Abstract: Recently, voice conversion has becoming the research hotspot, because of its widely application areas. However, the voice conversion technology is still immature. By the researching of existing voice conversion models, the voice conversion system based on the RBF neural network was designed, and the system simulation was implemented. During conversion, the unvoiced speech was excluded and the voiced speech was reserved. The LPC was the extracted from the source and target speech, then convert the LPC to LSP. The LPS was trained by RBF neural network after time-aligned. Obtained mapping function was used to convert the source LSP to target LSP, and synthesis the speech. Finally, the converted speech evaluated by ABX and MOS to test the tendency and quality of the speech.

Keywords: Voice conversion, RBF, Neutral network

1. Introduction

Voice conversion technology can be used to modifying one speaker’s voice (source) to sound like another speaker (target) without changing the language content. The voice conversion is widely used in personal artificial speaking devices, PC gamer and music, voice editing for films, repair of pathological voices. Voice conversion is usually realized by feature mapping, that the features can describe the speaker
individuality in the form of parameter vectors [1]. Recently, codebook mapping [2], Hidden Markov Model (HMM) [3] and Gaussian mixture model (GMM) [4] are be used to deal with the voice conversion. Due to the nonlinear and continuous feature of neural networks, they are one of the most promising tool to deal with voice conversion [5]. However, there isn’t any fully developed products in the market, because the voice conversion technology is immature, that is still in theoretical stage instead of application stage [6].

2. Related works

Voice conversion can be divided into 3 parts, extraction of source and target characteristic parameters, establish the conversion rules between source and target characteristic parameters, synthesis the speech with the rules. Review the literatures, in order to extract the speaker characteristic parameters, many speech feature representations have been developed, such as Linear Predictive Coefficients (LPC), Line Spectral Frequency (LSF), Mel Frequency Centrum Coefficient (MFCC), Linear Spectrum Pair (LSP), Formant Frequencies (FF), Licensed Professional Clinical Counselor (LPCC) and so on [7]. Nowadays, The LSP has been successfully used as the spectral feature in voice conversion. Due to several properties of LSP, which improved speech quality than other features [8].

3 Voice conversion system

3.1 The framework of voice conversion system

This article is focused on the conversion of the vocal tract spectral feature LSP, the block diagram of voice conversion is shown as Fig.1.
Fig.1 The block diagram of voice conversion system

From the Fig. 1, there is serval steps to conversi on the voice from the source to the target:
(1) Voice activity detection is used as the pretreatment of the voice conversion system;
(2) Extract the LPC of the source and target speech, and convert the LPC to LSP;
(3) Derive the time-aligned LSP from the parallel utterances of the source and target speakers;
(4) Use the RBF neural network to train the LSP and obtain the conversion rules;
(5) Extract the LPC of the source speech, and convert the LPC to LSP;
(6) Convert the source LSP to the target LSP based on the conversion rules;
(7) Convert the LSP to LPC, synthesis the speech with the converted LPC.

3.2 Pretreatment

Voice activity detection is necessary in almost speech signal processing, such as speech encoding, speech recognition, hands-free telephony and echo cancellation [9]. Voice activity detection can increase the treatment efficiency, and extract the effective voice clips form the whole the speech signal, without the interference of other unvoiced signal. The application of the voice activity detection can highly relive the subsequent treatment calculation pressure, improve the respond speed.

Short time energy is one of the ways to detect the voice activity, it can be used to calculate each frame of the speech signal. It is defined as Eq.1:

\[ E = \sum_{n=1}^{N} x^2(n) \]  

In which, E is the energy of the single frame, N is the frame length, x(n) is the amplitude values at point n.

Detection method: set a low threshold \( H1 \) and a high threshold \( H2 \), compare the short time energy with \( H1 \) and \( H2 \), if it bigger than \( H1 \) and bigger than \( H2 \) in a period of time with several frames, it can be considered as the start of the speech signal, it can be also used to detect the end of the speech signal, the voice activity detection of the speech signal is shown as Fig.2.

Fig.2 voice activity detection of the speech signal
3.3 Extraction of spectral parameters

LSP is one of the characteristic parameters of the speech signal spectral envelope, it is closely to the formant. It is widely used in the voice conversion, because of the quantization characteristics and interpolation characteristic [10].

The LSP is calculate by complex conjugate of dimensionality $p+1$ symmetric and antisymmetric polynomial. According to the linear prediction inverse filter $A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i}$, the dimensionality $p+1$ symmetric and antisymmetric polynomial is defined as Eq.2:

$$
P(z) = A(z) + z^{-(p+1)} A(z^{-1})$$
$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1})$$

$A(z)$ can be expressed as Eq.3:

$$A(z) = \frac{1}{2} [P(z) + Q(z)]$$

The roof of $A(z)$ are interleaved on the unit circle, the roof of $P(z)$ and $Q(z)$ are also alternately interleaved on the unit circle. Set $e^{\pm j\omega_i}$ as the zero point of $P(z)$, $e^{\pm j\theta_i}$ as the zero point of $Q(z)$, $P(z)$ and $Q(z)$ can expressed as Eq.4:

$$P(z) = (1 + z^{-1}) \prod_{t=1}^{p/2} (1 - 2 \cos \omega_i z^{-1} + z^{-2})$$

$$Q(z) = (1 + z^{-1}) \prod_{t=1}^{p/2} (1 - 2 \cos \theta_i z^{-1} + z^{-2})$$

The $\omega_i$ is LSP, the $\theta_i$ is the LSF. The continuous 16 frames LSP and LSF are shown as Fig.3.
3.4 Dynamic time warping

Dynamic Time Warping (DTW) is mainly used for alignment of voice feature parameters in time domain, DTW allows time-flexible alignment, it warps the feature parameters with measures the distance between voice parameters [11].

Even the same person speaks the same words in different time, the various features are different. DTW algorithm corresponds the phonemes of resource speaker and the target speaker, aligns the extracted parameters, so that the obtained conversion functions will be more accurate after training.

Set feature parameters of source speaker $S = [s_1, s_2, s_{m_1}]$, feature parameters of target speaker $T = [t_1, t_2, t_{n_1}], m \neq n$. It is defined as Eq.5:

$$D = \min_f \sum_{i=1}^{m_1} d(s(i), t(f))$$

The $f$ is time warping function, $d(s(i), t(f))$ is the distance between source and target parameters. When $D$ is minimum, the $f$ is the most suitable function. The time-flexible alignment generated by DTW is shown as Fig.4[12].

![Fig.4 The time-flexible alignment generated by DTW](image)

3.5 Radial basis function network training

The Radial basis function (RBF) is the class of neural networks, which capture a globally nonlinear mapping function in the context of voice conversion system [13].

The RBF network is composed of input layer, hidden layer and output layer. The input layer is responsible for the input of the training data, the hidden layer is responsible for non-linear transformation with hidden unit, the output layer is responsible for output the response to input signal. The structure of RBF neural network is shown as Fig.5[6].
Set the neuron nodes of 3 layer are respectively m, h and n.: $x = [x_1,x_2,...x_m]^T$ is the input vector; $z = [z_1,z_2,...z_n]^T$ is the expected output response; $b = [b_1,b_2,...b_n]^T \in \mathbb{R}$ is the threshold; $y = [y_1,y_2,...y_n]^T$ is the output value of network; $\Phi_1(\cdot)$ is the activation function of ith hidden node; $c_i = [c_{i1}, c_{i2},...,c_{im}]^T \in \mathbb{R}$ is the data center of the ith hidden node.

The activation function $\Phi_1(\cdot)$ of neural network can be the following forms(Eq.6~Eq.9):

1. Gaussian function
   $$\Phi(x) = \exp\left(-\frac{x^2}{\sigma^2}\right)$$  \hspace{1cm} (6)

2. Sigmoid function
   $$\Phi(x) = \frac{1}{1+\exp\left(\frac{x}{\sigma}\right)}$$  \hspace{1cm} (7)

3. Multiquadric function
   $$\Phi(x) = \sqrt{x^2 + \nu^2}$$  \hspace{1cm} (8)

4. Inverse multiquadric function
   $$\Phi(x) = \frac{1}{\sqrt{x^2 + \nu^2}}$$  \hspace{1cm} (9)

The jth output of RBF network can expressed as Eq.10:
$$y_j = \sum_{i=1}^{h} \omega_{ij} \Phi_i(\|x-c_i\|) + b_j, (j=1,2,...n)$$  \hspace{1cm} (10)

Where $\omega_{ij}$ is the connection weights of ith hidden node and jth output node, $b_j$ is the threshold of jth output node.
3.6 The LSP conversion based on RBF

During the training stage, the LSP of source speech and target speech are extracted, and the LSP aligned by DTW. Set the aligned source LSP as the input of the RBF network, the aligned target LSP as the output of the RBF network. The mapping function between source and target LSP obtained after the finished of training stage. The converted LSP and target LSP are shown as Fig.6.

![Fig.6 Converted LSP and target LSP](image)

4 Experiment and analysis

The experiment is based on Maltah, the corpus is consisting of phoneme-balanced sentences by 1 male and 1 female speakers. The converted voice is shown as Fig.7.

![Fig.7 Converted voice](image)
The aim of voice conversion is to make the converted voice sound like the target voice, so the evaluation of the voice conversion should base on the human hearing. The ABX and MOS can be used to evaluate the performance of speech quality and the speaker individuality [14].

4.1 ABX

In ABX, A represents the source speaker, B represents the target speaker, X represents transformed speech signal, also conducted using the same set of utterances and speakers. During the test, the listeners are asked to judge the X is closer to A or closer to B. ABX can evaluate the tendency of converted speeches. The results of the ABX test is shown in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>None of both</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female to male</td>
<td>27.3</td>
<td>41.4</td>
<td>31.6</td>
</tr>
<tr>
<td>Male to female</td>
<td>40.5</td>
<td>20.1</td>
<td>38.4</td>
</tr>
</tbody>
</table>

4.2 MOS

The testers are asked to judge the quality of voice in the scale of 1 to 5. The rating 5 represents that it is the excellent match between the transformed and target utterance, and rating 1 represents that it is the poor match between the transformed and target utterance. The other rating represents different levels of variation between 1 and 5. The obtained results is shown in Table 2.

<table>
<thead>
<tr>
<th></th>
<th>MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female to male</td>
<td>2.1</td>
</tr>
<tr>
<td>Male to female</td>
<td>1.6</td>
</tr>
</tbody>
</table>

5. Conclusions and future works

The ABX results show that, when converse female voice to male voice, the converted speech is closer to the target speech, and when converse female voice to female voice, the converted speech is closer to the source speech. The MOS result show that, the quality of the converted speech is poor, and should be increased in the future.

The RBF neural network was used for mapping the vocal tract characteristics of the source speaker from the target speaker. The LSP converted from the LPC are used to
represent the vocal tract characteristics. The converted speech is synthesized by the parameters, which mapping from the trained function.

Voice conversion has great value, but it still several problems, and can’t be commercialize. Researching for the new speech conversion model is very important for the quality of the speech. The new model can obtain the new and better feature according to the mechanism of utterance. The new feature parameter should represent the characterize of speech and tendency of voice personality.

Acknowledge

This research has been partially supported by JSPS KAKENHI Grant Number 15H01712.

Reference:

Xun Feng

He received the bachelor’s degree in 2014 from Nantong University, China. He is currently a master at the university of Tokushima. His research interests include voice signal treatment. He is a member of IEEE.

Hongjun Ni

He received the Ph. D degree from Shanghai Jiao Tong University in 2003, associate Professor of Nantong University. He was the visiting scholar of University of Toronto, Newcastle University, University of East Anglia, and University of Nevada. He has published 30 papers in SCI/EI.

Fuji Ren

Fuji Ren received his Ph. D degree in 1991 from the Faculty of Engineering, Hokkaido University, Sapporo, Japan. From 1991 to1994, he worked at CSK as a chief researcher. In 1994, he joined the Faculty of Information Sciences, Hiroshima City University, as an Associate Professor. Since 2001, he has been a Professor of the Faculty of Engineering, Tokushima University. His current research interests include Natural Language Processing, Artificial Intelligence, Affective Computing, Emotional Robot. He is a senior member of IEEE, a member of AAMT, IPSJ, IEICE, CAAI, IEEJ, Editor-in-Chief of International Journal of Advanced Intelligence, a vice president of CAAI, and a Fellow of The Japan Federation of Engineering Societies.