Computational Linguistics and Chinese Language Processing
Vol. 14, No. 4, December 2009, pp. 363-382

A Discrete-cepstrum Based Spectrum-envelope Estimation Scheme and Its Example Application of Voice Transformation

Hung-Yan Gu* and Sung-Feng Tsai*

Abstract

Approximating a spectral envelope via regularized discrete cepstrum coefficients has been proposed by previous researchers. In this paper, we study two problems encountered in practice when adopting this approach to estimate the spectral envelope. The first is which spectral peaks should be selected, and the second is which frequency axis scaling function should be adopted. After some efforts of trying and experiments, we propose two feasible solution methods for these two problems. Then, we combine these solution methods with the methods for regularizing and computing discrete cepstrum coefficients to form a spectral-envelope estimation scheme. This scheme has been verified, by measuring spectral-envelope approximation error, as being much better than the original scheme. Furthermore, we have applied this scheme to building a system for voice timbre transformation. The performance of this system demonstrates the effectiveness of the proposed spectral-envelope estimation scheme.

Keywords: Spectral Envelope, Discrete Cepstrum, Harmonic-plus-noise Model, Voice Timbre Transformation.

1. Introduction

Here, a spectral envelope means a magnitude-spectrum envelope. Various methods have been proposed to estimate the spectral envelope of a speech frame. For example, in LPC (linear prediction coding) based methods (O'Shaughnessy, 2000; Schwarz & Rodet, 1999), the frequency response of an all-pole model is used to approximate the spectral envelope of a speech frame. Nevertheless, the frequency response curve of an LPC all-pole model will usually go below the true envelope around speech formants, and go above the regions where

* Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, 43 Keelung Rd., Sec. 4, Taipei, Taiwan.
E-mail: {guhy, M9615069}@mail.ntust.edu.tw

spectrum magnitudes fall suddenly. This is illustrated in Figure 1 using a frame sliced from an utterance of /i/. Therefore, the mismatches between the LPC envelope curve and the true curve cannot be ignored.

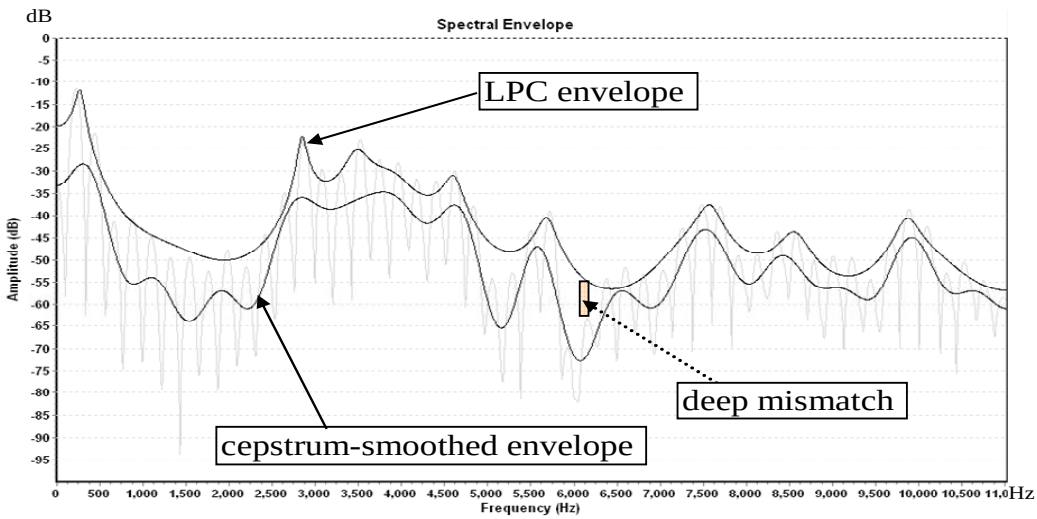


Figure 1. LPC and cepstrum smoothed spectral curves for a frame from /i/.

Besides LPC, several estimation methods are based on cepstrum analysis. The simplest one is to keep some leading cepstrum coefficients but truncate the remanded ones, *i.e.* replace them with zeros. Then, DFT (discrete Fourier transform) is used to transform the cepstrum coefficients back to the spectrum domain to obtain a smoothed spectrum curve. Nevertheless, such a smoothed spectrum curve is not really an envelope curve because it goes between the peaks and valleys of a DFT spectrum. One example is the lower smoothed curve in Figure 1. Therefore, a real cepstrum-based method to estimate a spectral envelope was proposed later by Imai and Abe (Imai & Abe, 1979; Robel & Rodet, 2005). They call this method true envelope estimation. In our opinion, this method is good but lacking in efficiency because a lot of computations are required. Similarly, the method proposed by Kawahara, Masuda-katsuse, and Cheveign (1999), STRAIGHT, is very accurate in its estimated spectral envelope. Nevertheless, it also requires a considerable number of computations and cannot be used to implement real-time systems currently. On the other hand, Galas and Rodet (1990) proposed the concept of discrete cepstrum and designed a feasible estimation method with this concept. Later, Cappé and Moulines (1996) improved this estimation method by adding a regularization technique to prevent unstable vibrating of the envelope curve from occurring. We think that estimating a spectral envelope with discrete cepstrum is a good approach if the feasibility and accuracy issues must be considered simultaneously. Therefore, we began to study the problems that will be encountered in practice.

As an overview, the spectral envelope estimation scheme proposed here is shown in Figure 2. When a speech frame is given, its fundamental frequency is first detected in the first block. If a frame is decided to be voiced, its estimated fundamental frequency will be used later in the block, "spectral peaks selection". Here, a method combining an autocorrelation function and AMDF (absolute magnitude difference function) is adopted to detect a frame's fundamental frequency (Kim *et al.*, 1998; Gu, Chang & Wu, 2004). Next, the frame is Hanning windowed and appended with zeros to form a sequence of 1,024 signal samples. This sequence is then transformed to frequency domain with FFT (fast Fourier transform) to obtain its magnitude spectrum. Given the magnitude spectrum, the block "spectral peaks selection" will determine which spectral peaks should be selected according to a method proposed here. After spectral peaks are selected, the frequency value of each selected peak is mapped to its target value with a frequency-axis scaling function proposed here. As the final step, the block "discrete cepstrum computation" will adopt an envelope-approximation criterion (Cappé & Moulines, 1996) to compute discrete cepstrum coefficients according to the selected and mapped spectral peaks.

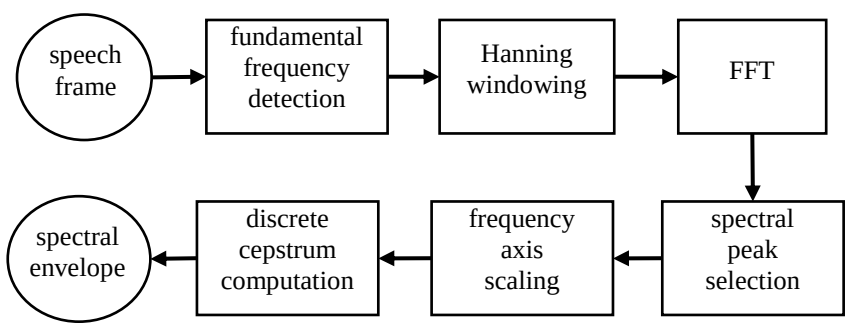


Figure 2. Main flow of the spectral-envelope estimation scheme.

In Figure 2, discrete-cepstrum computation is the main block, which already has been solved by other researchers (Cappé & Moulines, 1996). Nevertheless, the blocks, spectral-peak selection and frequency-axis scaling, still play important roles. When inappropriate peaks are selected or the frequency-axis is not scaled appropriately, the approximated spectral envelope will noticeably deviate from the true envelope. Therefore, we began to study these two blocks' problems, and the results are presented in Sections 3 and 4, respectively. As to discrete cepstrum, its computation and regularization will be reviewed in Section 2. In Section 5, the proposed scheme is practically evaluated by applying the scheme to build a voice timbre transformation system.

2. Spectral-envelope Estimation with Discrete Cepstrum

2.1 Discrete Cepstrum

The concept of discrete cepstrum was proposed by Galas and Rodet (1990). They adopted the least square criterion to a given set of spectral peaks to derive cepstrum coefficients. Such a derivation method is different from the conventional one. The conventional method transforms the logarithmic magnitude-spectrum with inverse DFT (IDFT) to get its cepstrum coefficients. In the conventional method, let the obtained cepstrum coefficients be c_0, c_1, …, c_{N-1} where N is the length of the signal sample sequence. According to these cepstrum coefficients, the original logarithmic magnitude-spectrum can be restored with DFT, *i.e.*

$$\log\left|X(k)\right| = \sum_{n=0}^{N-1} c_n e^{-j\frac{2\pi}{N}kn}, \quad 0 \le k \le N-1 \tag{1}$$

where $|X(k)|$, k=0, 1, …, N-1 represent the magnitude spectrum. Since $\log|X(k)|$ is even symmetric, *i.e.* $\log|X(k)| = \log|X(N-k)|$, the derived cepstrum coefficients are also even symmetric, $c_k = c_{N-k}$. Therefore, Equation (1) can be rewritten as:

$$\log\left|X(k)\right| = c_0 + 2\sum_{n=1}^{\frac{N}{2}-1} c_n \cos(\frac{2\pi}{N}kn) \; + \; c_{N/2}\cos(\pi k), \quad 0 \le k \le N-1 \,. \tag{2}$$

If most of the terms on the right side of Equation (2) are cancelled except the leading terms (*e.g.* p+1 terms), the magnitude spectrum computed, $\log S(f)$, would be a smoothed version of the original, $\log|X(f)|$. Here, the index variable, k, in Equations (1) and (2) is replaced with f in order to change the frequency scale from bins to the normalized frequency range from 0 to 1. Accordingly, $\log S(f)$ is computed as:

$$\log S(f) = c_0 + 2\sum_{n=1}^{p} c_n \cdot \cos(2\pi f n), \quad f = \frac{0}{N}, \frac{1}{N}, \frac{2}{N}, ..., \frac{N-1}{N} \,. \tag{3}$$

Based on Equation (3), some researchers have proposed to approximate the spectral envelope of $\log|X(f)|$ with $\log S(f)$. Nevertheless, the coefficients, c_n, in Equation (3) cannot be derived directly with IDFT. One derivation method proposed by Galas and Rodet is to define a set of envelope constraints and find the values of the coefficients, c_n, that can best satisfy the envelope constraints. In this manner, the derived coefficients, c_n, n=0, 1, …, p, are called the discrete cepstrum for $\log|X(f)|$.

The envelope constraints just mentioned are actually L pairs of (f_k, a_k) for L representative spectral peaks selected from the original spectrum $\log|X(f)|$. Here, f_k and a_k represent the frequency (already normalized to the value range from 0 to 1) and amplitude of the k-th spectral peak, respectively. Note that L is usually larger than the cepstrum order, p. Hence, a least-squares criterion is adopted to minimize the approximation errors between $S(f_k)$

and a_k, k=1, 2, …, L. That is, the approximation error computed as

$$\varepsilon = \sum_{k=1}^{L} \left| \log a_k - \log S(f_k) \right|^2 \tag{4}$$

is to be minimized. This equation can be rewritten in a matrix form

$$\varepsilon = (A - M \cdot C)^{\mathrm{T}} (A - M \cdot C) \tag{5}$$

where $A = [\log(a_1), \log(a_2), \ldots, \log(a_L)]^{\mathrm{T}}$, C is a column vector of (p+1) discrete cepstrum coefficients to be derived, *i.e.* $C = [c_0, c_1, \ldots, c_p]^{\mathrm{T}}$, and

$$M = \begin{bmatrix} 1 & 2\cos(2\pi f_1) & 2\cos(2\pi f_1 \cdot 2) & \cdots & 2\cos(2\pi f_1 \cdot p) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 2\cos(2\pi f_L) & 2\cos(2\pi f_L \cdot 2) & \cdots & 2\cos(2\pi f_L \cdot p) \end{bmatrix}$$

When the error ε of Equation (4) is minimized with the least-square criterion, the optimal values of the discrete cepstrum coefficients can be derived to be

$$C = (M^{\mathrm{T}} \cdot M)^{-1} \cdot M^{\mathrm{T}} \cdot A \tag{6}$$

That is, by just executing the operations of matrix inversion and multiplication, the values of the discrete cepstrum coefficients can be obtained.

2.2 Regularization of Discrete Cepstrum

According to Equations (5) and (6), it is seen that the discrete cepstrum coefficients are derived in the frequency domain with the least-square criterion. Nevertheless, such a derivation method may encounter a vital problem in practice. That is, the spectral envelope computed according to Equation (3) may vibrate radically and have very large approximation error at some frequencies slightly away from the selected spectral-peak frequencies, f_k. This is because the direct estimation method (*i.e.* Equation (6)) may sometimes be ill-conditioned. That is, slightly varying the frequency values of the detected spectral peaks may result in a very different spectral envelope curve being obtained. For example, look at the dash-lined spectral envelope in Figure 3. This curve is computed with 40 derived discrete cepstrum coefficients. Even though the first 7 spectral peaks are passed by this curve, the curve vibrates radically between two adjacent peaks. In practical applications, such a spectral envelope as the one in Figure 3 cannot be tolerated.

Therefore, Cappé and Moulines (1996) proposed a regularization technique to prevent such radical vibrations from occurring. To do this, they added a curve-sharpness penalty term to the approximation-error calculation equation, *i.e.* Equation (4), thereby making the approximation-error calculation equation:

$$\varepsilon = \sum_{k=1}^{L} \left| \log a_k - \log S(f_k) \right|^2 + \lambda \cdot R\big(S(f)\big) \tag{7}$$

where the function $R(\cdot)$ is intended to measure the sharpness of the spectral-envelope curve $S(f)$, and λ is a parameter to adjust the relative weight of the value returned by $R(\cdot)$. A typical function suggested for $R(\cdot)$ is

$$R\big(S(f)\big) = \int_0^{\pi} \left[\frac{d}{df} S(f) \right]^2 df \quad . \tag{8}$$

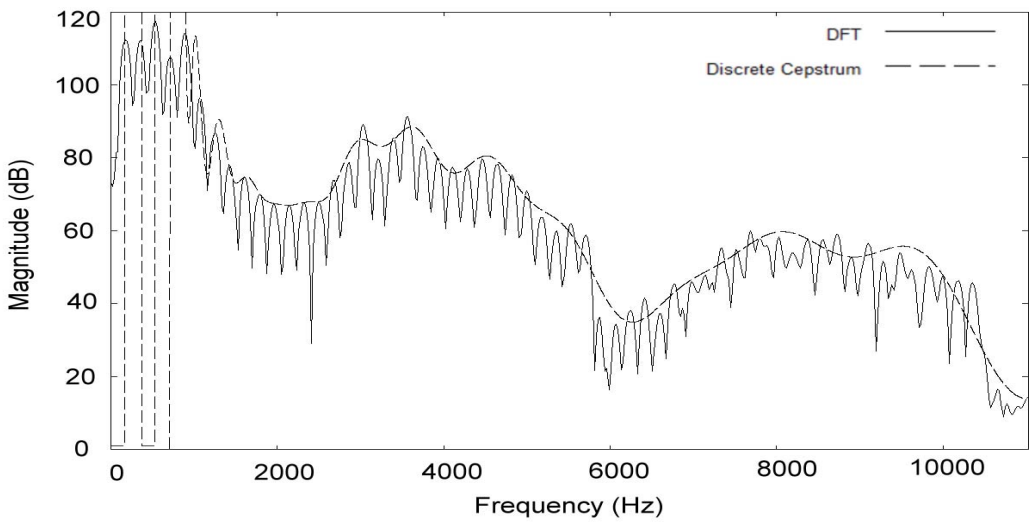


Figure 3. A spectral envelope computed with 40 non-regularized discrete cepstrum coefficients.

When the definition of $S(f)$ given in Equation (3) is taken into Equation (8), the following equation,

$$R\big(S(f)\big) = C^{\mathrm{T}} \cdot U \cdot C \quad , \quad U = 8\pi^2 \begin{bmatrix} 0 & & & 0 \\ & 1^2 & & \\ & & \ddots & \\ 0 & & & p^2 \end{bmatrix}, \tag{9}$$

can be derived (Cappé & Moulines, 1996; Stylianou, 1996). Then, the optimal solution that minimizes the error calculated in Equation (7) can be derived via:

$$\begin{aligned} &\varepsilon = (A - MC)^{\mathrm{T}} (A - MC) + \lambda C^{\mathrm{T}} UC \ ; \\ &\text{Let} \quad \frac{\partial \varepsilon}{\partial C} = -2 \cdot M^{\mathrm{T}} (A - MC) + 2 \cdot \lambda UC = 0 \ ; \\ &(M^{\mathrm{T}} M + \lambda U) \cdot C = M^{\mathrm{T}} A \ ; \\ &C = (M^{\mathrm{T}} M + \lambda U)^{-1} \cdot M^{\mathrm{T}} A \quad . \end{aligned} \tag{10}$$

According to empirical experience, the parameter λ is better set to a value around 0.0001. Thereafter, the ill-conditioning problem can be solved, and a regularized spectral envelope curve will be obtained. An example of a regularized spectral envelope is the dash-lined curve in Figure 4, where the solid-lined curve is the same as the one in Figure 3. Apparently, the phenomenon of radical vibration is not seen in this figure. Note that frequency axis scaling (to be discussed in Section 4.2) is already applied in addition to regularization to obtain the spectral envelope in Figure 4.

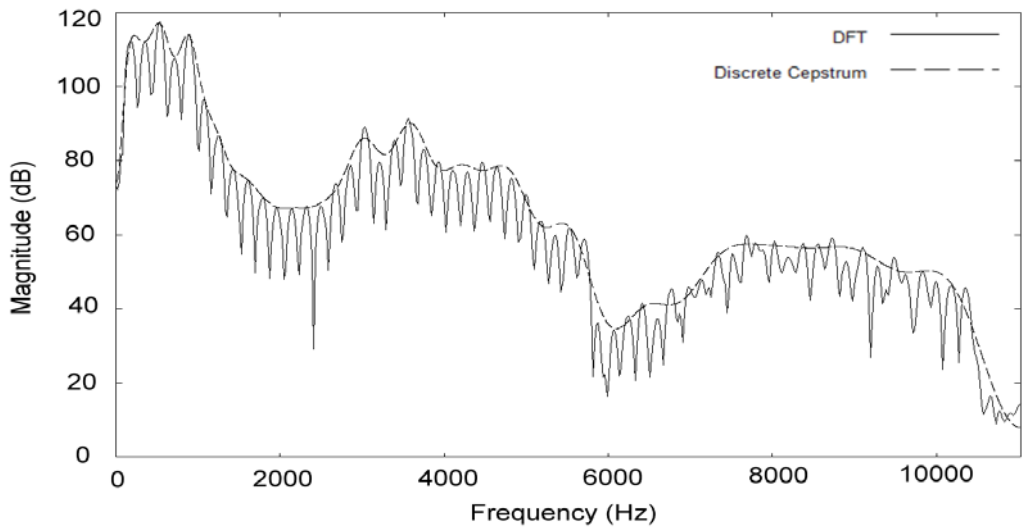


Figure 4. A spectral envelope computed with 40 regularized discrete cepstrum coefficients.

3. Selection of Spectral Peaks

Note that the discrete cepstrum coefficients are obtained by minimizing the squared errors between the selected spectral peaks, a_k, k=1, 2, …, L, and S(f). Therefore, selecting appropriate spectral peaks from a DFT spectrum is an important preprocessing step. The simplest selection method, *i.e.* locating and selecting all the spectral peaks on the spectrum as the final selected peaks, leads to the approximated spectral envelope being very bad and having a large approximation error. When such bad spectral envelopes are used to transform voice signals, the output obtained will suffer significant voice-quality degradation.

Therefore, we studied this problem and found that the concept of MVF (maximum voiced frequency) proposed in HNM (harmonic-plus-noise model) (Stylianou, 1996; Stylianou, 2005) is utilizable. The MVF of a DFT spectrum is searched by testing the sharpness of the spectral peaks one after another. After some low-frequency spectral peaks pass the test, eventually no more spectral peak can pass the test. Then, the frequency of the last spectral peak passing the test is defined to be the MVF. In this paper, we first detect if a signal frame is voiced or

unvoiced. If it is detected to be voiced, the frame is further searched for the MVF value, *fv*, through the searching method proposed by Stylianou (1996). According to *fv*, the DFT spectrum of the frame is split into the lower-frequency harmonic part and the higher-frequency noise part. Then, for the harmonic part, the first spectral peak of a frequency within the range ($0.5\times F_0$, $1.5\times F_0$), where F_0 is the detected fundamental frequency, is searched for. Let the obtained frequency and amplitude be f_1 and a_1. Next, the second spectral peak of a frequency within the range ($f_1+0.5\times F_0$, $f_1+1.5\times F_0$) is searched for, and the results will be the frequency f_2 and amplitude a_2. When going on in this manner, we can find the frequencies and amplitudes of the other spectral peaks within the harmonic part. Sometimes, it may occur that no spectral peak is found within a designated frequency range. In this situation, we will right shift the frequency range, *i.e.* adding $0.5\times F_0$, and try to find again.

For the noise part of a voiced frame, we think the searching method explained above for the harmonic part cannot be adopted. Note that the harmonic structure becomes obscure in the noise part, and the frequency gaps between adjacent peaks become randomly varied. As an example, inspect the DFT spectrum curve beyond 5,800Hz in Figure 4. Therefore, we adopt another method to find the spectral peaks for the noise part. In this method, a smooth spectral curve is obtained first by truncating the real-cepstrum coefficients outside the leading 30 ones, and transforming (via DFT) the resulting real-cepstrum sequence back to the spectrum domain. Then, each spectral speak within the noise part is located and its amplitude is checked. It will be selected if its amplitude is higher than the height of the smooth spectral curve at the peak's frequency. As for an unvoiced frame, the method just explained can still be applied. This is because such a frame's MVF can be directly set to 0Hz and its spectrum can be viewed as all in the noise part. When applying the spectral peak selection method explained above, we may obtain a typical result shown in Figure 5. In this figure, each occurrence of plus-sign, +, represents a selected spectral peak.

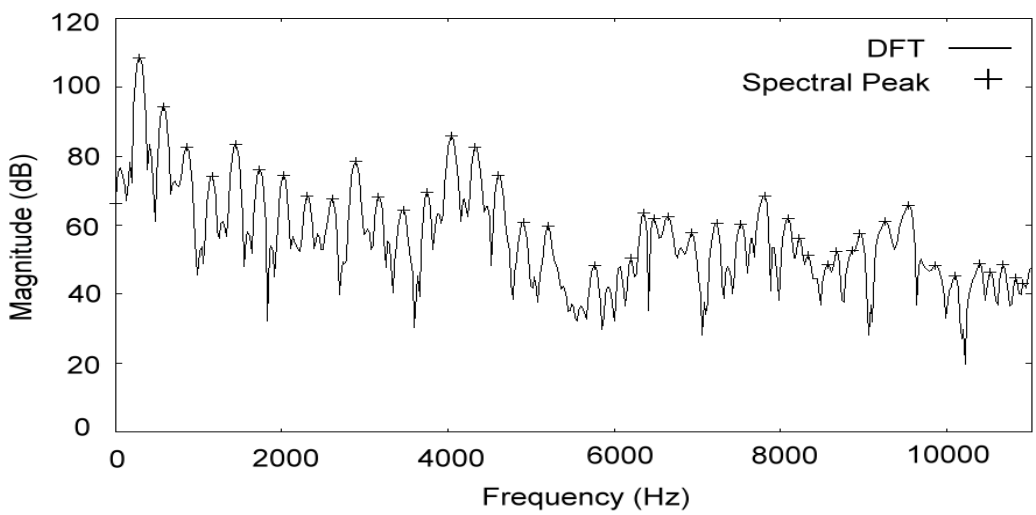


Figure 5. A typical result for spectral peak selection.

4. Order of Discrete Cepstrum and Frequency Axis Scaling

4.1 Order of Discrete Cepstrum

What value should be set for the parameter, *p*, for the order of a discrete cepstrum? If a smaller value (*e.g.* 10) is set, the approximated spectral envelope curve will be overly smooth, and the approximation error will become considerably large. On the other hand, when a larger value is set for *p*, the number of computation operations needed to solve Equation (10) will rise at a cubic rate. Nevertheless, sufficient accuracy in spectral-envelope approximation is very important to prevent quality-degradation and inconsistent-timbre from occurring. For setting the value of *p*, Shiga and King (2004) argue that *p* should have a value from the range (48, 64) to obtain a sufficiently accurate approximation of spectral envelope.

Here, we study the correlation between approximation errors and order numbers, *p*, experimentally. The approximation error is computed as

$$Es = \frac{1}{Nr}\sum_{t=0}^{Nr\text{-}1}\left[\frac{1}{L(t)}\sum_{k=1}^{L(t)}\left|20\log_{10} a_k^t - 20\log_{10} S(t, f_k)\right|\right] \tag{11}$$

where *Nr* is the total number of frames, a_k^t denotes the amplitude of the *k*-th spectral peak in the *t*-th frame, and S(*t*, *f*) represents the approximated spectral envelope for the *t*-th fame. Here, 375 Mandarin sentences, consisting of 2,925 syllables, were recorded from a male speaker and used as the testing data. According to our measurement results, the approximation error, *Es*, will decrease considerably as the order number *p* is increased from 5 to 30. Thereafter, *Es* will decrease only slightly as *p* is further increased to 50. This trend is illustrated by the curve in Figure 6. Therefore, we decide here to adopt 40 as the order number for *p* in order to obtain sufficiently accurate spectral-envelop approximation.

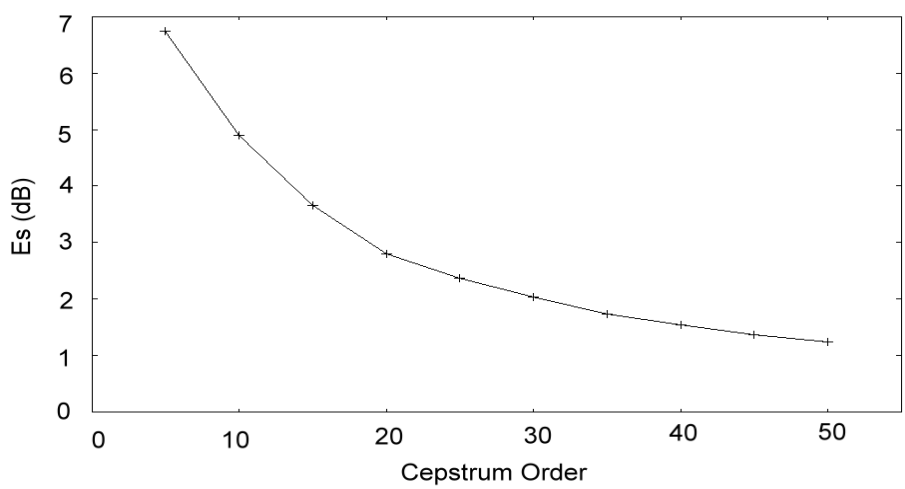
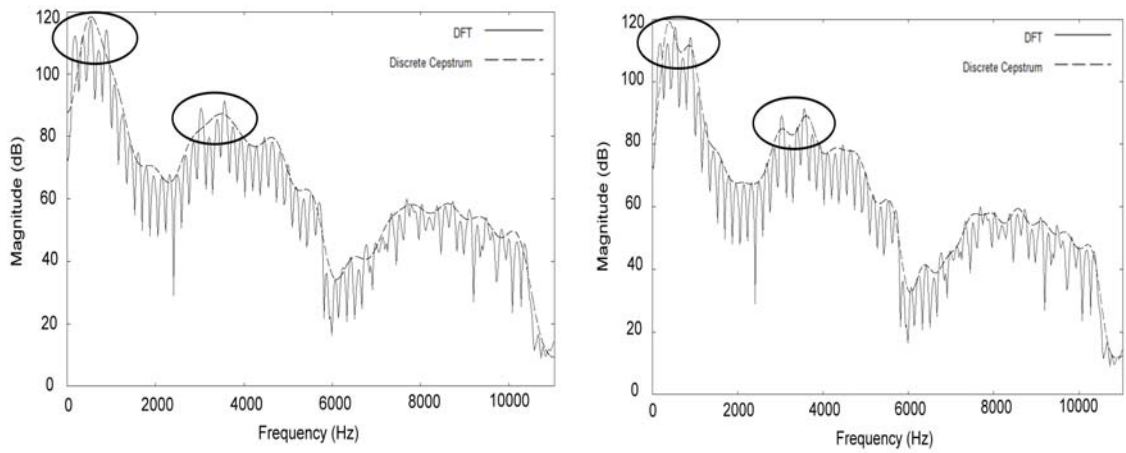


Figure 6. Approximation errors versus discrete cepstrum orders.

4.2 Frequency Axis Scaling

Through use of a larger order number, *e.g.* 40, the global approximation error of a frame's spectral envelope can now be controlled. Nevertheless, local approximation errors that are large and cannot be ignored can still be found. For example, the spectral envelope in Figure 7(a) is obtained by approximating with 30 discrete cepstrum coefficients, and it has two significant local approximation errors as circled. If the value of the order number p is increased to 40, the spectral envelope will become the one shown in Figure 7(b). Although the local approximation errors are somewhat reduced, they are still significant and cannot be ignored.



(a) Cepstrum order set to 30 (b) Cepstrum order set to 40

Figure 7. Spectral envelopes approximated in the linear frequency scale.

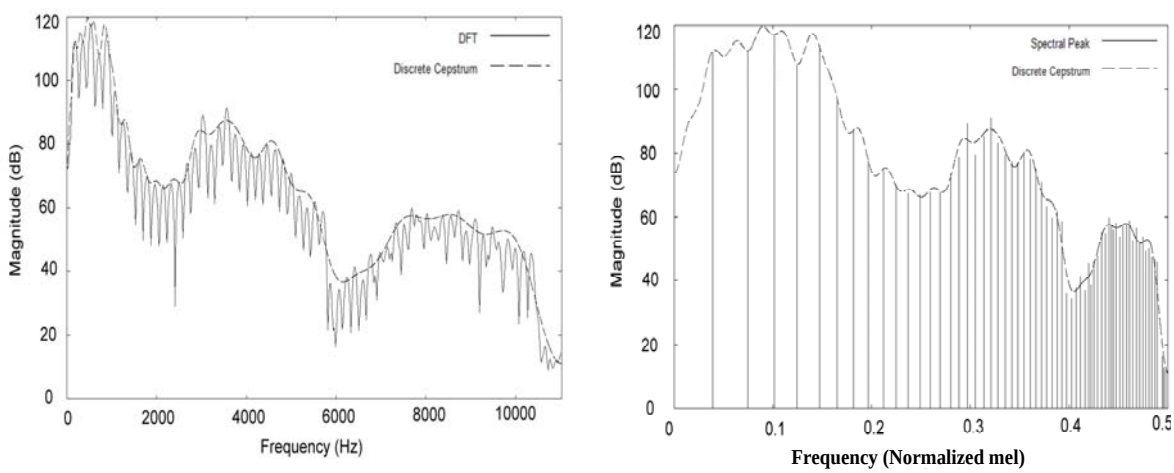
Consider the local approximation error around 1,000Hz, as seen in Figures 7(a) and 7(b). From the DFT spectrum, it is known that the pitch is low (*i.e.* it is a male's pitch) and the frequency gap between two adjacent spectral peaks is small. Under such a situation, the amplitudes of the third and fifth spectral peaks show rapid growth that is higher than the nearby peaks. Such rapid change of spectral envelope is very hard to approximate. To solve this situation, a conventional idea is to nonlinearly scale the frequency axis to enlarge the frequency gaps between low-frequency spectral peaks. Therefore, when a spectral peak of a frequency f_k is detected, its frequency value will be scaled to $\hat{f}_k$ according to the formula,

$$\hat{f}_k = \frac{1}{2} \cdot \frac{scl(f_k \times F_s)}{scl(0.5 \times F_s)} \quad , \tag{12}$$

where $scl(\cdot)$ represents a frequency-scale conversion function and F_s is the sampling frequency. After frequency scaling, the spectral peaks' frequencies and amplitudes are then taken into Equation (10) to compute the optimal discrete cepstrum coefficients. The step of replacing f_k

with $\hat{f}_k$ also implies that the computed discrete cepstrum coefficients must be used in the scaled frequency axis instead of the original axis. That is, for a linear and normalized frequency *f*, its envelope magnitude, *S*(*f*), should be computed by scaling *f* into $\hat{f}$ first then taking $\hat{f}$ into Equation (3).

For nonlinear frequency-axis scaling, mel and Bark frequency scales are the most famous (O'Shaughnessy, 2000). If the frequency conversion function, *scl*(·), adopted is a mel-frequency conversion, the spectral envelope shown in Figure 7(b) will be changed to the one shown in Figure 8(a). The major difference is that the lower-frequency spectral peaks in Figure 8(a) are now all passed by the approximated spectral envelope curve. Nevertheless, the local approximation error around 3,000Hz is still noticeable. In addition, the spectral envelope curve in Figure 8(a) shows much stronger vibration near the lower-frequency end than the curve in Figure 7(b). This stronger vibration can be seen in more detail when we redraw the approximated spectral envelope curve with a mel-frequency horizontal axis as shown in Figure 8(b). This phenomenon of over-vibration is thought to be due to the mel-frequency conversion that widens the frequency-scale at the low frequency end. According to the observed stronger vibration for mel-frequency conversion, we think much stronger vibration will occur if we adopt the Bark-frequency conversion for *scl*(·). This is because Bark-frequency conversion will have the frequency-scale at the low frequency end being widened more than that widened by the mel-frequency conversion.



(a) linear-frequency horizontal axis (b) mel-frequency horizontal axis

Figure 8. Spectral envelopes approximated in mel-frequency scale.

Therefore, we were motivated to design a frequency conversion function in the hope of eliminating the phenomenon of over-vibration at the low frequency end and of reducing the local approximation error around 3,000Hz. After several attempts at trying function-design

and inspecting the approximated spectral envelope curves, we finally found a better frequency conversion function:

$$scl(f) = \log(1 + \frac{f}{1{,}750}) \tag{13}$$

where f is in the unit Hz. This conversion function will have the scaled frequency value, $\hat{f}_k$, growing more slowly with f_k at the low frequency end when it is used as the $scl(\cdot)$ function for Equation (12). The three curves shown in Figure 9 are obtained by taking Bark, mel, and our frequency conversions, respectively, as the $scl(\cdot)$ function for Equation (12). From Figure 9, it can be seen that our frequency conversion, as given in Equation (13), can indeed grow the scaled frequency $\hat{f}$ more slowly with the linear frequency f. Via the frequency conversion function of Equation (13), the approximated spectral envelope in Figure 8(a) will become the one drawn in Figure 4. According to the spectral envelope obtained in Figure 4, it can be said that the frequency conversion function proposed can indeed eliminate the over-vibration phenomenon at the low frequency end, and reduce the local approximation error around 3,000Hz. The reducing of the local approximation error we think is due to the increased vibrating capability around 3,000Hz by using the proposed frequency conversion instead of the mel-frequency conversion.

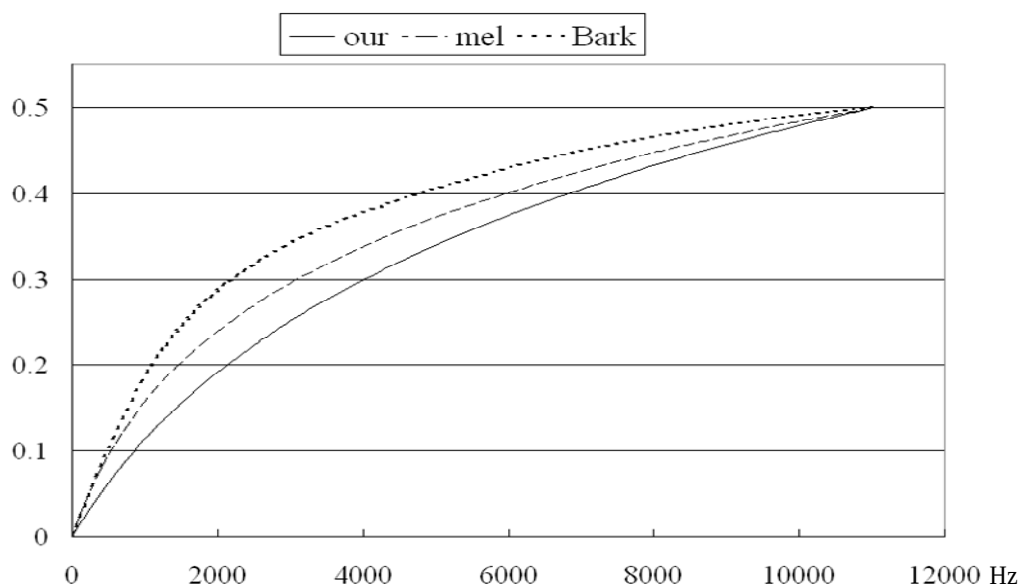


Figure 9. Three curves of scaled frequencies by using Bark, mel, and our frequency conversion functions, respectively.

4.3 Approximation Error Comparison

One may ask if our frequency conversion function is only better than mel-frequency conversion for certain signal frames. Therefore, we decided to compare the approximation errors of the two frequency conversions in the four frequency ranges, *i.e.* 0 ~ 2,000Hz, 0 ~ 4,000Hz, 0 ~ 6,000Hz, and 0 ~ 11,025Hz. Here, the approximation error is still measured by the formula in Equation (11). Nevertheless, the number of spectral peaks, L, is dynamically checked for each frame to ensure that only the spectral peaks of frequencies within the currently concerned frequency range are counted. Here, 375 Mandarin sentences, consisting of 2,925 syllables, as mentioned in Section 4.1 were used as the testing data. After all of the frames of the data are processed, the approximation errors measured in different frequency ranges and different discrete cepstrum orders are illustrated in Figure 10.

Inspecting the error curve in Figure 10, it can be seen that across the cepstrum order-numbers from 30 to 50, our frequency conversion and the mel-frequency conversion have almost same approximation errors in the frequency range, 0 ~ 2,000Hz. Nevertheless, in the other three frequency ranges, our conversion function will apparently obtain smaller approximation errors for different cepstrum-order numbers. This decreasing of approximation error becomes more apparent as the frequency range becomes wider.

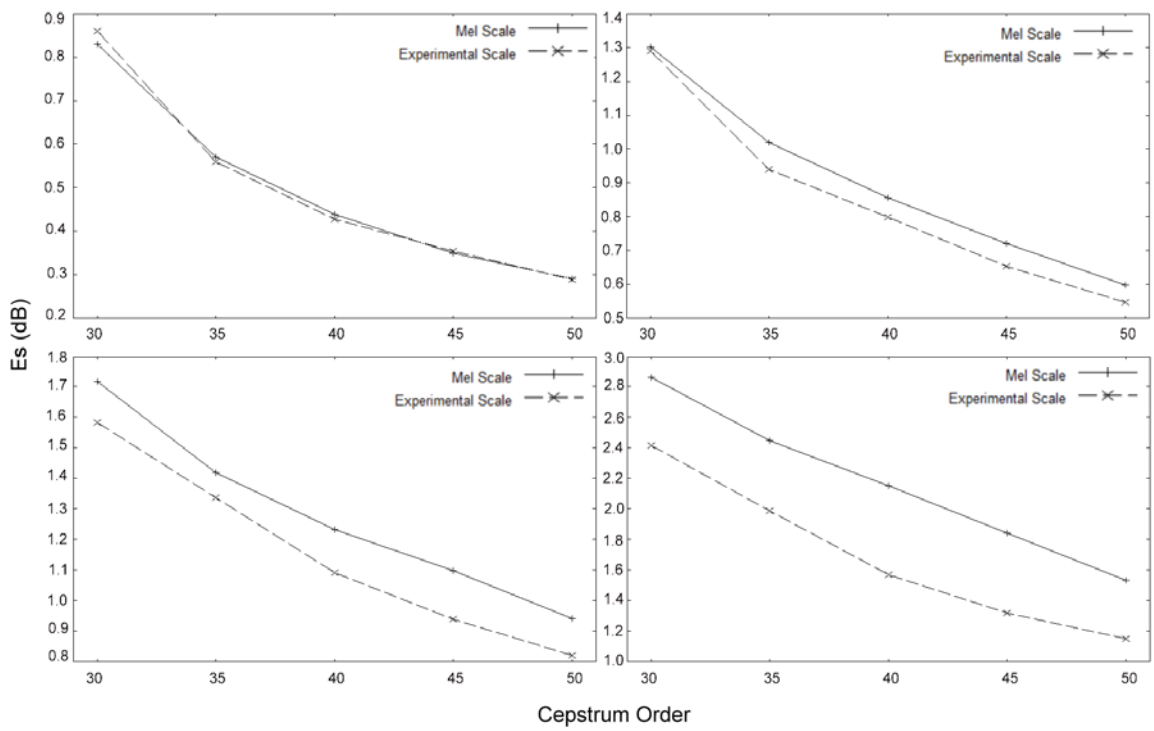


Figure 10. Approximation errors measured for our frequency conversion and mel-frequency conversion in the four frequency ranges: 0 ~ 2,000Hz (upper left), 0 ~ 4,000Hz (upper right), 0 ~ 6,000Hz (bottom left), and 0 ~ 11,025Hz (bottom right).

5. An Example Application: Voice Transformation

Here, voice transformation means to change the timbre of an input voice to a different timbre. For example, it could be changing the timbre of a female adult into the timbre of a male adult or a child. In the past, phase vocoder was a frequently used technique to transform voice timbre (Moore, 1990; Dolson, 1986). Nevertheless, the basic transformation method of phase vocoder cannot support independent control of spectral-envelope scaling and pitch shifting. Therefore, we decided to apply the technique of additive synthesis developed for computer music synthesis (Moore, 1990) and the signal model of HNM (harmonic-plus-noise model) (Stylianou, 1996; Stylianou, 2005). In other words, we will use the estimated spectral envelope to do spectral-envelope scaling. Then, we will place harmonic partials and noise sinusoids under the scaled spectral envelope according to the pitch shifting requirement. In this manner, spectral envelope scaling and pitch shifting can be performed independently.

We have practically implemented a voice transformation system. Its main processing flow is as shown in Figure 11. In this system, the input voice is first sliced into a sequence of frames. The frame width is 512 sample points (23.2ms) and the frame shift is 256 points (11.6ms) under the sampling frequency, 22,050Hz. For each frame, the processing flow shown in Figure 2 is executed to estimate its spectral envelope with 40 discrete-cepstrum coefficients. The other three blocks, "spectral envelope scaling," "pitch shifting," and "signal re-synthesizing," will be explained in the following subsections. We tested the processing speed of this system on a notebook computer with an Intel T5600 1.83GHz CPU, and found that it will consume 0.75 sec. of CPU time on average to transform 1 sec. of voice signal.

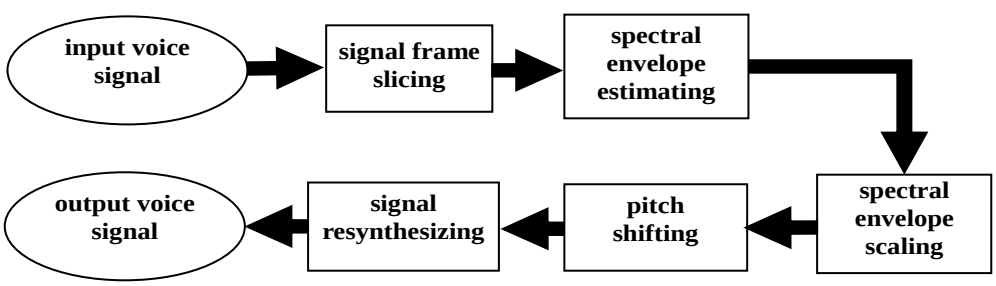
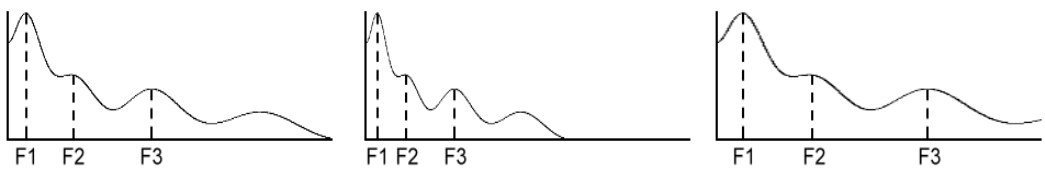


Figure 11. Main processing flow of the voice transformation system.

5.1 Spectral Envelope Scaling

Scaling of a spectral envelope can be performed in two possible directions. One direction is to shrink the spectral envelope to lower formant frequencies in order to obtain a male adult's timbre. The other direction is to extend the spectral envelope to raise formant frequencies in order to obtain a child's timbre. For example, inspect the spectral envelopes drawn in Figure 12. The curve drawn in Figure 12(a) represents the originally estimated spectral envelope,

Vo(*f*). If this spectral envelope is shrunk and the shrinking rate is 0.7, the resulting envelope will be the one drawn in Figure 12(b). Apparently, the formant frequencies, F1, F2, and F3, are all lowered. Let the spectral envelope in Figure 12(b) be denoted by *Vs*(*f*). Then, it is simple to derive that $Vs(f) = Vo(\frac{10}{7}f)$. On the other hand, if the spectral envelope in Figure 12(a) is extended and the extending rate is 10/7, the resulting envelope will be the one drawn in Figure 12(c). Apparently, the formant frequencies, F1, F2, and F3, are all raised. Let the spectral envelope in Figure 12(c) be denoted with *Ve*(*f*). Then, it can be derived that $Ve(f) = Vo(\frac{7}{10}f)$.



(a) original envelope (b) shrunk envelope (c) extended envelope

Figure 12. The scaling of an example spectral envelope.

5.2 Pitch Shifting

After a frame's spectral envelope is shrunk (or extended), we can use *Vs*(*f*) (or *Ve*(*f*)) to determine a new set of harmonic partials and noise sinusoids. Suppose that the original pitch frequency of the *i*-th frame is 180Hz and that we intend to tune its pitch to 250Hz. Although the original pitch must be used in the block "spectral envelope estimating" of Figure 11, it is not used for pitch shifting and signal re-synthesizing. This is because we just need *Vs*(*f*) (or *Ve*(*f*)) to determine the amplitudes of the new harmonic partials and noise sinusoids. For example, the new harmonic structure of a voiced frame may look like the one shown in Figure 13. According to a given MVF, we can place new harmonic partials into the frequency range below MVF, and place new noise sinusoids into the frequency range above MVF.

In detail, the frequencies of the new harmonic partials are set as f_1^i=250, f_2^i=500, f_3^i=750, *etc*. As to their amplitudes, the spectral envelope, *Vs*(*f*) (or *Ve*(*f*)), is evaluated at the targeted frequencies. That is, their amplitudes are set to $a_1^i = Vs(250)$, $a_2^i = Vs(500)$, $a_3^i = Vs(750)$, *etc*. Besides frequency and amplitude, the other parameter of a harmonic partial is phase. Nevertheless, the phase values of the harmonic partials are not a concern here because they will not be used in the signal re-synthesizing step. For the noise sinusoids, any two adjacent ones are placed 100Hz apart as shown in Figure 13. After being placed, each noise sinusoid's amplitude can be determined according its frequency position.

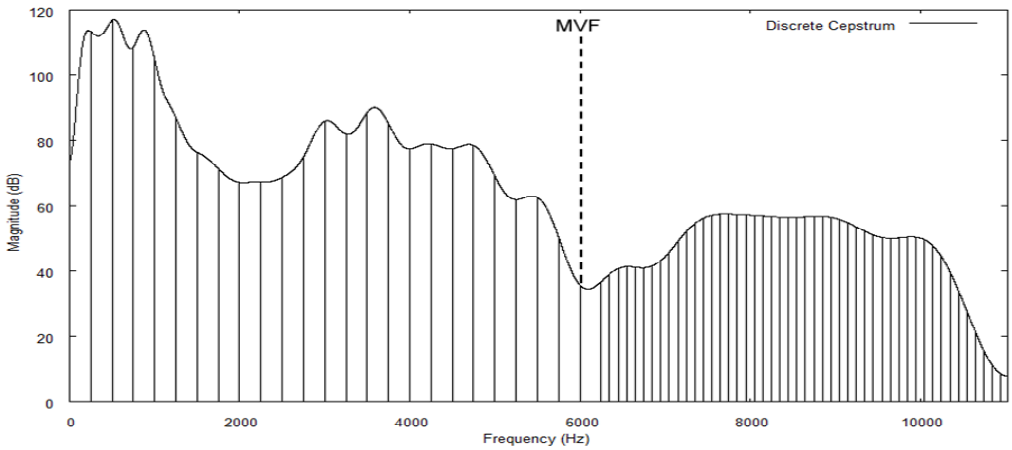


Figure 13. An example harmonic structure for a frame tuned to have the pitch, 250Hz.

5.3 Signal Re-synthesizing

Here, the signal model, HNM, proposed by Stylianou (1996) is adopted for signal re-synthesis. In HNM, the spectrum of a voice frame is divided into the lower-frequency harmonic part and the higher-frequency noise part. The frequency that the two parts are divided according to is called the MVF. In the original work (Stylianou, 1996), a method is provided to dynamically detect each frame's MVF. Here, to simplify the synthesis processing, we just use the static MVF value, 6,000Hz, across all voiced frames. As an example, the spectral envelope in Figure 13 is divided into the harmonic and noise parts according to the MVF, 6,000Hz.

Suppose the *i*-th and (*i*+1)-th frames are both voiced and have L^i and L^{i+1} harmonic partials, respectively, after pitch shifting. To synthesize a signal sample for the *t*-th sampling point between the *i*-th and (*i*+1)-th frames, we first derive the frequencies and amplitudes of the harmonic partials for this sampling point with linear interpolation. That is,

$$\begin{aligned} f_k(t) &= f_k^i + \frac{f_k^{i+1} - f_k^i}{N} t, \quad k = 1, 2, ..., L, \\ a_k(t) &= a_k^i + \frac{a_k^{i+1} - a_k^i}{N} t, \quad k = 1, 2, ..., L \end{aligned} \tag{14}$$

where *N* is the number of sampling points between two adjacent frames, and *L* is the larger one of L^i and L^{i+1}. Here, we directly set $a_k^i = 0$, $k = L^i + 1, \ldots, L^{i+1}$, if L^i is less than L^{i+1}. Then, the harmonic signal, $h(t)$, for the *t*-th sampling point is computed as

$$\begin{aligned} h(t) &= \sum_{k=1}^{L} a_k(t) \cdot \cos(\phi_k(t)), \quad 0 \le t < N, \\ \phi_k(t) &= \phi_k(t-1) + 2\pi \cdot f_k(t) / 22,050 \end{aligned} \tag{15}$$

where $\phi_k(t)$ denotes the accumulated phase on time t for the k-th harmonic partial and 22,050 is the sampling frequency. $\phi_k(-1)$ is equal to $\phi_k(N\text{-}1)$ of the last frame to keep continuity of phase. If $i = 0$, *i.e.* there is no last frame, the value of $\phi_k(-1)$ is set randomly.

To synthesize the noise signal for the t-th sampling point, we apply a method mentioned in Stylianou (1996). That is, synthesize the noise signal as the summation of the sinusoids whose frequencies are larger than MVF, fixed (not affected by the pitch frequency) and are 100Hz apart. The amplitudes of the sinusoids are, however, varied with time. Let KL = MVF / 100 and KU = 22,050 / 100. Then, the noise signal, $g(t)$, is synthesized as:

$$\begin{aligned} g(t) &= \sum_{k=KL}^{KU} b_k(t) \cdot \cos(\psi_k(t)), \quad 0 \le t < N \ , \\ \psi_k(t) &= \psi_k(t-1) + 2\pi \cdot k \cdot 100 / 22,050 \end{aligned} \tag{16}$$

where $b_k(t)$ and $\psi_k(t)$ denote the amplitude and accumulated phase, respectively, of the k-th sinusoid at the time point t. The value of $b_k(t)$ is obtained, with linear interpolation, with a formula similar to the one in Equation (15). Finally, the signal sample for the t-th sampling point is synthesized as $h(t)$ plus $g(t)$.

5.4 Perception Testing

To evaluate the performance of our voice transformation system, we first recorded three sentences each from a female adult and a male adult. For the female source voice, we set the envelope shrinking rate to 0.8 and set the pitch shifting rate to 0.6 in order to transform into a male timbre. For the male source voice, we set the envelope extending rate to 1.2 and set the pitch shifting rate to 2.1 in order to transform into a female timbre. The source voices and their transformed voices can be accessed at http://guhy.csie.ntust.edu.tw/dcc/vt.html.

We invited thirteen persons to participate in the perception tests. The first type of perception test conducted was for evaluating the timbre recognizability of the source and transformed voices. That is, each participant was asked how similar the timbre of a played voice was to a female (or a male). Each participant was requested to give a score between 1 and 5 to indicate how similar the heard timbre seemed. As a result, we obtained the averaged scores and standard deviations shown in Table 1. According to the average scores of the transformed timbres, *i.e.* 4.85 and 4.36, it can be said that the transformed voice from our system will have sufficiently high timbre recognizability. In addition, when comparing the score differences between the original and transformed voices, *i.e.* 0.10 (4.95-4.85) vs. 0.37 (4.73-4.36), we find that the female source voice will induce less recognizability degradation than the male source voice.

Table 1. Perception test results for timbre recognizability.

Source voice \ Timbre		Original voice	Transformed voice
Female	Avg. score	4.95	4.85
	Std. deviation	0.15	0.23
Male	Avg. score	4.73	4.36
	Std. deviation	0.45	0.48

The second type of perception test is for evaluating the voice qualities of the source and transformed voices. The same participants were asked what level the quality of a played voice was at. Each participant was requested to give a score between 1 and 5 to indicate the quality level of the played voice. After this type of perception test was conducted, we averaged the scores collected and computed their standard deviation. The results are shown in Table 2. According to the score differences, 0.67 (4.38 – 3.71) and 0.82 (4.00 – 3.18), it can be said that our system will inevitably induce a perceivable degradation of voice quality for the transformed voices no matter whether the source voice is uttered by a female or male. One of the possible reasons is that the pitch frequencies of some frames are wrongly detected, which causes their spectral envelopes to be approximated with noticeable errors.

Table 2. Perception test results for voice quality.

Source voice \ Timbre		Original voice	Transformed voice
Female	Avg. score	4.38	3.71
	Std. deviation	0.39	0.53
Male	Avg. score	4.00	3.18
	Std. deviation	0.74	0.72

6. Concluding Remarks

The concept of approximating spectral envelope with discrete cepstrum was proposed several years ago. There are, however, three problems that must be solved for practical implementation. The first problem is the regularization of the discrete cepstrum coefficients to prevent a radical vibrating envelope curve from occurring. This problem has been solved already by previous researchers. In this paper, we tried to solve the other two problems, *i.e.* selecting appropriate spectral speaks and finding a better frequency axis scale. For selecting spectral peaks, we apply the concept of HNM to divide a spectrum into the lower frequency

harmonic part and the higher frequency noise part. Then, we find the spectral peaks in the harmonic part according to the detected pitch frequency and screen the spectral peaks in the noise part according to a cepstrum smoothed spectral curve. As to the problem of frequency axis scaling, we found that the spectral envelope approximated via the mel or Bark-frequency conversion still has noticeable local approximation errors. Therefore, after some attempts at scaling-function design, we propose a better frequency conversion function that can reduce the local approximation errors significantly. Then, applying the solutions to the three problems, we construct a spectral envelope estimation scheme.

In addition, we built a voice transformation system on the proposed spectral envelope estimation scheme as an example application. This system follows the steps, spectral envelope estimating, spectral envelope scaling, pitch shifting, and signal re-synthesizing, to transform an input voice into an output voice that is of a very different timbre, *i.e.* the perceived gender and age of the voice can both be changed. To evaluate the performance of this system, we conducted perception tests. The averaged scores from 13 participants show that our system can indeed achieve the function of timbre transformation. In the future, we will apply the proposed spectral envelope estimation scheme to study another kind of voice transformation problem. That is, we will convert the voice of a specific person into the voice of another specific person.

Acknowledgments

This study is supported by National Science Council under the contract number, NSC 98-2221-E-011-116.

References

Cappé, O., & Moulines, E. (1996). Regularization techniques for discrete cepstrum estimation. *IEEE Signal Processing Letters*, 3(4), 100-102.

Dolson, M. (1986). The phase vocoder: A tutorial. *Computer Music Journal*, 10(4), 14-27.

Galas, T., & Rodet, X. (1990). An improved cepstral method for deconvolution of source filter systems with discrete spectra: Application to musical sound signals. *Proceedings of International Computer Music Conference*, Glasgow, Scotland.

Gu, H. Y., Chang, H. F., & Wu, J. H. (2004). A pitch-contour normalization method following Zhao's pitch scale and its application. *Proceedings of ROCLING XVI*, Taipei, Taiwan. 325-334. (in Chinese)

Imai, S., & Abe, Y. (1979). Spectral envelope extraction by improved cepstral method. *Electron. and Commun. in Japan*, 62-A(4), 10-17. (in Japanese)

Kawahara, H., Masuda-katsuse, I., & Cheveign, A. De. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an

instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. *Speech Communication*, 27, 187-207.

Kim, H. Y., *et al.* (1998). Pitch detection with average magnitude difference function using adaptive threshold algorithm for estimating shimmer and jitter. *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Hong Kong, China.

Moore, F. R. (1990). *Elements of computer music*. Prentice-Hall, Englewood Cliffs, NJ, U.S.A.

O'Shaughnessy, D. (2000). *Speech communications: human and machine*, IEEE Press, Piscataway, NJ, U.S.A.

Robel, A., & Rodet, X. (2005). Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation. *Proceedings of International Conference on Digital Audio Effects*, Madrid, Spain.

Schwarz, D., & Rodet, X. (1999). Spectral envelope estimation and representation for sound analysis-synthesis. *Proceedings of International Computer Music Conference*, Beijing, China.

Shiga, Y., & King, S. (2004). Estimating detailed spectral envelopes using articulatory clustering. *Proceedings of International Conference on Spoken Language Processing* (ICSLP2004), Jeju, Korea.

Stylianou, Y. (1996). *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.

Stylianou, Y. (2005). Modeling speech based on harmonic plus noise models. in *Nonlinear Speech Modeling and Applications*, eds. Chollet, G., *et al*., Springer-Verlag, Berlin, Germany.