An Automatic Voice Conversion Evaluation Strategy Based on Perceptual Background Noise Distortion and Speaker Similarity

Dong-Yan Huang[1], Lei Xie[2], Yvonne Siu Wa Lee[1], Jie Wu[2], Huaiping Ming[1], Xiaohai Tian[3] Shaofei Zhang[2], Chuang Ding[2], Mei Li[2], Quy Hy Nguyen[3], Minghui Dong[1], Eng Siong Chng[3], Haizhou LI[1]

[1]Institute for Infocomm Research, A*STAR, Singapore
[2] School of Computer Science, Northwestern Polytechnical University, Xi′an, China
[3] School of Computer Engineering, Nanyang Technological University (NTU),Singapore

{huang, swylee, mhdong, hli}@i2r.a-star.edu.sg
{flxie, jiewu, shaofeizhang, dingchuang, meili}@nwpu.edu.cn
{XHTian, Nguyen.QH, aseschng}@ntu.edu.sg

Abstract

Voice conversion aims to modify the characteristics of one speaker to make it sound like spoken by another speaker without changing the language content. This task has attracted considerable attention and various approaches have been proposed since two decades ago. The evaluation of voice conversion approaches, usually through time-intensive subject listening tests, requires a huge amount of human labor. This paper proposes an automatic voice conversion evaluation strategy based on perceptual background noise distortion and speaker similarity. Experimental results show that our automatic evaluation results match the subjective listening results quite well. We further use our strategy to select best converted samples from multiple voice conversion systems and our submission achieves promising results in the voice conversion challenge (VCC2016).

Index Terms: Voice conversion, objective measures, speech quality assessment, speaker similarity score, subjective listening tests.

1. Introduction

Human voice conveys rich information during communication such as language content and speaker individuality. The speaker individuality is characterized by timber, prosody and linguistic information. Voice conversion technique aims to modify the timber and prosody acoustic features of one speaker to make an impression that it was spoken by another speaker without changing the language content. The applications of voice conversion technique can be found in assistive technology, Text-To-Speech systems, emotion conversion, cross-language speaker conversion, and bandwidth extension for audio [1].

Various voice conversion approaches have been proposed since 1988 such as vector quantization (VQ) [2], Gaussian mixture model (GMM) [3, 4, 5], dynamic kernel partial least squares (PLS) [6, 7], the non-negative matrix factorization [8], as well as artificial neural networks, deep neutral networks, and Bidirectional Long Short-Term Memory (BLSTM) [9, 10, 11].

The evaluation of voice conversion methods, usually through subject listening tests, is time consuming and expensive. The method also requires considerably human efforts in scoring voice quality and the speaker similarity of the converted speech. Therefore, appropriate objective measures need to be developed. Some of objective measures are proposed for evaluating the distortions introduced by speech codecs and/or communication channels [12], by speech enhancement algorithms [13], and by speech synthesizers [14]. In speech coding, the perceptual evaluation of speech quality (PESQ) measure was optimized for speech processed through networks and quantization [12]. The distortions introduced by speech enhancement algorithms and synthesizers affect the speech signal itself and the background noise. The overall quality distortion depends mainly on the speech distortion [13]. Composite objective measures were proposed for the three subjective rating scales (speech distortion, background noise distortion, and overall quality distortion) [13, 14] since any conventional objective measures can not correlate highly with speech/noise distortions and overall quality [13, 14]. All objective measures are perceptual intrusive speech quality measures by measuring the similarity/distance between the original speech and the processed speech. In voice conversion, the converted speech is quite different from the original speech and the conversion function may lead to background noise distortion. Therefore, we investigated the perceptual background noise distortion to measure the voice quality of the converted speech.

The performance of voice conversion depends on not only the voice quality, but also the target speaker similarity of the converted speech. A fast i-vector tool is adopted for scoring the target speaker similarity of the converted speech from different voice conversion systems [15, 16]. In this paper, we propose an automatic voice conversion evaluation strategy to evaluate the performance of voice conversion system. 1) to use the perceptual background distortion noise to identify the ranking of different voice conversion systems; 2) to use a fast i-vector tool [15, 16] to calculate speaker similarity score to select the most target similar sample of the converted speech signals from different VC systems. With this strategy, our submission achieves promising results in the voice conversion challenge (VCC2016).

This paper is organized as follows. We describe briefly our automatic voice conversion evaluation strategy in Section 2; Section 3 presents voice conversion systems including voice conversion scheme, frame alignment, and the several mapping functions used in our system such as dynamic kernel partial least squares (DKPLS) and the postfilter of the modulation spectrum (DKPLS-MS), non-negative matrix factorization-based sparse representation of spectral with residual error compensation using linear spectral frequency (NMF-LSF), deep

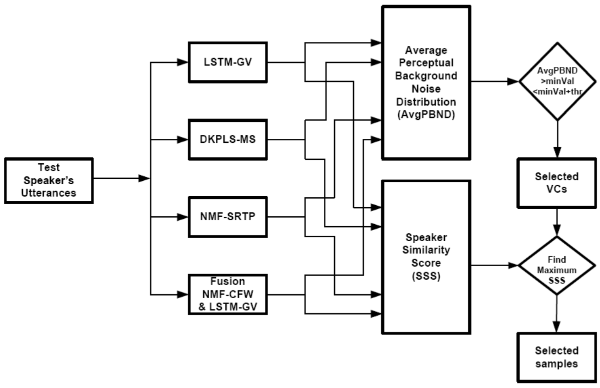

Figure 1: *Block diagram of automatic voice conversion evaluation scheme.*

bidirectional long short-term memory and the global variance (DBLSTM-GV), and fusion of NMF-FW (Frequency Warping) and DBLSTM-GV (Fusion); Section 4 describes the perceptual background distortion noise, and the fast computational i-vector in details ; Section 5 shows the experiments of our automatic evaluation results on development and test data sets in the voice conversion and discussion. Finally, the conclusion of the paper is given in Section 6.

2. System Overview

An automatic voice conversion evaluation scheme is proposed for selecting the best samples from different VC systems shown in Figure 1. The system consists of inputs, different voice conversion systems, average perceptual background noise distortion, speaker similarity score, and selection of the best sample from multiple voice conversion (VC) systems. For each converted speech from different VC systems, we can calculate the average perceptual background noise distortion [13] and speaker similarity score [15]. The system selects the most similar to target voice sample based on speaker similarity score from the VC systems chosen by the average perceptual background noise distortion.

3. Voice Conversion Systems

The voice conversion technique is to build up a relationship of acoustic features between source and target speakers. It can be formulated as a mapping function $\mathcal{F}(\cdot)$ between source speech $\mathbf{S}$ and target speech $\mathbf{Y}$

$$\hat{\mathbf{Y}} = \mathcal{F}(\mathbf{S}) \tag{1}$$

The typical voice conversion scheme composes of training and conversion processes shown in Fig. 2. During training process, the acoustic features related to the speaker identity are extracted from source and target speech signals. Next, each source acoustic feature is mated to the correspondent target feature by frame alignment method, to build a source-target transfer function. Finally, a mapping function is learned from the aligned source-target feature pairs. During conversion process, the mapping function is applied to the acoustic feature extracted from source speech to produce converted feature matrix. Then the converted feature matrix is passed to a synthesizer to reconstruct a speech signal.

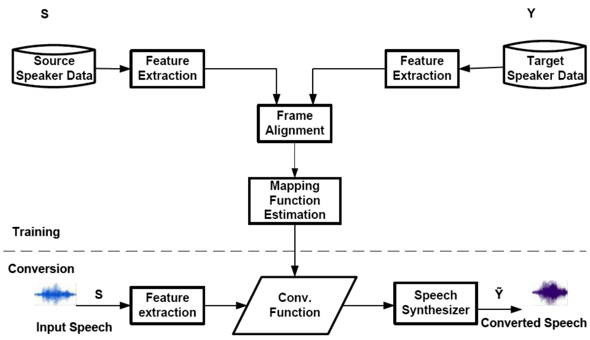

Figure 2: *Block diagram of voice conversion.*

3.1. Speech Model using TANDEM-STRAIGHT

In voice conversion, spectral and prosodic features are usually used for voice conversion. Spectral features represent spectral attributes that relate to voice timbre. Mel-cepstral coefficients (MCCs), linear predictor cepstral coefficients (LPCCs), and line spectral frequency are the popular spectral features to represent the spectral envelope for voice conversion. Prosodic features contain significant information of speaker individuality. Intonation, intensity, and duration are prosodic features. Intonation can be described by fundamental frequencies contour over a longer time, depicts the tones of syllables as well as the accent of a speaker. In this challenge, in our system, the vocoder TANDEM-STRAIGHT is used to extract 513-dimensional spectrum, aperiodicity components and F0. Other features needed in different individual VC systems can be extracted from these basic three acoustic features.

3.2. Frame Alignment

About the alignment between the source and target speaker, the corresponding frames are found by using a two-stage alignment process [17]. As the speech data in Voice Conversion Challenge 2016 is lively and expressive, this two-stage alignment is applicable and bring more accurate alignment over the typical alignment by dynamic time warping (DTW) alone. In the first stage, the speech signals of both the source and target speakers are recognized using a deep neural network (DNN)-based speech recognizer. Only speech signal pairs with identical recognized texts are used. With the phone boundaries in the recognition results, the start and end times of individual phones are known. In the second stage, based on these timing information, individual phone spectral segments of the source speech and the target speech are extracted and aligned by dynamic time warping (DTW). This eventually gives the sets of aligned feature vectors for modeling.

3.3. Mapping Function

Spectral conversion and prosodic mapping transfer the timber and prosodic information, respectively, from source speaker to target speaker. The spectral mapping methods can be divided into three categories: frequency warping, statistical, and unit-selection methods.

In statistical methods, the mapping function between source and target features is built through parametric models. During the conversion, they are taken as the conversion function to transfer source feature into target feature space. the mapping function can be built based on Vector Quantization

(VQ) [2], the Gaussian mixture model (GMM) [3, 4, 18, 5], variant partial least-squares regression [6, 7], artificial neural networks [9, 19, 10, 11], support vector regression [20]. Since the statistical methods learn the central tendency of speech features, it leads to over smoothing effects in converted speech [4, 19]. Frequency warping methods incorporate physical principle into the statistical methods by warping the frequency axis of the spectrum of the source speaker to mate that of the target speaker [21, 22, 23]. Inspired by the idea of unit-selection for speech synthesis, unite-selection methods are proposed that the original target speaker's feature vectors are used to construct converted speech [19, 24].

The prosodic features, such as fundamental frequency, intensity, and duration, are used for prosody mapping. The prosodic feature mapping methods have been developed such as normalization of the mean and variance (MVN) of prosodic features of source-target pairs, GMM-based mapping [25], piecewise linear transformation [24], and higher order polynomial [26]. We integrate the following VC systems in our proposed automatic VC evaluation strategy scheme

3.3.1. Voice Conversion based on DKPLS-MS

In this VC system, we use mel-cepstral coefficients (MCCs) as spectrum feature and apply DKPLS to MCCs for training the mapping function [6]. After obtaining the converted spectrum feature, the modulation spectrum (MS) postprocessing is adopted to improve the converted spectrum feature [5]. The modeling of F0 is to use CWT to decompose F0 into 5 temporal scale $\mathrm{F0}_{\mathrm{cwt}}$, then DKPLS is adopted to build the mapping function of $\mathrm{F0}_{\mathrm{cwt}}$ between source and target. The F0 is reconstructed by inverse CWT.

3.3.2. Voice Conversion based on NMF-LSF

Exemplar-based voice conversion reconstructs a speech spectrogram by a weighted linear combination of high-resolution spectra, called exemplars [8]. The spectrum, aperiodic component, and fundamental frequency (F0) are converted simultaneously [27]. A five-scale continuous wavelets transform (CWT) representation of F0 is used for pitch conversion. The compensation of residual errors uses a 20 order linear spectral frequency to model and source residual errors as excited signal.

3.3.3. Voice Conversion based on DBLSTM and GV

This system mainly uses Deep Bidirectional Long Short-Term Memory based Recurrent Neural Networks (DBLSTM-RNNs) [11] model while considering a global variance (GV) [4] feature of the converted spectra for voice conversion. This system uses neural nets of an average model as initial values of nets, where the average model is trained based on any extra large parallel data. The speech parameters, including mel-cepstral cofficients (MCCs), fundamental frequency (F0) and aperiodic component (AP), are converted separately. The MCCs are converted by the DBLSTM model, and then added the global variance feature. LogF0 is linearly converted, and AP is directly copied from source speech to synthesize the converted speech.

3.3.4. Voice Conversion based on Fusion of NMF-FW and DBLSTM-GV

Each individual method has its own pros and cons. To leverage the merits of these state-of-the-art conversion methods, a system fusion framework [28] is used. DBLSTM-GV and non-negative factorization (NMF)-based frequency warping (FW) [29] are chosen as the candidate systems. The spectral feature is first transformed by these two system separately. Then the system fusion is applied to the converted results. Because different features, spectrum and MCCs, will be used in these two systems, the converted MCCs of DBLSTM-GV will be transformed to spectrogram, then the system fusion will be applied to the converted spectrogram of two methods. As only voiced frames will be transformed in FW-based method, while the unvoiced frames are not modified, the fusion is applied to voiced frames only.

4. Performance Evaluation Metrics

According to our study on subjective listening tests on transformed voices since 2009 [30], there have been more than 200 students between the ages of 16 and 22 to participate different subjective tests. The study shows that people prefer firstly evaluating the voice quality (naturalness) of the modified speech. When natural transformed voice achieves enough good, they then would like to identify the similarity of the transformed voice to the target speech. Based on these studies, we propose an automatic voice conversion evaluation strategy: 1) to select VC systems based on perceptual background noise distortion at a range of pre-defined values; 2) to choose the most probable target speech samples from the selected VC systems based on speaker similarity score. In the following, we present briefly perceptual background noise and speaker similarity score.

4.1. Objective Voice Quality Measures

Several composite objective measures are proposed to evaluate the quality of enhanced speech along three dimensions: signal distortion, noise distortion, and overall quality [13]. They were acquired by linear combination of basic objective measures such as segmental SNR (segSNR) [31], weighted-slope spectral distance [32], and PESQ measure [12]. These objective measures are calculated by comparing extracted features from the enhanced speech and the original speech based on waveform, or spectral, or speech production model parameters, or internal representations of computational models of auditory processing. Inspired from this study, we propose to use perceptual background noise distortion as follows to evaluate the voice quality of the converted speech.

Different from speech enhancement, voice conversion aims to convert spectral of the source speaker speech to that of the target speaker speech. The distortion of the converted speech (speech distortion) is higher if the acoustic characteristics are quite different. It does not mean that the voice quality of the converted speech is poor. It is obvious that this objective measure is not suitable to predict the voice quality of the converted speech. In general, the source speech is available for voice conversion. The objective measure of the perceptual background noise distortion looks like a more appropriate measure for evaluating the voice quality of the converted speech. The larger the perceptual background noise distortion, the poorer the voice quality of the converted speech. The composite objective measure of perceptual background noise distortion is proposed as [13]

$$C = 1.634+0.478*PESQ-0.007*WSS+0.063*segSNR; \quad (2)$$

where segSNR is a time-domain segmental SNR measure for computing the average signal-to-noise of processed signal [31].

The weighted spectral slope (WSS) measure is a per-frame measure in decibels and is estimated as follows [32]

$$WSS(j) = K_{spl}(K - \hat{K}) + \sum_{k=1}^{25} w_a(k)(S(k) - \hat{S}(k))^2 \quad (3)$$

where $K, \hat{K}$ are related to overall sound pressure level of the original and converted speech, and K_{spl} is a parameter which can be varied to increase overall performance.

Perceptual evaluation of speech quality (PESQ) [12] may be the most popularly used objective measurefor speech quality assessment of speech coders and is computed by as follows [12]:

$$PESQ = a_0 + a_1 D_{ind} + a_2 A_{ind} \quad (4)$$

where $a_0 = 4.5, a_1 = -0.1$, and $a_2 = -0.0309$.

4.2. Speaker Similarity Score

The purpose of voice conversion is to make the converted speech sound like the target speech. Another objective measure of voice conversion is to predict the target speaker similarity of the converted speech. We use a speaker verification system based on I-vectors, which formulate the speaker verification problem in the total variability space. Given an utterance, the speaker- and channel-dependent representation of the Gaussian mixture model (GMM) super-vector using joint factor analysis (JFA) is defined as

$$M = m + T\omega \quad (5)$$

where m is the speaker- and channel-independent GMM or universal background model (UBM) supervector, T is a rectangular matrix of low rank total variability, and ω refers to i-vector having a standard normal distribution $N(0, I)$ [33].

Suppose a speech utterance with L frame feature sequence $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_T\}$ and an GMM UBM model γ, with $\gamma_c = \{P_c, \mu_c, \Sigma_c\}, c = 1, \cdots, C$, the i-vector ω can be estimated by computing the zero-order and centered first-order Baum-Welch statistics on the UBM, respectively

$$N_c = \sum_{t=1}^{L} P(c|\mathbf{o}_t, \gamma_c) \quad (6)$$

$$\mathbf{F}_c = \sum_{t=1}^{L} P(c|\mathbf{o}_t, \gamma_c)(\mathbf{o}_t - \mu_c) \quad (7)$$

where $c \in [1, C]$ is the index of GMM component and $P(c|\mathbf{o}_t, \gamma_c)$ is the posterior probability for $\mathbf{o}_t$ on mixture component c. μ_c is the mean of UBM mixture component c. Given a speech utterance μ, the corresponding i-vector can be calculated by the equation

$$\omega = (I + T^t \Sigma^{-1} N(\mu) T)^{-1} T^t \Sigma^{-1} \mathbf{F}_c(\mu) \quad (8)$$

where t is transpose operation.

In this total variability space, the length normalization method [34] is applied to improve the performance of i-vector. After length normalization, cosine distance scoring is used for i-vector modeling. The cosine kernel between two i-vectors ω_{target} and $\omega_{conv.}$ is defined as follows:

$$score = \frac{\langle \omega_{target}, \omega_{conv.} \rangle}{||\omega_{target}|| ||\omega_{conv.}||} \quad (9)$$

5. Experiments

5.1. VCC 2016 Challenge

The task of the challenge is to develop voice conversion systems with the 162 parallel utterance pairs in each pair as training data. For all pairs of the source and target speakers, there are 25 conversion systems to be developed in total (i.e., 5 sources by 5 targets). For conversion stage, each participant team is required to convert 54 utterance samples of each source speaker' voice into individual target speaker's voice with the developed 25 conversion systems. The total 1,350 converted voice samples (54 utterances times 25 speaker pairs) will be generated.

Our experiments are based on training data set and testing data set, respectively. In order to develop 25 voice mapping functions for all pairs, the 168 training sentences are divided into 138 utterances as sub-training set and randomly selected 30 utterances as development data set. For testing data set, the total 168 training data are used for estimated 25 voice mapping functions, the 54 sentences of each source speaker are taken as testing data to generate converted samples.

5.2. Systems Setting

In our proposed automatic voice conversion evaluation scheme, TANDEM-STRAIGHT was used to extract 513-dimensional spectrum, aperiodicity components and F0. The acoustic features, such as MCCs, energy, and duration, can be extracted from these basic acoustic features.

The frame alignment information was obtained by performing two-stage alignment [17]. In the first stage, the speech signals of both the source and target speakers are recognized using a deep neural network (DNN)-based speech recognizer. Only speech signal pairs with identical recognized texts are used. With the phone boundaries in the recognition results, the start and end times of individual phones are known. In the second stage, based on these timing information, individual phone spectral segments of the source speech and the target speech are extracted. The corresponding 39-dimension MFCCs are extracted from the spectrum. The frame alignment information was obtained by performing DTW on the MFCC feature sequence for all the VC systems in this paper.

The four voice conversion systems, such as DKPLS-MS, NMF-SRSP, DBLSTM-GV, and fusion of NMF-CFW and DBLSTM-GV, are incorporated in to the proposed evaluation scheme. They are summarized as follows:

- **DKPLS-MS:** The dynamic kernel are formed using 3 adjacent frames. The number of the latent PLS is set to 5 and the scaling term for the dynamic kernel is set to 10. The continuous wavelet transform (CWT) is used to decompose $F0$ into 5 temporal scale representations. DKPLS is adopted to establish mapping function of $F0_{cwt}$ between source and target. An inverse CWT is used to reconstruct the $F0$.
- **NMF-LSF:** This system converts spectrum, aperiodicity components, energy contour and $F0_{cwt}$ jointly under the exemplar-based voice conversion framework. The residual error compensation is carried out by a 24 order LPC and source residual errors of LPC are taken as excited signal. Finally, a duration conversion is conducted on all the converted features.
- **DBLSTM-GV:** This system uses Deep Bidirectional Long Short-Term Memory based Recurrent Neural Networks (DBLSTM-RNNs) [11] model while consider-

ing a global variance (GV) [4] feature of the converted spectra for voice conversion. In order to obtain optimal solution with limited data set, an extra large parallel database is trained on the trained DBLSTM-RNNs to obtain initial values. The VCC 2016 training data aligned by alignment-index-identical-phones-only. The acoustic signals are sampled at 16kHz with mono channel, windowed by 25ms. The frame shift is 5ms. 49-dimensional Mel-Cepstral Coefficient (MCC) is used in these two systems. Both in two systems, the number of units in each layer is [49 96 128 96 49] respectively. We train the networks using a C++ CUDA-enabled machine learning library named RECURRENT with the learning rate of $1.0 * 10^{-5}$ and a momentum of 0.9.

- **Fusion of NMF-CFW and DBLSTM-GV:** The DBLSTM-based approach and sparse representation based FW [29] are used to convert the spectral feature, respectively. Then the system fusion is applied on the converted samples from both systems, only on the voiced part.

For evaluating perceptual background noise distortion of individual VC systems, the objective measures PESQ, WSS, and segSNR need to be estimated. PESQ is estimated according to ITU-T recommendation P.862 [12] using the equation (4). WSS is estimated using Eq. (3), the maximum value and local maximum value of K and $\hat{K}$ are 20 and 1 dB, respectively. The frames with segmental SNR in the range of -10 to 35 dB were considered in the average.

For speaker similarity score, the acoustic features used in the experiments consists of 18-dimensional mel frequency cepstral coefficients (MFCC). Delta and delta-delta features were appended giving rise to 54-dimensional feature vector. We used gender-independent UBM consisting of 256 mixtures with full covariance matrices. The total variability matrix T was trained using TIMIT data. The rank of the matrix T = 200.

5.3. Result Comparisons

In order to show the effectiveness of proposed automatic evaluation strategy for voice conversion based on perceptual background noise distortion and speaker similarity score, we perform the analysis of the matching between the objective measures and subjective listening scores on the training data set. Then we further apply our strategy to select best converted samples from multiple voice conversion systems on the testing data set and report our submission results in the voice conversion challenge 2016.

5.3.1. Training Data Set

We split the 168 training utterances into 138 utterances for training and 30 utterances for testing, which are randomly selected from 168 training utterances. Due to the rich expression and variation of speaking style of speakers, the four source-target pairs are selected for subjective listening tests. The subjective voice quality and similarity of the converted speech use a five-point scale and a four-point scale given by VCC 2016 organizers.

A five-point scale of noise distortion introduced by different voice conversion: 1) completely unnatural; 2) mostly unnatural; 3) equally natural and unnatural; 4) mostly natural; 5) completely natural. The higher the score, the better the voice quality.

Table 1: Subjective listening tests on the voice quality of converted speech

	SF1-TF1	SF1-TM3	SM2-TF1	SM2-TM3	Avg
DKPLS	2.84	2.92	2.63	3.37	2.94
NMF	3.11	2.50	2.51	3.18	2.83
BLSTM	3.34	3.18	3.00	3.51	3.26
Fusion	3.35	2.86	2.46	3.62	3.07
AVCES	3.35	3.21	2.93	3.43	3.23

Table 2: Objective perceptual background noise distortion of converted speech

	SF1-TF1	SF1-TM3	SM2-TF1	SM2-TM3	Avg
DKPLS	0.89	0.62	0.31	1.54	0.84
NMF	1.79	0.52	0.24	1.38	0.98
BLSTM	1.29	0.47	0.03	1.38	0.80
Fusion	1.47	0.57	0.32	1.46	0.96
AVCES	0.89	0.47	0.03	1.38	0.69

Table 3: Subjective listening tests on the target speaker similarity of converted speech

	SF1-TF1	SF1-TM3	SM2-TF1	SM2-TM3	Avg
DKPLS	2.55	2.57	2.78	2.44	2.59
NMF	2.68	2.79	2.84	2.71	2.76
BLSTM	2.62	2.60	2.85	2.58	2.66
Fusion	2.67	2.73	2.90	2.62	2.73
AVCES	2.63	2.61	2.76	2.58	2.65

The target speaker similarity score of the converted speech uses a four-point scale of similarity given by different voice conversion: 1) same, absolutely sure; 2) same, not sure; 3) different, not sure; 4) different, absolutely sure. The lower the score, the higher the speaker similarity.

The subjective listening tests are conducted to evaluate the performance of different voice conversion methods such as DKPLS-MS (DKPLS), NMF-LSF (NMF), DBLSTM-GV (BLSTM), Fusion of NMF-CFW-DBLSTM-GV (Fusion), and proposed fusion method - automatic voice conversion evaluation strategy (AVCES).

Table 1 shows that the subjective voice quality results of converted samples from different voice conversion systems. Although the subjective perceptual score of BLSTM (DBLSTM-GV)-based VC method is slightly better than our proposed automatic voice conversion evaluation method, the objective perceptual background noise distortion can be taken as a fast objective measure for evaluating voice quality of voice conversion systems and can be used to differentiate the performance of different voice conversion systems. Table 2 shows the results of objective perceptual background noise distortion of different voice systems. The results selected by this objective measure match well the subjective voice quality results.

Table 3 show the results of subjective speaker similarity of different VC systems - the lower the score, the more similar the converted speech. We observe that the similarity of converted speech by DKPLS-MS-based is the best of the four VC systems [15].

Table 4 shows the results of objective speaker similarity score of different VC systems and our proposed method - the higher the similarity ratio, the more similar the converted speech. The speaker verification system can distinguish the performance of similarity of different VC systems [15]. However, the accuracy of speaker similarity ratio of the system needs to be further improved.

Table 4: Objective speaker similarity scores of converted speech (%) for different VC systems

	SF1-TF1	SF1-TM3	SM2-TF1	SM2-TM3	Avg
DKPLS	24.4	21.8	24.1	23.0	23.3
NMF	26.0	19.5	23.5	17.4	21.6
BLSTM	27.8	21.3	24.0	20.0	23.2
Fusion	24.3	22.3	23.2	23.0	23.0
AVCES	24.4	21.3	24.0	17.4	23.8

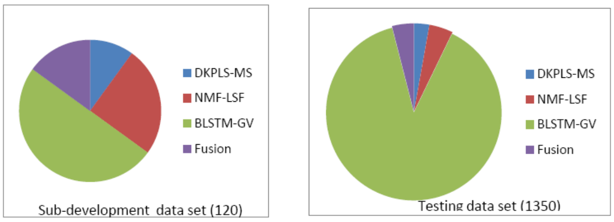

Figure 3: *The results selected from four VC systems by AVCES for sub-development set (left) and the testing data set (right)*

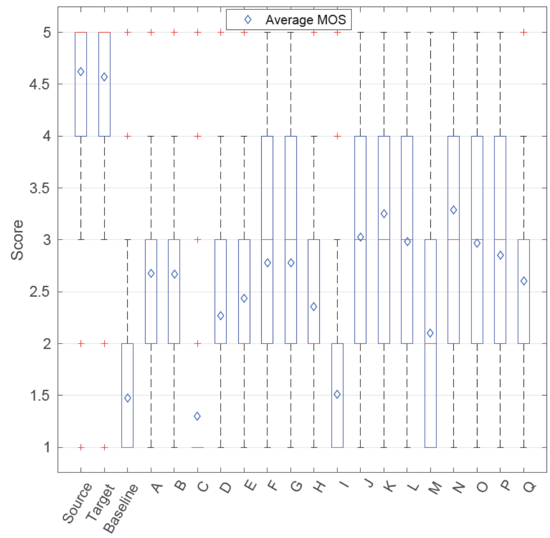

Figure 4: *The VCC 2016 MOS results*

5.3.2. Testing Data Set

For testing data set, the total 168 training data are used for estimated 25 voice mapping functions, the 54 sentences of each source speaker are taken as testing data to generate converted samples.

The pie chart shows the percentage of audio samples selected from the different VC systems for the sub-development data set (left) and for the testing data set (right), respectively in Fig. 3. The selected results are submitted to VCC 2016 organizers. Figures 4 and 5 are the results of MOS and similarity tests in VCC 2016, respectively. In the figures, the results of our system correspond to the letter L. First, for naturalness, the results of our system are moderate according to Figure 4 (MOS: 2.98). Compared the performance of our system with that of baseline, our system achieved better performance than the baseline system (similarity 65.50 %).

We take the naturalness score and similarity score as two dimensions and plot in Fig. 6. The proposed method achieves a

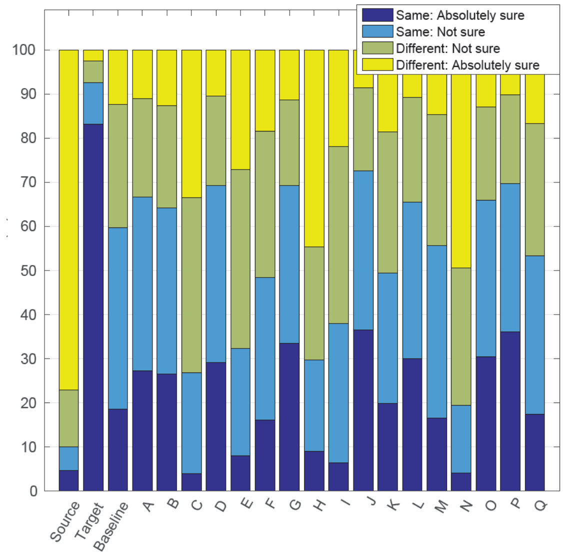

Figure 5: *The VCC 2016 similarity results*

promising result in the voice conversion challenge 2016.

6. Conclusion

In the voice conversion challenge 2016, we have developed different voice conversion systems. However, the most accurate evaluation of voice conversion approaches, usually through the subject listening tests, is time-consuming and expensive. In order to select optimal samples from different VC systems, we proposed an automatic voice conversion evaluation strategy based on perceptual background noise distortion and speaker similarity ratio. The simulation results showed that the proposed automatic evaluation method is able to give reliable ranking of the different voice conversion systems, which fits well the subjective listening results. That leads to promising results in the voice conversion challenge (VCC2016).

7. References

[1] J. Nurminen, H. Silen, and V. Popa, "Voice conversion," *Speech Enhancement, Modeling and Recognition-Algorithms and Applications*, pp. 69–94, 2012.

[2] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *International Conference on Acoustics, Speech, and Signal Processing, 1988*, Apr 1988, pp. 655–658 vol.1.

[3] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, Mar 1998.

[4] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2222–2235, 2007.

[5] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura, "Modulation spectrum-constrained trajectory training algorithm for gmm-based voice conversion," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, 2015, pp. 4859–4863.

[6] E. Helander, H. Silen, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression,"

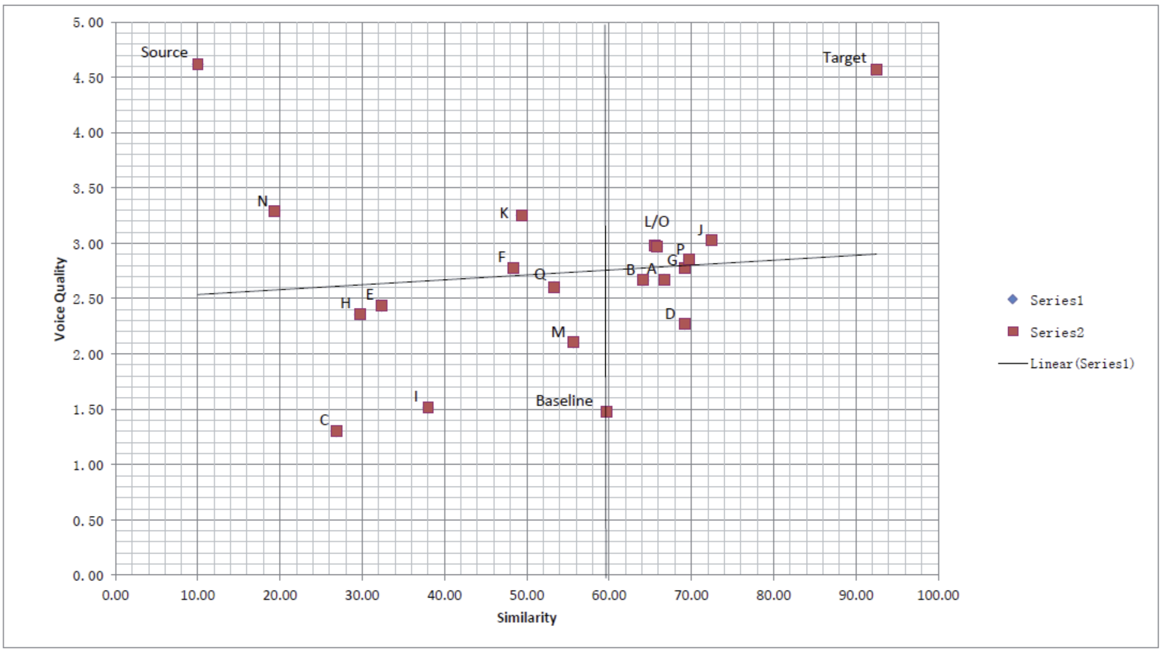

Figure 6: *The VCC 2016 final results.*

IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 3, pp. 806–817, March 2012.

[7] D.-Y. Huang, M. Dong, and H. Li, "A dynamic gaussian process for voice conversion," in *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on*, July 2013, pp. 1–4.

[8] Z. Wu, T. Virtanen, E. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 22, pp. 1506–1521, 2014.

[9] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 3893–3896.

[10] L. H. Chen, Z. H. Ling, L. J. Liu, and L. R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1859–1872, Dec 2014.

[11] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 4869–4873.

[12] "Perceptual evaluation of speech quality (pesq), an objective method for end-to-end speech quality assessment of 3.1 khz handset telephony (narrow-band) networks and speech codecs," nternational Telecommunications Union, Geneva, Switzerland, 2001.

[13] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan 2008.

[14] D.-Y. Huang, "Prediction of perceived sound quality of synthetic speech," in *In: Proc. APSIPA ASC.*, Xi'an, China, 2011.

[15] L. Xu, K. A. Lee, H. Li, and Z. Yang, "Rapid computation of i-vector," in *Proc. Odyssey*, 2016.

[16] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, p. 788798, 2011.

[17] S. W. Lee, Z. Wu, M. Dong, X. Tian, and H. Li, "A comparative study of spectral transformation techniques for singing voice synthesis," in *Proc. Interspeech*, 2014, pp. 2499–2503.

[18] H. Benisty, D. Malah, and K. Crammer, "Modular global variance enhancement for voice conversion systems," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, Aug 2012, pp. 370–374.

[19] F. Zhu, Z. Fan, and X. Wu, "Voice conversion using conditional restricted boltzmann machine," in *Signal and Information Processing (ChinaSIP), 2014 IEEE China Summit International Conference on*, July 2014, pp. 110–114.

[20] P. Song, Y. Q. Bao, L. Zhao, and C. R. Zou, "Voice conversion using support vector regression," *Electronics Letters*, vol. 47, no. 18, pp. 1045–1046, September 2011.

[21] H. Mizuno and M. Abe, "Voice conversion based on piecewise linear conversion rules of formant frequency and spectrum tilt," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, vol. i, Apr 1994, pp. I/469–I/472 vol.1.

[22] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 922–931, July 2010.

[23] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1313–1323, May 2012.

[24] D. T. Chappell and J. H. L. Hansen, "Speaker-specific pitch contour modeling and modification," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2, May 1998, pp. 885–888 vol.2.

[25] Z. Inanoglu, "Transforming pitch in a voice conversion framework," 2003.

[26] B. Gillett and S. King, "Transforming F0 contours," in *8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland, September 1-4, 2003*, 2003.

[27] H. Ming, D.-Y. Huang, L. Xie, S. Zhang, M. Dong, and H. Li, "Exemplar-based sparse representation of timbre and prosody for voice conversion," 2016.

[28] X. Tian, Z. Wu, S. W. Lee, N. Q. Hy, M. Dong, and E. Chng, "System fusion for high-performance voice conversion," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015, pp. 2759–2763.

[29] X. Tian, Z. Wu, S. W. Lee, N. Q. Hy, E. Chng, and M. Dong, "Sparse representation for frequency warping based voice conversion." in *ICASSP*. IEEE, 2015, pp. 4235–4239.

[30] D.-Y. Huang, E. P. Ong, S. Rahardja, M. Dong, and H. Li, "Transformation of vocal characteristics: A review of literature," *International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering*, vol. 3, no. 12, pp. 77 – 85, 2009.

[31] J. Hansen and B. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. ICSLP*, Sydney, Australia, 1998, pp. 81–84.

[32] D. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: A first step," in *Proc. IEEE ICASSP*, 1982, pp. 1278–1281.

[33] N. Dehak, "Discriminative and generative approches for long- and short-term speaker characteristics modeling: Application to speaker verification," Ph.D. dissertation, École de Technologie Supérieure, Montreal, 2009.

[34] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, 2011, pp. 249–252.