LINEAR FILTERING OF A FEATURE VECTOR SEQUENCE FOR SPEECH RECOGNITION

K. Katagishi[1], H. Singer[1], K. Aikawa[2], and S. Sagayama[1]

[1] ATR Interpreting Telephony Research Labs.
[2] ATR Auditory and Visual Perception Research Labs.

ABSTRACT - This paper provides a new interpretation of the so-called "delta-cepstrum". We show that it can be regarded as linear filtering of a cepstrum vector sequence, and this basic idea is then extended to a general class of linear filters. Two possible cases, i.e. scalar and matrix coefficients, are considered and tested using phoneme recognition.

INTRODUCTION

The statics and dynamics of speech spectra have been proved very powerful in enhancing performance of speech recognition systems.

The delta-cepstrum was first proposed to capture the dynamic characteristics of speech for the purpose of speaker identification (Sagayama and Itakura, 1979), then used in DP matching-based speech recognition in combination with the original cepstrum (Furui, 1986), and now is extensively being used in HMM-based speech recognition systems.

This paper provides a new interpretation for the delta-cepstrum, namely that it can be regarded as a linear filtered cepstrum vector sequence, and this basic idea is extended to a general class of linear filters. This paper shows that linear filtering of a cepstrum vector sequence can provide better feature parameters for speech recognition than conventional parameters like cepstrum and delta-cepstrum.

NEW INTERPRETATION FOR DELTA-CEPSTRUM AND DYNAMIC CEPSTRUM

Linear filtering with scalar coefficients

The delta-cepstrum was proposed to capture the dynamic characteristics for a sequence of cepstra and derived as the slope of the weighted least squares fit regression line. Let's denote a LPC-based cepstrum coefficient vector at time τ and a delta-cepstrum at time t by $c(t)$ and $\Delta c(t)$ respectively. Then, $\Delta c(t)$ is given as $\boldsymbol{a}$ which satisfies

$$\min_{\boldsymbol{a},\boldsymbol{b}} \sum_{\tau=-M}^{M} w(\tau)\, ||\boldsymbol{a}\tau + \boldsymbol{b} - c(t+\tau)||^2 , \qquad (1)$$

where $\boldsymbol{b}$ denotes the intersection of the y-axis and the regression line. The quantity which expresses the regression window length in frame units is given by $(2M+1)$. If the weighting factor $w(\tau)$ is an even function, $\Delta c(t)$ is derived as follows:

$$\Delta c(t) = \frac{\sum_{\tau=-M}^{M} \tau w(\tau) c(t-\tau)}{\sum_{\tau=-M}^{M} \tau^2 w(\tau)}. \qquad (2)$$

Let's denote $h(\tau)$ by

$$h(\tau) = \frac{\tau w(\tau)}{\sum_{\tau_0=-M}^{M} {\tau_0}^2 w(\tau_0)}, \qquad (3)$$

then Eq.(2) can be rewritten as

$$\Delta c(t) = \sum_{\tau=-M}^{M} h(\tau)c(t-\tau). \tag{4}$$

Eq.(4) shows that the delta-cepstrum can be expressed by a linear combination of a partial cepstrum sequence and is calculated by a convolution of $h(\tau)$ and $c(\tau)$. Furthermore, the filter coefficients $h(\tau)$ are given as a scalar quantity. Therefore, the delta-cepstrum can be interpreted as the output from a linear filter with a cepstrum sequence as input. Frequency characteristics of the filter are decided by two parameters, i.e., the weighting factor $w(\tau)$ and the window length $(2M+1)$. For example, suppose that a triangular window is used as the window and that frame period and window length are 5ms and 100ms, respectively, then the frequency characteristic of $h(\tau)$ is shown as the solid line in Fig.3(a). The mean FFT log-spectrum for a cepstrum sequence is shown in Fig. 1. The Delta-cepstrum is interpreted as the output from a bandpass filter with a passband between around 0Hz and 20Hz (solid line in Fig.3(a)) with a cepstrum sequence as input.

For speech recognition, both, cepstrum and delta-cepstrum, are often used as feature parameters. These two kinds of feature parameters can be interpreted as outputs from two filters with different frequency characteristics and a cepstrum sequence as input. More concretely, cepstrum and delta-cepstrum are outputs from an allpass filter and from a bandpass filter, respectively.

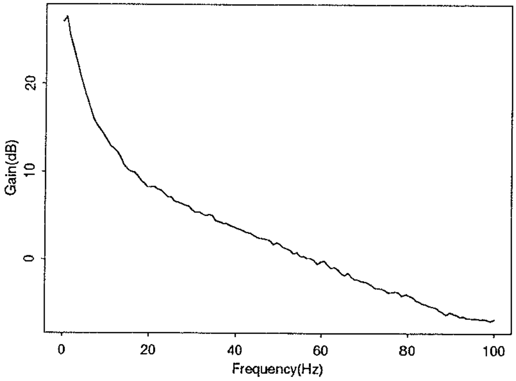

Figure 1: Example for long term spectrum of a cepstrum sequence (quefrency=3)

Linear filtering with matrix coefficients

Recent research on auditory perception reports that the forward masking pattern becomes more widespread over the frequency axis as the masker-signal interval increases (Miyasaka,1983). A novel matrix liftering methodology is derived for simulating the time-frequency masking characteristics. The masking mechanism is abstracted into a masking model in which the masking pattern at the current time is calculated as the sum of the preceding spectra smoothed by time-dependent low-quefrency-pass lifters. The cutoff quefrency of the lifter shifts lower as a function of the time interval between the current and the preceding spectra. The Dynamic Cepstrum is obtained by subtracting the masking level from the current cepstrum.

Let's denote the k-th order Dynamic Cepstrum at time t by $b_k(t)$, then the overall operation for calculating the Dynamic Cepstrum is described as

$$b_k(t) = c_k(t) - m_k(t), \tag{5}$$

$$m_k(t) = \alpha \sum_{\tau=1}^{N_k} c_k(t-\tau)\beta^{\tau-1}, \tag{6}$$

$$N_k = \begin{cases} \min((q_0 - k)/\nu + 1, N), & |k| \le q_0 \\ 0, & |k| > q_0, \end{cases} \tag{7}$$

where $c_k(t)$ is the k-th order cepstrum at time t and $m_k(t)$ denotes the masking pattern. The parameter τ indicates the time interval between the current and the preceding spectra. The masking window length N_k limits the duration of the masking effect and is dependent on the cepstral order k. Initial masking decay α, masking decay β, cutoff quefrency decay ν, cutoff frequency decay q_0 and maximal time window N are experimentally optimized parameters. Let's denote a Dynamic Cepstrum at time t by $c_{\mathrm{Dy}}(t)$, then it can be expressed as follows:

$$c_{\mathrm{Dy}}(t) = \sum_{\tau=0}^{N_{\max}} \boldsymbol{D}(\tau)c(t-\tau), \tag{8}$$

where $\boldsymbol{D}(\tau)$ is a diagonal matrix and $N_{\max} = \max(N_1, N_2, ..., N_k)$. Eq.(8) shows that the Dynamic Cepstrum can be expressed by a linear combination of a partial cepstrum sequence and is calculated by a convolution of $\boldsymbol{D}(\tau)$ and $c(\tau)$. Furthermore, the filter coefficients are given in matrix form. Therefore, the Dynamic Cepstrum can be interpreted as the output from a linear filter with a cepstrum sequence as input.

GENERATION OF FEATURE PARAMETERS BY LINEAR FILTERING OF A CEPSTRUM SEQUENCE

Feature parameters by linear filters with scalar coefficients

The concept of the conventional delta-cepstrum can be easily extended, if $h(\tau)$ is not chosen according to Eq.(3) but according to some other criterion. Namely, if an input signal into multiple filters is given by a cepstrum sequence, the output signal from each filter with different frequency characteristics can be used as a new feature parameter. This idea is similar to a subband division. Fig. 2 shows the concept. In our experiment, the number of linear filters N is set to 2 and the following three types of filter characteristics (Type I, Type II and Type III) are considered.

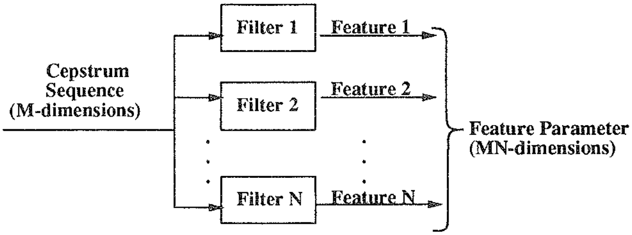

Figure 2: Concept of generating new feature parameters for speech recognition

1. Type I : Cepstrum and delta-cepstrum
   Frequency characteristics for this type are shown in Fig. 3(a). The output from the filter represented by the solid line is the conventional delta-cepstrum and the output from the filter represented by the dotted line is the cepstrum itself.

2. Type II : Feature parameters based on a general band division
   The frequency band can be simply divided into two subbands as shown in Fig.3(b). New feature

parameters from each band are then generated. In this paper, a lowpass filter (dotted line) and a bandpass filter (solid line) are proposed. The lowpass filter is realized by using a Hamming window function $h_{\mathrm{low}}(\tau)$ in the cepstral domain. If we denote the Fourier transform of $h_{\mathrm{low}}(\tau)$ by $H(\omega)$, then the function $h_{\mathrm{band}}(\tau)$ is chosen as the inverse Fourier transform of $H(\omega)$ shifted by ω_0 on the frequency axis. Therefore, the function $h_{\mathrm{band}}(\tau)$ can be expressed as

$$h_{\mathrm{band}}(\tau) = 2(0.54 + 0.46 \cos \frac{2\pi\tau}{2M+1}) \sin(\omega_0 \tau). \tag{9}$$

3. Type III : Feature parameters based on an expansion of Type I
   Frequency characteristics for this type are shown in Fig.3(c). This type can be regarded as an expansion of Type I as the frequency range of the bandpass filter is also covered by the lowpass filter. As is understood from some experiments later, outputs from two filters with this type will provide better feature parameters for speech recognition than conventional parameters.

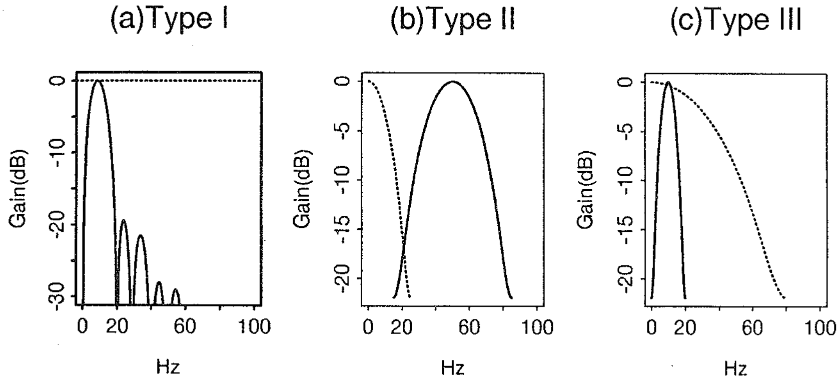

Figure 3: Frequency characteristics for each filter type

Feature parameters by linear filters with matrix coefficients

The Dynamic Cepstrum is calculated by using Eq.(8). Here, parameters α, β, ν, N and q_0 have been optimized in preliminary experiments (Aikawa,Kawahara and Tohkura,1992) and are shown in Table 3.

RECOGNITION PERFORMANCE

As described in the previous section, the form of the filter coefficients in linear filtering of a cepstrum sequence can be divided into scalar form and matrix form. This section describes phoneme recognition performance for each form.

Scalar coefficient linear filter

The new feature parameters obtained by linear filtering of a cepstrum sequence were extracted under the analysis condition as shown in Table 1. Phoneme HMM's were trained with isolated words and tested using 24 phonemes extracted from phrase-by-phrase utterances.

First, the cepstrum and the delta-cepstrum obtained by linear filters with frequency characteristics of Type I were evaluated in speaker-dependent phoneme recognition experiments using two male and two female speakers. The results are shown in Table 2.

Table 1: Analysis conditions

sampling frequency	12kHz, 16bit
pre-emphasis	$1 - 0.98z^{-1}$
window length	20 ms (Hamming window)
frame shift	5 ms
LPC analysis order	16
LPC cepstrum order	16
Δ window length	100 ms

Second, new feature parameters obtained by linear filters with frequency characteristics of Type II were evaluated. The following two cases of Type II are considered here. One is the case of using both a lowpass filter with cutoff frequency around 20Hz and a bandpass filter with a passband between around 10Hz and 80Hz. The other is the case of using both a lowpass filter with cutoff frequency around 80Hz and a bandpass filter with a passband between around 60Hz and 100Hz. However, for these two cases of Type II, no improvement in recognition performance was found.

Finally, new feature parameters obtained by linear filters with frequency characteristics of Type III were evaluated in using the same four speakers. Table 2 shows recognition error rate for the four speakers. For this type, a lowpass filter with cutoff frequency around 80Hz and a bandpass filter with a passband between around 0Hz and $(800/(2M+1))(Hz) : M = 17, 18..., 23)$ are used. Here, $(800/(2M+1))$ is a zero point of the mainlobe for $H(\omega - 4\pi/(2M+1))$.

The results show that combination of the lowpass filter with cutoff frequency around 20Hz and the bandpass filter with a passband between around 0Hz and 20Hz gives lower recognition error rate, and that the average recognition error rate for the four speakers is reduced from 12.2% to 10.2%.

Table 2: Phoneme recognition error rates (Japanese 24 phonemes) (%)

window type	type I	type III						
window length(2M+1)	21	35	37	39	41	43	45	47
speaker MAU	19.36	—	13.98	13.58	14.45	14.54	—	—
speaker MHT	11.38	—	—	—	11.43	11.30	11.70	11.56
speaker FSU	7.59	7.00	6.86	6.80	6.47	6.58	—	—
speaker FTK	10.41	9.38	9.56	11.98	14.53	—	—	—

Matrix coefficient linear filter

The Dynamic Cepstrum was extracted under the analysis condition as shown in Table 3. For obtaining the dynamic feature of a log-power contour, a Dynamic Power is defined as the scalar case of the Dynamic Cepstrum (Aikawa, Kawahara and Tohkura, 1992). Phoneme HMM's for 1 male speaker were trained with the odd numbered words of a Japanese database of 5240 common words and tested using 24

Table 3: Optimal parameters for the Dynamic Cepstrum

initial masking decay α	0.25
masking decay β	0.5
cutoff quefrency decay ν	1.0
cutoff quefrency q_0	6
max. time window N	4

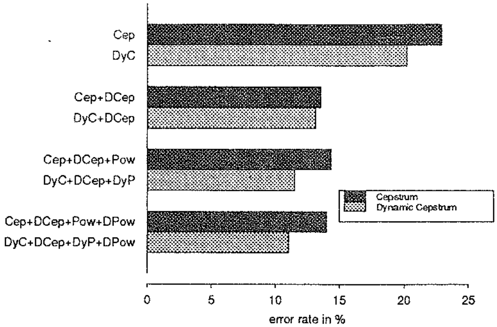

Figure 4: Recognition performance for Dynamic Cepstrum

phonemes extracted from a database of short phrases.

Fig. 4 shows recognition result for this experiment. Dynamic Cepstrum is performing between 2 % and 4 % better than the cepstrum also in combinations with other features like delta-cepstrum (DCep), delta-power(DPow), and Dynamic Power (DyP). It must however be noted, that the cepstrum combined with delta-cepstrum performs still better than the Dynamic Cepstrum alone.

CONCLUSION

In this paper, the concept of the conventional delta-cepstrum was extended, based on a new interpretation that the delta-cepstrum is obtained by linear filtering of a cepstrum vector sequence. It was found that the delta-cepstrum can be calculated by a convolution of the filter coefficients and the cepstrum, and that the filter coefficients are given in scalar form. Furthermore, it was also found that the Dynamic Cepstrum can also be obtained by linear filtering of a cepstrum vector sequence and that the filter coefficients are then given in matrix form.

In the case that the filter coefficients were given in scalar form, the average recognition error rate in a Japanese 24 phoneme recognition experiment for four speakers was reduced from 12.2% to 10.2%. For matrix form filter coefficients, the Dynamic Cepstrum was performing between 2 % and 4 % better than the cepstrum also in combination with other features.

These results show that linear filtering of a cepstrum vector sequence can provide better feature parameters for speech recognition than conventional parameters.

REFERENCES

Aikawa,K, Kawahara,H and Tohkura,Y (1992) *Dynamic Cepstrum Reflecting Time Frequency Masking Characteristics and Its Application to Speech Recognition*, SP92-43, IEICE (in Japanese).

Furui,S (1986) *Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum*, ASSP-34, 1, pp.52-59.

Miyasaka,E (1983) *Spatio-temporal characteristics of masking of brief test-tone pulses by a tone-burst with abrupt switching transients*, JASJ, Vol.39, 9, pp.614-623 (in Japanese).

Sagayama,S and Itakura,F (1979) *On Individuality in a Dynamic Measure of Speech*, ASJ, pp.589-590 (in Japanese).