Voice Conversion based on Joint Pitch and Spectral Transformation with Component Group-GMM

Jianchun MA, Wenju LIU

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
95. East of Zhong Guan Cun Road, Haidian District, Beijing, 100080, Mailbox 2728
Email: jcma@nlpr.ia.ac.cn, lwj@nlpr.ia.ac.cn

Abstract- Spectral and pitch are two most important features in voice conversion which including a majority of speaker identity information. Some researchers use the GMM (Gaussian Mixture Model) to model the joint spectral and pitch. But these two features have the discrepancy of unit and meaning, so should do some processing before training the model. In this paper, a new framework CG-GMM (Component-Group GMM) is used for the joint pitch and spectral transformation. Experiments are setup and compared with the previous approach of voice conversion. The converted speeches indicate satisfactory speech quality and speaker identifiability. Meanwhile the speaking style is much like to the target speaker.

I. INTRODUCTION

Voice conversion is a technology which modify a speaker (source speaker) voice to be perceived as if another speaker (target speaker) had uttered it [1]. Various applications are available that range from TTS (Text-To-Speech) customization, building a multi-speaker database from one-speaker corpus, education tools for foreign language learning, medical aids to help the people with speech impairments.

In voice conversion, spectral and pitch are two most important features which including a majority of speaker identity information. Current most voice conversion focus on transforming spectral features, just do a simple linear modification on pitch. There are many ways to transform the spectral features, such as mapping codebooks [2], discrete transformation function [3], artificial neural networks [4], GMM (Gaussian Mixture Model) [5]. The statistical method of GMM has been shown efficient and robust than other methods above. However only transforming the spectral is not sufficient for correctly depict the target speaker's voice. The pitch which describe the prosody feature related with the speaking style is also very important to speaker's identity. So precise pitch transformation is needed.

Conventional methods use simple linear conversion with mean values of pitch to the pitch conversion. Moreover, we can also apply the GMM-based conversion to the pitch conversion. These methods have poor performance which ignoring the correlation between pitch and spectrum. Research has been proved that some spectral characteristics of glottal waveform are dependent on the pitch [6]. Taoufik En-Najjary [7] provided a method transforming joint pitch and spectral, using GMM to train the joint feature mapping

function. Problems come into being, these two features have different unit and contribution for the model. Face the first problem, the element whose magnitude is bigger will predominate the effect of clustering, the authors did a log scaled normalization to the pitch values. Also finding the relative importance of spectral and pitch information was not discussed.

In this paper, we use a new framework Component-Group GMM [8] to model the jointed spectral and pitch. By splitting of elements of feature vector (according to different unit and meaning) into element-group and mixture components into component-group, the raw data of one element-group are used to train model parameters of corresponding component-group, which insulate the direct interaction of feature element-group (with different unit), and mixing coefficients are optimized among all component-group globally, which introduces the competition among feature element-group (to represent different contribution). ML (Maximum Likelihood) estimation is used to estimate the parameters of CG-GMM, and EM (Expectation-Maximum) algorithm is applied to optimize the ML estimates of the model parameters iteratively.

The rest of paper is organized as follows. The next section gives the outline of new voice conversion method, while the section III represents the experiments and evaluation results.

II. STRAIGHT AND CG-GMM

2.1 STRAIGHT

STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) is a high quality analysis-synthesis framework which is proposed by Kawahare et al. [9]. It is a channel VOCODER based on advanced F0 adaptive procedures. The procedures are grouped into three subsystems; a source information extractor, a smoothed time-frequency spectrum representation extractor, and a synthesizing engine consisting of an excitation source and a time varying filter.

2.2 Component-Group GMM [8]

At first, we introduce the familiar statistical Gaussian Mixture Model. The GMM has been used in many pattern recognition techniques, such as text-independent speaker

recognition which gives good performance. The GMM assume that the probability distribution of the observed feature vector can be represented as M component density functions weighted by M mixing coefficients.

$$p(x | \lambda) = \sum_{i=1}^M \alpha_i p_i(x) = \sum_{i=1}^M \alpha_i N(x; u_i, \Sigma_i) \quad (1)$$

where $p_i(x)$ denote the i 'th D-dimensional normal distribution with mean vector u_i and covariance matrix Σ_i defined by

$$p_i(x) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x-u_i)' \Sigma_i^{-1} (x-u_i)\right\} \quad i=1, \dots, M \quad (2)$$

In GMM, all component density function has the same variate number D. GMM can be seen as a particular CG-GMM, it is a one element-group and one component-group CG-GMM. We describe the model as follows: Elements of an original D-dimensional feature vector X can be splitted into P element-groups (according to different meaning or unit), such as $X = [x_1, x_2, \dots, x_p]$. The dimensions of the element-groups $x_1, x_2, \dots, x_p$ are $d_1, d_2, \dots, d_p$ which are not

necessarily the same but must satisfy $\sum_{i=1}^p d_i = D$. Meanwhile,

M components of GMM for the features are also splitted into P groups. One component-group is corresponding for one element-group. And the amounts of components included in these component-groups are $m_1, m_2, \dots, m_p$, which satisfy

$\sum_{i=1}^p m_i = M$. Thus we obtain the Component-Group GMM's mixture density function as:

$$p(X | \lambda) = \sum_{i=1}^p \sum_{j=1}^{m_i} \alpha_{ij} p_{ij}(x_i) \quad (3)$$

The variate number of a component density $p_{ij}(x_i)$ is dependent on which component-group (the i 'th group) it belongs to and equal to the dimension (d_i) of the corresponding element-group (x_i).

$$p_{ij}(x_i) = \frac{1}{(2\pi)^{d_i/2} |\Sigma_{ij}|^{1/2}} \exp\left\{-\frac{1}{2}(x_i - u_{ij})' \Sigma_{ij}^{-1} (x_i - u_{ij})\right\} \quad i=1, \dots, p \quad j=1, \dots, m_i \quad (4)$$

2.3 Initialise for CG-GMM

K-means clustering is used to initialize the mean vector, variance matrix, and mixing coefficient. It is different from GMM, the element-group of feature vector are clustered independently for the corresponding component-group, so we need not care of the elements having different unit.

2.4 ML estimation using EM algorithm

Like the GMM, we use the ML estimation to estimate the parameter of the CG-GMM, and EM algorithm is applied to fit the mixture model parameter iteratively.

Assume that the model $\lambda = \{\alpha_{ij}, u_{ij}, \Sigma_{ij}\}$ will be trained with a T-frame feature vector sequence $X = \{x^{(t)} | t=1, \dots, T\}$. Then the incomplete-data log-likelihood expression is given by:

$$\begin{aligned} \log(L(\lambda | X)) &= \log \prod_{t=1}^T p(x^{(t)} | \lambda) \\ &= \sum_{t=1}^T \log \left(\sum_{i=1}^p \sum_{j=1}^{m_i} \alpha_{ij} p_{ij}(x_i^{(t)}) \right) \end{aligned} \quad (5)$$

which is maximized due to the definition of the maximum-likelihood estimation. Using EM algorithm, we can get the re-estimation of the parameters as follows(the detail deduction please see [8]):

$$\begin{aligned} \hat{\alpha}_{ij} &= \frac{1}{T} \sum_{t=1}^T \gamma_{ij}^{(t)} = \frac{1}{T} \gamma_{ij} \\ \hat{u}_{ij} &= \sum_{t=1}^T \frac{\gamma_{ij}^{(t)}}{\gamma_{ij}} x_i^{(t)} \\ \hat{\Sigma}_{ij} &= \sum_{t=1}^T \frac{\gamma_{ij}^{(t)}}{\gamma_{ij}} x_i^{(t)} x_i^{(t)'} \end{aligned} \quad (6)$$

Where:

$$\begin{aligned} \gamma_{ij}^{(t)} &= \frac{\alpha_{ij} p_{ij}(x_i^{(t)})}{\sum_{i=1}^p \sum_{j=1}^{m_i} \alpha_{ij} p_{ij}(x_i^{(t)})} = \frac{\alpha_{ij} p_{ij}(x_i^{(t)})}{p(x^{(t)})} \\ \gamma_{ij} &= \sum_{t=1}^T \gamma_{ij}^{(t)} = \sum_{t=1}^T \frac{\alpha_{ij} p_{ij}(x_i^{(t)})}{p(x^{(t)})} \end{aligned} \quad (7)$$

2.5 Setting up the mapping function

Now we use the CG-GMM training to set up the mapping function for spectral and pitch features. Let $a = [x, y]$ denote a frame of spectral for the source and target speakers, $b = [p, q]$ denote a frame of pitch for the source and target speaker, $A = [a_1, a_2, \dots, a_N]$ represent there are N frames spectral pairs, same as $B = [b_1, b_2, \dots, b_N]$ represent N frame pitch pairs. The combination of spectral and pitch $z = [a, b]$ is used to estimate the CG-GMM parameters.

In the training process, the joint feature z is splitted into two element-groups which represent spectral group and pitch group respectively, and the component-group also splitted to two groups, the amount of components are m_1, m_2 . After training, we can get the parameters $(\alpha_{1j}, u_{1j}, \Sigma_{1j} | j=1, \dots, m_1)$ and $(\alpha_{2j}, u_{2j}, \Sigma_{2j} | j=1, \dots, m_2)$ for the two component-groups. Then the mapping function can be written as:

$$\begin{aligned} \hat{y} &= F(x) = E[y | x] \\ &= \sum_{j=1}^{m_1} h_j(x) [u_j^y + \Sigma_j^{xy} (\Sigma_j^{xx})^{-1} (x - u_j^x)] \\ \hat{q} &= F(p) = E[q | p] \end{aligned} \quad (8)$$

$$= \sum_{j=1}^{m_2} h_j(p) [u_j^q + \Sigma_j^{pq} (\Sigma_j^{pp})^{-1} (p - u_j^p)] \quad (9)$$

where
$$h_j(x) = \frac{\alpha_{1j} N(x; u_j^x, \Sigma_j^{xx})}{\sum_{j=1}^{m_1} \alpha_{1j} N(x; u_j^x, \Sigma_j^{xx})} \quad (10)$$

$$h_j(p) = \frac{\alpha_{2j} N(p; u_j^p, \Sigma_j^{pp})}{\sum_{j=1}^{m_2} \alpha_{2j} N(p; u_j^p, \Sigma_j^{pp})} \quad (11)$$

are the posterior probabilities of given input vector x and p . The mean vector and covariance matrix of two component-groups can written as:

$$u_{1j} = \begin{bmatrix} u_j^x \\ u_j^y \end{bmatrix}, u_{2j} = \begin{bmatrix} u_j^p \\ u_j^q \end{bmatrix}, \quad \Sigma_{1j} = \begin{bmatrix} \Sigma_j^{xx} & \Sigma_j^{xy} \\ \Sigma_j^{yx} & \Sigma_j^{yy} \end{bmatrix}, \Sigma_{2j} = \begin{bmatrix} \Sigma_j^{pp} & \Sigma_j^{pq} \\ \Sigma_j^{qp} & \Sigma_j^{qq} \end{bmatrix} \quad (12)$$

III. EXPERIMENT AND EVALUATION RESULTS

3.1 Experiments

The training data consist of 400 short utterances of 2 speakers (one male and one female) from the 863DB synthesis database of Mandarin speech, which sampled at 16KHz. Another 20 utterances are used for testing.

After STRAIGHT analysis, we get the spectrum and pitch of each frame. To reduce the dimension of the feature vectors, Mel-scale DCT is performed on the STRAIGHT spectrum to get the 20 dimensions of feature vector, and only the lowest 19 dimensions are kept, the first dimension DCT feature corresponds to the signal power is kept as the value of the

source speaker.

S/U/V (silence/unvoiced/voiced) classification is done to each frame. We can classify that whether the frame is voiced or not from the pitch value, because the not voiced frame's pitch is zero. If the frame is not voiced, we compute ZCR (zero crossing rate) and EN (energy) of that frame to decide the frame is silence or unvoiced. In converting procedure we only transform the unvoiced and voiced frame speech, the silence frame is kept as the speech of the source speaker.

DTW (dynamic time wrapping) algorithm is used to align the source and target speech feature (spectral) streams, which already discard the silence frames. The pairs of aligned speech frames where one speaker's speech was voiced and the other speaker's was unvoiced were rejected from the training sets. The different classes of voicing property usually occurs at boundaries of Initials and Finals in the Mandarin speech, they will lead to classify error, and result in the negative influence on the converted speech quality and speaker identifiability. The remaining time aligned data comprise 30,000 spectral vectors.

Then the paired spectral feature vectors are split into two datasets as voiced and unvoiced. The voiced spectral vectors combined with the pitch are inputted to the CG-GMM training.

The training and converting procedures are indicated in Figure 1 and Figure 2 below. For unvoiced frames, only the spectral features are transformed, so we can use the GMM to training the mapping function, and the transformation function is similar to equation (8) and (9).

3.2 Evaluation results

We make experiments to compare three different methods of voice conversion.

Fig.1. Training procedure

Fig.2. Converting procedure

Method-1 use GMM training to transform the spectral features combined with a simple linear modification of pitch, which can be written as:

$$\hat{f}_0 = \frac{u_t}{u_s} \times f_0 \quad (13)$$

where $f_0, \hat{f}_0$ denote log scale F_0 of source and converted speech, and u_s, u_t denote mean log scale F_0 of source and target speech.

Method-2 use the method described by [7] which use GMM to transform the joint pitch and spectral features. Method-3 is our method using CG-GMM to transform the joint pitch and spectral features.

In Method-1 and Method-2 the mixture components are set to 64. In Method-3, the spectral and pitch component-group is set to 64 and 16. We just do male-to-female conversion for these three methods.

3.2.1 Objective evaluation

To evaluate the pitch transformation performance of these three methods, we use the Normalized Pitch Distortion (NPD) [7] defined as:

$$NPD = \sqrt{\frac{\sum_{n=1}^N (F_{0,n}^y - \hat{F}_{0,n}^y)^2}{\sum_{n=1}^N (F_{0,n}^y - F_{0,n}^x)^2}} \quad (14)$$

For spectral transformation, the Normalized Spectral Distortion (NSD) [7] is defined:

$$NSD = \frac{\sum_{n=1}^N \|c_{y,n} - \hat{c}_{y,n}\|_2}{\sum_{n=1}^N \|c_{y,n} - c_{x,n}\|_2} \quad (15)$$

TABLE I
NORMAL PITCH DISTORTION COMPARE

Method	Method-1	Method-2	Method-3
NPD	0.23	0.17	0.16

TABLE II
NORMAL SPECTRAL DISTORTION COMPARE

Method	Method-1	Method-2	Method-3
NSD	0.54	0.50	0.48

From the result above, we can see that the method we proposed has lowest NPD and NSD. It manifest that CG-GMM is better than GMM to model the joint features of spectral and pitch. We can also see the spectral and the pitch are correlated. Modify them globally improve the overall performance of voice conversion.

3.2.2 Subjective evaluation

Subjective evaluation experiments is performed to investigate the speaker identifiability between these three methods. ABX test was performed and 10 listeners participated in the perceptual tests. The listeners are requested to decide whether utterance A or B is closer to X in term of speaker identity. A is converted speech using Method-3, B is converted speech using Method-1 or Method-2. X is the target speech. The results that showing in figure 3 indicate our method give best speaker identifiability. It sounds most likely to the target speaker, particularly at speaking style.

Fig.3. Results for speaker identification evaluation

IV. CONCLUSIONS

A new framework of CG-GMM is used to transforming joint of spectral and pitch features. CG-GMM can model the feature elements of different meaning or unit. Compared with the previous approach of voice conversion, the results show this method has better performance. The converted speeches indicate satisfactory speech quality and speaker identifiability. Meanwhile the speaking style is much like to the target speaker.

ACKNOWLEDGMENT

The authors would like to thank the Hideki Kawahara for kindly providing the STRAIGHT codes. This work is supported by the Natural Science Foundation of China under grant No.60172055,60121302 and Beijing Nature Foundation under grant No.4042025.

REFERENCES

- [1] E. Moulines, Y. Sagisaka, "Voice conversion: state of the art and perspectives[J]", *Speech Communication*, 16(2): 125-126, 1995.
- [2] M. Abe, S. Nakanura, and K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization", *Proceeding of ICASSP*, pp: 655-658, 1988.
- [3] H. Valbret, E. Moulines, and J.P. Tabach, "Voice transformation using PSOLA technique", *Speech communications*, vol.11, pp: 175-187, 1995.
- [4] M. Narendranath, H.A. Murthy, S. Rajendran, B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks", *Speech Communication*, vol.16, pp: 207-216, 1995.
- [5] Y. Stylianou, O. Cappe, E. Moulines, "Continuous probabilistic transform for voice conversion", *IEEE Tran. Speech and Audio proc.*, vol.6 pp: 131-142, 1998.
- [6] C. D'Alessandro, B. Doval, "Voice quality modification for emotional speech synthesis", *Proceeding of Eurospeech*, pp:1653-1656, 2003.
- [7] Taoufik En-Najjary, Olivier Roec and Thierry Chonavel, "A Voice Conversion Method Based on Joint Pitch and Spectral Envelope Transformation", *Proceeding of ICSLP*, 2004.

- [8] Yang Liu, Zhiqiang Wand and Bo Xu, "Text-Independent Speaker Identification Using Feature-Splitting Method and Component-Group GMM", *The third Biometric Recognition Conference*, China, 2002.
- [9] H. Kawahara, I. Masuda-katsuse, A. De Cheveign, "Restructuring Speech Representations Using a Pitch-adaptive Time-frequency Smoothing and an Instantaneous Frequency-based F0 Extraction: Possible Role of Repetitive Structure in Sounds", *Speech Communication*, vol.27, pp: 187-207, 1999.