UDC [621.372.54::681.325.5]:534.782
Electronics and Communications in Japan, Vol. 66-A, No. 2, 1983

Mel Log Spectrum Approximation (MLSA) Filter for Speech Synthesis

Satoshi Imai, Kazuo Sumita* and Chieko Furuichi, Regular Members

Research Laboratory of Precision Machinery and Electronics,
Tokyo Institute of Technology, Yokohama, Japan 227

SUMMARY

The spectral envelope of speech can be represented efficiently by the log magnitude spectrum on the nonlinear frequency scale, which is close to mel scale (called mel-log spectrum). The mel cepstrum defined by its Fourier coefficients is also considered to have a suitable property as the parameter to represent the spectral envelope. So far, no satisfactory filter has been reported for the synthesis approximating the mel-log spectrum. This paper presents a method of constructing the mel-log spectrum approximation (MLSA) filter, which has a relatively simple structure and a low coefficient sensitivity, together with a design example of MLSA filter for speech synthesis. The transfer function of MLSA filter is represented by Padé approximation, which approximates the exponential of the transfer function of the filter (so-called basic filter). Since the transfer function of the basic filter is represented by a polynomial with the transfer function of the first-order all-pass filter as the variable, it is necessary in the realization of the filter to delete from the feedback loop the path without a delay. By the construction method of MLSA filter shown in this paper, the path without delay can easily be deleted from the feedback loop in the MLSA filter. The obtained MLSA filter is of relatively simple structure and has low coefficient sensitivity. The quantization characteristics of the coefficient are also satisfactory.

1. Introduction

Since the spectral envelope of the speech is generally of pole-zero type and the sensation to the sound level is logarithmic, the spectral envelope can well be represented by the log spectrum. The cepstrum is defined by the Fourier coefficient of the log spectrum, and is matched to the characteristics of the hearing sensation while retaining the correspondence to the original spectrum. The sensitivity of the log spectral envelope to the cepstrum is low and the cepstrum has a good quantization property.

In the log spectral envelope, the cepstrum has low distortion in the interpolation along the time axis [1]. The extraction of the cepstrum is relatively easy, and can be used as the parameter in the pole-zero model. In the speech synthesis, the log magnitude approximation filter (LMA filter) can be used, which applies the cepstrum directly to the filter coefficients. However, the order of the cepstrum required in speech synthesis with satisfactory quality, generally, is slightly larger than that of the parameter in the linear prediction method.

By considering the hearing characteristics in regard to the sound level, it seems more desirable to represent the speech spectral envelope by the log magnitude spectrum on a nonlinear frequency scale such as mel scale (called mel-log spectrum). This will lead to a more efficient representation than the usual log spectrum, and by using the mel cepstrum defined by its Fourier coefficients, the spectral envelope can be represented by the order which is lower than that in usual cepstrum representation.

Until now, however, no satisfactory filter has been reported which can approximate the mel-log spectrum for speech synthesis. It is considered that the log magnitude spectrum can better be approximated by recursive filter than by nonrecursive filter, and it is a natural consequence that the mel-log spectrum is approximated by a recursive

*Presently with Central Res. Lab., Toshiba Co.

ISSN0424-8368/83/0002-0010$7.50/0

filter. As the recursive filter realizing a frequency response on the nonlinear frequency scale such as mel scale, a construction method has been proposed of a nonlinear frequency linear prediction synthesis filter by Strube [3]. His method, however, is difficult to apply to the realization of mel-log spectrum approximation (MLSA) filter.

The transfer function of MLSA filter can be represented by a rational function with the transfer function of a first-order, all-pass filter as the variable. Consequently, it may contain a path without a delay in the feedback loop of the filter, which must be deleted. In Strube's method, to delete the path without a delay from the feedback loop, the denominator polynomial of the transfer function, with the transfer function of the first-order all-pass filter as the variable, is a polynomial with the transfer function of the first-order low-pass filter with the same parameter as the all-pass filter as the variable. In this method, the sensitivity of the log magnitude response to the filter coefficient becomes very large, and at the same time, the coefficient values exhibit a wide distribution. For those reasons, it is almost impossible to apply the method to MLSA filter.

This paper presents a construction method for MLSA filter, resulting in a relatively simple structure and low coefficient sensitivity. A design example is shown for the MLSA filter for speech synthesis. In the proposed construction method, in contrast to Strube's method, the path without a delay is deleted from the feedback loop by essentially modifying the denominator polynomial of the transfer function, which uses the transfer function of the first-order, all-pass filter as the variable, into a polynomial in which the terms not containing the unit delay variable as a factor and the terms containing the variable as a first-order factor are separated.

2. Mel-Log Spectrum and Mel Cepstrum

2.1 Approximate representation of mel-scale

One of the frequency scales that can well approximate the mel scale, leading to easy modelling of the spectrum, is as follows. Consider an all-pass filter for which the transfer function $H^{(\alpha)}(z)$ and the frequency response $H^{(\alpha)}(e^{j\Omega})$ ($\Omega = \omega \Delta t$, ω: radian frequency, Δt: unit delay in the digital filter) are given by

$$H^{(\alpha)}(z) = (z^{-1} - \alpha) / (1 - \alpha z^{-1}) \qquad (1)$$

$$H^{(\alpha)}(e^{j\Omega}) = \exp(-j \beta_\alpha(\Omega)) \qquad (2)$$

The phase response is given by

$$\beta_\alpha(\Omega) = \tan^{-1} \frac{(1-\alpha^2)\sin\Omega}{(1+\alpha^2)\cos\Omega - 2\alpha} \qquad (3)$$

Then the nonlinear frequency scale is defined as

$$\tilde{\Omega} = \beta_\alpha(\Omega) \qquad (4)$$

In the phase response $\beta_\alpha(\Omega)$ of this all-pass filter, the unit delay of the digital filter Δt is set as 100 μs, 125 μs (=(8 kHz)$^{-1}$), or 156.25 μs (=(6.4 kHz)$^{-1}$), and α in the transfer function $H^{(\alpha)}(z)$ of Eq. (1) is set as 0.35, 0.31, or 0.28. Then the error to the mel scale is less than 2.5% of the full-scale value which is satisfactory. When α = 0, it is the same as the linear scale. The parameter α of the all-pass filter is called the frequency compression parameter. The scale defined by Eq. (4) includes the linear scale as a special case and can well represent the nonlinear frequency scale such as mel scale. It is called simply mel frequency scale in this paper.

2.2 Mel-log spectrum

Letting the usual log spectrum on the linear frequency scale be $G_0(\Omega)$, the log magnitude spectrum $G_\alpha(\tilde{\Omega})$ on the mel-frequency scale $\tilde{\Omega}$ is given by

$$G_\alpha(\tilde{\Omega}) = G_0(\beta_\alpha^{-1}(\tilde{\Omega})) \qquad (5)$$

$$\beta_\alpha^{-1}(\tilde{\Omega}) = \tan^{-1} \frac{(1-\alpha^2)\sin\tilde{\Omega}}{(1+\alpha^2)\cos\tilde{\Omega} + 2\alpha} \qquad (6)$$

where $\beta_\alpha^{-1}(\tilde{\Omega})$ is the inverse function of the phase response $\beta_\alpha(\Omega)$ of the all-pass filter given by Eq. (3).

The log magnitude spectrum $G_\alpha(\tilde{\Omega})$ on the mel-frequency scale $\tilde{\Omega}$ is simply called mel-log spectrum. The mel-log spectrum representing the spectral envelope of the speech waveform is called mel-log spectral envelope.

2.3 Mel cepstrum

Mel cepstrum $c_\alpha(m)$ is defined by the Fourier coefficients of the mel-log spectrum $G_\alpha(\tilde{\Omega})$ as

$$g_\alpha(m) = \frac{1}{2\pi} \int_{-\pi}^{\pi} G_\alpha(\tilde{\Omega})\, e^{jm\tilde{\Omega}}\, d\tilde{\Omega} = g_\alpha(-m) \qquad (7)$$

as

$$c_\alpha(m) = \begin{cases} 2\, g_\alpha(m) & (m > 0) \\ g_\alpha(m) & (m = 0) \\ 0 & (m < 0) \end{cases} \qquad (8)$$

Consequently, the mel-log spectrum $G_\alpha(\tilde{\Omega})$ can be represented by the mel spectrum $c_\alpha(m)$ as

$$G_\alpha(\tilde{\Omega}) = \sum_{m=0}^{\infty} c_\alpha(m) \cos(m\tilde{\Omega}) \tag{9}$$

3. Direct Approximation of Mel-Log Spectrum

3.1 Frequency response on mel-frequency scale

Let the transfer function $\tilde{z}^{-1} = H^{(\alpha)}(z)$ of Eq. (1) be

$$\tilde{z}^{-1} = H^{(\alpha)}(z) \tag{10}$$

Then if there exists the relation

$$H_\alpha(\tilde{z}) = H(z) \tag{11}$$

for any transfer function $H(z)$, it follows from Eqs. (1), (2) and (3) that

$$H_\alpha(e^{j\tilde{\Omega}}) = H(e^{j\Omega}) \tag{12}$$

In other words, $H_\alpha(e^{j\tilde{\Omega}})$ represents the frequency response on the mel-frequency scale $\tilde{\Omega}$, and $H_\alpha(\tilde{z})$ represents the transfer function of the system with frequency response $H_\alpha(e^{j\tilde{\Omega}})$ on $\tilde{\Omega}$.

3.2 Filter with exponential transfer function

Assume that the transfer function $H_\alpha(\tilde{z})$ of the digital filter can be represented, using the transfer function $F_\alpha(\tilde{z})$ of the digital filter called basic filter, as

$$H_\alpha(\tilde{z}) = \exp(F_\alpha(\tilde{z})) \tag{13}$$

If the basic filter is stable, its transfer function $F_\alpha(\tilde{z})$ does not have a pole on or outside of the unit circle on the $\tilde{z}$-plane. Consequently, the transfer function $H_\alpha(\tilde{z})$ of Eq. (13) takes a finite value in that region. Since the inside, circumference, and outside of the unit circle on the $\tilde{z}$-plane correspond, respectively, to the inside, circumference and outside of the unit circle of the z-plane, it follows that if the basic filter is stable, the filter of Eq. (13) with exponential transfer function is always stable and is a minimum-phase system.

The basic filter is arbitrary as far as it is stable. In the following, a nonrecursive-type filter is considered in which the transfer function $F_\alpha(\tilde{z})$ can be represented as

$$F_\alpha(\tilde{z}) = \sum_{m=0}^{M} c_\alpha(m) \tilde{z}^{-m} \tag{14}$$

In such a case, the log magnitude response $\ln|H_\alpha(e^{j\tilde{\Omega}})|$ of the filter with exponential transfer function on the mel-frequency scale is given by Eqs. (13) and (14) as

$$\ln|H_\alpha(e^{j\tilde{\Omega}})| = \sum_{m=0}^{M} c_\alpha(m) \cos(m\tilde{\Omega}) \tag{15}$$

Letting $c_\alpha(m)$ in Eqs. (14) and (15) be the mel cepstrum corresponding to the desired mel-log spectral envelope, it follows from the property of the Fourier series that the log magnitude response of the filter on the mel-frequency scale is the best approximation in the sense of least mean-square error to the desired mel-log spectral envelope. It is actually impossible directly to realize the digital filter with the transfer function of Eq. (13). However, by approximating the transfer function $\exp(F_\alpha(\tilde{z}))$ of exponential type by a polynomial or a rational function of the transfer function $F_\alpha(\tilde{z})$ of the basic filter, the digital filter can be realized. The system with the transfer function of Eq. (13) or its adequate approximation is called mel-log spectrum approximation (MLSA) filter.

4. Construction of MLSA Filter

4.1 Padé-approximation of exponential function

MLSA filter is realized by approximating the transfer function $\exp(F_\alpha(\tilde{z}))$ of exponential type by a rational function of the transfer function $F_\alpha(\tilde{z})$ of the basic filter. The complex exponential function $\exp(w)$ can be approximated by the modified Padé-approximation of (L, L)-th order:

$$R_L(w) = \frac{P_L(w)}{P_L(-w)} \qquad \left(\begin{matrix} L = 1, 2, \cdots; \\ |w| \leq \bar{r}_L \end{matrix}\right) \tag{16}$$

$$P_L(w) = P_L^{\mathrm{I}}(w) = 1 + \sum_{l=1}^{L} p_{L,l}^{\mathrm{I}} w^l \tag{17}$$

$$p_{L,l}^{\mathrm{I}} = \frac{1 - \lambda_{L,l}}{l!} \binom{L}{l} \Big/ \binom{2L}{l} \tag{18}$$

The coefficients $\lambda_{L,l}$ $(L = 1,2,3,4;\ 1 \leq l \leq L)$ are selected so that the approximation error on the logarithmic function scale

$$E_L(\bar{r}_L) = \max_{|w| \leq r_L} |\ln R_L(w) - \ln(\exp w)| \tag{19}$$

is minimized [2]

$$\lambda_{1,1} = 0.13 \times 10^{-1}\ \bar{r}_1^{\,4} \qquad (\bar{r}_1 < W_1 = 2.00) \tag{20}$$

$$\left.\begin{aligned}\lambda_{2,1} &= 0.44\times10^{-5}\,\bar{r}_2^{\,8}\\ \lambda_{2,2} &= 0.11\times10^{-2}\,\bar{r}_2^{\,4}\end{aligned}\right\}\quad(\bar{r}_2<W_2=3.48) \tag{21}$$

$$\left.\begin{aligned}\lambda_{3,1} &= 0.40\times10^{-9}\,\bar{r}_3^{\,12}\\ \lambda_{3,2} &= 0.74\times10^{-6}\,\bar{r}_3^{\,8}\\ \lambda_{3,3} &= 0.30\times10^{-3}\,\bar{r}_3^{\,4}\end{aligned}\right\}\quad(\bar{r}_3<W_3=4.75) \tag{22}$$

$$\left.\begin{aligned}\lambda_{4,1} &= 0.49\times10^{-14}\,\bar{r}_4^{\,16}\\ \lambda_{4,2} &= 0.42\times10^{-10}\,\bar{r}_4^{\,12}\\ \lambda_{4,3} &= 0.75\times10^{-7}\,\bar{r}_4^{\,8}\\ \lambda_{4,4} &= 0.11\times10^{-3}\,\bar{r}_4^{\,4}\end{aligned}\right\}\quad(\bar{r}_4<W_4=6.25) \tag{23}$$

where W_L ($L=1,2,3,4$) represents the bounds for the variable for which the modified Padé approximation has neither pole nor zero.

In this paper, the modified Padé approximation is called simply Padé approximation. The approximation error $E_L(\bar{r}_L)$, on the logarithmic function scale, of the Padé approximation of (L, L)-th order is bounded by [2]

$$E_L(\bar{r}_L)\lessapprox\frac{2L!\,(L+1)!}{(2L)!\,(2(L+1))!}\,\bar{r}_L^{2L+1}\quad(\bar{r}_L<W_L) \tag{24}$$

To suppress the approximation error to ε or less, the upper bound $\bar{r}_L$ of the variable should be set as

$$\bar{r}_L\lessapprox\left(\frac{(2L)!\,(2(L+1))!\,\varepsilon}{2L!\,(L+1)!}\right)^{\frac{1}{2L+1}} \tag{25}$$

For example, if the approximation error is to be suppressed to 0.023 (0.2 dB) or less, the upper bound $\bar{r}_L$ of the variable should be set so as not to exceed 0.65 ($L=1$), 1.75 ($L=2$), 3.03 ($L=3$) or 4.37 ($L=4$).

The numerator polynomial $P_L(w)$ of the Padé approximation $R_L(w)$ can be written as

$$P_L(w)=P_L^{\mathrm{I}}(w)=1+\sum_{l=1}^{L}p_{L,1}^{\mathrm{I}}\,p_{L,2}^{\mathrm{I}}\cdots p_{L,l}^{\mathrm{I}}\,w^l \tag{26}$$

$$p_{L,l}^{\mathrm{I}}=\frac{L-l+1}{2L-l+1}\,\frac{1-\lambda_{L,l}}{1-\lambda_{L,l-1}}\quad(\lambda_{L,0}=0) \tag{27}$$

or

$$P_L(w)=P_L^{\mathrm{II}}(w)=1+\sum_{l=1}^{L}p_{L,1}^{\mathrm{II}}\,p_{L,2}^{\mathrm{II}}\cdots p_{L,l}^{\mathrm{II}}\,2^{-\frac{(l-1)(l+2)}{2}}\,w^l \tag{28}$$

$$p_{L,l}^{\mathrm{II}}=\frac{2^l\,(L-l+1)}{2L-l+1}\,\frac{1-\lambda_{L,l}}{1-\lambda_{L,l-1}}\quad(\lambda_{L,0}=-1) \tag{29}$$

These representations are more suitable from the viewpoint of the accuracy in computation. Padé approximation $R_L(w)$, in which the numerator polynomial is represented as $P_L^J(w)$ ($J=\mathrm{I},\mathrm{II},\mathrm{III}$), is written as $R_L^J(w)$ ($J=\mathrm{I},\mathrm{II},\mathrm{III}$).

4.2 Transfer function of MLSA filter.

To realize a MLSA filter, the transfer function $\exp(F_\alpha(\tilde{z}))$ of exponential type is approximated by the Padé approximation of Eq. (16). The resulting transfer function $H_\alpha(\tilde{z})$ of the MLSA filter is given by

$$H_\alpha(\tilde{z})=R_L\left(F_\alpha(\tilde{z})\right) \tag{30}$$

If the maximum magnitude of the basic filter

$$r_L=\max_{\tilde{\omega}}\left|F_\alpha(e^{j\tilde{\omega}})\right| \tag{31}$$

does not exceed the upper bound $\bar{r}_L$ for the variable range for which the Padé approximation is effective, the log magnitude response of the MLSA filter $\ln|H_\alpha(e^{j\tilde{\omega}})|$ is given by

$$\begin{aligned}\ln\left|H_\alpha(e^{j\tilde{\omega}})\right| &= F_\alpha(e^{j\tilde{\omega}})+e_L\\ &\simeq F_\alpha(e^{j\tilde{\omega}})\end{aligned} \tag{32}$$

$$|e_L|\le E_L(\bar{r}_L) \tag{33}$$

where $E_L(\bar{r}_L)$ is the approximation error of Pade approximation given by Eq. (24).

When the transfer function $F_\alpha(\tilde{z})$ of the basic filter is specified by the speech mel cepstrum $c_\alpha(m)$ through Eq. (14) and if the amplitude $|F_\alpha(e^{j\tilde{\omega}})|$ of the basic filter does not exceed the upper bound $\bar{r}_L$ for the variable range to make the Padé approximation effective, the log magnitude response $\ln|H_\alpha(e^{j\tilde{\omega}})|$ of the MLSA filter is a good approximation to the mel-log spectral envelope of the speech.

4.3 Modification of transfer function of basic filter

The transfer function of $F_\alpha(\tilde{z})$ ($=F(z)$) of the basic filter given by Eq. (14) contains terms which do not have the unit-delay variable z^{-1} as the factor. As itself, MLSA filter of transfer function $R_L(F_\alpha(\tilde{z}))$ of Eq. (30) cannot be realized.

Applying Strube's method (called method I) to the transfer function $F_\alpha(\tilde{z})$ ($=F(z)$) of the basic filter for the MLSA filter, Eq. (14) is modified as [3]

$$F_\alpha(\tilde{z}) = F_\alpha^{\mathrm{I}}(\tilde{z}) = F(z)$$
$$= b_\alpha^{\mathrm{I}}(0) + \sum_{m=1}^{M} b_\alpha^{\mathrm{I}}(m) \left(\frac{z^{-1}}{1-\alpha z^{-1}}\right)^m \quad (34)$$

$$b_\alpha^{\mathrm{I}}(m) = (1-\alpha^2)^m \sum_{k=m}^{M} (-\alpha)^{k-m} \binom{k}{m} c_\alpha(k) \quad (35)$$

In this way, the term which does not contain the unit-delay variable z^{-1} as a factor can be separated from the transfer function of the basic filter so that the feedback loop of the MLSA filter with the transfer function of Eq. (30) does not contain a path without a delay. As will be described later, it is difficult to apply the basic filter, obtained by method I to MLSA filter. In this paper, two methods of effective modification are proposed for the transfer function of the basic filter.

By Eqs. (1) and (10), the unit-delay variable z^{-1} can be represented, by the transfer function $\tilde{z}^{-1}$ of the all-pass filter, as

$$z^{-1} = \frac{\alpha + \tilde{z}^{-1}}{1+\alpha\tilde{z}^{-1}} \quad (36)$$

Consequently, separating the transfer function $F_\alpha(\tilde{z})\,(=F(z))$ (of the basic filter into terms containing z^{-1} as a factor and the terms not containing z^{-1}, corresponds to the separation of $F_\alpha(\tilde{z})$ into the terms containing $(\alpha+\tilde{z}^{-1})$ as a factor and the terms not containing it.

Method I represents the transfer function $F_\alpha(\tilde{z})$ of the basic filter as a polynomial of $(\alpha+\tilde{z}^{-1})$. As method II, one may modify the transfer function $F_\alpha(\tilde{z})$ of the basic filter as

$$F_\alpha(\tilde{z}) = F_\alpha^{\mathrm{II}}(\tilde{z})$$
$$= B_\alpha^{\mathrm{II}}(0) + (\alpha+\tilde{z}^{-1}) \sum_{m=1}^{M} B_\alpha^{\mathrm{II}}(m)\,\tilde{z}^{-(m-1)}$$
$$= \frac{b_\alpha^{\mathrm{II}}(0)}{1-\alpha} + \frac{z^{-1}}{1-\alpha z^{-1}} \sum_{m=1}^{M} b_\alpha^{\mathrm{II}}(m)\tilde{z}^{-(m-1)} \quad (37)$$

Comparing the coefficients of Eqs. (14) and (37), it is seen that the coefficient $b_\alpha^{\mathrm{II}}(m)$ of the filter can be obtained from the mel cepstrum $c_\alpha(m)$ as

$$b^{\mathrm{II}}(m) = \begin{cases} (1-\alpha^2)\,B_\alpha^{\mathrm{II}}(m) & (1 \leqq m \leqq M) \\ (1-\alpha)\,B_\alpha^{\mathrm{II}}(m) & (m=0) \end{cases} \quad (38)$$

$$B_\alpha^{\mathrm{II}}(M+1) = 0 \quad (39)$$

$$B_\alpha^{\mathrm{II}}(m) = c_\alpha(m) - \alpha B_\alpha^{\mathrm{II}}(m+1)$$
$$(m = M, M-1, \cdots, 1, 0) \quad (40)$$

As method III, the transfer function $F_\alpha(\tilde{z})$ of the basic filter is modified as

$$F_\alpha(\tilde{z}) = F_\alpha^{\mathrm{III}}(\tilde{z})$$
$$= b_\alpha^{\mathrm{III}}(0) + \frac{\alpha+\tilde{z}^{-1}}{1+\alpha\tilde{z}^{-1}} \sum_{m=1}^{M+1} b_\alpha^{\mathrm{III}}(m)\,\tilde{z}^{-(m-1)}$$
$$= b_\alpha^{\mathrm{III}}(0) + z^{-1} \sum_{m=1}^{M+1} b_\alpha^{\mathrm{III}}(m)\,\tilde{z}^{-(m-1)} \quad (41)$$

Comparing coefficients of Eqs. (14) and (41), the coefficient $b_\alpha^{\mathrm{III}}(m)$ of the filter is obtained from the mel cepstrum $c_\alpha(m)$ as

$$b_\alpha^{\mathrm{III}}(M+1) = \alpha c_\alpha(M) \quad (42)$$

$$b_\alpha^{\mathrm{III}}(m) = c_\alpha(m) + \alpha\,(c_\alpha(m-1) - b_\alpha^{\mathrm{III}}(m+1))$$
$$(m = M, M-1, \cdots, 3, 2) \quad (43)$$

$$b_\alpha^{\mathrm{III}}(1) = (c_\alpha(1) - \alpha\, b_\alpha^{\mathrm{III}}(2)) / (1-\alpha^2) \quad (44)$$

$$b_\alpha^{\mathrm{III}}(0) = c_\alpha(0) - \alpha\, b_\alpha^{\mathrm{III}}(1) \quad (45)$$

In the calculation of the coefficient of the basic filter $b_\alpha^K(m)\,(K = \mathrm{I}, \mathrm{II}, \mathrm{III};\ m = 0, 1, \cdots, M_K;\ M_{\mathrm{I}} = M_{\mathrm{II}} = M,\ M_{\mathrm{III}} = M+1)$ from the mel cepstrum $c_\alpha(m)$, the amount of computations is $(M+1)^2/2$ multiplication-addition in method I, $2(M+1)$ multiplications and $(M+1)$ additions in method II, and $(M+3)$ multiplications and $2M$ additions in method III. Also, it is possible to perform the computation in method I by a recursive formula as in methods II and III, but the amount of computation does not change much.

Usually, the order M of the speech mel cepstrum is set between 8 and 15. Consequently, methods II and III require less computation than method I. When the fixed-point arithmetic is used, where addition and subtraction are easier than multiplication, method III is the simplest.

5. Coefficient Sensitivity of MLSA Filter

5.1 Variation of log magnitude due to coefficient error

Consider the case where the transfer function $F_\alpha(\tilde{z})$ of the basic filter is given by

$$F_\alpha(\tilde{z}) = F_\alpha^K(\tilde{z}) \quad (K = \mathrm{I}, \mathrm{II}, \mathrm{III}) \quad (46)$$

and the Padé approximation $P_L(w)$ approximating the exponential function $\exp(w\,F_\alpha(\tilde{z}))$ of $F_\alpha(\tilde{z})$ is represented as

$$R_L(w) = R_L^J(w) \quad (J = \mathrm{I}, \mathrm{II}, \mathrm{III}) \quad (47)$$

Then consider the variation in the log magnitude response characteristic due to the error in the coefficients of MLSA filter.

Padé approximation has the coefficients $p^J_{L,l}$ ($J=\mathrm{I},\mathrm{II},\mathrm{III}$; $L=1,2,3,4$; $1\leqq l\leqq L$), and the transfer function of the basic filter has the coefficients $b^K_\alpha(m)$ ($K=\mathrm{I},\mathrm{II},\mathrm{III}$; $m=0,1,\cdots,M_K$; $M_\mathrm{I}=M_\mathrm{II}=M$, $M_\mathrm{III}=M+1$). Let the errors in $p^J_{L,l}$ and $b^K_\alpha(m)$ be $\Delta p^J_{L,l}$, and $\Delta b^K_\alpha(m)$, respectively. Then the variation ΔS^{JK} of the log magnitude of MLSA filter is given by

$$\Delta S^{JK} \simeq \sum_{l=1}^{L}\left(\frac{\partial}{\partial p^J_{L,l}}\ln|R^J_L(F^K_\alpha(e^{j\tilde{\Omega}}))|\right)\Delta p^J_{L,l} + \sum_{m=0}^{M_K}\left(\frac{\partial}{\partial b^K_\alpha(m)}\ln|R^J_L(F^K_\alpha(e^{j\tilde{\Omega}}))|\right)\Delta b^K_\alpha(m) \quad (J,K=\mathrm{I},\mathrm{II},\mathrm{III},\ M_\mathrm{I}=M_\mathrm{II}=M,\ M_\mathrm{III}=M+1) \tag{48}$$

The discussion of the variation of Eq. (48) is meaningful only in the range where the exponential type transfer function $\exp(F_\alpha(\tilde{z}))$ can well be approximated by Padé approximation $R_L(F_\alpha(\tilde{z}))$, and the following discussion is made only for the case where the logarithmic error of Padé approximation is 0.023 (0.2 dB) or less, i.e., the variable does not exceed the upper bound $\bar{r}_L$ given by Eq. (25).

5.2 Sensitivity to Padé approximation coefficients

Padé approximation $R_L(\omega)$ is characterized by its numerator polynomial $P_L(\omega)$, and $P_L(\omega)$ is represented by Eqs. (17), (26) or (28). The maximum sensitivity $S(p^J_{L,l})$ of the log magnitude of Padé approximation to the coefficient $p^J_{L,l}$ ($J=\mathrm{I},\mathrm{II},\mathrm{III}$) of numerator polynomials is given by

$$S(p^J_{L,l}) = \max_{|w|\leqq\bar{r}_L}\left|\frac{\partial}{\partial p^J_{L,l}}\ln|R^J_L(w)|\right| \quad (w=F_\alpha(e^{j\tilde{\Omega}})) \tag{49}$$

Calculating numerically the maximum sensitivity $S(p^J_{L,l})$ of the log magnitude of Padé approximation to the coefficient $p^J_{L,l}$ ($J=\mathrm{I},\mathrm{II},\mathrm{III}$), Table 1 is obtained. For MLSA filter, the amount of computation in filtering is almost the same, independently of the type of Padé approximation $R^J_L(w)$ ($J=\mathrm{I},\mathrm{II},\mathrm{III}$). Consequently, it is seen from Table 1 that the Padé approximation $R^\mathrm{II}_L(w)$ is the best suited since the coefficient sensitivity is the lowest.

Table 1. Maximum Sensitivity $S(p^J_{L,l})$ ($J=\mathrm{I},\mathrm{II},\mathrm{III}$) of Logarithmic Amplitude of Padé Approximation $R^J_L(w)$ to Coefficient $p^J_{L,l}$

$S(p^J_{L,l})$	L \ l	1	2	3	4
$S(p^\mathrm{I}_{L,l})$	1	1.5			
	2	5.5	6.7		
	3	18.6	50.1	168.7	
	4	55.7	236.0	1153.9	4380.2
$S(p^\mathrm{II}_{L,l})$	1	1.5			
	2	4.4	3.3		
	3	11.1	18.1	16.8	
	4	24.7	68.5	92.1	50.6
$S(p^\mathrm{III}_{L,l})$	1	1.5			
	2	4.4	0.8		
	3	11.1	4.5	2.1	
	4	24.8	17.7	11.8	3.4

5.3 Sensitivity to basic filter coefficients

The maximum sensitivity $S(b^K_\alpha(m))$ of the log magnitude of MLSA filter to the coefficient $b^K_\alpha(m)$ ($K=\mathrm{I},\mathrm{II},\mathrm{III}$) of the basic filter is given by

$$\begin{aligned} S(b^K_\alpha(m)) &= \max_{|w_K|\leqq\bar{r}_L}\left|\frac{\partial}{\partial b^K_\alpha(m)}\ln|R^J_L(w_K)|\right| \\ &\leqq \max_{|w_K|\leqq\bar{r}_L}\left|\frac{\partial}{\partial b^K_\alpha(m)}\ln|R^J_L(w_K)| + j\frac{\partial}{\partial b^K_\alpha(m)}\arg R^J_L(w_K)\right| \\ &= \max_{|w_K|\leqq\bar{r}_L}\left|\frac{\partial}{\partial b^K_\alpha(m)}\ln R^J_L(w_K)\right| \\ &\simeq \max_{|w_K|\leqq\bar{r}_L}\left|\frac{\partial w_K}{\partial b^K_\alpha(m)}\right| = \bar{S}(b^K_\alpha(m)) \\ &\quad (w_K=F^K_\alpha(e^{j\tilde{\Omega}})) \end{aligned} \tag{50}$$

The maximum sensitivity $S(b^K_\alpha(m))$ of the log magnitude of MLSA filter does not exceed the maximum sensitivity $\bar{S}(b^K_\alpha(m))$ of the logarithmic frequency response; $\bar{S}(b^K_\alpha(m))$ ($K=\mathrm{I},\mathrm{II},\mathrm{III}$) are given by Eqs. (34), (37), or (41), (50) as

$$\bar{S}(b^\mathrm{I}_\alpha(m)) = \max_{\tilde{\Omega}}\left|\frac{\alpha+e^{-j\tilde{\Omega}}}{1-\alpha^2}\right|^m = \frac{1}{(1-\alpha)^m} \tag{51}$$

$$\bar{S}(b^\mathrm{II}_\alpha(m)) = \max_{\tilde{\Omega}}\left|\frac{\alpha+e^{-j\tilde{\Omega}}}{1-\alpha^2}\right|\cdot|e^{-j(m-1)\tilde{\Omega}}| = \frac{1}{1-\alpha} \tag{52}$$

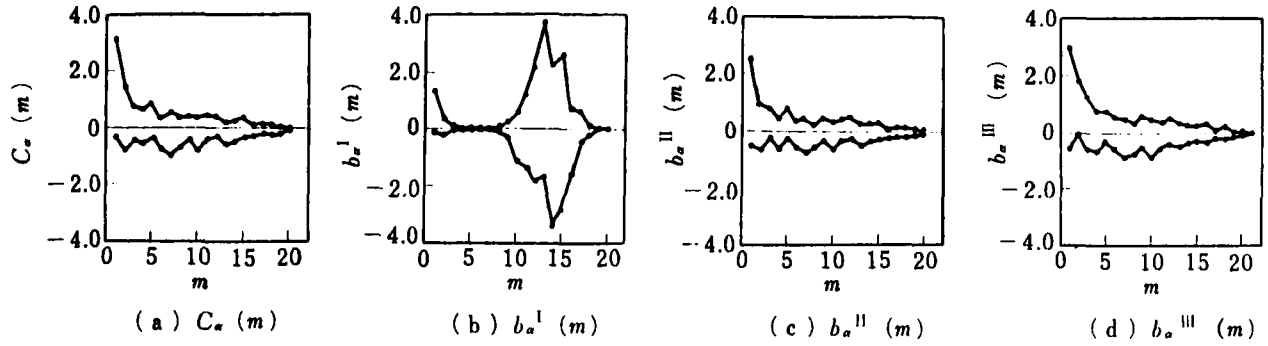


Fig. 1. Maximal and minimal values of mel cepstrum $c_\alpha(m)$ and filter coefficients $b_\alpha^K(m)$ for utterance "nambu dewa higashi no kaze" of a male speaker ($\alpha=0.4$, $K=\mathrm{I},\mathrm{II},\mathrm{III}$).

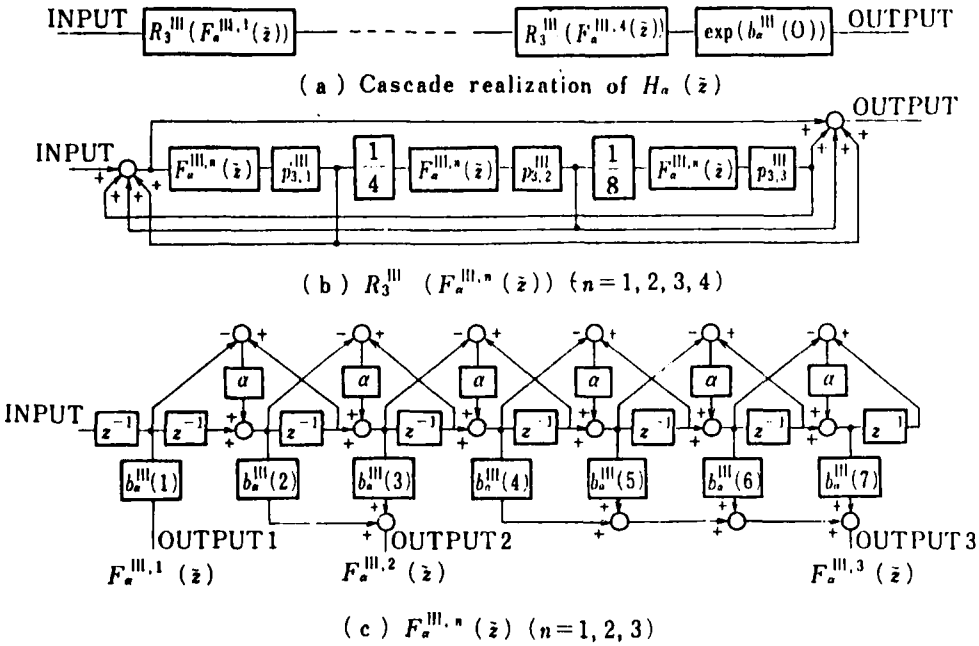


Fig. 2. MLSA filter for speech synthesis: (a) cascade realization; (b) partial filter $R_3^{III}(F_\alpha^{III,n}(\tilde{z}))$; (c) basic filter $F_\alpha^{III,n}(\tilde{z})$.

$$\bar{S}(b_\alpha^{III}(m))=\max_{\tilde{\Omega}}\left|\frac{\alpha+e^{-j\tilde{\Omega}}}{1+\alpha e^{-j\tilde{\Omega}}}\right|\cdot|e^{-j(m-1)\tilde{\Omega}}|=1 \tag{53}$$

Restricting the magnitude response of the basic filter so that the logarithmic error of Padé approximation is 0.023 (0.2 dB) or less, the maximum sensitivity $\bar{S}(b_\alpha^K(m))$ of the logarithmic frequency response of MLSA filter is calculated numerically for the frequency compression parameter of 0.2 and 0.4. Then, it is seen that Eqs. (51) and (52) coincide within the error of 2%. The values of the coefficients $\bar{S}(b_\alpha^{I}(m))$ can be determined when the frequency compression parameter α is small and the order m of the coefficient $b_\alpha^{I}(m)$ is small. Moreover, in that case, they are considerably large.

The maximum sensitivity $S(b_\alpha^K(m))$ of the log magnitude response to the basic filter coefficient $b_\alpha^K(m)$ $(K=\mathrm{I},\mathrm{II},\mathrm{III})$ of MLSA filter does not exceed the maximum sensitivity $\bar{S}(b_\alpha^K(m))$ of the logarithmic frequency response. Thus, one may conclude that the basic filter

by method III, for which $\bar{S}(b_\alpha^{\mathrm{K}}(m))$ is the smallest, has the best quantization characteristic for the coefficient.

6. Design Example of MLSA Filter

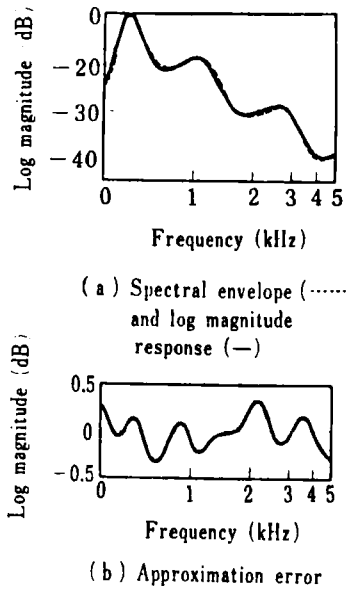
An MLSA filter for speech synthesis is designed. The filter for speech synthesis must approximate naturally well the speech spectral envelope. Other requirements are that the filter coefficients should easily be derived from the parameters representing the spectral envelope, it should have good quantization and time interpolation characteristics of the spectral envelope parameters or filter coefficients, and the filter should have a simple structure with less filtering computation. If an adequate design is made, the MLSA filter satisfies almost all these requirements.

When the speech spectral envelope is represented by the mel-log spectrum, the mel cepstrum can be used as the parameter to represent the spectral envelope, which, however, cannot be used directly as the coefficient of the MLSA filter. The coefficients of the basic filter in MLSA filter, rather than the mel cepstrum, can be used in speech synthesis in a more direct and suitable way.

To apply the filter coefficients as the parameter of the speech spectral envelope, the data compression ratio must be high in its utilization. Consequently, it is required that the sensitivity of the log magnitude of MLSA filter to the filter coefficients is low and the distributions of the filter coefficients are narrow. Figure 1 shows the case of frequency compression parameter α of 0.4, where a male utters "Nanbu dewa higashi no kaze (East wind in southern part)" and the maxima and the minima of the mel cepstrum $c_\alpha(m)$ and the filter coefficient $b_\alpha^{\mathrm{K}}(m)$ $(\mathrm{K}=\mathrm{I},\mathrm{II},\mathrm{III})$ are shown. The sampling frequency is 10 kHz and the speech is digitized with the word length of 12 bits. The frequency analysis is made using 256 point Blackman window and the frame period of 5 ms. The condition for the analysis is the same in the following.

As is shown in Fig. 1, the maximum and the minimum of the filter coefficients $b_\alpha^{\mathrm{II}}(m)$ and $b_\alpha^{\mathrm{III}}(m)$ are almost the same as those of the mel cepstrum $c_\alpha(m)$. Considering the coefficient sensitivity discussed in section 5, it is seen that the most suitable filter coefficient to use is $b_\alpha^{\mathrm{III}}(m)$. The filter coefficient $b_\alpha^{\mathrm{I}}(m)$ has wider distribution and larger coefficient sensitivity (which is not practical).

The mel cepstrum $c_\alpha(m)$ is given as in Eq. (9), by the cosine Fourier coefficients of the mel-log spectral envelope. Consequently, the sensitivity of the mel-log spectral envelope to the mel cepstrum $c_\alpha(m)$ is 1, which is the same as the filter coefficient $b_\alpha^{\mathrm{III}}(m)$. Thus, one may consider that the data compression ratio of speech by $b_\alpha^{\mathrm{III}}(m)$ is nearly the same as that of $c_\alpha(m)$.



(a) Spectral envelope (·····) and log magnitude response (—)

(b) Approximation error

Fig. 3. Mel-log spectral envelope and log magnitude response of 8-bit coefficient MLSA filter, and approximation error for a vowel portion of utterance /na/ of a male speaker ($\alpha = 0.4$, $M = 12$).

As Padé approximation $R_L(w)$ to represent the transfer function of MLSA filter, $R_L^{\mathrm{III}}(w)$ is the best suited with the smallest coefficient sensitivity as was shown in section 5. Thus, the MLSA filter for speech synthesis can be represented as a system with the transfer function, in which the exponential-type function $\exp(F_\alpha^{\mathrm{III}}(\tilde{z}))$ is approximated by Padé approximation $P_L^{\mathrm{III}}(w)$.

To apply Padé approximation $P_L^{\mathrm{III}}(w)$ in the range of good approximation, the transfer function $F_\alpha^{\mathrm{III}}(\tilde{z})$ of the basic filter given by Eq. (41) is partitioned into the sum of $b_\alpha^{\mathrm{III}}(0)$ and $F_\alpha^{\mathrm{III},n}(\tilde{z})$ $(n=1,2,\cdots,N)$ as

$$F_\alpha^{\mathrm{III}}(\tilde{z}) = b_\alpha^{\mathrm{III}}(0) + \sum_{n=1}^{N} F_\alpha^{\mathrm{III},n}(\tilde{z}) \qquad (54)$$

Consequently, the transfer function $H_\alpha(\tilde{z})$ of MLSA filter is given by

$$H_\alpha(\tilde{z}) = \exp\left(b_\alpha^{\mathrm{III}}(0)\right) \prod_{n=1}^{N} R_{L(n)}^{\mathrm{III},n}\left(F_\alpha^{\mathrm{III},n}(\tilde{z})\right) \qquad (55)$$

In the above, care is taken so that the maximum of the frequency response of the term $F_a^{\mathrm{III},n}(\tilde{z})$ obtained by partitioning the transfer function $F_a^{\mathrm{III}}(\tilde{z})$ of the basic filter:

$$r_{L(n)} = \max_{\tilde{\Omega}} | F_a^{\mathrm{III},n}(e^{j\tilde{\Omega}}) | \qquad (56)$$

does not exceed the upper bound $\bar{r}_L$ for which the Padé approximation $R_L(\omega)$ is effective.

Let the terms $F_a^{\mathrm{III},n}(\tilde{z})$ ($n = 1,2,3,4$) obtained by partitioning the transfer function $F_a^{\mathrm{III}}(\tilde{z})$ of the basic filter of Eq. (41) be

$$F_a^{\mathrm{III},1}(\tilde{z}) = z^{-1}\, b_a^{\mathrm{III}}(1) \qquad (57)$$

$$F_a^{\mathrm{III},2}(\tilde{z}) = z^{-1} (b_a^{\mathrm{III}}(2)\,\tilde{z}^{-1} + b_a^{\mathrm{III}}(3)\,\tilde{z}^{-2}) \qquad (58)$$

$$F_a^{\mathrm{III},3}(\tilde{z}) = z^{-1} (b_a^{\mathrm{III}}(4)\,\tilde{z}^{-3} + b_a^{\mathrm{III}}(5)\,\tilde{z}^{-4} + b_a^{\mathrm{III}}(6)\,\tilde{z}^{-5} + b_a^{\mathrm{III}}(7)\,\tilde{z}^{-6}) \qquad (59)$$

$$F_a^{\mathrm{III},4}(\tilde{z}) = z^{-1} (b_a^{\mathrm{III}}(8)\,\tilde{z}^{-7} + b_a^{\mathrm{III}}(9)\,\tilde{z}^{-8} + \cdots + b_a^{\mathrm{III}}(M+1)\,\tilde{z}^{-M}) \quad (M \leqq 20) \qquad (60)$$

The frequency compression parameter α is set between 0 and 0.4, and male speech "Nanbu dewa higashi no kaze (East wind in southern part)" and female speech "Dare ka kite kudasai (Somebody should come)" are analyzed. In the worst condition, where quantization width is 0.5 so that the filter coefficient $b_a^{\mathrm{III}}(m)$ is quantized very roughly, the maximum magnitude $r_{L(n)}$ ($n = 1,2,3,4$) of the term $F_a^{\mathrm{III},n}(m)$ of the transfer function $F_a^{\mathrm{III}}(z)$ is obtained as

$$r_{L(n)} \lessapprox \begin{cases} 3.00 & (n = 1,2) \\ 2.75 & (n = 3,4) \end{cases} \qquad (61)$$

The value of Eq. (61) is in the range of good approximation for the third-order Padé approximation, and the required transfer function $H_a(\tilde{z})$ of the MLSA filter is given, for $F_a^{\mathrm{III},n}(\tilde{z})$ of Eqs. (57) to (60), as

$$H_a(\tilde{z}) = \exp(b_a^{\mathrm{III}}(0)) R_3^{\mathrm{III}}(F_a^{\mathrm{III},1}(\tilde{z})) R_3^{\mathrm{III}}(F_a^{\mathrm{III},2}(\tilde{z})) \times R_3^{\mathrm{III}}(F_a^{\mathrm{III},3}(\tilde{z})) R_3^{\mathrm{III}}(F_a^{\mathrm{III},4}(\tilde{z})) \qquad (62)$$

The MLSA filter with the transfer function $H_a(\tilde{z})$ of Eq. (62) can be constructed as in Fig. 2. The frequency compression parameter α can practically be set as 2 to 3 bits. Setting the word length of the coefficient $p_{L,l}^{\mathrm{III}}$ of Padé approximation $R_L^{\mathrm{III}}(\omega)$ as 7 bits, and the maximum word lengths of the basic filter coefficient $b_a^{\mathrm{III}}(m)$ as 8 bits including the sign bit, the approximation error in the log magnitude of MLSA filter is 0.5 dB or less for all frames in male "Nanbu dewa higashi no kaze," and female "Dare ka kite kudasai." When the basic filter coefficients are not quantized, the approximation error is 0.25 dB or less. The mel spectral envelope of speech is obtained from the improved cepstral method [1]. An example of the approximate mel-log spectral envelope by 8-bit MLSA filter is shown in Fig. 3.

7. Conclusions

This paper presented a construction method for mel-log spectral approximation (MLSA) filter with relatively simple structure and low coefficient sensitivity, together with a design example for MLSA filter for speech synthesis. The transfer function of MLSA filter is given as a rational function with the transfer function of a first-order, all-pass filter as the variable. Consequently, the path without delay must be deleted from the feedback loop in the realization of the filter. Using the construction method for MLSA filter shown in this paper, the path without delay can easily be deleted from the feedback loop. The obtained MLSA filter has a relatively simple structure and low coefficient sensitivity.

Although not stated in this paper, it is verified that the proposed MLSA filter is very useful in synthesis by analysis and synthesis by rule of speech. This will be reported in another apper.

Acknowledgement. A part of this work was supported by the Scientific Research Grant, Min. Education, General B 57460116.

REFERENCES

1. Kitamura and S. Imai. Spectrum distortion and quality of synthesized speech in cepstrum vocoder, Trans. (A) I.E.C.E., Japan, J65-A, 5, pp. 478-484 (May 1982).
2. Imai. S. Logarithmic amplitude approximation (LMA) filter, Trans. (A) I.E.C.E., Japan, J63-A, 12, pp. 886-893 (Dec. 1980).
3. Strube, H. W. Linear prediction on a warped frequency scale, J. Acoust. Soc. Am., 68, 4, pp. 1071-1076 (Oct. 1980).