CEPSTRAL ANALYSIS OF SPEECH

The most popular model of speech production views speech signals as consisting of two components, representing an excitation source (either quasiperiodic pulses or random noise) and vocal tract resonance. In order to study the nature of speech and to develop speech processing technologies in various ways, it is desirable to separate these two components. Cepstral analysis (or homomorphic analysis) is a procedure which can satisfy this demand. The word *cepstrum* was created by reversing the first four letters of the word *spec-trum*. In general, if two spectrally different components are combined additively, it is more or less possible to separate them by linear filtering. The two components of speech, the excitation source and the vocal tract response, are sufficiently different in their spectral features (i.e., rapidly varying component vs. slowly varying component) that these two components can be separated as follows. Let $S(\omega)$ and $H(\omega)$ be the spectra of the excitation source and the vocal tract resonance, respectively. The speech spectrum represented by their product $X(\omega) = S(\omega)H(\omega)$ is transformed into a sum by a logarithmic transformation:

$$\log(|X(\omega)|) = \log(|S(\omega)|) + \log(|H(\omega)|) \tag{1}$$

Because of the sufficient difference in the spectral features of the two components, $\log(S(\omega))$ and $\log(H(\omega))$, are linearly separable.

CEPSTRAL ANALYSIS

The inverse discrete Fourier transform (IDFT), c_n, of $\log|X(\omega)|$ is called the cepstrum of speech signal $x(t)$ and is represented as

$$c_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log(|X(\omega)|)\cos(\omega n)\,d\omega \quad \text{for } n = 0, 1, \ldots, N \tag{2}$$

The cepstrum is normally real because the spectrum $|X(\omega)|$ is symmetrical against the origin of the frequency axis.

In the cepstrum, the cepstra of the excitation source and the vocal tract resonance are additively combined since the IDFT is a linear operation. Thus, as is illustrated in Fig. 1, the cepstral analysis of a speech signal is a series of operations consisting of windowing, DFT, absolute operation, logarithmic transformation, and IDFT. An input speech segment (Fig. 2), a log spectrum [Fig. 3(b)], and the resulting cepstrum (Fig. 4) are depicted. When plotting the cepstrum, the ordinate is called the *quefrency* (created from *frequency*) instead of time. The slowly varying component of the log spectrum, corresponding to vocal tract resonance, is represented by the low-quefrency component of the cepstrum. In contrast, the rapidly varying component, corresponding to the excitation source, is represented by the high-quefrency component. Note that a strong peak component is observed at a quefrency equal to the pitch period of the input speech signal.

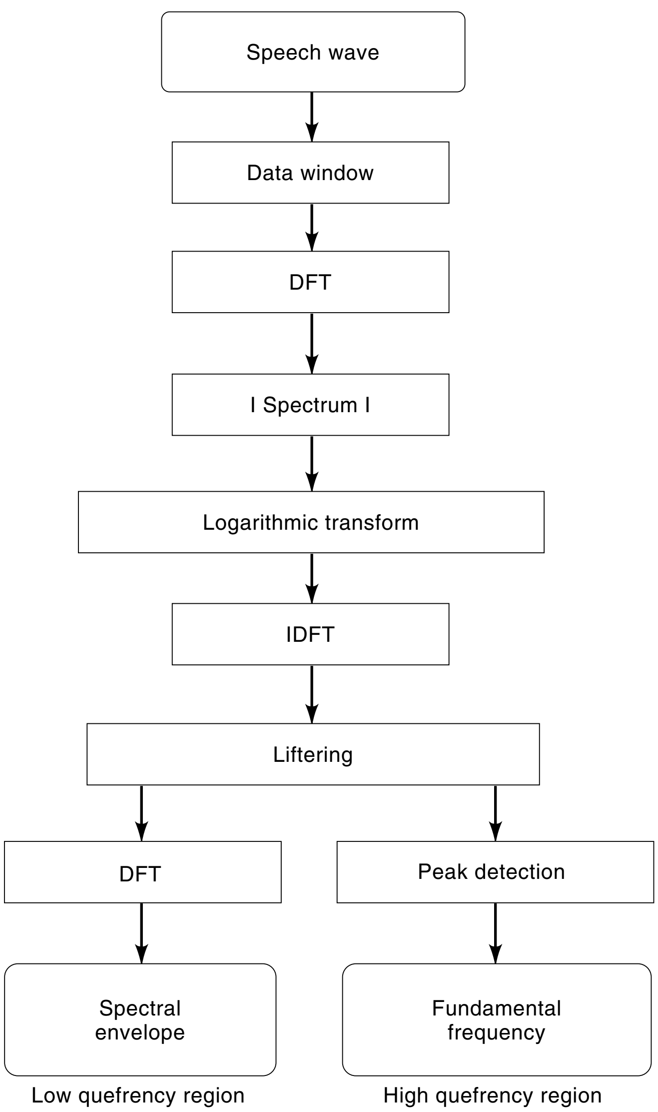

Figure 1. Cepstral analysis.

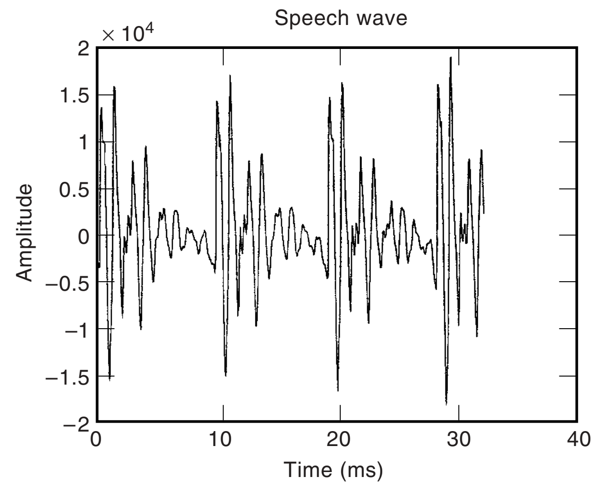

Figure 2. Speech wave.

Spectral Envelope and Pitch Extraction

A low-order cepstral coefficient represents a slowly varying component of the log spectrum. Taking only low quefrency components of a cepstrum yields the spectral envelope.

In the case of voiced speech sounds, the excitation source is quasi-periodic with the fundamental frequency (*pitch frequency* or F_0) of the vocal cord vibration. Since the excitation source component can be separated from the vocal tract resonance in the cepstrum, cepstral analysis is a valuable tool for pitch extraction as well as formant analysis. Again the excitation source corresponds to the rapidly varying component, i.e., the high-quefrency cepstrum component. As is shown in Fig. 4, the fundamental frequency component appears as a strong peak in the high-quefrency component. The peak location in terms of quefrency is equal to the pitch period ($1/F_0$). Therefore, automatically picking the peak of the cepstrum within the possible pitch range of quefrency is a

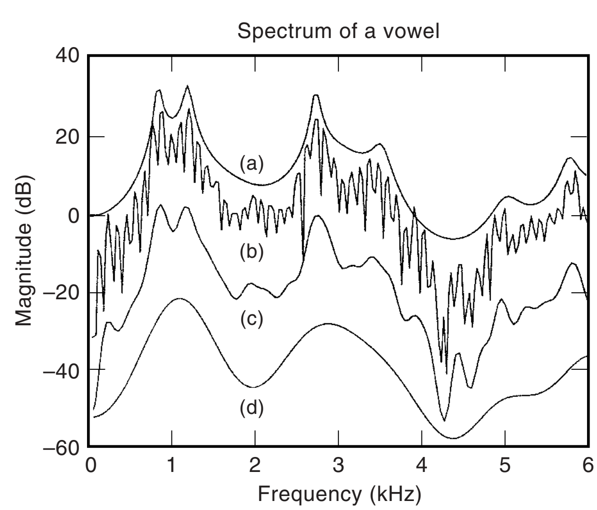

Figure 3. Spectrum of a vowel sampled at a frequency of 12 kHz: (a) LPC spectral envelope (the order of LPC analysis is 16); (b) FFT spectrum; (c) FFT spectrum obtained by truncating cepstrum at quefrency = 48; (d) LPC spectrum obtained by truncating cepstrum at quefrency = 12.

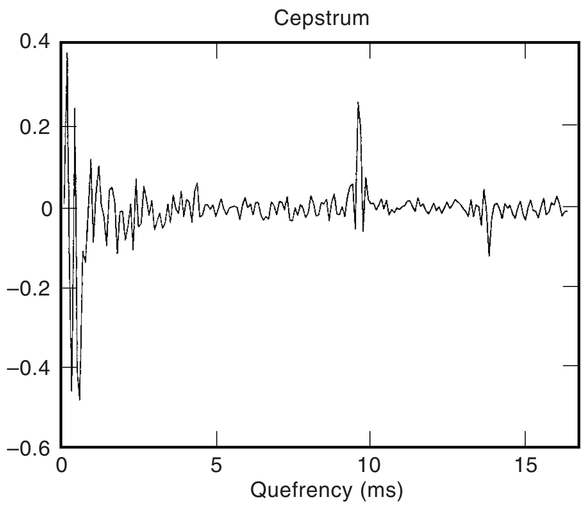

Figure 4. Cepstral coefficients.

viable pitch extraction method. Such strong cepstrum peaks are indicators of voiced speech sounds and they are not observed in the case of unvoiced portions of speech. The value of the peak indicates the periodicity of the speech signal and can be used for voiced-unvoiced decision purposes.

Linear Predictive Coding Cepstrum

Linear predictive coding (LPC) analysis is an alternative to cepstral analysis. The LPC cepstrum is a parameter which has basic properties similar to the cepstrum and can be derived in a computationally efficient manner.

In the linear prediction model, the transfer function $H(z)$ of the vocal tract is represented by an all-pole transfer function with p poles as

$$H(z) = \frac{1}{1 - \sum_{n=1}^{p} \alpha_n z^{-n}} \tag{3}$$

where α_n, $n = 1, 2, \ldots, p$ are LPC coefficients and z is the usual z-transform variable. By considering the power-series expansion of the logarithmic transfer function with powers z^{-1}, $\log(H(z))$ is described by the LPC cepstrum coefficients c_n as

$$\log H(z) = \sum_{n=1}^{p} c_n z^{-n} \tag{4}$$

All the poles of $H(z)$ must be inside the unit circle. After substituting $H(z)$ from Eq. (3) into Eq. (4), derivative operations for both sides of Eq. (4) eventually lead to the following simple relationship between the cepstral coefficients c_n and LPC coefficients α_n

$$c_1 = -\alpha_1 \tag{5}$$

$$c_n = \begin{cases} -\alpha_n - \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) \alpha_k c_{n-k} & 1 < n \le p \\ -\sum_{k=1}^{p} \left(1 - \frac{k}{n}\right) \alpha_k c_{n-k} & p < n \end{cases} \tag{6}$$

Thus, the cepstrum coefficients can be recursively derived from the LPC coefficients.

If Eq. (3) is represented by

$$H(z) = \frac{1}{A(z)} \tag{7}$$

$$A(z) = 1 - \sum_{n=1}^{p} \alpha_n z^{-n} \tag{8}$$

$$= \prod_{k=1}^{p} (1 - q_k z^{-1}) \tag{9}$$

the log-spectrum can be represented as

$$\log(1/A(z)) = \sum_{n=1}^{p} \frac{1}{n} R_n z^{-n} \tag{10}$$

$$R_n = \sum_{l=1}^{p} q_l^n \tag{11}$$

The variable denoted as R_n is called the root-power sum. By equating equal powers of z^{-1} between Eq. (10) and Eq. (4), the simple relation

$$c_n = \frac{1}{n} R_n \tag{12}$$

is derived (1).

A direct relation between predictor coefficients and cepstrum is derived as

$$c_n = \sum \frac{((\sum_{i=1}^{p} k_i) - 1)!}{\prod_{i=1}^{p} (k_i!)} \prod_{i=1}^{p} (-\alpha_i)^{k_i} \tag{13}$$

where the sum is taken over all combinations of k_i to meet the condition

$$\sum_{i=1}^{p} i k_i = n \tag{14}$$

The root-power sums can be computed nonrecursively by the direct relations between cepstrum and LPC coefficients (1).

Spectral Distance Measure

A variety of spectral distance measures and distortion measures have been proposed for speech processing. These are used for measuring the spectral difference between two spectra. The spectral distance measures have been widely applied in template matching-based speech recognition.

The Euclidean distance of two spectra $S^{(f)}(\omega)$ and $S^{(g)}(\omega)$ is given by

$$D_{\mathrm{CEP}} = \frac{1}{2\pi} \int_{-\pi}^{\pi} (\log(S^{(f)}(\omega)) = \log(S^{(g)}(\omega)))^2 \, d\omega \tag{15}$$

$$= \sum_{n=-\infty}^{\infty} (c_n^{(f)} - c_n^{(g)})^2 \tag{16}$$

D_{CEP} is called the cepstral distance, and c_0 displays a wider dynamic range than other coefficients. Therefore, the following weighted distance measure can be used

$$D_{\mathrm{CEP}} = r(c_0^{(f)} - c_0^{(g)})^2 + 2(1-r)\sum_{n=1}^{N}(c_n^{(f)} - c_n^{(g)})^2 \qquad (17)$$

where r is a balancing weight. Spectral distance between spectra can be used for representing spectral difference, however, the log spectral distance provides better speech recognition performance when applied as a spectral matching measure in speech recognition.

LIFTERING

An operation to separate slowly varying spectral components from rapidly varying ones begins with linear filtering of the spectrum. The filter in the spectral domain is called the *lifter*. The word *lifter* was created by reversing the first three letters of the word filter.

Let $S(\omega)$ be a spectrum and $H(\omega)$ be a lifter. A liftered spectrum $Q(\omega)$ is given by

$$\log(Q(\omega)) = \int_{-\infty}^{\infty} \log(S(\omega - \lambda))H(\lambda)\,d\lambda \qquad (18)$$

In digital speech processing, a speech signal is obtained not as a continuous function of time, but as a sampled time series. When the sampling frequency is f_s, the bandwidth of a speech signal is limited to

$$f_{\max} = f_s/2 \qquad (19)$$

If $Q(\omega)$ and $S(\omega)$ are represented in a logarithmic magnitude scale, these spectra are represented by cosine expansions

$$\log(Q(\omega)) = \sum_{n=-\infty}^{\infty} b_n \cos(\omega n) \qquad (20)$$

$$\log(S(\omega)) = \sum_{n=-\infty}^{\infty} c_n \cos(\omega n) \qquad (21)$$

where b_n and c_n are cepstral coefficients. The lifter impulse response can also be represented by a cosine expansion.

$$H(\omega) = \sum_{n=-\infty}^{\infty} l_n \cos(\omega n) \qquad (22)$$

By using the orthogonality of cosine functions, the following equation is derived:

$$b_n = l_n c_n \qquad (23)$$

Lifter weights l_n correspond to the transfer function of each cosine component in the log spectrum.

Applying a low-pass lifter yields the slowly varying cepstrum component corresponding to the vocal tract resonance. DFT computation of the low-pass liftered cepstrum produces the so-called spectrum envelope [Fig. 3(c)] consisting of the slowly varying spectrum component. This operation is called *cepstral smoothing.* Since the spectral envelope is sufficiently smooth, local spectral peaks characterized by the formant frequencies and bandwidths corresponding to the vocal tract resonances can be well determined by applying an automatic formant tracking algorithm.

Truncation of Cepstrum

Truncating cepstrum c_n at $n = \nu$, is equivalent to multiplying it by the lifter weight

$$l_n = \begin{cases} 1 & |n| \le \nu \\ 0 & |n| > \nu \end{cases} \qquad (24)$$

The spectrum corresponding to the truncated cepstrum is smoothed. As is shown in Fig. 3, the LPC spectrum (a) has sharper peaks than that produced by the truncated LPC cepstrum (d).

Spectral Slope Distance

The frequency derivative of the log spectrum is given by

$$\frac{d\log(S(\omega))}{d\omega} = \sum_{n=-\infty}^{\infty} nc_n \cos(n\omega) \qquad (25)$$

A distance measure can then be defined to measure the Euclidean distance between two spectral slope functions (2,3). The spectral slope distance between two spectra $S^{(f)}$ and $S^{(g)}$ is given by

$$D_{SS} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\frac{d\log(S^{(f)}(\omega))}{d\omega} - \frac{d\log(S^{(g)}(\omega))}{d\omega}\right)^2 d\omega \qquad (26)$$

$$= \sum_{n=-\infty}^{\infty}\left(nc_n^{(f)} - nc_n^{(g)}\right)^2 \qquad (27)$$

Weighted Cepstrum

To measure distance in a multidimensional space, variance-normalized distance measures such as the Mahalanobis distance provide good performance in general pattern recognition. A weighted cepstrum is such a variance-normalized parameter if each cepstral coefficient is regarded as an independent parameter (4,5).

$$c_n^{(W)} = \frac{c_n}{\sigma_n} \qquad (28)$$

where σ_n^2 is the variance of the nth ceptral coefficient. In case of actual speech cepstrum data, $1/\sigma_n$ is approximately proportional to order n, so the weighted cepstrum is approximately given by

$$c_n^{(W)} = w_n c_n \qquad (29)$$

$$w_n = \begin{cases} n & \text{if } n < n_s \\ n_s & \text{otherwise} \end{cases} \qquad (30)$$

If $n_s = \infty$, the weighted cepstrum is equal to the root-power sum shown in Eq. (11). The weighting function w_n saturates at $n = n_s$ to suppress excess weighting of high-order cepstral coefficients. The Euclidean distance between two weighted spectra can be defined as a spectral distance measure.

Group Delay Spectrum

The group delay spectrum provides a spectrum wherein peaks are emphasized more than the power spectrum (6). Let $H(z)$ be the transfer function of an all-pole filter representing a speech spectrum as

$$H(z) = \frac{1}{1+\sum_{n=1}^{p}\alpha_n z^{-n}} \tag{31}$$

$$= \prod_{n=1}^{p}\frac{1}{1-(z_n/z)} \tag{32}$$

$$= \prod_{n=1}^{p} H_n(z) \tag{33}$$

Let the gain and the phase be $A_n(\omega)$ and $\phi_n(\omega)$, respectively; each term of the all-pole model is given by

$$H_n(e^{j\omega}) = A_n(\omega)e^{j\phi_n(\omega)} \tag{34}$$

The group delay spectrum $T_n^G(\omega)$ is defined by the frequency derivative of the phase

$$T_n^G(\omega) = \sum_{n=1}^{p} -\frac{d\phi_n(\omega)}{d\omega} \tag{35}$$

Cepstral coefficients for a group delay spectrum are derived as

$$g_n = nc_n \tag{36}$$

The original group delay spectrum formulation excessively emphasizes high-order cepstral coefficients. Thus, the following generalized weighting function is used for practical speech recognition.

$$g_n = w_n c_n \tag{37}$$

$$w_n = n^s \exp\left(-\frac{n^2}{2\tau^2}\right) \quad (s \ge 0) \tag{38}$$

This representation is called the smoothed group delay spectrum. τ and s are parameters that control the smoothness of the spectrum.

Bandpass Lifter

The above approaches can be generalized as liftering by lifter weights represented as

$$b_n = w_n c_n \tag{39}$$

The lifter weight

$$w_n = 1 + \frac{\nu}{2}\sin\left(\pi\frac{n}{\nu}\right) \qquad (1 \le n \le \nu) \tag{40}$$

is a good choice to increase speech recognition accuracy (7).

TEMPORAL ANALYSIS OF CEPSTRUM

Delta-Cepstrum

The auditory system is sensitive to temporal changes in sound features. Research effort has been focused on simulating that function of the auditory system. Delta-cepstrum is one such dynamic feature parameter (8) and is defined as the slope of the linear fitting curve of the cepstral time series:

$$c_n'(i) = \frac{\sum_{l=-L}^{L} c_n(i+l)lw(l)}{\sum_{l=-L}^{L} l^2 w(l)} \tag{41}$$

where i denotes frame number of the cepstral time series. Delta-cepstrum picks up the trend in the window $-L \le l \le L$ of a cepstral time series. The original delta-cepstrum is derived based on linear regression against the scatter of cepstral values in the time axis.

The formulation of the delta-cepstrum in Eq. (41) can be translated into a filter for cepstral time series. The delta-cepstrum shows bandpass filtering characteristics. A triangular weighting function

$$w(l) = \begin{cases} \dfrac{L-l}{L} & \text{if } l \ge 0 \\ \dfrac{L+l}{L} & \text{if } l < 0 \end{cases} \tag{42}$$

gives a bandpass transfer function that is better in eliminating side-lobes than a uniform weighting expression derived from the original formulation of regression. The frequency of the temporally varying component in the cepstral time series is called the *modulation frequency*.

A delta-cepstral distance can be defined by a Euclidean distance like the cepstral distance. Let D_{CEP} be a cepstral distance and $D_{\Delta\mathrm{CEP}}$ be the delta-cepstral distance, the combinational distance can be defined as

$$D_{\mathrm{CEP}+\Delta CEP} = rD_{\mathrm{CEP}} + (1-r)D_{\Delta\mathrm{CEP}} \tag{43}$$

where r is a balancing weight between the cepstral distance and the delta-cepstral distance. A typical value of r for automatic speech recognition is 0.05.

Dynamics-Emphasized Cepstrum

Dynamics-emphasized cepstrum is a spectral representation composed of instantaneous and transitional feature parameters. As discussed in Reference (9), the formulas for calculating dynamics-emphasized cepstrum are given by

$$d_n(i) = c_n(i) + 8\Delta c_n(i) - 8\Delta^2 c_n(i) \tag{44}$$

$$\Delta c_n(i) = \frac{\sum_{l=-3}^{3} lc_n(i+l)}{\sum_{l=-3}^{3} l^2} \tag{45}$$

$$\Delta^2 c_n(i) = \frac{\sum_{l=-3}^{3}(l^2-4)c_n(i+l)}{\sum_{l=-3}^{3}(l^2-4)^2} \tag{46}$$

where $\Delta^2 c_n(i)$ denotes the n-th coefficient of the second-order delta-cepstrum at time i.

RASTA

RASTA is another filter for cepstral time series and performs as a bandpass filter (10). RASTA achieves a lower resonance modulation-frequency by multiplying a temporal integration term by a delta-cepstrum. The z transform of RASTA is given by

$$H(z) = 0.1\frac{z^{-2}(2z^2 + z - z^{-1} - 2z^{-2})}{z^{-4}(1 - 0.98z^{-1})} \tag{47}$$

The numerator is a delta-cepstrum. The denominator contributes to temporal integration. The terms z^{-2} and z^{-4} do not affect the gain of the modulation-frequency transfer function.

Bandpass and Highpass Filters

BPF (bandpass filter) and HPF (highpass filter) formulations have also been proposed for filtering the modulation-frequency component in the cepstral time sequence (11). A bandpass filter can be formulated for speech recognition as

$$H(z) = \frac{\kappa \sum_{l=0}^{L}(l - \frac{L-1}{2})z^{-l}}{1 - \rho z^{-1}} \tag{48}$$

and a highpass filter as

$$H(z) = 1 - \frac{\sum_{l=1}^{L} \beta^l z^{-l}}{\sum_{l=1}^{L} \beta^l} \tag{49}$$

where κ, β, and ρ are constants for determining the transfer functions.

FM Neuron Model

A biological FM neuron detects unidirectional frequency change. A model has been proposed to simulate the function of the FM neuron (12). The FM neuron model is formulated as a time and frequency derivative of a spectral time series and extracts formant movement. When a spectral sequence is given by $S(\omega, t)$, the output of the FM neuron model can be given by

$$F(\omega, t) = \frac{\partial^2 \log(S(\omega, t))}{\partial\omega\partial t} \tag{50}$$

The cepstrum coefficients for the FM neuron output are given by

$$f_n = n\frac{\partial c_n(t)}{\partial t} \tag{51}$$

A delta-cepstrum can be used for practical time-derivative operation on a cepstral time series. When the center frequency of a spectral peak decreases with time, the FM neuron outputs a negative response. When the center frequency increases with time, the FM neuron outputs a positive response.

Masked Cepstrum

The filter coefficients are identical among all cepstral orders in the above cepstral analysis. A spectrotemporal spectral representation is achieved by a two-dimensional matrix lifter based on the auditory masking effect (13). In an auditory system the target sound is suppressed by a masker sound. The masked spectrum is given by the spectrum reduced by the masking level. The masking level is a function of the frequency difference between the masker and the signal and the elapsed time after the masker is given to the auditory system. A masked cepstrum is derived from the masked spectrum. Let $M(\lambda, \tau)$ be the two-dimensional impulse response that simulates the auditory masking function, the masked spectrum is given by

$$\log(Q(\omega, t)) = \sum_{i=0}^{L} \int_{-\infty}^{\infty} \log(S(\omega - \lambda, t + i\Delta t))M(\lambda, i\Delta t)\, d\lambda \tag{52}$$

where $i\Delta t$ is the elapsed time after the end of the masker sound, Δt is the frame shift, and λ denotes frequency difference.

The masked cepstrum is represented by the cosine expansion coefficients of the masked spectrum. Let l_n and c_n be the inverse Fourier transform of the masking impulse response and the log spectrum, respectively, the inverse Fourier transform of Eq. (52) is given by

$$b_n(t) = \sum_{i=0}^{N} l_n(i)c_n(t + i\Delta t) \tag{53}$$

This lifter has two different aspects as an order-dependent temporal filter for cepstral time series, and an elapsed-time-dependent lifter. The former acts as a high modulation-frequency-pass filter for cepstral time series. The latter acts as a low quefrecncy-pass lifter for the spectrum at a given elapsed time.

BIBLIOGRAPHY

1. M. R. Schroeder, Direct (nonrecursive) relations between cepstrum and predictor coefficients, *IEEE Trans. Acoust. Speech Signal Process.,* **ASSP-29**: 297–301, 1981.
2. D. Klatt, Prediction of perceived phonetic distance from critical-band spectra: A first step, *Proc. ICASSP82,* pp. 1278–1281, 1982.
3. B. A. Hanson and H. Wakita, Spectral slope distance measures with linear prediction analysis for word recognition in noise, *IEEE Trans. Acoust. Speech Signal Process.,* **ASSP-35**: 968–973, 1987.
4. Y. Tohkura, A weighted cepstral distance measure for speech recognition, *IEEE Trans. Acoust. Speech Signal Process.,* **ASSP-35**: 1414–1422, 1987.
5. K. K. Paliwal, On the performance of the frequency-weighted cepstral coefficients in vowel recognition, *Speech Commun.,* pp. 151–154, 1982.
6. F. Itakura and T. Umezaki, Distance measure for speech recognition based on the smoothed group delay spectrum, *Proc. ICASSP87,* pp. 1257–1260, 1987.
7. B.-H. Juang, L. R. Rabiner, and J. G. Wilpon, On the use of bandpass liftering in speech recognition, *IEEE Trans. Acoust. Speech Signal Process.,* **ASSP-35**: 947–954, 1987.
8. S. Furui, Speaker-independent isolated word recognition using dynamic features of speech spectrum, *IEEE Trans. Acoust. Speech Signal Process.,* **ASSP-34**: 52–59, 1986.
9. S. Furui, Speaker-independent isolated word recognition based on dynamics-emphasized cepstrum, *Trans. IECE Jpn.,* **E 69** (12): 1310–1317, 1986.
10. H. Hermansky et al., Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP), **Proc. Eurospeech 91,** pp, 1367–1370, 1991.

11. B. A. Hanson and T. H. Applebaum, Subband or cepstral domain filtering for recognition of Lombard and channel-distorted speech, *Proc. ICASSP93,* **2**: 79–82, 1993.
12. K. Aikawa and S. Furui, Spectral movement function and its application to speech recognition, *Proc. ICASSP88,* pp. 223–226 15.S5.11, 1988.
13. K. Aikawa et al., Cepstral representation of speech motivated by time-frequency masking: An application to speech recognition, *J. Acoust. Soc. Am.,* **100** (1): 603–614, 1996.

YOH'ICHI TOHKURA
KIYOAKI AIKAWA
NTT Laboratories