ represents a d-variate Gaussian probability density function, with mean vector μ_i, covariance matrix Σ_i and mixture weights w_i with $\Sigma_{i=1}^{M} w_i = 1$. During the speaker enrollment, one GMM model was built for each speaker using her/his enrollment speech segment. In the experiments, we took the GMM-based system as our baseline for performance comparison. The GMM mixture number was set to 32 and only diagonal covariance was used. In fact, we tried many different mixtures in our experiments, from 8 to 1024, and the 32 mixtures GMM obtained the best and most stable results. Given a set of acoustic features of each target speaker, the parameters of the target speaker GMM model was estimated using the maximum likelihood (ML) criterion with the popular expectation-maximization algorithm. More details of the GMM model training can be found in [1].

During the testing, given the d-dimensional feature vectors $X = (x_1, x_2, ..., x_T)$ of test utterance with T frames, we computed the log-likelihood score on each target speaker model as:

$$\Lambda(X) = \frac{1}{T} logp(X|\lambda) = \frac{1}{T} \sum_{t=1}^{T} log \sum_{i=1}^{M} w_i g(x_t|\mu_i, \Sigma_i) \quad (2)$$

These log-likelihood scores were then used to compare with a threshold to give the final speaker verification decision.

DTW-based system: DTW is a dynamic programming technique to compute the distance between two sequences. It has been taken as a sequence-matching algorithm and it has been widely used to tackle the text-dependent speaker verification [1,27]. It attempts to match the enrollment and test templates of feature vectors. In [28], it has been shown that a DTW-based ASV system can outperform model-based systems (relevance MAP and i-vector) in content-mismatch condition. In this study, the DTW algorithm was investigated for short-duration ASV systems.

X-vector based system: With the rapid development of deep neural network technology in speaker recognition, many types of DNN embeddings have been proposed, such as the DNN-Ivector [29], the D-vector [5], J-vector [30], C-vector [31], S-vector [32], etc. All of these vectors were derived from DNN models to map a variable-length speech utterance to a fixed-dimensional space for speaker modeling. Given these types of vectors, a simple cosine distance measure or a probabilistic linear discriminant analysis (PLDA) [33] was then normally applied to these vectors as the final decision function for speaker verification.

In this work, the state-of-the-art deep neural network based ASV system was the X-vector based system. The X-vector is a type of deep neural network (DNN) embedding that has been recently proposed to map variable-length utterances to fixed-dimensional embeddings [3,4]. It has been proved to be very effective and the X-vector based system is the state-of-the-art dominant technique for both text-independent and text-dependent speaker verification. We used the same embedding DNN architecture as in [4]. To build an X-vector based ASV system, we needed to train an X-vector extractor first, and then, given each speech segment, we extracted an X-vector from the well-trained extractor to represent the speaker identity. During the testing, the simple cosine distance between the X-vector of each test speech and the one of target speaker enrollment speech was computed as the final verification decision score. Normally, to improve the robustness and generalization ability of these X-vectors, a PLDA backend [4,33] was normally applied to these X-vectors.

Figure 4 shows the detail architecture of the X-vector extractor model. The first five layers of the network worked at the frame level, with a time-delay architecture [34]. Suppose t is the current time step. At the input, we spliced together frames at $\{t-2, t-1, t, t+1, t+2\}$. The next two layers splice dtogether the output of the previous layer at times $\{t-2, t, t+2\}$ and $\{t-3, t, t+3\}$, respectively. The next two layers also operate dat the frame-level, but without any added temporal context. In total, the frame-level portion of the network had a temporal context of $t-8$ to $t+8$ frames. Layers varied in size, from 512 to 1536, depending on the splicing context used. The statistics pooling layer received the output of the final frame-level layer as input, aggregated over the input segment, and computed its mean and standard deviation. These segment-level statistics were concatenated together and passed to two additional hidden layers with dimension 512 and 300 (either of which could be used to compute embeddings; in this study, the X-vector b was taken as our X-vectors) and finally the softmax output layer. The X-vector extractor network was trained to classify training speakers using a multi-class cross entropy objective function. Please refer to the work of Snyder [3] to obtain more details of the extractor training.

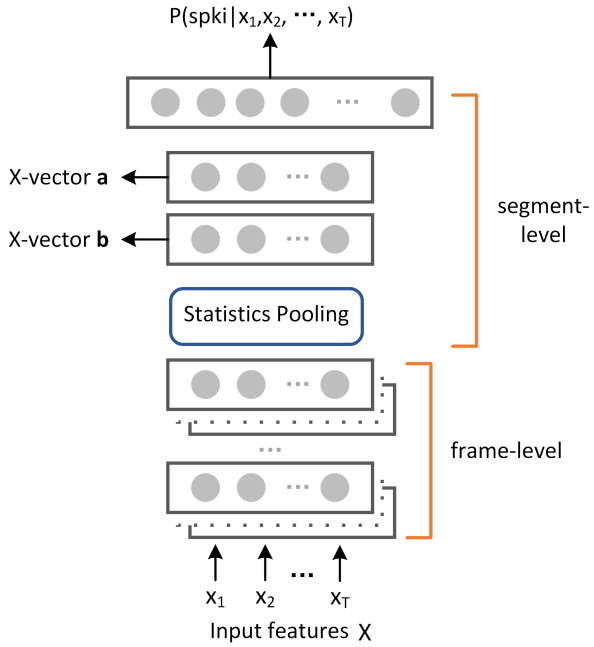

Figure 4. Diagram of the X-vector extractor.

As shown in Figure 4, the X-vector extractor is also a DNN model [4]; it can be challenging to collect substantial quantities of labeled singing data for training a good X-vector extractor. Our RSS corpus is not enough. Moreover, to avoid the overfitting of X-vector extractor on the specific ASV task, the speaker overlap among X-vector extractor training data, the target speaker training and testing data is normally not allowed. Therefore, as all of the speakers in RSS were taken as the target speakers in our ASV systems, we could not use the RSS to train or pre-train the X-vector extractor model. Researchers normally used in-domain large-scale datasets with similar acoustic characteristics as the target speaker training and test data to train the extractor. However, to obtain a preliminary performance using X-vector system for our task, we chose to use the out-of-domain data for the extractor network training. We used the open-source speech corpus "AISHELL-2" [35] to train our X-vector extractor. It has 1000 h of clean read-speech data with 1991 speakers. Then, we extracted one X-vector for each 1 s and 3 s segments in our RSS corpus using this extractor. We computed the log-likelihood ratio between X-vectors of target speaker and testing speaker as the decision score, using a probabilistic linear discriminant analysis (PLDA-based) [33] backend. The PLDA was trained using 178 h of "AISHELL-1" corpus with 400 speakers [36].

5. Experimental Results

In this section, we present experimental results for the short-duration gender-independent as well as gender-dependent ASV tasks: the 1 s-1 s-GD and 1 s-1 s-GI tasks as well as the 3 s-3 s-GD and 3 s-3 s-GI tasks. The performances are reported in terms of equal error rate (EER) [1], a verification error measure that gives the accuracy at decision threshold for which the probabilities of false rejection (miss) and false acceptance (false alarm) are equal. The probability of false rejection is the ratio of the number of false rejection (target speaker is misclassified as non-target speaker) divided by the total number of target test trials. The probability of false acceptance is the ratio of the number of false acceptance (non-target speaker is misclassified as target speaker) divided by the total number of non-target test trials.

In Table 3, we examine the difference between Mandarin singing and normal reading speech for ASV 1 s-1 s task using three different systems, the conventional GMM-based, DTW-based and the state-of-the-art DNN X-vector based systems. From preliminary results in Table 3, it is clear that the GMM-based and DTW-based systems achieved almost the same overall performances. Moreover, we

found that the GMM and DTW based systems obtained big EER difference between using singing and reading speech, both for the gender-independent trials and the female gender-dependent trials. These preliminary results show that the voiceprint information extracted from the singing speech was more discriminative than the one extracted from reading speech, especially for the female speaker discrimination. By comparing EERs obtained from singing and reading speech on gender-dependent tasks, we obtained a relative 23.2% EER reduction for female trials. For gender-independent task, a relative 26.1% EER reduction was obtained for the overall test trials.

Table 3. EER% comparison on 1 s-1 s text-dependent ASV tasks, using different systems.

System	Task	Gender	Reading	Singing
GMM-based	1 s-1 s-GD	Male	1.0	1.0
		Female	4.3	3.3
		All	2.5	2.2
	1 s-1 s-GI	All	2.3	1.7
DTW-based	1 s-1 s-GD	Male	1.0	1.0
		Female	4.3	3.3
		All	2.5	**2.2**
	1 s-1 s-GI	All	2.3	**1.8**
X-vector based	1 s-1 s-GD	Male	1.7	1.7
		Female	0.0	0.0
		All	0.5	0.5
	1 s-1 s-GI	All	0.8	0.8

In addition, compared with GMM-based and DTW-based systems, the state-of-the-art X-vector based system achieved much better overall results on the "GD" and "GI" tasks. However, we can see that the X-vector based system obtained the exact same results for both the ASV tasks using reading and singing speech. From the X-vector based systems, we did not see any performance difference between using these two different speaking styles. The main reason is that both the model training of X-vector extractor and PLDA backend need a large amount of labeled data. In this study, they were trained using the out-of-domain "AISHELL" corpus with only reading style speech because the RSS corpus is very limited. As mentioned in the X-vector based system description in Section 4, the RSS corpus cannot be used for X-vector extractor training in general. The acoustic properties learned by the extractor and PLDA deviated far from the singing speech; they were biased to the reading speech. The biased models make the speaker discrimination of singing incapable of being reflected. We will re-validate the X-vector based systems in our future works, when enough singing speech samples are available to train an unbiased deep X-vector extractor.

In fact, we used the X-vector based system in this study to only show preliminary results for speaker verification based on singing speech, even the training data of X-vector extractor was biased to the reading-speech. Therefore, it is unfair to compare the X-vector based system with the conventional GMM-based and DTW-based systems, because the amount of data we used for X-vector based system was much higher than the amount we used for the GMM-based and DTW-based system (only using the target speaker enrollment speech for model training), and this is the main reason for the better performance on some conditions in Table 3.

In our experiments on 3 s-3 s tasks, it is interesting to observe that all of the GMM-based, DTW-based and X-vector based systems obtained zero EERs on both the 3 s-3 s-GD and 3 s-3 s-GI ASV tasks. That is, all of the test trials were correctly detected. This indicates that, when the enrollment speech duration reaches 3 s, the speaker identity information can be accurately captured for our text-dependent ASV task. This may also due to the quiet speech recording environment and the limited number of speakers in our tasks (only 20 speakers).

From the extraordinary results on 3 s-3 s tasks, we may think that, in a real-world system with limited speakers, when the enrollment speech duration is long enough, e.g., 3 s, the speaker identity information can be accurately captured, even we use very simple classifiers such as GMM and DTW. However, in most real-world applications, the background noise or speech recording and transmission channel mismatch between training and testing will degrade the ASV system performance significantly. There are many research works focused on the speech enhancement and channel compensation issues in ASV [1,37]. However, our motivation is to see the difference between singing and normal reading speech in ASV, thus we recorded the RSS corpus under a very quiet office environment to eliminate other interference such as background noise, etc. It is normal for us to obtain zero EER on a text-dependent ASV task under our speech recording setups. Therefore, the extraordinary results on our 3 s-3 s tasks does not indicate that we can obtain exactly the same good results under most real-world ASV applications, except for under the same data recording and setups with our RSS corpus.

In X-vector based systems, the other interesting thing we found was that the NCCF feature brought no performance gains; we obtained the same EERs on both 1 s-1 s and 3 s-3 s tasks from the 60-dimensional MFCCs and the same 60-dimensional MFCCs plus one-dimensional NCCF feature. This observation may also be affected by the fact that the X-vector extractor was trained from out-of-domain corpus. Anyway, all of these observations will be further validated in our future works when many singing and reading speech samples are recorded.

6. Conclusions

This study attempted to investigate the short-time text-dependent ASV between the singing and normal reading speech. A new corpus was released to the public for pure research. Detail comparison and analysis between MFCC and NCCF feature were presented. Preliminary experiments were performed on both 1 s-1 s and 3 s-3 s speaker verification conditions, either for gender-dependent or gender-independent tasks. We found that the speaker identity information extracted from the singing speech was more distinguishable than the one extracted from reading speech. Furthermore, it is worth noting that, in our current corpus, all of the speakers are very familiar with the singing text and original songs, thus the melody of these original songs may guide the speaker to sing with a similar singing style. This similar singing style may also reduce the discriminative information between different speakers. The data amount of our current RSS corpus is limited because creating a high-quality speech database for thousands of hours takes a very long time and a huge amount of money. Because until now there is no publicly released corpus designed for our purpose, we chose to put our preliminary works in this study first. We hope that creating and releasing the RSS would promote the research work in this field. Since, in this study, the performance improvements were only observed on the limited RSS dataset, we cannot guarantee their generalization and significance. Therefore, in our future works, we will focus on recording a larger corpus on reading, singing and humming speech using speakers' personalized password texts. They may sing and hum using their own styles. All observations obtained from this study will be examined in our future larger database. New features related to the singing and humming speech will also be considered.

Author Contributions: Y.S. mainly built the GMM, DTW systems and wrote this paper. J.Z. collected the RSS corpus and revised this paper. Y.L. was the supervisor of this research. Y.L. advised about the X-vector implementation and experimental analysis. H.M. was the co-supervisor of this research.

Funding: This work was funded by the Project 61701306 supported by National Natural Science Foundation of China.

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

References

1. Kinnunen, T.; Li, H. An overview of text-independent speaker recognition: From features to supervectors. *Speech Commun.* **2010**. *52*, 12–40.
2. Dehak, N.; Kenny, P.J.; Dehak, R.; Dumouchel, P.; Ouellet, P. Front-end factor analysis for speaker verification. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 788–798.
3. 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; pp. 999–1003.
4. 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, 2018; pp. 5329–5333.
5. 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.
6. Wu, Z.; Evans, N.; Kinnunen, T.; Yamagishi, J.; Alegre, F.; Li, H. Spoofing and countermeasures for speaker verification: A survey. *Speech Commun.* **2015**, *66*, 130–153.
7. Kinnunen, T.; Sahidullah, M.; Delgado, H.; Todisco, M.; Evans, N.; Yamagishi, J.; Lee, K.A. The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In Proceedings of the Interspeech, 2017; pp. 2–6.
8. Yingke, Z.; Ko, T.; Snyder, D.; Mak, B.; Povey, D. Self-attentive speaker embeddings for text-independent speaker verification. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 3573–3577.
9. Heigold, G.; Moreno, I.; Bengio, S.; Shazeer, N. End-to-end text-dependent speaker verification. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5115–5119.
10. Zhang, C.; Yu, C.; Hansen, J.H. An investigation of deep-learning frameworks for speaker verification anti-spoofing. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 684–694.
11. Lai, C.I.; Abad, A.; Richmond, K.; Yamagishi, J.; Dehak, N.; King, S. Attentive Filtering Networks for Audio Replay Attack Detection. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton , UK, 12–17 May 2019; pp. 6316–6320.
12. Richardson, F.; Reynolds, D.; Dehak, N. Deep neural network approaches to speaker and language recognition. *IEEE Signal Process. Lett.* **2015**, *22*, 1671–1675.
13. Bhattacharya, G.; Alam, M.J.; Kenny, P. Deep Speaker Embeddings for Short-Duration Speaker Verification. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 1517–1521.
14. Jin, M.; Kim, J.; Yoo, C.D. Humming-based human verification and identification. In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19–24 April 2009; pp.1453–1456.
15. Patil, H.A.; Jain, R.; Jain, P. Identification of speakers from their hum. In *Lecture Notes in Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2008; Volume 5246, pp. 461–468.
16. Patil, H.A.; Parhi, K.K. Novel Variable Length Teager Energy based features for person recognition from their hum. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 14–19 March 2010; pp. 4526–4529.
17. Amino, K.; Arai, T. Speaker-dependent characteristics of the nasals. *Forensic Sci. Int.* **2009**, *185*, 21–28.
18. Patil, H.A.; Madhavi, M.C.; Chhayani, N.H. Person Recognition using Humming, Singing and Speech. In Proceedings of the 2012 International Conference on Asian Language Processing, Hanoi, Vietnam, 13–15 November 2012; pp. 149–152.
19. Mehrabani, M.; Hansen, J.H. Singing speaker clustering based on subspace learning in the GMM mean supervector space. *Speech Commun.* **2013**, *55*, 653–666.
20. Nwe, T.L.; Li, H. Exploring vibrato-motivated acoustic features for singer identification. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *15*, 519–530.
21. Nwe, T.L.; Li, H. On fusion of timbre-motivated features for singing voice detection and singer identification. In Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, 31 March–4 April 2008; pp. 2225–2228.

22. Ghahremani, P.; BabaAli, B.; Povey, D.; Riedhammer, K.; Trmal, J.; Khudanpur, S. A pitch extraction algorithm tuned for automatic speech recognition. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 2494–2498.
23. Mehrabani, M.; Hansen, J.H. Dimensionality analysis of singing speech based on locality preserving projections. In Proceedings of the Interspeech, Lyon, France, 25–29 August 2013; pp. 2910–2914.
24. De Cheveigné, A.; Kawahara, H. YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.* **2002**, *111*, 1917–1930.
25. 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 2011 Workshop on Automatic Speech Recognition and Understanding, Hilton Waikoloa Village, Big Island, HI, USA, 11–15 December 2011; pp. 4–7.
26. Kumar, K.; Kim, C.; Stern, R.M. Delta-spectral cepstral coefficients for robust speech recognition. In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011, pp. 4784–4787.
27. Huang, X.; Acerd, A.; Hon, H.-W. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*; Prentice Hall PTR: May 2001. Available online: http://www.worldcat.org/isbn/0130226165 (accessed on 2 June).
28. Dey, S.; Madikeri, S.; Ferras, M.; Motlicek, P. Deep neural network based posteriors for text-dependent speaker verification. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5050–5054.
29. Lei, Y.; Scheffer, N.; Ferrer, L.; McLaren, M. A novel scheme for speaker recognition using a phonetically-aware deep neural network. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 1695–1699.
30. Chen, N.; Qian, Y.; Yu, K. Multi-task learning for text-dependent speaker verification. In Proceedings of the 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015; pp.185–189.
31. Liu, Y.; Qian, Y.; Chen, N.; Fu, T.; Zhang, Y.; Yu, K. Deep feature for text-dependent speaker verification. *Speech Commun.* **2015**, *73*, 1–13.
32. Li, X.; Wu, X. Modeling speaker variability using long short-term memory networks for speech recognition. In Proceedings of the 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015; pp.1086–1090.
33. Prince, S.J.; Elder, J.H. Probabilistic linear discriminant analysis for inferences about identity. In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007; pp. 1–8.
34. Peddinti, V.; Povey, D.; Khudanpur, S. A time delay neural network architecture for efficient modeling of long temporal contexts. In Proceedings of the 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015; pp. 3214–3218.
35. Du, J.; Na, X.; Liu, X.; Bu, H. AISHELL-2: Transforming Mandarin ASR Research Into Industrial Scale. *arXiv* **2018**, arXiv:1808.10583.
36. 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, Korea, 1–3 November 2017; pp. 1–5.
37. Mosner, L.; Matejka, P.; Novotny, O.; Cernocky, J.H. Dereverberation and Beamforming in Far-Field 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. 5254–5258.