applied sciences

MDPI

Article

Noise-Robust Voice Conversion Using High-Quefrency Boosting via Sub-Band Cepstrum Conversion and Fusion

Xiaokong Miao, Meng Sun *, Xiongwei Zhang * and Yimin Wang

Lab of Intelligent Information Processing, Army Engineering University, Nanjing 210007, China; miao_xk@163.com (X.M.); vivhappyrom@163.com (Y.W.)

* Correspondence: sunmengccjs@gmail.com (M.S.); xwzhang9898@163.com (X.Z.)

Received: 11 November 2019; Accepted: 19 December 2019; Published: 23 December 2019

Featured Application: In this paper, we proposed a method of noise-robust voice conversion using high-quefrency boosting via sub-band cepstrum conversion and fusion. This method could improve the quality and similarity of the converted voice compared to other methods, even with the noisy inputs of source speakers. Furthermore, it could be used in many applications, such as speech enhancement, personalized text-to-speech (TTS) synthesis, and emotion conversion. This algorithm also has potential important applications in speaker spoofing, speaking assistance, singing voice processing, and so on.

Abstract: This paper presents a noise-robust voice conversion method with high-quefrency boosting via sub-band cepstrum conversion and fusion based on the bidirectional long short-term memory (BLSTM) neural networks that can convert parameters of vocal tracks of a source speaker into those of a target speaker. With the implementation of state-of-the-art machine learning methods, voice conversion has achieved good performance given abundant clean training data. However, the quality and similarity of the converted voice are significantly degraded compared to that of a natural target voice due to various factors, such as limited training data and noisy input speech from the source speaker. To address the problem of noisy input speech, an architecture of voice conversion with statistical filtering and sub-band cepstrum conversion and fusion is introduced. The impact of noises on the converted voice is reduced by the accurate reconstruction of the sub-band cepstrum and the subsequent statistical filtering. By normalizing the mean and variance of the converted cepstrum to those of the target cepstrum in the training phase, a cepstrum filter was constructed to further improve the quality of the converted voice. The experimental results showed that the proposed method significantly improved the naturalness and similarity of the converted voice compared to the baselines, even with the noisy inputs of source speakers.

Keywords: voice conversion; BLSTM; statistical filtering; sub-band cepstrum; noise robustness

1. Introduction

Voice conversion (VC) is a technique that converts the characteristics of a source speaker to those of a target speaker while maintaining the linguistic contents of the input speech [1–3]. VC has been used in many applications of speech processing, such as singing voice processing [4], personalized text-to-speech (TTS) synthesis [5], speaker spoofing [6], speaking assistance [7], speech enhancement, and emotion conversion [8–10]. VC may be categorized into two categories: one with parallel corpus (i.e., both source and target speakers utter the same sentences) and the other with non-parallel corpus (i.e., the target speakers utter different sentences from the source speaker) tasks according to whether

the source and target speakers speak the same texts. Many existing approaches that have yielded conversion results with both high quality and high similarity are based on parallel data, such as Gaussian mixture models (GMM) [2,11], frequency warping (FW) [12–14], deep neural networks (DNN) [15–17], non-negative matrix factorization (NMF) [18,19], and so on. This paper also focuses on VC with parallel training data.

In the Voice Conversion Challenge 2018 (VCC 2018) [20], the methods based on bidirectional long short-term memory (BLSTM) and GMM achieved excellent results, especially the strong baseline based on GMM, which ranked second among all the submitted algorithms. In this competition, the system *N10* submitted by the University of Science and Technology of China (USTC) and iFlytek Research achieved the best results for the parallel HUB sub-tasks. The HUB task was a sub-task of VCC 2018, which was designed to evaluate the voice conversion algorithms on a parallel corpus [20]. The system was trained on a large amount of data that consisted of hundreds of hours of noisy speech and dozens of hours of clean English speech. The trend reflected in VCC 2018 indicated the future directions of the research on VC, such as alleviating the reliance on a large amount of training data, improving the similarity between converted voices and target voices, considering the noise resistance of the conversion system, and making the system real-time. Speech recorded in daily life is more or less polluted by various kinds of noises, so it is difficult to get clean speech directly. Hence, the conversion of noisy speech plays an important role when putting VC into practical usage.

In view of the problems mentioned above, in order to ensure the high similarity between the converted voice and the target voice, and to improve the noise robustness of the system, especially with limited training data, in this study, we investigated a noise-robust VC using high-quefrency boosting via sub-band cepstrum conversion and fusion with a BLSTM neural network. First, two different filtering methods were introduced to suppress noises in the preprocessing stage and the post-processing stage, respectively. One was the low-pass filtering of time-domain waveforms to remove part of the high frequency noises from recording, and the other was statistical filtering of Mel-cepstral coefficients (MCEPs) to reduce the interference of noises on the converted coefficients. Second, in the process of feature extraction, the feature dimension of MCEPs was extended, and two different BLSTMs were used to convert the sub-band cepstrum. The motivation for doing this was that, based on the experience of our previous experiments, when the voice is contaminated by noises, the high-quefrency components of speech seem to be vulnerable to interference. Meanwhile, in order to alleviate the problem of parameter discontinuity when concatenating a converted sub-band cepstrum, a weighted fusion scheme was proposed. Finally, the converted waveform was generated using a vocoder after statistical filtering. Experiments and subjective evaluations demonstrated that the converted voices of the proposed method held clearer pronunciations and better speaker similarity than those from the BLSTM and GMM baselines on both English and Mandarin datasets. Experiments also demonstrated that the proposed method maintained a good performance for noisy input speech.

The remainder of this paper is organized as follows: Section 2 mainly presents related work and recent progress. Section 3 describes the proposed approach. The setup of experiments and analysis of results are discussed in Section 4. Finally, the conclusions and future work are presented in Section 5.

2. Related Work

Like conventional machine learning recipes, a VC system usually consists of two phases: a training phase and a conversion phase. During the training phase, the input voice is first decomposed into acoustic features, such as fundamental frequency (F0), spectral envelope, and aperiodic components [14], and conversion functions are subsequently estimated to bridge the acoustic features obtained from the parallel corpus of the source speaker and the target speaker. During the conversion phase, the conversion function is applied on features extracted from the new input voice [10]. Finally, a converted speech waveform is generated from the converted acoustic features by implementing a vocoder [21]. In this paper, two benchmark methods are taken as baselines: one is the GMM-based VC method from VCC 2018 and the other is a BLSTM-based VC method in Sun et al. [15].

2.1. GMM-Based VC

In VCC 2018, two different baseline methods (i.e., a traditional VC system based on a GMM [2], and a vocoder-free VC system based on a differential Gaussian mixture model (DIFFGMM) [22]) were officially released, which were used by the participants to improve their work on the tasks. However, few submissions beat these competitive baselines.

GMM is a statistical model for feature mapping. As the most popular baseline, Stylianou et al. proposed a conversion method with a Gaussian mixture model (GMM) that accomplished feature mapping via soft clustering [1]. However, the performance of the conversion was still insufficient. The quality of the converted voice is deteriorated by some factors, e.g., spectral movement with inappropriate dynamic characteristics caused by the frame-by-frame conversion process and over-smoothing of converted spectra [23,24]. To address the problems of the time-independent mapping and over-smoothing, Toda et al. proposed a conversion method based on the maximum likelihood estimation of a spectral parameter trajectory [2]. This method considered the global variance (GV) of the converted spectra over a time sequence as a novel feature, which was able to capture the characteristics of the parameter trajectory. This idea effectively complemented the conventional frameworks of statistical conversion. The feature conversion part of a VC system can be denoted using

$$y = f(x), \tag{1}$$

where $f(\cdot)$ is a mapping function. The vectors x and y are features from the source and target speakers, respectively. Given the sequential nature of speech, the features are sequences of frame-level vectors, i.e.,

$$x = \left[x_1^{\mathrm{T}}, x_2^{\mathrm{T}}, \cdots, x_i^{\mathrm{T}}, \cdots, x_I^{\mathrm{T}}\right]^{\mathrm{T}}, \tag{2}$$

$$y = \left[y_1^{\mathrm{T}}, y_2^{\mathrm{T}}, \cdots, y_i^{\mathrm{T}}, \cdots, y_I^{\mathrm{T}}\right]^{\mathrm{T}}, \tag{3}$$

where T denotes the transposition of the vector and x_i and y_i are row vectors with features of the ith frame.

In Toda's approach, $2D$-dimensional joint static and dynamic feature vectors, $X_i = \left[x_i, \Delta x_i\right]$ and $Y_i = \left[y_i, \Delta y_i\right]$ are composed of D-dimensional static feature vectors x_i and y_i and D-dimensional dynamic feature vectors Δx_i and Δy_i, respectively. Their joint probability density function modeled using GMM is given as Equation (4),

$$P(X_i, Y_i|\lambda) = \sum_{m=1}^{M} \alpha_m N\left(\begin{bmatrix} X_i \\ Y_i \end{bmatrix}; \begin{bmatrix} \mu_m^{(X)} \\ \mu_m^{(Y)} \end{bmatrix}, \begin{bmatrix} \sum_m^{(X\,X)} & \sum_m^{(X\,Y)} \\ \sum_m^{(Y\,X)} & \sum_m^{(Y\,Y)} \end{bmatrix}\right), \tag{4}$$

where $N(\mu,\Sigma)$ denotes the normal distribution with mean μ and covariance matrix Σ. m is the index of the Gaussian mixture. M is the total number of Gaussian mixtures. λ is a GMM parameter set consisting of the mixture weight α_m, the mean vector μ_m, and the covariance matrix Σ_m of the mth mixture. The GMM is trained using joint vectors of X_i and Y_i in the parallel data set, which have been automatically aligned to each other by dynamic time warping (DTW) [4]. The detailed steps can be found in References [2,25].

DIFFGMM is a differential Gaussian mixture model. The DIFFGMM is analytically derived from the GMM (in Equation (4)) used in the conventional VC. Let $D_i = \left(d_i^T, \Delta d_i^T\right)^T$ denote the static and dynamic differential feature vector, where $d_i = y_i - x_i$ and $\Delta d_i = \Delta y_i - \Delta x_i$. The $2D$-dimensional joint static and dynamic feature vector between the source and the differential features is given as:

$$\begin{bmatrix} X_i \\ D_i \end{bmatrix} = \begin{bmatrix} X_i \\ Y_i - X_i \end{bmatrix} = A \begin{bmatrix} X_i \\ Y_i \end{bmatrix}, \tag{5}$$

$$A = \begin{bmatrix} I & 0 \\ -I & I \end{bmatrix}, \tag{6}$$

where A is a transformation matrix that transforms the joint feature vector between the source and target features into that of the source and difference features. I denotes the identity matrix. By applying the transformation matrix to the traditional GMM in Equation (4), the DIFFGMM is derived and given as follows:

$$P(X_i, D_i|\lambda) = \sum_{m=1}^{M} \alpha_m N\left(\begin{bmatrix} X_i \\ D_i \end{bmatrix}; \begin{bmatrix} \mu_m^{(X)} \\ \mu_m^{(D)} \end{bmatrix}, \begin{bmatrix} \sum_m^{(X\,X)} & \sum_m^{(X\,D)} \\ \sum_m^{(D\,X)} & \sum_m^{(D\,D)} \end{bmatrix}\right). \tag{7}$$

For a more detailed conversion process of VC and DIFFGMM and the training process of parallel VC based on GMM, please refer to Figures 1 and 2 in Kobayashi and Toda [25].

2.2. BLSTM-Based VC

BLSTM is an improvement of the bidirectional recurrent neural network (RNN), which can model a certain amount of contextual information with cyclic connections and map the whole history of previous inputs to each output in principle [10]. However, conventional RNNs can access only a limited range of context because of the gradient explosion or vanishing over time in long-range contextual transmission. An effective way to overcome this problem is to introduce long short-term memory (LSTM), which takes advantage of specially designed memory cells that store information in a linear storage unit for many temporal steps and can learn the optimal amount of contextual information related to regression tasks [26].

Sun et al. investigated the use of deep bidirectional long short-term memory (DBLSTM) for VC [15]. Ming et al. also used DBLSTM to model timbre and prosody for emotional VC [27]. The DBLSTM network architecture is a combination of bidirectional RNNs and LSTM memory blocks, which is able to learn long-range contextual relations in both forward and backward directions. By stacking multiple hidden layers, a deep network architecture is built to capture a high-level representation of voice features [27]. Hence, it was adopted in this study to accurately depict acoustic features.

3. Proposed Method

3.1. Overall Architecture

Figure 1 is the overall framework of the proposed method. The main contributions of the proposed system are as follows: (1) two filtering methods are introduced into the conversion network and (2) an approach with sub-band cepstrum conversion and fusion is proposed. In this system, the input time domain signal was filtered through a temporal filter, and then acoustic features were extracted by WORLD [28], including Mel-cepstral coefficients (MCEPs), F0, and aperiodic components, which were treated with their corresponding recipes separately, as presented below. MCEPs (except for the energy feature) were converted by the proposed sub-band cepstrum conversion and fusion network. LogF0 was converted by equalizing the mean and the standard deviation of the source and target voice, which is a widely used method in VC. The aperiodic component was directly copied to perform as the corresponding parameters of the converted speech since previous research has shown that converting the aperiodic component does not make a statistically significant contribution toward improving the synthesized speech [15,29]. After training the conversion model, the mean and variance of all target MCEPs and the converted MCEPs were calculated, and the target and conversion matrices T_1 and T_2 in Figure 1 were computed and stored, respectively. The detailed construction process of T_1 and T_2 is presented in Section 3.3. During the conversion phase, a global variance consistency filter was constructed using T_1 and T_2 for MCEPs filtering. The log-linear transform was applied to F0 conversion as follows:

$$p_t{}^{(Y)} = \frac{p_t{}^{(X)} - u^{(X)}}{\sigma^{(X)}} \times \sigma^{(Y)} + u^{(Y)}, \tag{8}$$

where $p_t^{(X)}$ and $p_t^{(Y)}$ are the original logF0 and the converted logF0, respectively. $u^{(X)}$ and $u^{(Y)}$ are the means, and $\sigma^{(X)}$ and $\sigma^{(Y)}$ are the standard deviations of logF0's of the training data for the source and the target speakers, respectively [2].

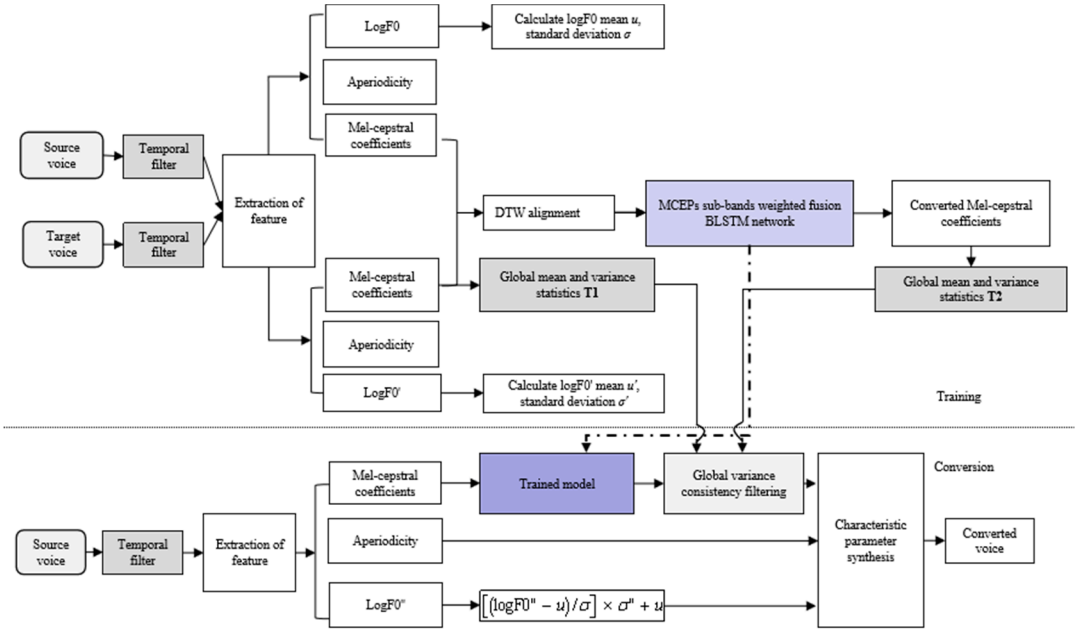


Figure 1. Voice conversion flow chart of sub-band cepstrum weighted fusion bidirectional long short-term memory (BLSTM). DTW: dynamic time warping.

3.2. High-Quefrency Boosting via Sub-Band Cepstrum Conversion and Fusion

The sub-band cepstrum conversion and weighted fusion is designed to improve the noise robustness of the whole conversion system. Figure 2 shows the sub-band fusion BLSTM network for MCEPs. BLSTM can be created by stacking multiple hidden layers. In the proposed model, two BLSTMs (i.e., BLSTM1 and BLSTM2 in Figure 2) were used, and both of them were composed of three hidden layers.

The motivation for designing a sub-band cepstrum conversion system was that when voice is contaminated by strong noises, the high quefrency components of cepstrum failed to be recovered accurately, given our previous experiences. Simply training a conversion network may not be sufficient to recover the full-band cepstrum accurately, which could be overcome by a converging cepstrum in the sub-bands. In the proposed model, BLSTM1 was used for training and converting a full-band cepstrum, because full-band MCEPs are a high-dimensional cepstrum, which contains a large amount of data, such that the training data is sufficient and is of benefit to training appropriate network parameters and making the network convergent. An additional BLSTM, BLSTM2, was designed to convert and generate the high-quefrency part of the MCEPs by mapping the full-band information of the source MCEPs with the high-quefrency band of the target MCEPs. Finally, the two networks were fused. By applying the fusion coefficients extracted in the training phase to the conversion phase, the sub-band cepstral parameters were smoothly connected, and then the discontinuity in the process of cepstrum splicing was alleviated.

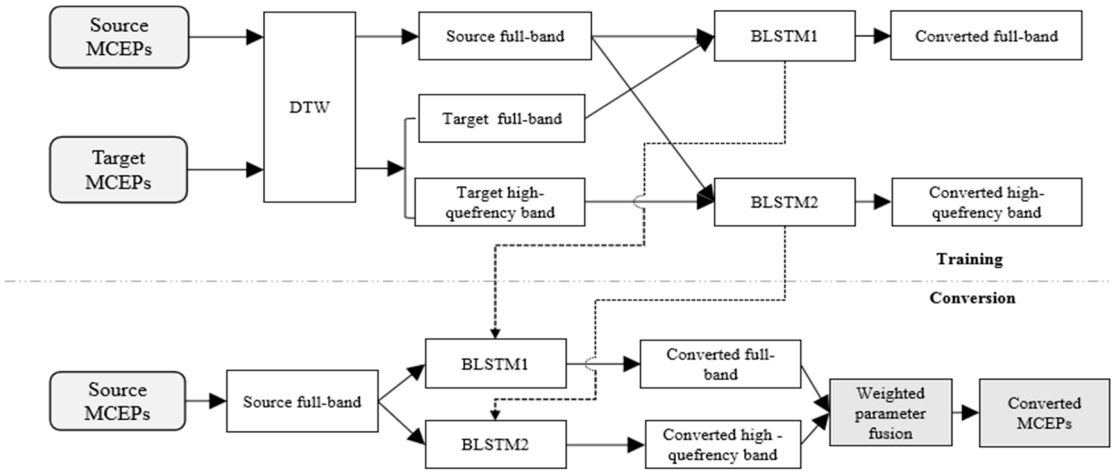


Figure 2. The flowchart of voice conversion with sub-band cepstrum conversion by BLSTMs. MCEPs: Mel-cepstral coefficients

As shown in Figure 3, the trained BLSTM1 and BLSTM2 were used to convert MC_{full} to $MC_{full}^{\#}$ and MC_{full} to $mcep_{high2}$, respectively. The $mcep_{high1}$ and $mcep_{low1}$ were from $MC_{full}^{\#}$ and the $mcep_{high2}$ was converted from MC_{full} by BLSTM2. α was the weight parameter used to perform sub-band cepstrum fusion.

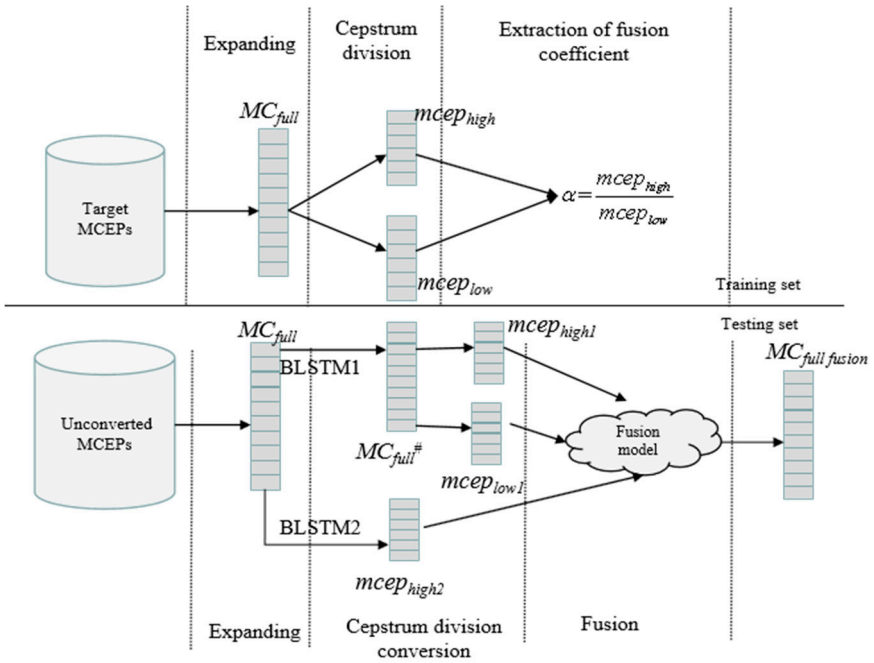


Figure 3. The diagram of the cepstrum-weighted fusion process, where the MC represents the full band MCEPs in different cases.

In a sub-band cepstrum conversion, the full-band cepstrum is divided into two parts with equal numbers of cepstral coefficients: one represents the low-quefrency sub-band and the other represents the high-quefrency sub-band. The two sub-bands are subsequently treated in different ways. When the converted sub-band cepstrum is ready, the two parts are fused by a weight parameter alpha (i.e., α in Figure 3). The specific algorithm in the fusion process is given as:

$$\begin{aligned} MC_{high\ fusion} &= \alpha \times mcep_{high1} + (1-\alpha) \times mcep_{high2} \\ MC_{low1} &= mcep_{low1} \\ MC_{full\ fusion} &= [MC_{low1}; MC_{high\ fusion}] \end{aligned} \quad . \tag{9}$$

3.3. Time–Frequency Domain Filtering

Filtering is a common denoising method in signal processing. In this study, a global variance consistency filtering model was introduced into the BLSTM conversion network. A low-pass filter was used to remove high-frequency recording noises from electrical devices. After the MCEPs conversion, the converted cepstrum was filtered again through the global variance consistency filter constructed using T_1 and T_2 in Figure 1, which is believed to be good at removing the artificial noised generated by the conversion process. The details of conducting the global variance consistency filtering are given as follows.

(1) The mean and variance of each one-dimensional MCEPs of the target sentences was calculated using Equation (10):

$$\begin{cases} \overline{x}_{i_{tar}} = \frac{1}{(N \times M)} \sum_{n=1}^{N} \sum_{m=1}^{M} x_i \\ {\sigma^2}_{i_{tar}} = \frac{1}{(N \times M)} \sum_{n=1}^{N} \sum_{m=1}^{M} (x_i - \overline{x}_{i_{tar}})^2 \end{cases} (i = 1, 2, 3, \cdots, T), \tag{10}$$

where N, M, and T represent the number of target sentences in the training phase, the number of frames per sentence, and the dimension of MCEPs, respectively. x_i denotes the ith dimensional MCEPs, $\overline{x}_{i_{tar}}$ and ${\sigma^2}_{i_{tar}}$ represent the mean and variance of MCEPs in the ith dimension of all the training sentences, respectively. *tar* represents the information from the target speaker.

(2) Equation (10) was also used to calculate the mean and variance of all the converted MCEPs during the training phase and the MCEPs to be converted during the conversion phase. $\overline{x}_{i_{con}}$ and ${\sigma^2}_{i_{con}}$ represent the mean and variance in ith dimension of the converted MCEPs, respectively. *con* represents the information of the converted voice. The vector $\overline{y}$ was constituted of each dimensional MCEP of sentences to be converted. The vector ${\sigma_y}^2$ was constituted of the variance of each dimensional MCEP of sentences to be converted.

(3) A global variance consistency filter was constructed to obtain the primary filtered data, as shown in Equation (11):

$$\hat{y} = \sqrt{\left(\frac{\overline{x}_{tar}}{\overline{x}_{con}}\right)} \times (y - \overline{y}) + \overline{y}, \tag{11}$$

where $\overline{x}_{tar}$ denotes the vector consisting of means of each dimension of the MCEPs of target sentences. $\overline{x}_{con}$ denotes the vector consisting of the means of each dimension of MCEPs of the converted sentences. y denotes the MCEP vector of the sentences to be converted. $\hat{y}$ is the output after the initial filtering.

(4) As in Equation (12), α was set to adjust for the final filtering data:

$$\widetilde{y} = \alpha \times \hat{y} + (1 - \alpha) \times y, \tag{12}$$

where $\widetilde{y}$ is the final MCEPs after filtering.

4. Experiments and Results

4.1. Experimental Setup and Implementation Details

The experiments were conducted on an English dataset CMU ARCTIC [30] and a Mandarin dataset THCHS-30 [31]. Babble noises were added to the utterances of all the source speakers with 15 dB signal-to-noise ratio (SNR) to evaluate the noise robustness of the VC methods. The voice signals of CMU ARCTIC were resampled to 16 kHz with a mono channel. The window length was 25 m and the frame shift was 5 m. Four speakers were chosen: two male speakers, bdl and jmk, and two female speakers, slt and clb. Both intra-gender and inter-gender conversions were conducted between the following pairs of speakers: bdl to jmk (male to male, M-M), clb to slt (female to female, F-F), clb to jmk (female to male, F-M), and bdl to slt (male to female, M-F). The voice signals of THCHS-30 were also resampled to 16 kHz. A male speaker (A8) and two female speakers (A2 and A23) were chosen and A2 to A23 (F-F), A8 to A2 (M-F) were taken as the intra-gender and inter-gender conversions, respectively. In the experiments on both databases, 200 utterances were randomly selected for training, and another 15 utterances of each speaker were randomly chosen for evaluation, where there was no overlap between the training and the evaluation utterances.

In our work, WORLD was used as the vocoder to analysis and synthesis voice signals. For voice conversion purposes, in order to avoid underfitting and reduce the training complexity of the data, low dimensional MCEPs were usually adopted as spectral features. However, thanks to the sub-band cepstrum processing scheme proposed in this paper, our method was able to process relatively high-dimensional MCEPs without training difficulty. The high-dimensional representation contained many details of speakers, which should be retained and utilized, especially in the cases with limited training data. In the following experiments, 129-dimensional MCEPs were calculated from the spectrum using a Speech Signal Processing Toolkit (SPTK), where the 128-dimensional MCEPs (except the energy feature. i.e., MC_{full} in Figure 3) were divided into the 64-dimensional low-quefrency MCEPs and the 64-dimensional high-quefrency sub-bands (i.e., $mcep_{low}$ and $mcep_{high}$ in Figure 3) for subsequent conversion models [32].

The models for cepstrum conversion in this paper included two BLSTM networks, i.e., BLSTM1 and BLSTM2 for low- and high-quefrency bands, respectively. BLSTM1 had three hidden layers, and the number of units in each layer was [128 256 128], respectively. Given the low fluctuation of parameters of the high-quefrency bands, BLSTM2 took a relatively small units, and the number of units in each layer was [64 128 64], respectively. A dropout layer with parameter 0.5 was introduced to avoid overfitting. The learning rate was adjusted automatically from 1.0×10^{-3} to 1.0×10^{-5}.

Three competitive systems of VC were chosen as baselines for comparison:

GMM: A conventional GMM-based voice conversion system, which was one of the baseline approaches in VCC 2018 [21].

DIFFGMM: A strong baseline for VCC 2018.

BLSTM: A BLSTM with three hidden layers for voice conversion from Sun et al. [15].

4.2. Experimental Results

Objective and subjective tests were conducted on the data of speakers from CMU ARCTIC and THCHS-30. The objective metric for evaluation was Mel-cepstral distortion (MCD) [15]:

$$\text{MCD [dB]} = \frac{10}{ln\ 10}\sqrt{2\sum_{i=1}^{I}(m_i^{con} - m_i^{tar})^2}, \tag{13}$$

where m_i^{con} and m_i^{tar} are the MCEPs of the converted features and target features, respectively [22].

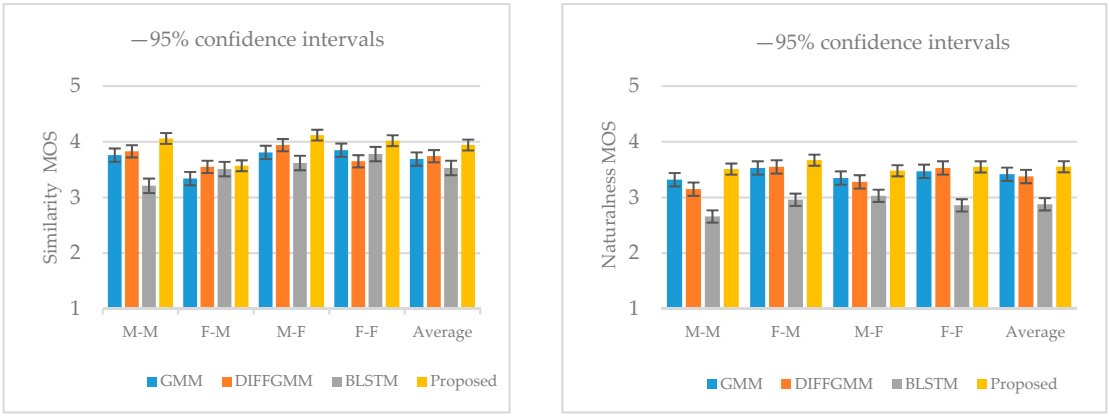
The results of the VC methods on CMU ARCTIC are given in Table 1, where smaller values indicate a lower distortion and a better performance on the conversion. The values of the source–target row are the MCD values of the source voice and the target voice without performing the conversion.

According to the results in Table 1, the proposed method achieved a lower MCD than the competitive baselines in all the cases of conversion. This validated the efficacy of the proposed method.

Table 1. Mean Mel-cepstral distortion (MCD) of all non-silent frames of the voice conversion (VC) methods on the CMU ARCTIC dataset (15dB babble added to source). GMM: Gaussian mixture model, DIFFGMM: differential GMM.

	MCD (dB)			
Methods	**M-M (bdl-jmk)**	**F-M (clb-jmk)**	**M-F (bdl-slt)**	**F-F (clb-slt)**
Source–target	10.682	12.763	10.755	9.896
GMM	9.084	10.732	8.836	9.635
DIFFGMM	8.513	10.214	9.253	9.716
BLSTM	8.210	9.401	7.485	8.032
Proposed	**8.013**	**8.874**	**7.180**	**7.822**

To provide a subjective evaluation of the proposed method, the mean opinion score (MOS) test on the naturalness and similarity of the converted voices was also conducted. Figure 4 shows the results of a subjective evaluation for different methods on the CMU ARCTIC dataset. In the MOS tests, listeners were asked to compare the four utterances of each set with the target voice and to rank the naturalness and similarity of the converted voices on a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). The subjective evaluation results are illustrated in Figure 4. The proposed approach significantly outperformed the baseline approaches in terms of speech naturalness and speaker similarity of the converted speech for both cross-gender and intra-gender conversions. By comparing Table 1 and Figure 4, it was found that BLSTM obtained a higher objective performance than GMM and DIFFGMM but failed to outperform them on subjective evaluations.



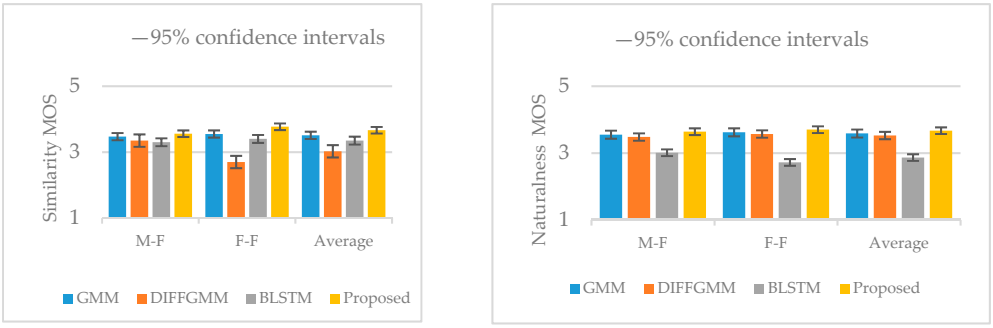
(**a**) (**b**)

Figure 4. Results of the subjective evaluation for different methods on CMU ARCTIC (15 dB babble added to source): (**a**) mean opinion score (MOS) for similarity with 95% confidence intervals and (**b**) MOS for naturalness with 95% confidence intervals.

In order to conduct an extensive comparison, experiments of cross-gender and intra-gender conversions were also performed on the Mandarin dataset THCHS-30. The results on the objective and subjective evaluations are given in Table 2 and Figure 5, respectively. A similar performance was observed as that from CMU ARCTIC.

Table 2. Mean values of the MCD of all non-silent frames of the VC methods on the THCHS-30 dataset (15 dB babble added to source).

MCD (dB)		
Methods	**Cross-Gender (A8–A2)**	**Intra-Gender (A2–A23)**
Source–target	11.873	11.507
GMM	10.326	8.794
DIFFGMM	10.207	8.958
BLSTM	6.964	7.421
Proposed	**6.786**	**7.155**



(**a**) (**b**)

Figure 5. Results of subjective evaluation for different methods on THCHS-30 (15 dB babble added to source): (**a**) MOS for similarity with 95% confidence intervals and (**b**) MOS for naturalness with 95% confidence intervals.

In order to intuitively verify the noise robustness of different methods, the prediction accuracy of different methods for high-quefrency sub-band of the cepstrum was evaluated first. Furthermore, the high-quefrency sub-band Mel-cepstrum distortion (HQMCD) of the converted voice and the target voice was calculated. The conversion results of different methods for noisy speech are shown in Table 3. At the same time, in order to show the noise robustness of the proposed method in a straightforward way, the spectrogram of the converted voices obtained using the involved methods was extracted, as shown in Figure 6.

Table 3. Mean high-quefrency sub-band Mel-cepstrum distortion (HQMCD) for different methods on CMU ARCTIC (15 dB babble add to source).

HQMCD (dB)	
Methods	**Cross-Gender (bdl-slt)**
Source–target	1.131
GMM	0.946
DIFFGMM	1.166
BLSTM	0.912
Proposed	**0.713**

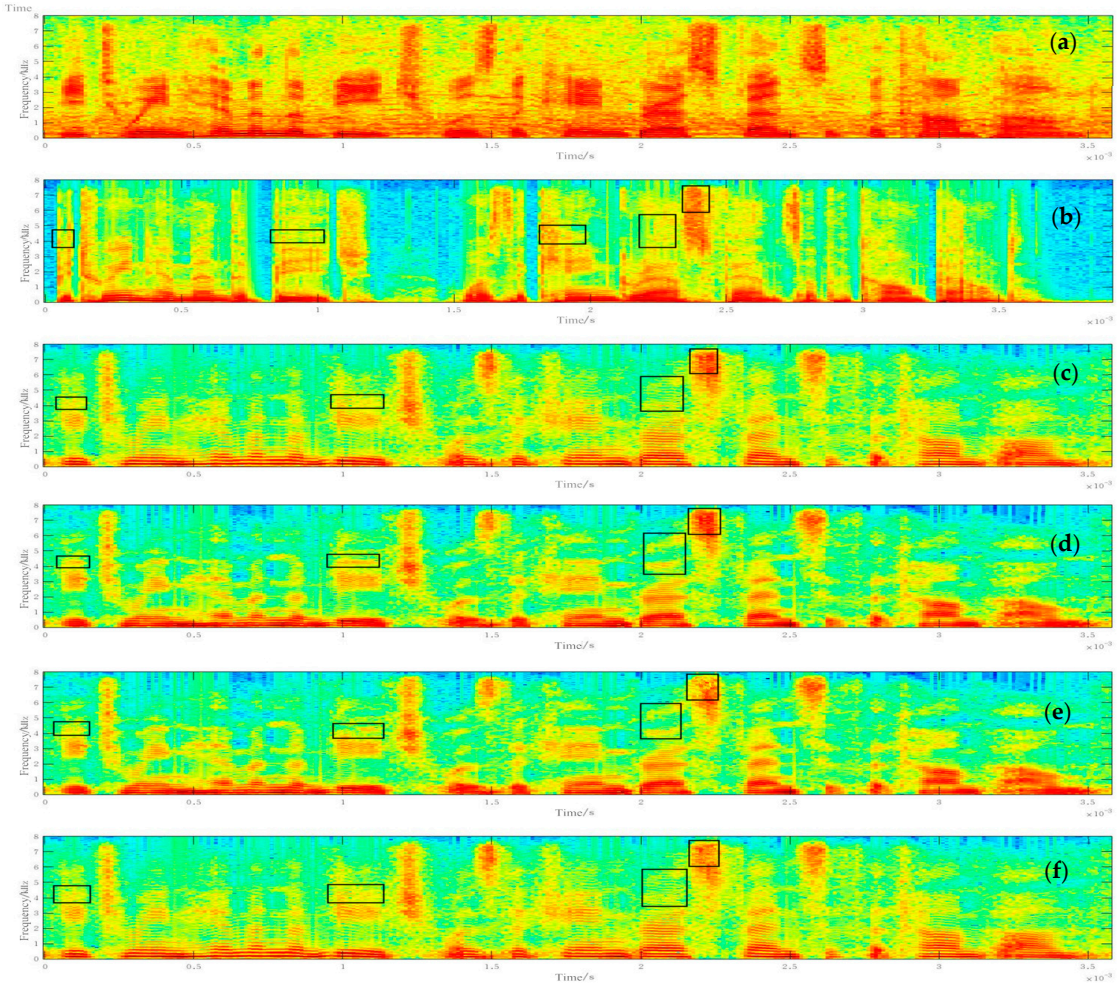


Figure 6. Spectrogram of the methods of voice conversion: (**a**) source voice with babble noise at 15 dB, (**b**) clean target voice, (**c**) BLSTM conversion, (**d**) the proposed method, (**e**) DIFFGMM conversion, and (**f**) GMM conversion.

By reading the results in Table 3 and the boxes in Figure 6, it was found that the proposed method was good at capturing the details of the spectral features of speakers, which finally contributed to its relatively better performance than the baselines. For readers' reference, audio samples for the subjective listening tests are accessible via: https://github.com/miaoxk/Demo.

5. Conclusions

In this paper, a method for noise-robust voice conversion using high-quefrency boosting via sub-band cepstrum conversion and fusion using BLSTM and statistical filtering was proposed. A divide-and-conversion scheme of high-dimensional cepstral coefficients and signal filtering methods were applied to improve the noise robustness of the system by comparing it with several recently proposed methods of voice conversion on the CMU ARCTIC and THCHS-30 databases. It was found that the converted voice obtained by the proposed method showed a relatively high naturalness and similarity. Both objective and subjective experiments confirmed the effectiveness of the proposed method. Meanwhile, the proposed method was able to improve the robustness of the voice conversion system, especially with the noisy inputs of source speakers.

Due to the increase of the network size and the expansion of the feature dimension, the time required for conversion of the proposed method was slightly longer than those of the baselines.

As future work, we would like to study how to improve the network structure to further improve the robustness of the system and to accelerate the conversion to make it real-time.

Author Contributions: Conceptualization, X.M. and M.S.; methodology, X.M.; validation, X.M. and M.S.; formal analysis, Y.W.; investigation, M.S.; data curation, X.M.; writing—original draft preparation, X.M. and M.S.; writing—review and editing, M.S. and X.Z.; supervision, M.S. and X.Z.; project administration, M.S. and X.Z.; funding acquisition, M.S. and X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Natural Science Foundation of Jiangsu Province (grant number BK20180080) and the National Natural Science Foundation of China (grant number 61471394).

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

References

1. Stylianou, A.Y.; Cappe, O.; Moulines, E. Continuous probabilistic transform for voice conversion. *IEEE Trans. SAP* **1998**, *6*, 131–142. [CrossRef]
2. Toda, T.; Black, A.W.; Tokuda, K. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Trans. ASLP* **2007**, *15*, 2222–2235. [CrossRef]
3. Kurita, Y.; Kobayashi, K.; Takeda, K.; Toda, T. Robustness of Statistical Voice Conversion based on Direct Waveform Modification against Background Sounds. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 684–688.
4. Kobayashi, K.; Toda, T.; Neubig, G.; Sakti, S.; Nakamura, S. Statistical Singing Voice Conversion with Direct Waveform Modification based on the Spectrum Differential. In Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014; pp. 2514–2518.
5. Kain, A.; Macon, M.W. Spectral voice conversion for text-to-speech synthesis. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seattle, WA, USA, 12–15 May 1998; pp. 285–288.
6. Wu, Z.; Li, H. Voice conversion and spoofing attack on speaker verification systems. In Proceedings of the APSIPA, Kaohsiung, Taiwan, 29 October–1 November 2013; pp. 1–9.
7. 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]
8. Deng, L.; Acero, A.; Jiang, L.; Droppo, J.; Huang, X. High-performance robust speech recognition using stereo training data. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Salt Lake City, UT, USA, 7–11 May 2001; pp. 301–304.
9. Niwa, J.; Yoshimura, T.; Hashimoto, K.; Oura, K.; Nankaku, Y.; Tokuda, K. Statistical voice conversion based on wavenet. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5289–5293.
10. Miao, X.K.; Zhang, X.W.; Sun, M.; Zheng, C.Y.; Cao, T.Y. A BLSTM and WaveNet based Voice Conversion Method with Waveform Collapse Suppression by Post-processing. *IEEE Access* **2019**, *7*, 54321–54329. [CrossRef]
11. Helander, E.; Virtanen, T.; Nurminen, J.; Gabbouj, M. Voice conversion using partial least squares regression. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *18*, 912–921. [CrossRef]
12. Erro, D.; Alonso, A.; Serrano, L.; Navas, E.; Hernáez, I. Towards physically interpretable parametric voice conversion functions. In Proceedings of the 6th Advances in Nonlinear Speech Processing International Conference, Mons, Belgium, 19–21 June 2013; pp. 75–82.
13. Tian, X.; Wu, Z.; Lee, S.W.; Hy, N.Q.; Chng, E.S.; Dong, M. Sparse representation for frequency warping based voice conversion. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; pp. 4235–4239.
14. Kawahara, H.; Masuda-Katsuse, I.; Cheveigne, A. Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. *Speech Commun.* **1999**, *27*, 187–207. [CrossRef]
15. 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 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; pp. 4869–4873.

16. Nguyen, H.Q.; Lee, S.W.; Tian, X.; Dong, M.; Chng, E.S. High quality voice conversion using prosodic and high-resolution spectral features. *Multimed. Tools Appl.* **2016**, *75*, 5265–5285. [CrossRef]
17. Desai, S.; Black, A.W.; Yegnanarayana, B.; Prahallad, K. Spectral mapping using artificial neural networks for voice conversion. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2010**, *18*, 954–964. [CrossRef]
18. Takashima, R.; Takiguchi, T.; Ariki, Y. Exemplar-based voice conversion using sparse representation in noisy environments. *IEICE Trans. Inf. Syst.* **2013**, *96*, 1946–1953. [CrossRef]
19. Wu, Z.; Virtanen, T.; Chng, E.S.; Li, H. Exemplar-based sparse representation with residual compensation for voice conversion. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 1506–1521.
20. Lorenzo-Trueba, J.; Yamagishi, J.; Toda, T.; Saito, D.; Villavicencio, F.; Kinnunen, T.; Lin, Z.H. The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods. In Proceedings of the Odyssey 2018 The Speaker and Language Recognition Workshop, Les Sables d'Olonne, France, 26–29 June 2018.
21. Dudley, H. Remaking Speech. *J. Acoust. Soc. Am.* **1939**, *11*, 169–177. [CrossRef]
22. Kobayashi, K.; Toda, T.; Nakamura, S. F0 transformation techniques for statistical voice conversion with direct waveform modification with spectral differential. In Proceedings of the SLT, San Diego, CA, USA, 13–16 December 2016; pp. 693–700.
23. Toda, T.; Saruwatari, H.; Shikano, K. Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum. In Proceedings of the ICASSP, Salt Lake City, UT, USA, 7–11 May 2001; pp. 841–844.
24. Toda, T.; Black, A.W.; Tokuda, K. Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter. In Proceedings of the ICASSP, Philadelphia, PA, USA, 18–23 March 2005; Volume 1, pp. 9–12.
25. Kobayashi, K.; Toda, T. sprocket: Open-source voice conversion software. In Proceedings of the Odyssey 2018 The Speaker and Language Recognition Workshop, Les Sables d'Olonne, France, 26–29 June 2018.
26. Martin, W.; Angeliki, M.; Nassos, K.; Björn, S.; Shrikanth, N. Analyzing the memory of BLSTM Neural Networks for enhanced emotion classification in dyadic spoken interactions. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 4157–4160.
27. Ming, H.; Huang, D.; Xie, L.; Wu, J.; Dong, M.; Li, H. Deep Bidirectional LSTM Modeling of Timbre and Prosody for Emotional Voice Conversion. In Proceedings of the Interspeech, San Francisco, CA, USA, 8–12 September 2016.
28. Morise, M.; Yokomori, F.; Ozawa, K. WORLD: A vocoderbased high-quality speech synthesis system for real-time applications. *IEICE Trans. Inf. Syst.* **2016**, *99*, 1877–1884. [CrossRef]
29. Ohtani, Y.; Toda, T.; Saruwatari, H.; Shikano, K. Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation. In Proceedings of the ICSLP, Pittsburgh, PA, USA, 17–21 September 2006.
30. Kominek, J.; Black, A.W. The CMU Arctic speech databases. In Proceedings of the Fifth ISCA Workshop on Speech Synthesis, Pittsburgh, PA, USA, 14–16 June 2004.
31. Wang, D.; Zhang, X.W. THCHS-30: A Free Chinese Speech Corpus. *arXiv* **2015**, arXiv:1512.01882v2.
32. Tian, X.H.; Chng, E.S.; Li, H.Z. A Speaker-Dependent WaveNet for Voice Conversion with Non-Parallel Data. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 201–205.