A Discrete-cepstrum Based Spectral-envelope Estimation Scheme with Improvements

Hung-Yan Gu and Sung-Feng Tsai
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
Taipei, Taiwan
e-mail: {guhy, M9615069}@mail.ntust.edu.tw

***Abstract*—Approximating spectral envelope with regularized discrete-cepstrum coefficients was proposed by previous researchers. In this paper, we study two problems encountered in practice when adopting this approach. The first is which spectral peaks should be selected, and the second is what 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 method 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, to be much better than the original scheme. In addition, we have applied this scheme to build a voice-timbre transformation system. This system demonstrates that the proposed estimation scheme is very effective.**

Keywords-spectral envelope; discrete cepstrum; spectral peak; frequency axis scaling; voice timbre transformation

I. INTRODUCTION

Here, a spectral envelope is meant a magnitude-spectrum envelope. Some methods for estimating a spectral envelope had been proposed previously. For example, in LPC based methods [1, 2], 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 all-pole model is usually not accurate enough. Besides LPC, a cepstrum based method for estimating a spectral envelope was proposed by Imai and Abe [3, 4]. They call this method, true envelope estimation. As our opinion, this method is good but lack efficiency because a lot of computations are required. Similarly, the method, STRAIGHT, proposed by Kawahara, *et al*. [5], is very accurate in its estimated spectral envelope. Nevertheless, it also requires a large amount of computations and cannot be used to implement real-time systems currently. On the other hand, Galas and Rodet proposed the concept of discrete cepstrum [6], and designed a feasible estimation method with this concept. Later, Cappé and Moulines improve this estimation method by adding a regularization technique to prevent unstable vibrating of envelope curve from occurring [7]. 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 Fig. 1. 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 latter in the block, "spectral peaks selection". Here, a method combining autocorrelation function and AMDF is adopted to detect a frame's fundamental frequency [1, 8]. Next, the frame is Hanning windowed, and appended with zeros to form a signal segment of 1,024 samples. This segment is then transformed to frequency domain with FFT to obtain its magnitude spectrum. Then, this magnitude spectrum is inputted to the block "spectral peaks selection" to select spectral peaks 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" adopts an envelope-approximation criterion [7] to compute discrete cepstrum coefficients (DCCs) according to the selected and mapped spectral peaks.

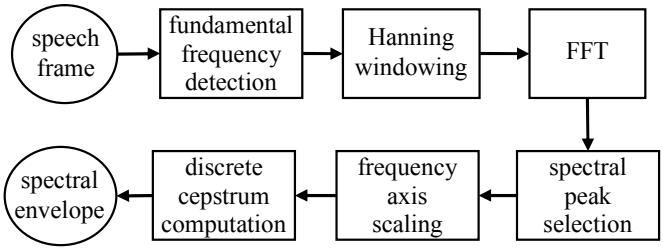


Fig. 1. Main flow of the spectral-envelope estimation scheme.

In Fig. 1, discrete-cepstrum computation is the major block, and it is already solved by other researchers [7]. Nevertheless, the blocks, spectral-peak selection and frequency-axis scaling, still play important roles. When inappropriate peaks are selected or frequency-axis is not scaled appropriately, the approximated spectral envelope will noticeably deviate from the true envelope. Therefore, we studied these two blocks' problems here, and the results are presented in Section 3 and 4, respectively. As to discrete cepstrum, its computation and regularization will be briefly reviewed in Section 2. In Section 5, the proposed scheme is practically verified by applying the scheme to build a voice timbre transformation system.

II. DISCRETE CEPSTRUM BASED SPECTRAL ENVELOPE

A. Discrete Cepstrum

To obtained cepstrum coefficients c_0, c_1, …, c_{N-1}, where N is the length of a signal frame, the conventional method is to transform the logarithmic magnitude-spectrum, $\log|X(k)|$, with

inverse DFT (IDFT). Then, the logarithmic magnitude-spectrum can be computed with the cepstrum coefficients as

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

If most terms at the right side of (1) are cancelled except the leading 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 (1) 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

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

Based on (2), some researchers proposed to approximate the spectral envelope of log$|X(f)|$ with log $S(f)$. Nevertheless, the coefficients, c_n, in (2) 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 between 0 and 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. In matrix form, the optimal values of the DCCs is derived by previous researchers [6, 7] to be

$$C = (M^{\mathrm{T}} \cdot M)^{-1} \cdot M^{\mathrm{T}} \cdot A \tag{3}$$

where $A = [\log(a_1), \log(a_2), \ldots, \log(a_L)]^{\mathrm{T}}$, $C = [c_0, c_1, \ldots, c_p]^{\mathrm{T}}$, and

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

B. Regularization of Discrete Cepstrum

When (3) is used to derived DCCs, the spectral envelope computed according to (2) may vibrate radically and have very large approximation error at some frequencies slightly away the selected spectral-peak frequencies, f_k. This is because the direct estimation method (i.e. Eq. (3)) 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. Therefore, Cappé and Moulines proposed a regularization technique to prevent such radical vibrations from occurring [7]. They add a curve-sharpness penalty term, i.e.

$$R(S(f)) = \int_0^{\pi} \left[\frac{d}{df} S(f) \right]^2 df \,, \tag{4}$$

to the approximation error calculation equation, and the resulted equation for deriving DCCs becomes

$$C = (M^{\mathrm{T}} M \; + \; \lambda U)^{-1} \cdot M^{\mathrm{T}} A \quad . \tag{5}$$

where λ is a weighting parameter (suggested value is around 0.0001), and

$$U = 8\pi^2 \begin{bmatrix} 0 & & & 0 \\ & 1^2 & & \\ & & \ddots & \\ 0 & & & p^2 \end{bmatrix}. \tag{6}$$

III. Selection of Spectral Peaks

DCCs are derived by minimizing the summation of 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. Consider a simplest selection method, i.e. locate and select all the spectral peaks on the spectrum as the final selected peaks. In this case, the approximated spectral envelope would be very bad and of 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) [9, 10] 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, it will eventually occur that 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, by using the searching method proposed by Stylianou [10]. 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 let the results be f_2 (frequency) and a_2 (amplitude). 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, 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. For an example, inspect the DFT spectrum curve beyond 5,800Hz in Fig. 2. Therefore, we adopt another method to find the spectral peaks for the noise part. In this method, a smoothed spectral curve is computed first by truncating the real-cepstrum coefficients outside the leading 30 ones, and transforming (via DFT) the resulted real-cepstrum

sequence back to the spectrum domain. Then, each spectral speak within the noise part of $\log|X(f)|$ is located and checked again its amplitude. It will be selected if its amplitude is higher than the height of the smoothed spectral curve at the peak's frequency. As for an unvoiced frame, the method just explained can still be applied. This is because such frame's MVF can be directly set to 0Hz, i.e. its spectrum is viewed as all in the noise part. When applying the spectral peak selecting method explained above, we may obtain a typical result as shown in Fig. 2. In this figure, each occurrence of plus-sign, +, represents a selected spectral peak.

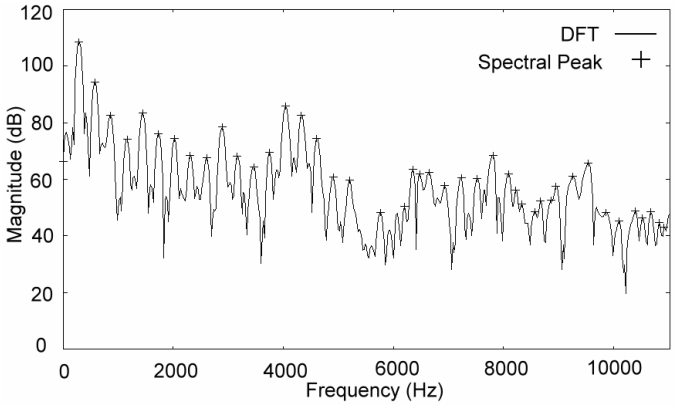


Fig. 2. A typical result for spectral peak selection.

IV. Frequency Axis Scaling

A. Frequency-scale Conversion

By using larger order number, *e.g.* 40, the global approximation error of a frame's spectral envelope can be under control. Nevertheless, local approximation errors that are large enough and cannot be ignored may still be found. For example, the spectral envelope in Fig. 3 is obtained by approximating with 40 DCCs, and has two significant local approximation errors as circled.

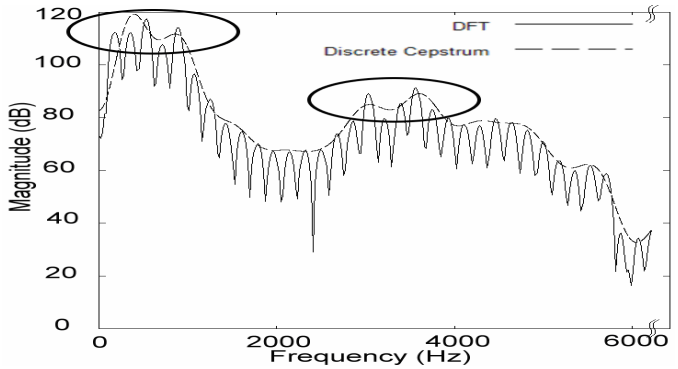


Fig. 3. Envelope approximated with 40 DCCs in linear freq. scale.

Consider the local approximation error around 1,000Hz as seen in Fig. 3. From the DFT spectrum, it is known the fundamental frequency is low and the frequency gap between two adjacent spectral peaks is small. Under such situation, the amplitudes of the third and fifth spectral peaks show rapidly growing 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 lower-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,

$$\hat{f}_k = \frac{1}{2} \cdot \frac{scl(f_k \times F_s)}{scl(0.5 \times F_s)} \quad , \tag{7}$$

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 took into (5) to compute the optimal DCCs. The step of replacing f_k with $\hat{f}_k$ also implies that the computed DCCs must be used in the scaled frequency axis instead of the original axis. That is, for a linear frequency f, its envelope magnitude, $S(f)$, should be computed by scaling f into $\hat{f}_k$ first and then taking $\hat{f}_k$ into (2).

For nonlinear frequency-axis scaling, mel and Bark frequency scales are the most famous ones [1]. If the frequency-scale conversion function, $scl(\cdot)$, adopted is a mel-frequency conversion, the spectral envelope shown in Fig. 3 will be changed to the one shown in Fig. 4. The major difference is that the lower-frequency spectral peaks in Fig. 4 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 Fig. 4 shows much stronger vibration near the lower-frequency end than the one in Fig. 3. 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 strong vibration for mel-frequency conversion, we think much stronger vibration will occur if we adopt the Bark-frequency conversion for $scl(\cdot)$. This is because Bark-frequency conversion will have the frequency-scale at the low frequency end being widen more than that widened by the mel-frequency conversion.

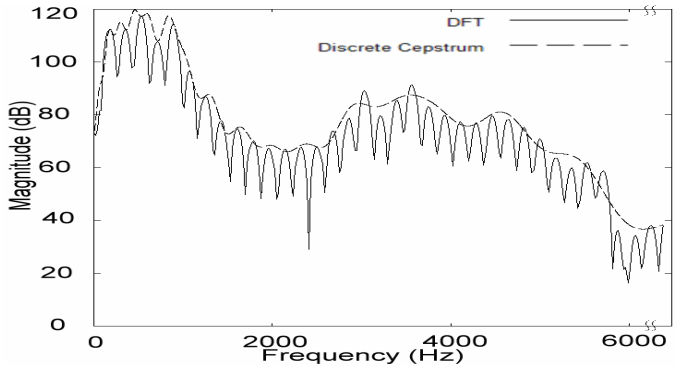


Fig. 4. Envelope approximated in mel freq. scale with 40 DCCs.

B. New Frequency Conversion Function

Therefore, we were motivated to design a frequency conversion function in the hope to eliminate the phenomenon of over vibration at the low frequency end and to reduce the local approximation error around 3,000Hz. After several times of trying function design and inspecting the approximated spectral envelope curves, we finally found a better frequency conversion function,

$$scl(f) = \log(1 + \frac{f}{1{,}750}) \quad , \tag{8}$$

where f is in the unit Hz. This conversion function will have the scaled frequency value, $\hat{f}$, growing more slowly with f at the low frequency end when it is used as the $scl(\cdot)$ function for (7).

The three curves shown in Fig. 5 are obtained by taking Bark, mel, and our frequency conversions as the $scl(\cdot)$ function for (7), respectively. From Fig. 5, it can be seen that our frequency conversion as given in (8) can indeed grow the scaled frequency $\hat{f}$ more slowly with the linear frequency f. By using the frequency conversion function of (8), the approximated spectral envelope in Fig. 3 will become the one drawn in Fig. 6. According to the spectral envelope in Fig. 6, 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.

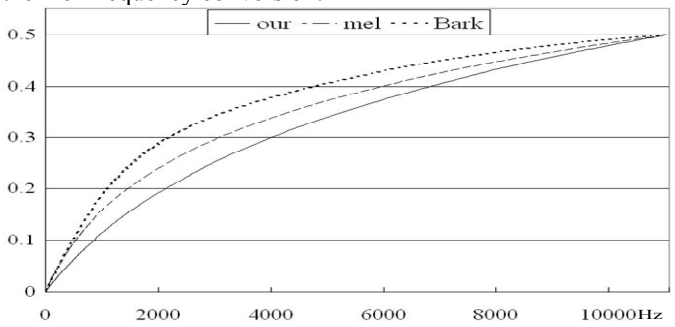


Fig. 5. Three curves of scaled frequencies by using Bark, mel, and our frequency conversion functions, respectively.

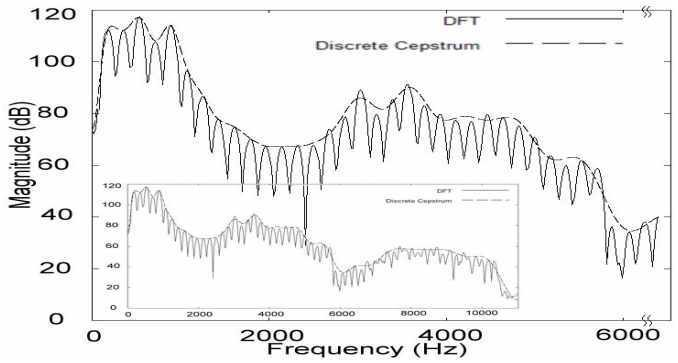


Fig. 6. Envelop approximated in our freq. scale with 40 DCCs.

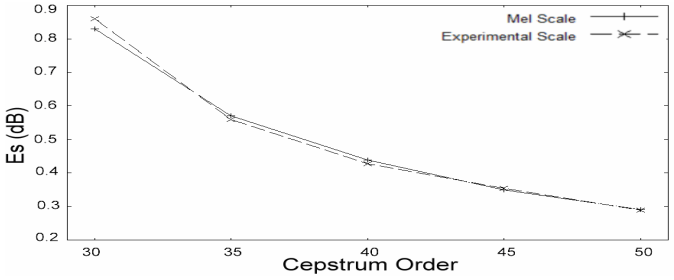
C. *Approximation Error Comparsion*

It may be queried whether our frequency conversion function is just better than mel-frequency conversion for certain signal frames. Therefore, we decide to compare the approximation errors of the two frequency conversions in the three frequency ranges, i.e. 0 ~ 2,000Hz, 0 ~ 4,000Hz, and 0 ~ 6,000Hz. Here, approximation error is measured with the formula,

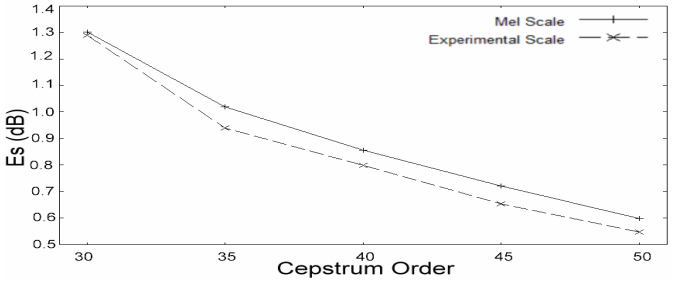
$$Es = \frac{1}{Nr}\sum_{t=0}^{Nr-1}\left[\frac{1}{L(t)}\sum_{k=1}^{L(t)}\left|20\cdot\log_{10}a_k^t - 20\cdot\log_{10}S(t,f_k)\right|\right] \tag{9}$$

where Nr is the total number of signal frames, and $L(t)$ is the number of spectral peaks, for the t-th frame, dynamically determined 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 recorded from a male are used as the testing speech. After all frames of the testing speech are processed, the approximation errors measured in different frequency ranges and different discrete cepstrum orders are illustrated in Fig. 7.

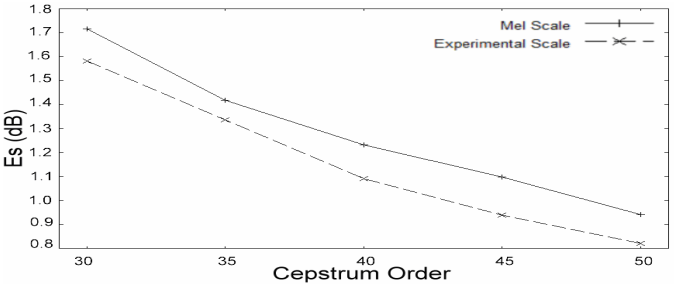
Inspecting the error curve in Fig. 7, it can be seen that across the cepstrum-order numbers from 30 to 50, our frequency conversion and the mel-frequency conversion have almost same approximation errors in the frequency range, 0 ~ 2,000Hz. Nevertheless, in the other two frequency ranges, our conversion function will apparently obtains smaller approximation errors for different cepstrum-order numbers. This decreasing of approximation error becomes more apparent as the frequency range becomes wider.



(a) Frequency range: 0 ~ 2,000Hz



(b) Frequency range: 0 ~ 4,000Hz



(c) Frequency range: 0 ~ 6,000Hz

Fig. 7. Approximation errors measured for our frequency conversion and mel-frequency conversion in the three frequency ranges.

V. AN EXAMPLE APPLICATION: TIMBRE TRANSFORMATION

Here, voice transformation is meant to change the timbre of an input voice to a different timbre. For example, change the

timbre of a female adult into the timbre of a male adult or a child. In the past, phase vocoder is a frequently used technique to transform voice timbre [11]. Nevertheless, the basic transformation method of phase vocoder cannot support independent control of spectral-envelope scaling and pitch shifting. Therefore, we based on the proposed spectral envelope estimation scheme to study and implement a voice transformation system. This system's main processing flow is as shown in Fig. 8. In this system, the inputted voice is first slicing 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 Fig. 1 is executed to estimate its spectral envelope with 40 DCCs. Then, the estimated spectral envelope is used to do spectral-envelope scaling and pitch shifting. Next, the signal model of HNM is used to re-synthesize speech signals. About the processing speed of this system, we had tested it on a notebook computer with Intel T5600 1.83GHz CPU, and found that in average it will consume 0.75 sec. of CPU time to transform 1 sec. of voice signal.

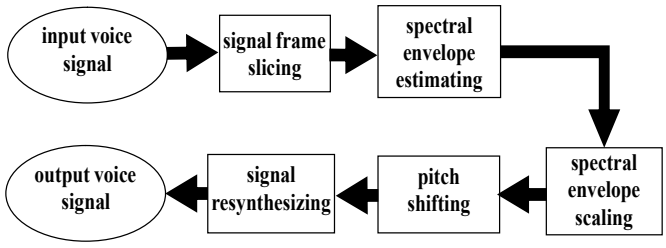


Fig. 8. Main processing flow of the voice transformation system..

To evaluate the performance of the voice-timbre transformation system, 3 sentence utterances were recorded from a male and a female, respectively. Then, the female's utterances were transformed to obtain a male's timbre. Similarly, the male's utterances were transformed to obtain a female's timbre. The source voices and their transformed voices can be downloaded from the web page, http://guhy.csie.ntust.edu.tw/dcc/vt.html. Thirteen persons are invited to listen to the source and transformed sentences for evaluating timbre recognizability. That is, each participant was asked how close the timbre of a played voice is like a female (or a male), and requested to give a score between 1 and 5. As a result, the average scores obtained are 4.84 for the source voices and 4.60 for the transformed voices. Therefore, the transformed voices from our system have sufficiently high timbre-recognizability.

VI. Concluding Remarks

There are three problems that must be solved for practically implementing a discrete-cepstrum based spectral envelope estimation scheme. The first problem is the regularization of the discrete cepstrum coefficients. This problem was already solved by previous researchers. In this paper, we had tried to solve the other two problems, *i.e.* selecting appropriate spectral speaks and finding a better frequency conversion function. For selecting spectral peaks, we apply the concept of HNM to divide a spectrum into lower-frequency harmonic part and higher- frequency noise part. Then, 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 real-cepstrum smoothed spectral curve. As to the problem of frequency axis scaling, we found that the spectral envelope approximated by using the mel or Bark-frequency conversion still has noticeable local approximation errors. Therefore, 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.

To verify the proposed spectral envelope estimation scheme, we had built a voice-timbre transformation system. This system transforms an input voice into an output voice that is of a very different timbre, i.e. the felt gender and age of the voice can both be changed. After perception tests, the average scores from 13 participants show that our system can indeed achieve the function of voice-timbre transformation.

Acknowledgment

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

References

[1] D. O'Shaughnessy, *Speech Communications: Human and Machine*, IEEE Press, Piscataway, NJ, 2000.

[2] D. Schwarz and X. Rodet, "Spectral envelope estimation and representation for sound analysis-synthesis", *International Computer Music Conference*, Beijing, China, pp. 351-354, Oct. 1999.

[3] S. Imai and Y. Abe, "Spectral envelope extraction by improved cepstral method", *Electron. and Commun. in Japan*, Vol. 62-A, No. 4, pp. 10–17, 1979. (in Japanese)

[4] A. Robel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation", *International Conference on Digital Audio Effects*, Madrid, Spain, pp. 1-6, September 2005.

[5] H. Kawahara, I. Masuda-katsuse, and A. De Cheveign, "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*, Vol. 27, pp. 187-207, 1999.

[6] T. Galas and X. Rodet, "An improved cepstral method for deconvolution of source filter systems with discrete spectra: Application to musical sound signals", *International Computer Music Conference*, Glasgow, Scotland, pp. 82-44, 1990.

[7] O. Cappé and E. Moulines, "Regularization techniques for discrete cepstrum estimation", *IEEE Signal Processing Letters*, Vol. 3, No. 4, pp. 100-102, 1996.

[8] Kim, H. Y., et al., "Pitch Detection with Average Magnitude Difference Function Using Adaptive Threshold Algorithm for Estimating Shimmer and Jitter", *20-th Annual Int. Conf. of IEEE Medicine and Biology Society*, pp. 3,162-3,164, 1998.

[9] Y. Stylianou, "Modeling speech based on harmonic plus noise models", in *Nonlinear Speech Modeling and Applications*, eds. G. Chollet et al., Springer-Verlag, Berlin, pp. 244-260, 2005.

[10] Y. Stylianou, *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.

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