# Improving Segmental GMM Based Voice Conversion Method with Target Frame Selection

*Hung-Yan Gu, Sung-Fung Tsai*

National Taiwan University of Science and Technology, Taipei

guhy@mail.ntust.edu.tw, m9615069@mail.ntust.edu.tw

## Abstract

In this paper, the voice conversion method based on segmental Gaussian mixture models (GMMs) is further improved by adding the module of target frame selection (TFS). Segmental GMMs are meant to replace a single GMM of a large number of mixture components with several voice-content specific GMMs each consisting of much fewer mixture components. In addition, TFS is used to find a frame, of spectral features near to the mapped feature vector, from the target-speaker frame pool corresponding to the segment class as the input frame belongs to. Both ideas are intended to alleviate the problem that the converted spectral envelopes are often over smoothed. To evaluate the performance of the two ideas mentioned, three voice conversion systems are constructed, and used to conduct listening tests. The results of the tests show that the system using the two ideas together can obtain much improved voice quality. In addition, the measured variance ratio (VR) values show that the system with the two ideas adopted also obtains the highest VR value.

**Index Terms**: voice conversion, GMM, frame selection, discrete cepstral coefficient, variance ratio

## 1. Introduction

The GMM based voice conversion method was introduced by Stylianou [1]. Thereafter, many researches had tried to improve this method by considering some relevant issues. The issues considered include spectral over-smoothing found in the converted spectrums [2-5], spectral discontinuities between some adjacently converted frames [2, 3, 5], prosody conversion [6, 7], and other minor issues.

Although previous researchers had already proposed their methods to improve voice-conversion performances, these issues, however, need more investigations in order to have various kinds of solution methods to satisfy different requirements by different application developers. Possible requirements include (a) voice quality first with acceptable similarity, (b) voice similarity first with acceptable quality, (c) voce quality compromised with implementation cost, etc.

We know that the issue, spectral over-smoothing, had been tackled with at least two kinds of methods, global variance (GV) [4, 5] and dynamic frequency warping (DFW) [2, 3]. Additionally, the methods based on DFW are intended to remedy a weak point of the GV based methods, i.e. the correlation between the source and target parameters is low [3], which causes decreased timbre similarity.

In this paper, we also study the issue, spectral over-smoothing, but with a different approach, segmental GMMs plus target frame selection. The advantages of our approach include (a) simpler in concept, (b) easier to implement (hence saving efforts or money), (c) compromised processing-time latency (e.g. 30 frames) between DFW (1 frame) and GV (utterance level), (d) effective for improving the converted-voice quality (the signal quality of the converted voice).

Here, we will use multiple segmental GMMs to alleviate the problem of over-smoothed converted spectrum. In addition, we notice that Mandarin is a syllable prominent language since we select Mandarin Chinese here to study voice conversion. Therefore, we treat each syllable of a training sentence as one segment if the syllable has no initial consonant or has just unvoiced initial consonant, or as two segments (i.e. the voiced initial consonant plus the syllable final) if the syllable is started with a voiced consonant. Next, each segment is grouped to one of the 39 classes, including 4 classes of voiced initial consonants (i.e. /m/, /n/, /l/, /r/) and 35 classes of syllable finals. In Mandarin Chinese, a syllable final is a vowel nucleus consisting of one to three vowels plus a possible nasal ending. For each of the 39 classes, a corresponding GMM will be trained from the segments grouped to. After training, the 39 GMMs are used for on-line voice conversion. Nevertheless, there is a problem that must be solved beforehand. The problem is how the right class that an input frame belongs to can be picked out? For this problem, we had previously developed an automatic selection algorithm based on dynamic programming (DP) [8].

Furthermore, we extend segmental GMMs by adding one more processing step, i.e. frame selection, to further alleviate the problem of spectral over smoothing. By frame selection, each converted feature vector is replaced with a real (i.e. not converted) feature vector analyzed from a target frame (target-speaker frame) in order to improve the converted-voice quality. In fact, the idea of frame selection is proposed previously by Dutoit, *et al.* [9]. Also, a variation of frame selection has been proposed by Wu, *et al.* [16]. In the paper [9], the feature vector of a source frame is mapped with a conventional GMM, and then a target frame is searched, in terms of the mapped feature vector, with a DP based algorithm. Here, we map the feature vector of a source frame with a segmental GMM, and then search for a target frame with a developed DP algorithm. The two steps, spectral mapping and frame selection, are not independent. A better spectral mapping method would help the module, frame selection, to find out a more appropriate target frame. We have built an on-line voice conversion system that cascades the two steps. By using this system, we have conducted listening tests. The details of the system and the listening tests are presented in the following sections.

## 2. Conversion procedure

The procedure studied here to convert a voice signal is as the processing flow drawn in Fig. 1. When a spoken sentence with unknown content is inputted, it is first sliced into a sequence of frames with the frame width, 512 sample points, and frame shift, 110 points (5 ms), under the sampling rate 22,050 Hz. Then, the pitch frequency of each frame is detected in the left flow of Fig. 1 with the method based on both autocorrelation

function and AMDF function [10]. When a frame is detected to be unvoiced, the four gray colored blocks in Fig. 1 are bypassed directly, which means that pitch adjusting is not needed and the spectral parameters, discrete cepstral coefficients (DCC), are not converted. On the other hand, when a frame is detected to be voiced, its pitch is simply converted with the equation,

$$q_t = \mu^{(y)} + \frac{\sigma^{(y)}}{\sigma^{(x)}}(p_t - \mu^{(x)}), \quad (1)$$

where $p_t$ is the detected pitch frequency, $\mu^{(x)}$ and $\sigma^{(x)}$ are the average and standard deviation of the source speaker's pitch frequencies, and $\mu^{(y)}$ and $\sigma^{(y)}$ are the average and standard deviation of the target speaker's pitch frequencies.

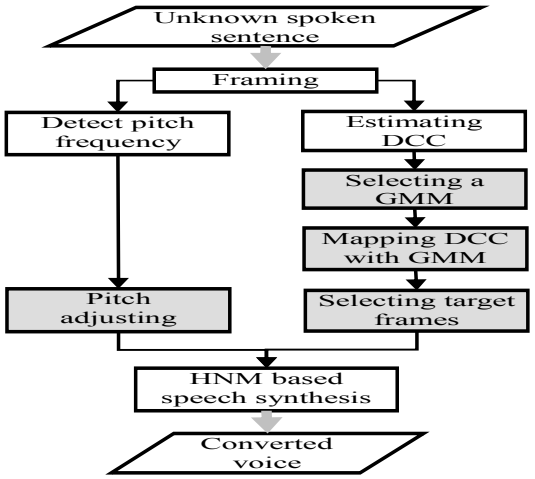


Figure 1: *Processing flow for converting voice signal.*

In the right flow of Fig. 1, the input frames are analyzed one after another to estimate each frame's DCC. Nevertheless, in the block, "Selecting a GMM", we propose a selection algorithm that processes every 30 successive voiced frames in a batch. With this algorithm, the correct GMM (or its nearby GMM sometimes) can be picked out from the 39 GMMs for each frame. Then, in the block, "Mapping DCC with GMM", the selected GMM is used to map the DCC. In this block, a sequence of voiced frames bounded with left and right unvoiced frames are processed in a batch.

Next, in the block, "selecting target frames", the sequence of voiced frames is also processed in a batch with another developed DP algorithm. Similarly, spectral continuity between adjacently selected target frames must also be considered besides the spectral matching distance between the DCC of the converted input frame and the DCC of a target frame. Finally, in the jointed block, "HNM based speech synthesis", speech signals are re-synthesized using a harmonic plus noise model (HNM) based method [11, 12].

## 2.1. Estimating DCC

Several methods have been proposed for estimating a signal frame's spectral envelope. The method, STRAIGHT, is very accurate in its estimated spectral envelope but it requires a large amount of computations and cannot be used to implement a real-time system currently. Therefore, in this study, we adopt the spectral envelope estimation method, discrete cepstrum [13, 14], and use the estimated DCC as the spectral features. For each signal frame, the DCC estimation scheme proposed in a previous work [14] is used to calculate 40 DCC. In that scheme, a mel-like frequency scale is adopted. The estimated DCC of each target frame are stored with its frame-sequence number to one of the 39 target-frame pools according to the segment class that this frame belongs to.

## 2.2. Selecting a GMM

Since the content of the input speech is unknown, which of the 39 segmental GMMs should be selected for mapping each input frame's DCC becomes a problem that must be solved. In general, this is a problem of speech recognition. Nevertheless, it is not so serious because some frames assigned with incorrect but nearby GMMs are tolerable.

Here, we use the 39 segmental GMMs trained to take the role of HMM (hidden Markov model) usually used for speech recognition. In addition, we notice that it is rare for a person to utter more than 2 different segments (syllables) within a very short time interval, e.g. 150 ms, under an ordinary speaking rate. Therefore, we decide to select GMMs for every 30 successive voiced frames (spanning 150ms of time) in a batch. Actually, we have experimented to inspect the differences in the numbers of segments selected when setting the batch length to 20, 30, and 40, respectively. It is found that the batch length, 30, is a better choice. Then, only one or two of the 39 GMMs will be picked out for a batch of 30 voiced frames. We have previously developed a DP based algorithm that selects one or two GMMs according to the criterion of maximum likelihood. The details of the DP based selection algorithm are referred to our previous work [8].

## 2.3. Mapping DCC with GMM

A conventional GMM based mapping function [1] is

$$y = F(x;\mu,\Psi) = \sum_{m=1}^{M} \gamma(m)\cdot\left(\mu_m^{(y)} + \Psi_m^{(yx)}\cdot\left(\Psi_m^{(xx)}\right)^{-1}\cdot(x-\mu_m^{(x)})\right), \quad (2)$$

$$\gamma(m) = \frac{w_m \cdot N(x;\mu_m^{(x)},\Psi_m^{(xx)})}{\sum_{m=1}^{M} w_m \cdot N(x;\mu_m^{(x)},\Psi_m^{(xx)})},$$

where $x$ denotes a spectral feature vector of the source speaker, $y$ denotes the converted feature vector for the target speaker, $M$ is the number of Gaussian mixture components, $w_m$ is the weight of the $m$-th mixture component, and $\mu$ and $\Psi$ represent the sets of mean vectors and covariance matrices, respectively. In this paper, we have experimented to measure average cepstral distances between converted DCC and target DCC for some $M$ values (varied from 8 to 16). The results show that the value of $M$ is not significant to the measured average distance. Therefore, we set the value of $M$ to 8 for each segmental GMM finally.

## 2.4. Selecting target frames

Let $y_1$, $y_2$, …, $y_T$ be a sequence of converted DCC vectors obtained from mapping with GMM. Notice that each vector, $y_t$, of the sequence may be somehow distorted during the mapping from $x_t$ to $y_t$. To improve the quality of the converted voice, we are thus motivated, by Dutoit, *et al*. [9], to replace $y_t$ with a real (not converted) DCC vector, $z_t$, analyzed from a

target frame belonging to the segment class indexed as $I(t)$. To select a frame, $z_t$, from a pool of frames corresponding to a segment class, we should consider not only the matching distance, $dist(y_t, z_t)$, but also the connection distance, $dist(z_{t-1}, z_t)$, in order to prevent spectral discontinuity from occurring. Besides the connection distance adopted in the work by Dutoit, *et al*. [9], we add another term of dynamic-spectral distance to reflect a dynamic spectral change, $\Delta y_t = y_t - y_{t-1}$, between a pair of adjacently converted frames, to its corresponding pair of real target frames, $\Delta z_t = z_t - z_{t-1}$. This dynamic-spectral distance is useful to slightly improve the quality of the converted speech according to our experiments. Consequently, we have developed another DP based algorithm to do target frame selection.

As the first step, for each converted DCC vector, $y_t$, $K$ target-frame DCC of the least distances to $y_t$ are found by fully searching the frame pool corresponding to the segment class indexed as $I(t)$. Here, $K$ is set to 24 according to the results of the experiments measuring VR (variance ratio defined in (5)) values with $K$ varied from 12 to 36. Next, let $Q(t, i)$ denote the best cumulated distance from time 1 to $t$ and the index of the target-frame DCC selected at time $t$ be $i$, i.e. the $i$-th frame of the $K$ found frames for replacing $y_t$. Then, the recursive formula,

$$Q(t,i) = \min_{0\le j<K} \begin{bmatrix} Q(t-1,j) + \alpha \cdot dist\left(z_{t-1}^{j}, z_t^{i}\right) + \\ \alpha \cdot dist\left(y_t - y_{t-1}, z_t^{i} - z_{t-1}^{j}\right) \end{bmatrix} + dist\left(y_t, z_t^{i}\right), \quad (3)$$

is used to execute dynamic programming, where $\alpha$ is a weighting factor for both connection and dynamic-spectral distances, and $z_t^i$ denotes the $i$-th target-frame DCC candidate of the $K$ found candidates at time $t$ for replacing $y_t$. Here, $\alpha$ is set to 1.5 according to the results of the experiments measuring VR values with $\alpha$ varied from 0.25 to 6. As mentioned in the work [9], a trick to obtain more natural spectral connection is to dynamically reset the value of $\alpha$ to 0 if $z_{t-1}^j$ and $z_t^i$ are checked to be adjacent frames coming from a same utterance. This trick is also adopted here, and is extended by accepting the case that $z_{t-1}^j$ and $z_t^i$ come from a same utterance and have just one another frame in between. Finally, at time $T$, the minimum cumulated distance $W(T)$ is computed as

$$W(T) = \min_{0\le j<K} \left[ Q(T, j) \right]. \quad (4)$$

In terms of (3) and (4), the minimum cumulated distance can be obtained. Also, the sequence of target-frame indices from time 1 to $T$ can be backtracked. Then, the real target-frame DCC, $z_t^i$, corresponding to the indices backtracked are taken to replace the converted-frame DCC, $y_t$, $t$=1, 2, …, $T$. Here, $T$ is the time length of a sequence of voiced frames.

### 2.5. HNM based speech synthesis

In HNM, the spectrum of a voiced frame is split into two parts, i.e. lower-frequency harmonic part and higher-frequency noise part. The boundary frequency between the two parts is termed the maximum voiced frequency (MVF) [11]. To simplify the processing of speech signal synthesis, the MVF values of voiced frames are all fixed to 6,000 Hz in this study.

The DCC coefficients sent to the block, HNM based speech synthesis, of Figure 1 are inversely transformed to obtain a curve of spectral envelope. According to the spectral envelopes of successive frames, speech signal is synthesized with an HNM based scheme. That is, harmonic signal and noise signal are separately synthesized and then added to obtain the final speech signal. The synthesis processing with HNM will not be detailed here because the details can be found from the previous works [8, 14].

## 3. Experimental evaluations

For evaluating the conversion method proposed here, we have constructed three voice conversion systems, named SOG, SLG, and SLGF, respectively. In the system SOG (system using original GMM for mapping), a single GMM of 128 mixture components are trained with 350 parallel sentences. Then, the mapping function, (2), is used to convert the DCC of each input frame. In the system SLG (system using selected GMM), we trained 39 segmental GMMs instead of one single GMM. The number of mixture components for each segmental GMM is set to 8. Then, the method presented in Section 2.2 is used to select segmental GMM. As to the system SLGF (adding target frame selection to the system SLG), the method presented in Section 2.4 is used to select target-frame DCC vectors to replace the converted DCC vectors.

### 3.1. Voice quality tests

For voice quality tests, 3 converted voice files are prepared first, which are named VXA (converted by the system SOG), VXB (converted by the system SLG), and VXC (converted by the system SLGF). These 3 files can be accessed at the web page, http://guhy.csie.ntust.edu.tw/VoiceConv/. Here, 15 persons (undergraduate or graduate students) are invited to listen to the voice files and give relative scores. The 3 files are played in the order AX where A is fixed to VXA and X is randomly selected from VXB and VXC. Each time that two files, AX, are played, the participant is requested to give a score. The score range is defined from 1 to 5. The score 5 (1) means the quality of X is much better (worse) than A, the score 4 (2) means the quality of X is slightly better (worse) than A, and the score 3 means the quality of X cannot be distinguished from that of A.

Table 1.. *Average scores for voice quality tests.*

| | | SOG vs SLG | SOG vs SLGF |
|---|---|---|---|
| MA=>MB | AVG | 3.73 | 4.33 |
| | (STD) | (0.59) | (0.62) |
| MA=>FA | AVG | 3.53 | 3.93 |
| | (STD) | (0.52) | (0.59) |

After listening tests, the scores given by the 15 persons are collected to compute average scores and standard deviations for the two systems, SLG and SLGF, respectively, when compared with the system SOG. The results are as those values listed in Table 1. From Table 1, it can be found that the average scores for voice conversion from MA (male speaker A) to MB (male speaker B) are about 0.3 better than the average scores for voice conversion from MA to FA (female speaker A). This indicates that the quality of the converted voice from different genders is harder to improve. In addition,

when the average scores of the two systems, SLG and SLGF, are compared, it can be found that the scores of SLGF are both better than those of SLG. Therefore, the idea of cascading automatic segmental GMM selection with automatic target frame selection can indeed help to improve the quality of the converted voice.

### 3.2. Cepstral distance and variance ratio

There are 25 remaining parallel sentences that are not used in the training stage. Hence, the 25 sentences uttered by the source speaker, MA, are fed to the three systems to obtain their corresponding converted sentences, respectively. Then, a geometric distance of DCC is measured between each voiced frame of the converted sentences and its corresponding frame in the target sentences according to the saved DTW alignment data. Next, the measured distances are averaged across all voiced frames. As a result, the average distances obtained for the three systems are as those listed in Table 2. From this table, it is seen that the system SOG obtains the smallest average distances. Nevertheless, the results of listening tests show that the system SOG is worse than SLGF in voice quality. Therefore, the average cepstral distances measured are inconsistent with the results of the listening tests. Such a situation, i.e. inconsistency between cepstral distance and voice quality, is also reported in several works by others [3, 4, 15]. Therefore, another objective measure, variance ratio (VR), is used in [3, 4], which is consistent in general with the converted-voice quality. As explained in [3], VR is a more global indicator of the ability of the transformation method to mimic realistic variations in the converted voice, and VR would be one in the case of perfect transformation.

Table 2.. *Average distances for the three systems.*

| | SOG | SLG | SLGF |
|---|---|---|---|
| MA=>MB | 0.5440 | 0.6312 | 0.6889 |
| MA=>FA | 0.5237 | 0.5252 | 0.5951 |

The formula of variance ratio adopted here is

$$VR = \sum_{v=1}^{V} \frac{N_v}{NT} \cdot \left( \frac{1}{D} \cdot \sum_{d=1}^{D} \frac{(\hat{\sigma}_v^d)^2}{(\sigma_v^d)^2} \right), \quad (5)$$

where $V$ is the number of segment classes ($V$=39 here), $N_v$ is the number of voiced test frames belonging to the $v$-th class, $NT$ is the total number of voiced test frames, $D$ is the dimensionality of DCC ($D$=40 here), $\hat{\sigma}_v^d$ and $\sigma_v^d$ are the standard deviations of the $d$-th dimension for the converted and target DCC, respectively. According to (5), we measure the values of VR for the three systems, SOG, SLG, and SLGF, by using the 25 parallel sentences. As a result, the measured values are as listed in Table 3. From Table 3, it can be seen that the average VR value, 0.2318, of SLG is greater than the one, 0.2003, of SOG, and the average VR value, 0.6021, of SLGF is much greater than the other two systems' values. Therefore, the measured VR values are consistent with the perceived qualities of the three systems' converted voices.

Table 3.. *VR values measured for the three systems.*

| | SOG | SLG | SLGF |
|---|---|---|---|
| MA=>MB | 0.2223 | 0.2578 | 0.6248 |
| MA=>FA | 0.1783 | 0.2058 | 0.5793 |
| Average | 0.2003 | 0.2318 | 0.6021 |

## 4. Conclusions

The results of the listening tests show that the system SLGF obtains higher average score, 4.13, than the average score, 3.63, of the system, SLG. Therefore, SLGF is the best in voice quality among the three systems, SOG, SLG, and SLGF. In addition, the measured VR values are 0.6021 for SLGF, 0.2318 for SLG, and 0.2003 for SOG. This indicates that the system SLGF has much better converted-voice quality than the other two systems. Therefore, the approach that combines segmental GMM with target frame selection can indeed help to improve the performances of GMM based voice conversion methods. As to the measured cepstral distances, the system SOG has the smallest average distance. Nevertheless, according to our observations, the smaller average distance is obtained in terms of over-smoothed converted spectral envelopes. Therefore, the system SOG (also SLG) suffers the degraded voice quality and timbre similarity.

Since converted-voice quality is the focus of this study, listening tests for timbre similarity are not conducted yet. As the next step, we will conduct listening tests for timbre similarity. In addition, we may continue to study the issue whether the number of the candidate frames in each of the segmental frame-pools will influence the voice quality of the final converted voice.

## 5. Acknowledgement

This study is supported by National Science Council of Taiwan under the contract number, MOST 103-2221-E-011-131.

## 6. References

[1] Stylianou, Y., Cappe, O., and Moulines, E., "Continuous Probabilistic Transform for Voice Conversion", IEEE Trans. Speech and Audio Processing, 6(2): 131-142, 1998.

[2] Erro, D., Moreno, A., and Bonafonte, A., "Voice Conversion Based on Weighted Frequency Warping", IEEE Trans. Audio, Speech, and Language Processing, 18(5): 922-931, 2010.

[3] Godoy, E., Rosec, O., and Chonavel, T., "Voice Conversion Using Dynamic Frequency Warping with Amplitude Scaling, for Parallel or Nonparallel Corpora", IEEE Trans. Audio, Speech, and Language Processing, 20(4): 1313-1323, 2012.

[4] Benisty, H., and Malah, D., "Voice conversion using GMM with enhanced global variance", in Proceedings of INTERSPEECH, Florence, Italy, 669-672, 2011.

[5] Toda, T., Black, A. W., and Tokuda, K., "Voice Conversion Based on Maximum-likelihood Estimation of Spectral Parameter Trajectory", IEEE Trans. Audio, Speech, and Language Processing, 15(8): 2222-2235, 2007.

[6] Wu, C. H., Hsia, C. C., Lee, C. H., and Lin, M. C., "Hierarchical Prosody Conversion Using Regression-based Clustering for Emotional Speech Synthesis", IEEE Trans. Audio, Speech, and Language Processing, 18(6): 1394-1405, 2010.

[7] Wu, Z. Z., Kinnunen, T., Chng, E. S., and Li, H. Z., "Text-Independent F0 transforma-tion with non-parallel data for voice conversion", in Proceedings of INTERSPEECH, Makuhari, Chiba, Japan, 1732-1735, 2010.

[8] Gu, H. Y. and Tsai, S. F., "An improved voice conversion method using segmental GMMs and automatic GMM selection",

Int. Congress on Image and Signal Processing (CISP2011), Shanghai, China, 2395-2399, 2011.

[9] Dutoit, T., Holzapfel, A., Jottrand, M., Moinet, A., Perez, J., and Stylianou, Y., "Towards a voice conversion system based on frame selection", Int. Conf. Acoustics, Speech, and signal Processing, Honolulu, Hawaii, 513-516, 2007.

[10] Kim, H. Y., Lee, J. S., Sung, M. W., Kim, K. H., and Park, K. S., "Pitch detection with average magnitude difference function using adaptive threshold algorithm for estimating shimmer and jitter," 20-th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society, Hong Kong, China, 3162-3165, 1998.

[11] Stylianou, Y., Harmonic plus Noise Models for Speech, Combined with Statistical Methods, for Speech and Speaker Modification, Ph.D. thesis, Ecole Nationale Supèrieure des Télécommunications, Paris, France, 1996.

[12] Gu, H. Y. and Liao, H. L., "Mandarin singing-voice synthesis using an HNM based scheme", Journal of Information Science and Engineering, 27(1): 303-317, 2011.

[13] Cappé, O. and Moulines, E., "Regularization techniques for discrete cepstrum estimation", IEEE Signal Processing Letters, 3(4): 100-102, 1996.

[14] Gu, H. Y. and Tsai, S. F., "A discrete-cepstrum based spectrum-envelope estimation scheme and its example application of voice transformation", Int. Journal of Computational Linguistics and Chinese Language Processing, 14(4): 363-382, 2009.

[15] Hwang, H. T., Tsao, Y., Wang, H. M., Wang, Y. R., and Chen, S. H., "Exploring mutual information for GMM-based spectral conversion", in Proceedings of ISCSLP, Hong Kong, China, 50-54, 2012.

[16] Wu, Z. Z., Virtanen, T., Kinnunen, T., Chng, E. S., and Li, H. Z., "Exemplar-based unit selection for voice conversion utilizing temporal information", in Proceedings of INTERSPEECH, Lyon, France, 3057-3061, 2013.