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

Author : Hung-Yan Gu and Sung-Fong Tsai

Professor: 陳嘉平

Reporter: 吳國豪

Outline

- Introduction
- Spectral-envelope Estimation with Discrete Cepstrum
- Selection of Spectral Peaks
- Order of Discrete Cepstrum and Frequency Axis Scaling
- Application of Spectral Envelope Estimation: Voice Transformation
- Perception Testing

Introduction

- **Approximating spectral envelope** with regularized discrete cepstrum coefficients was proposed by previous researchers.

- We study two problems encountered in practice when adopting this approach to estimate spectral envelope.

  - The first is which **spectral peaks** should be selected.

  - the second is what **frequency axis scaling function** should be adopted.

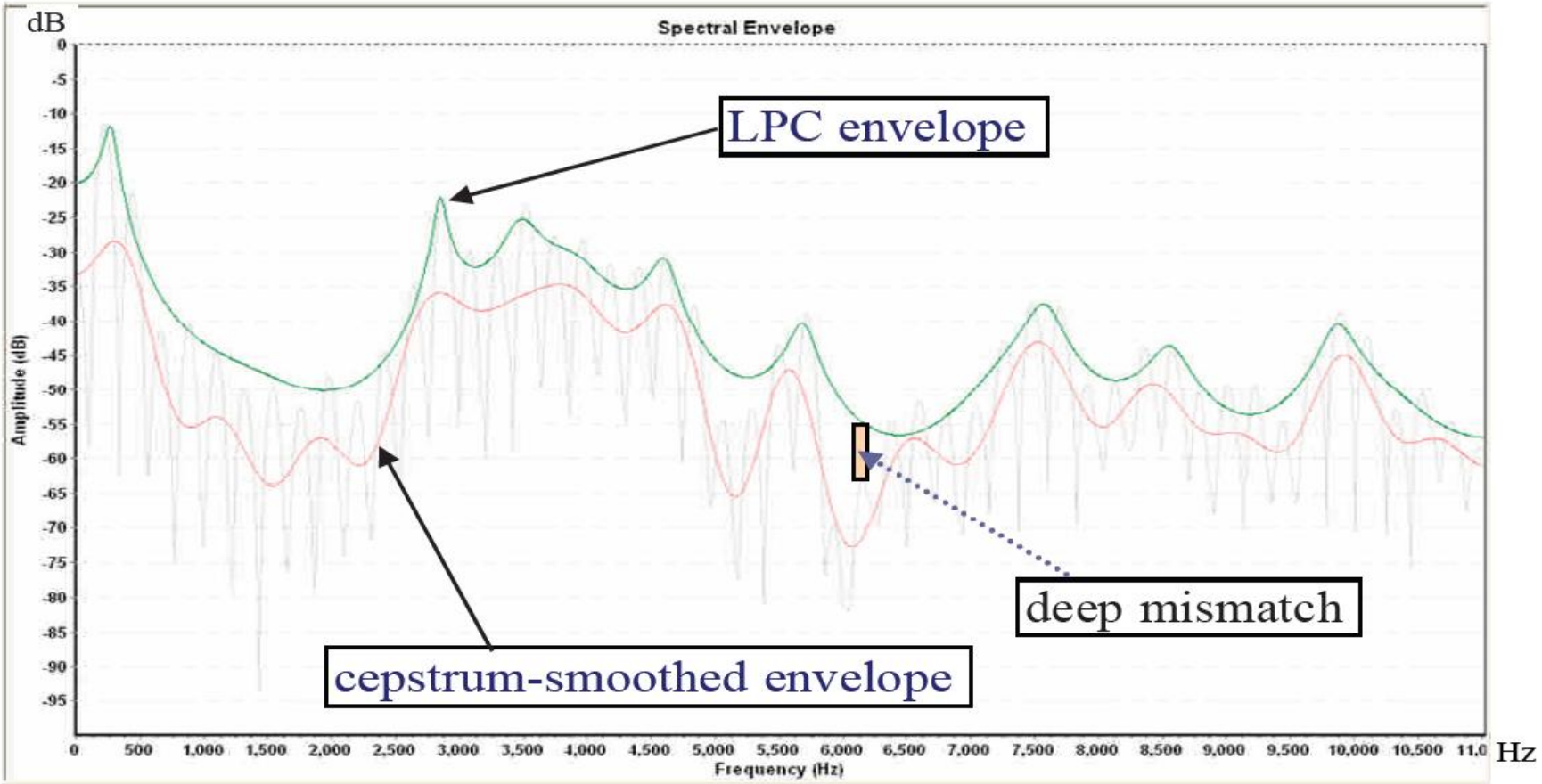


Fig. 1. LPC and cepstrum smoothed spectral curves for a frame from /i/.

Introduction

- We combine two solution methods with the methods for regularizing and computing discrete cepstrum coefficients to form a spectral envelope estimation scheme.

- Furthermore, we have applied this scheme to build a system for **voice timbre transformation**.

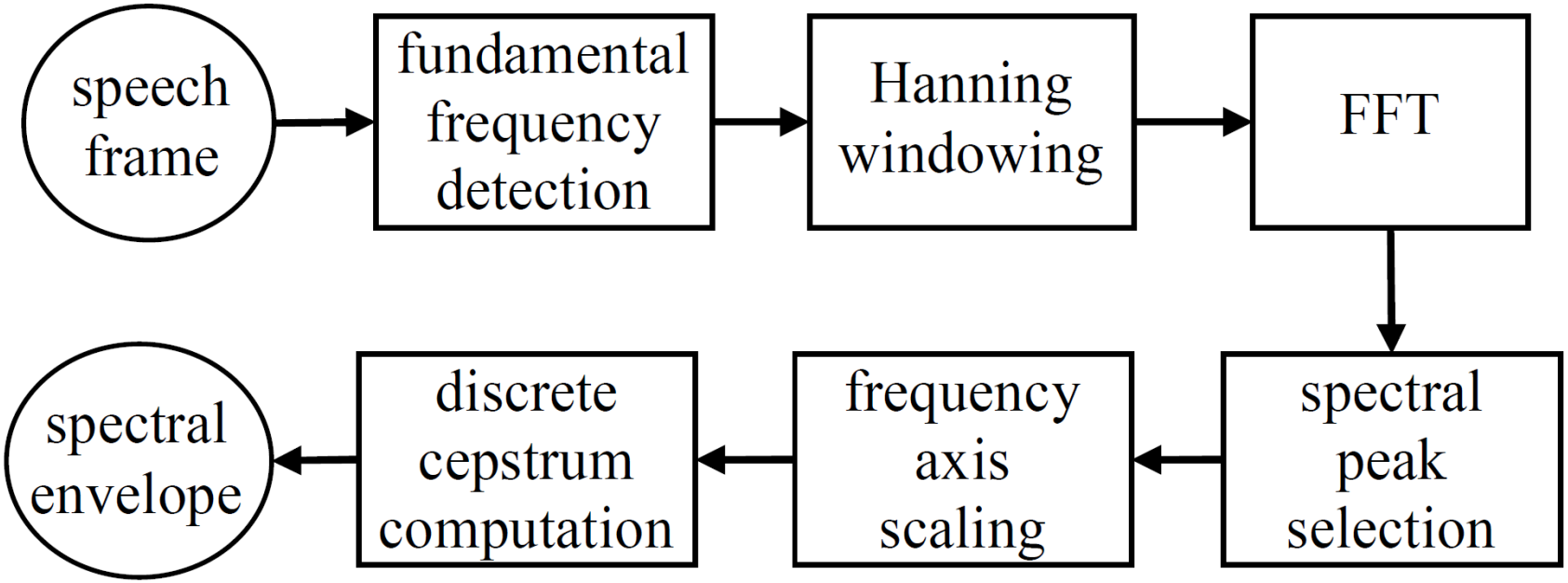


Fig. 2. Main flow of the spectral-envelope estimation scheme.

Spectral-envelope Estimation with Discrete Cepstrum

- The conventional method is transforming the logarithmic magnitude-spectrum with inverse DFT (IDFT) to get its **cepstrum coefficients**. Let the obtained cepstrum coefficients are $c_0, c_1, \ldots, c_{N-1}$ 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,.

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

Spectral-envelope Estimation with Discrete Cepstrum

- Since $\log|X(k)|$ is even symmetric, i.e. $\log|X(k)| = \log|X(N-k)|$ the derived cepstrum coefficients are also even symmetric, $c_k = c_{N-k}$.

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

- The magnitude spectrum computed, $\log|S(f)|$, would be a smoothed version of the original, $\log|X(k)|$.

$$\log S(f) = c_0 + 2\sum_{n=1}^{p} c_n \cos(2\pi f n)$$

Spectral-envelope Estimation with Discrete Cepstrum

- The envelope constraints just mentioned are actually *L* pairs of (f_k, a_k) for *L* representative spectral peaks selected from the original spectrum $\log|X(k)|$.

- The 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

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

Spectral-envelope Estimation with Discrete Cepstrum

- This equation can be rewritten in a matrix form

$$\varepsilon = (A - M \cdot C)^T (A - M \cdot C)$$

$$A = [\log(a_1), \log(a_2), ..., \log(a_L)]^T \quad C = [c_0, c_1, ..., c_p]^T$$

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

- The optimal values of the discrete cepstrum coefficients can be derived to be

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

Regularization of Discrete Cepstrum

- The approximation errorcalculation equation becomes

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

$$R(S(f)) = \int_0^{\pi} \left[\frac{d}{df} S(f) \right]^2 df = C^T \cdot U \cdot C$$

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

- The optimal solution that minimizes the error calculated can be

$$C = (M^T \cdot M + \lambda \cdot U)^{-1} \cdot M^T \cdot A$$

Selection of Spectral Peaks

- If it is detected to be voiced, the frame is further searched for the **MVF** (maximum voiced frequency) frequency, *fv.*

- According to *fv*, the DFT spectrum of the frame is split into the lower frequency harmonic part and the higher-frequency noise part.

Selection of Spectral Peaks

- 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 frequency f_2 and amplitude a_2.

Selection of Spectral Peaks

- The harmonic structure becomes obscure in the noise part, and the frequency gaps between adjacent peaks become **randomly varied**.

- Each spectral speak within the noise part is located and checked again its amplitude. It will be selected if its amplitude is **higher** than the height of the smooth spectral curve at the peak's frequency.

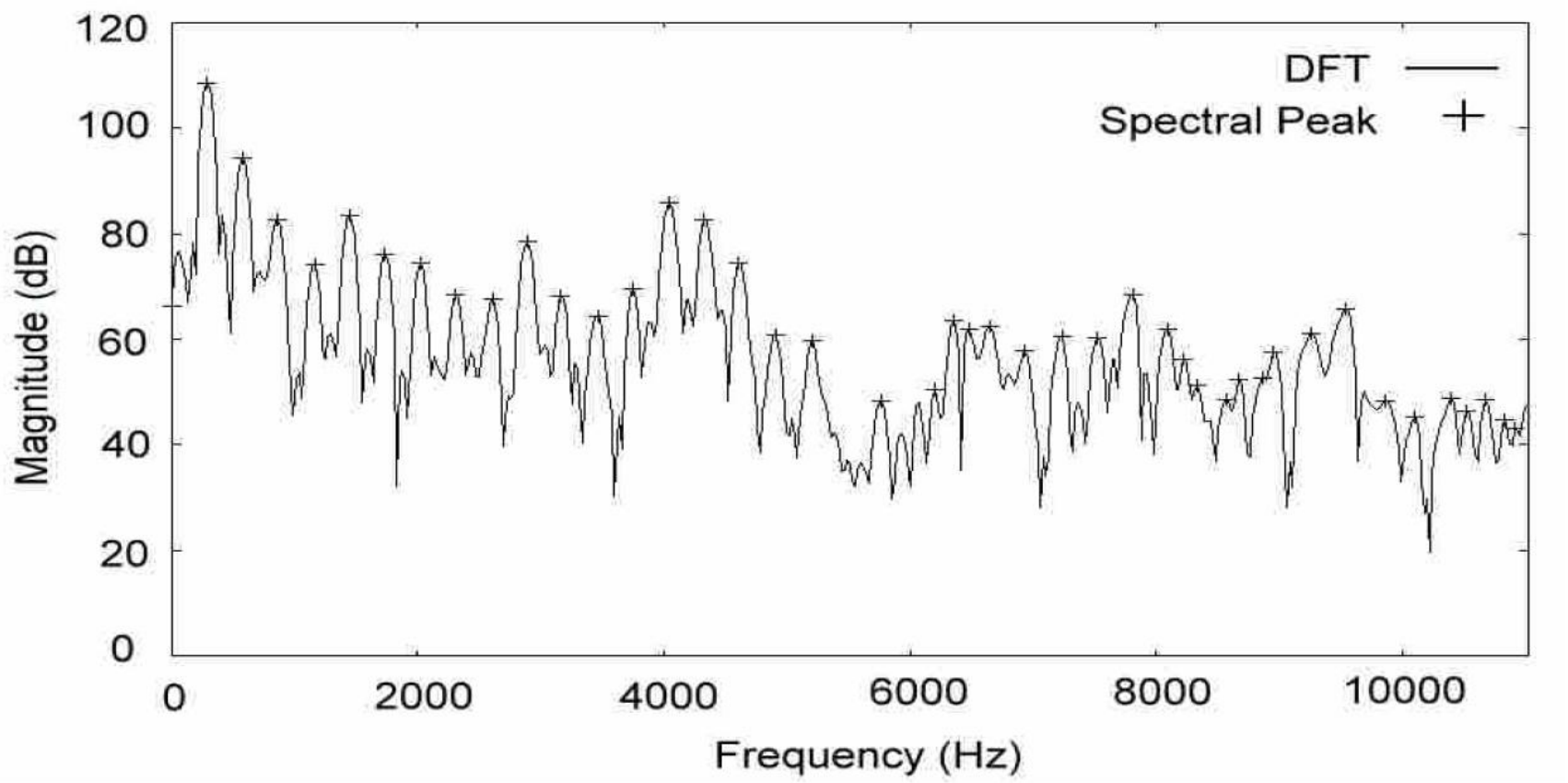


Fig. 5. A typical result for selecting spectral peaks.

Order of Discrete Cepstrum and Frequency Axis Scaling

- What value should be set to the parameter, p, for the order of discrete cesptrum?
- The approximation error is computed here as

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

  - Nr is the total number of frames
  - a_k^t denotes the amplitude of the k-th spectral peak in the t-th frame
  - $S(t, f_k)$ represents the approximated spectral envelope for the t-th fame.

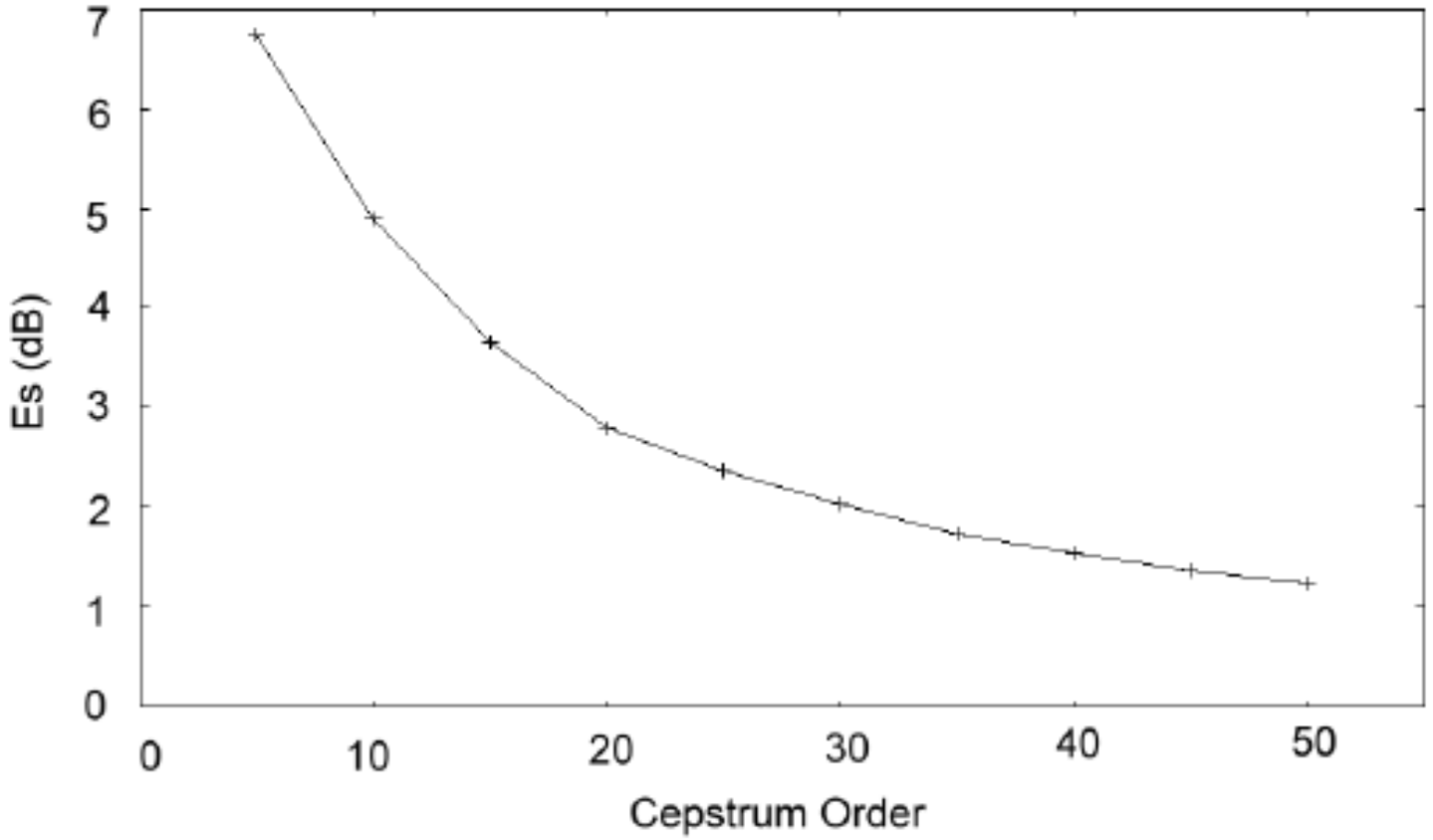
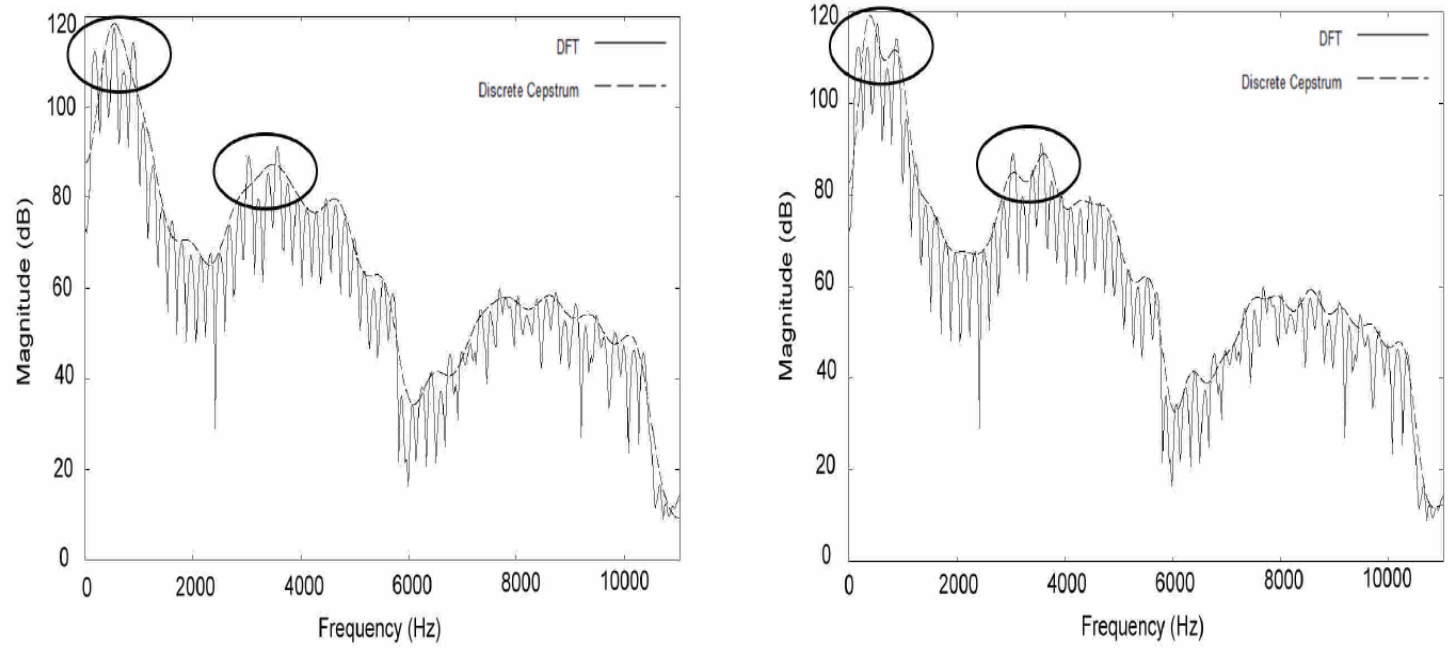


Fig. 6. Approximation error versus discrete cepstrum order.

Order of Discrete Cepstrum and Frequency Axis Scaling

- A conventional idea is **to nonlinearly scale** the frequency axis to enlarge the frequency gaps between **low-frequency spectral peaks**.



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

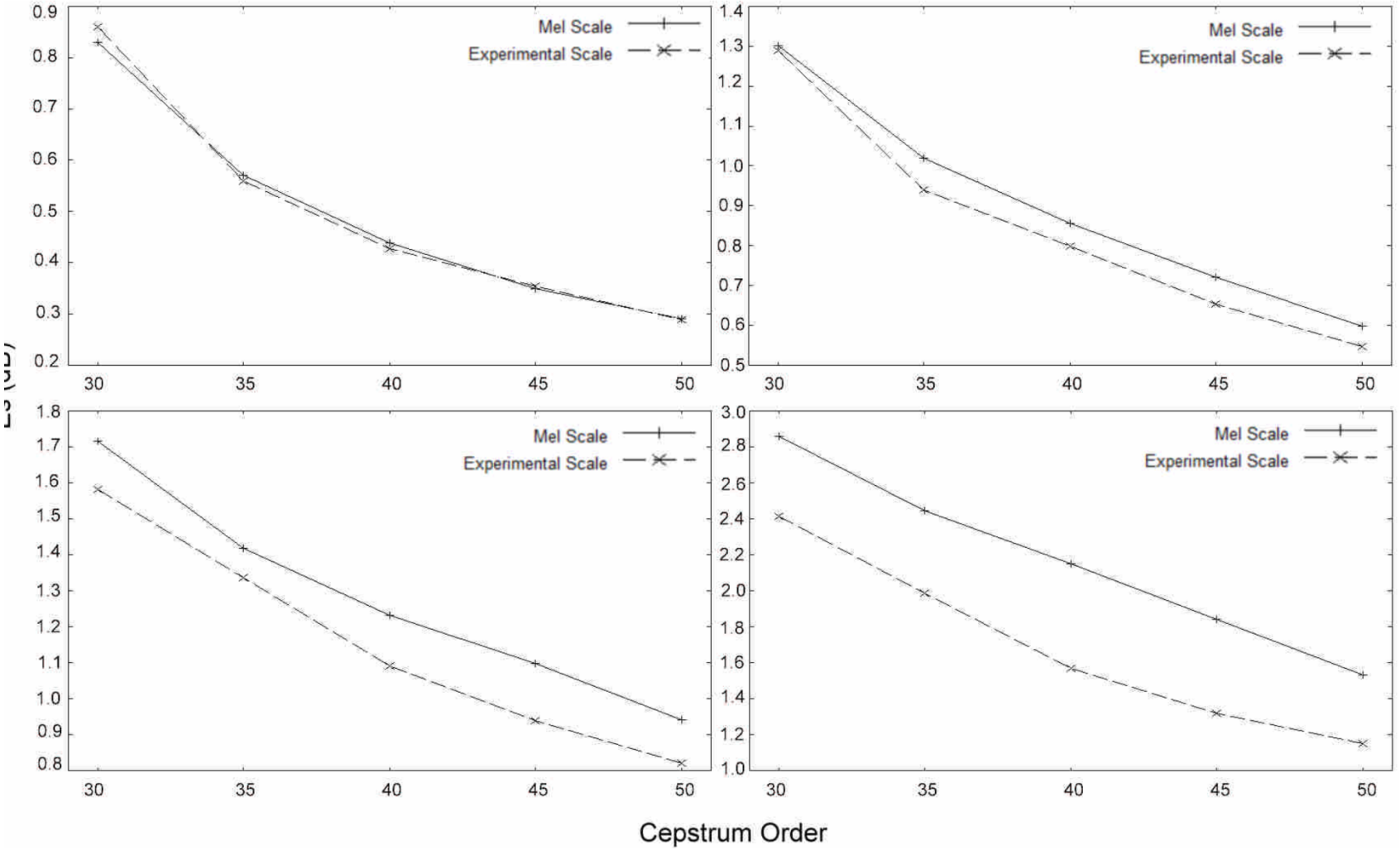
Fig. 7. Spectral envelopes approximated in the linear frequency scale.

Order of Discrete Cepstrum and Frequency Axis Scaling

- Therefore, we were motivated to design a frequency-scale conversion function in the hope to eliminate the phenomenon of over vibration at the low frequency end.

$$scl(f) = \log(1 + \frac{f}{1750})$$

- We decide to compare the approximation errors of the two scale conversion functions in the four frequency ranges, i.e. 0 ~ 2,000Hz, 0 ~4,000Hz, 0 ~ 6,000Hz, and 0 ~ 11,025Hz.


Mel Scale
Experimental Scale
Cepstrum Order

Voice Transformation

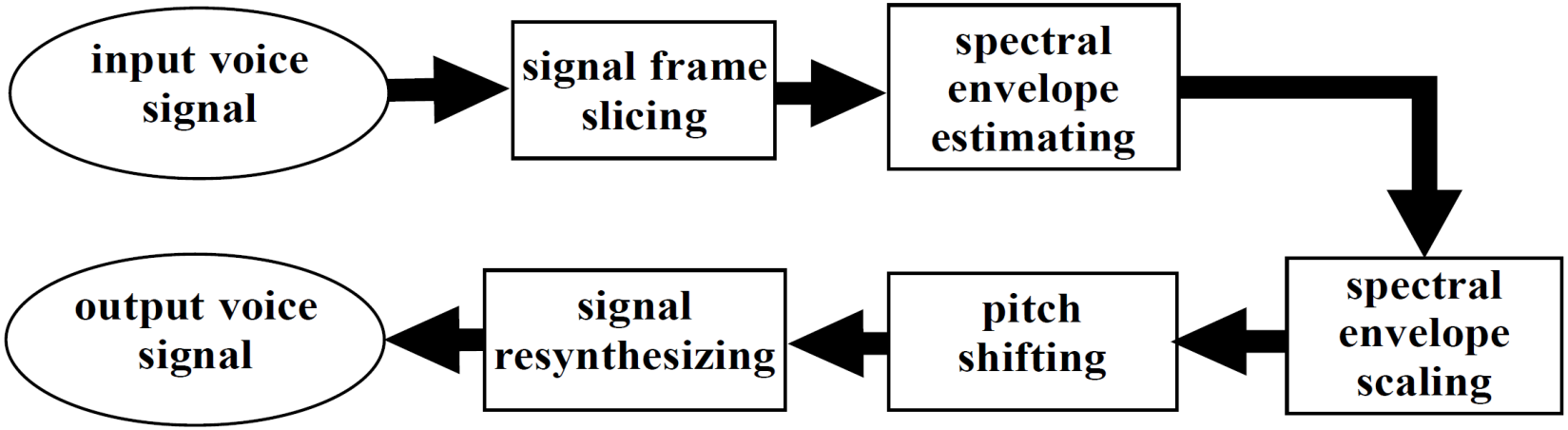


Fig. 10. Main processing flow of the voice transformation system.

Voice Transformation

- **Voice transformation** is meant to change the timbre of an input voice to a different timbre.

- We decide to apply **the technique of additive synthesis developed for computer music synthesis and the signal model of HNM** (harmonic-plus-noise model).

- In this manner, **spectral envelope scaling** and **pitch shifting** can be performed independently.

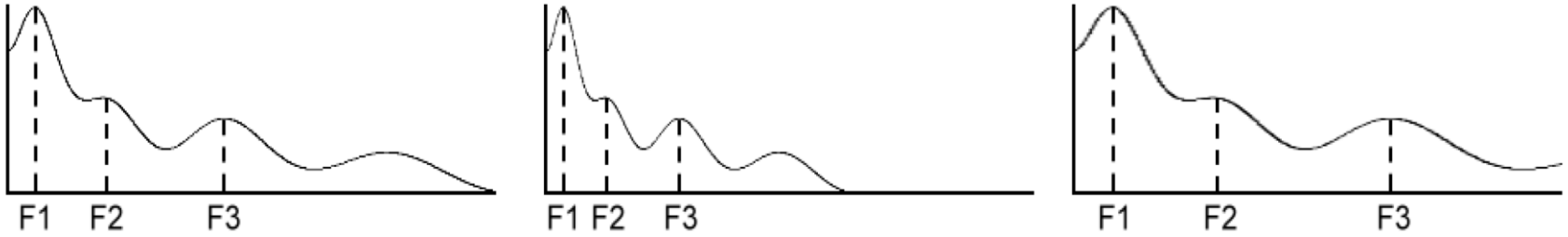
Voice Transformation

- **Scaling of a spectral envelope** can be performed in two possible directions, **shrink** or **extend** the spectral envelope.

  Shrink: $Vs(f) = Vo(\frac{10}{7} f)$ **Extend:** $Ve(f) = Vo(\frac{7}{10} f)$

- **Pitching Shifting :** Suppose the original pitch frequency of the *i*-th frame is 180Hz, and we intend to tune its pitch to 250Hz. Then, the frequencies of the new harmonic partials are apparently, $f_1^i = 250,\ f_2^i = 500,\ f_3^i = 750.$

  $a_1^i = Vs(250), a_2^i = Vs(500), a_3^i = Vs(750)$



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

Fig. 11. The scaling of an example spectral envelope.

Voice Transformation

- Here, the signal model, HNM, 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 is divided according to is called the **MVF**.

Voice Transformation

- 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,

$$f_k(t) = f_k^i + \frac{f_k^{i+1} - f_k^i}{N} t, k = 1,2..., L$$

$$a_k(t) = a_k^i + \frac{a_k^{i+1} - a_k^i}{N} t, k = 1,2..., L$$

  - *N* is the number of sampling points between two adjacent frames
  - *L* is the larger one of L^i and L^{i+1}

Voice Transformation

- Then, the harmonic signal, *h(t)*, for the *t*-th sampling point is computed as

$$h(t) = \sum_{k=1}^{L} a_k(t) \cdot \cos(\phi_k(t)), 0 \le t < N$$

$$\phi_k(t) = \phi_k(t-1) + 2\pi \cdot f_k(t) / 22050$$

  - where $\phi_k(t)$ denotes the accumulated phase till the time t for the *k*-th harmonic partial

Voice Transformation

- To synthesize the noise signal as the summation of the sinusoids whose frequencies are larger than MVF, fixed and are 100Hz apart. Let *KL* = MVF / 100 and *KU* = 22050 / 100. Then, the noise signal, *g(t)*, is synthesized as

$$g(t) = \sum_{k=KL}^{KU} b_k(t) \cdot \cos(\psi_k(t)), 0 \le t < N$$

$$\psi_k(t) = \psi_k(t-1) + 2\pi \cdot k \cdot 100/22050$$

$b_k(t)$ and $\psi_k(t)$ denote, respectively, the amplitude and accumulated phase of the *k*-th sinusoid at the time point *t*.

- Finally, the signal sample for the *t*-th sampling point is synthesized as *h(t)* plus *g(t)*.

Perception Testing

- We first recorded three sentences respectively 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 have a male timbre be transformed.

- For the **mal**e source voice, we set the envelope extending rate to 1.2 and set the pitch shifting rate to 2.1 in order to have a female timbre be transformed.

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

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