Acoustic Model-Based Filter Structure for Synthesizing Speech Signals

Il-Taek Lim and Byeong Gi Lee
Department of Electronics Engineering
Seoul National University, Seoul, 151-742, Korea
E-mail: blee@alliant.snu.ac.kr FAX: +82 (2) 885-4459

ABSTRACT This paper proposes a filter structure suitable for speech synthesis applications. We first derive the lossy pole-zero model by employing the wave digital filter(WDF) adaptor formula, and by converting the fixed termination value -1 into a loss factor $\mu_0^c \in (-1, 1)$. Then we discuss how to determine the reflection We employ the Durbin's method in estimating the numerator polynomial of the lossy pole-zero transfer function from the given speech sound, and then apply the step-down algorithm on the numerator to extract the reflection coefficients of the closed-termination tract. For determining the reflection coefficients of the other parts we employ a pre-calculated pole-estimator polynomial.

1. INTRODUCTION

The linear predictive coding(LPC) technique widely used in the area of speech analysis and synthesis is connected with the all-pole type digital filters. It is derived from the acoustic tube modeling of human vocal tract, and assumes that during the pronunciation the velum is closed and the sound wave proceeds only through the oral tract. Since the existence of the nasal tract is ignored in its assumption, the resulting all-pole type transfer functions can not properly handle some speech signals such as the nasal sounds. Therefore a considerable amount of efforts has followed for the pole-zero type modeling of speech signals [1], [2]. All these methods, however, are common in that they are *signal-based,* not *acoustic model-based.*

Recently, a new pole-zero modeling was reported acoustic model-based, which brought about a pole-zero type generalization of the all-pole type LPC modeling [3], [4]. In this approach, the nasal tract as well as the oral tract was taken into consideration, and an acoustic model-based pole-zero type transfer function was derived, which includes the LPC all-pole filter as a special case. The derived pole-zero type transfer function, however, has a limitation caused by the losslessness assumption. It assumed a lossless model, thus neglecting acoustic phenomena such as friction, viscosity, and heat conduction within the vocal tract. As a consequence, the numerator of the transfer function is forced to be symmetric, which limits the degree of freedom in practical modeling. Therefore a lossy pole-zero modeling is called for that can remove such limitation, yet preserving the characteristics of an acoustic model-based generalization. In this paper we will work on this lossy pole-zero modeling, deriving a pole-zero filter which is most suitable for speech synthesis applications.

2. REVIEW OF LOSSLESS POLE-ZERO MODELING

We first briefly review the lossless pole-zero modeling for reference in discussing the lossy modeling. We denote by $u_m(x,t)$ and $p_m(x,t)$ respectively the volume velocity and the acoustic pressure of point x within section m at time t. These continuous-time signals are then sampled into discrete-time signals. The configuration of the generalized vocal tract model is as shown in Fig. 1. It consists of three branches $--$ the *pharynx*, the *nasal tract*, and the *oral tract.* In the figure the *oral tract* is drawn with closed termination to indicate that it forms an oral cavity when pronouncing the nasal sounds. The three branches respectively consist of $L+1$, M, and N sections. A_m denotes the cross-sectional area of section m; U_m^+ and U_m^- denote the z-transformed volume velocity of section m respectively in the forward and the backward directions; and the superscript "c" indicates that the corresponding variables and constants belong to the "closed" tract. We introduce the following definitions[1]

$$\begin{bmatrix} C^{+}_{N-1}(z) \\ C^{-}_{N-1}(z) \end{bmatrix} = \begin{bmatrix} 1 & \mu^{c}_{N-1} \\ \mu^{c}_{N-1}z^{-1} & z^{-1} \end{bmatrix} \cdots \begin{bmatrix} 1 & \mu^{c}_{1} \\ \mu^{c}_{1}z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} 1 \\ -z^{-1} \end{bmatrix}, \tag{1}$$

$$\begin{aligned} P(z) &\equiv C^{+}_{N-1}(z) - (1-\sigma)C^{-}_{N-1}(z), & \text{(2a)} \\ Q(z) &\equiv -\sigma C^{+}_{N-1}(z), & \text{(2b)} \\ R(z) &\equiv \sigma C^{-}_{N-1}(z), & \text{(2c)} \\ S(z) &\equiv (1-\sigma)C^{+}_{N-1}(z) - C^{-}_{N-1}(z), & \text{(2d)} \end{aligned}$$

$$\sigma \equiv \frac{A^{c}_{N-1}}{A_{M-1} + A^{c}_{N-1}.} \tag{3}$$

Then, according to [4],

$$H(z) = \frac{X^{+}_{0}(z)}{X^{+}_{M+L}(z)} = (1-\sigma)\frac{B(z)}{A(z)},. \tag{4}$$

where

$$\begin{aligned} A(z) \;=\; & \begin{bmatrix} 1 & \mu_{M+L} \end{bmatrix} \begin{bmatrix} 1 & \mu_{M+L-1} \\ \mu_{M+L-1}z^{-1} & z^{-1} \end{bmatrix} \cdots \begin{bmatrix} 1 & \mu_{M} \\ \mu_{M}z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} P(z) & Q(z) \\ R(z) & S(z) \end{bmatrix} \\ & \cdot \begin{bmatrix} 1 & \mu_{M-1} \\ \mu_{M-1}z^{-1} & z^{-1} \end{bmatrix} \cdots \begin{bmatrix} 1 & \mu_{1} \\ \mu_{1}z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} 1 \\ z^{-1} \end{bmatrix}, \end{aligned} \tag{5}$$

$$B(z) = C^{+}_{N-1}(z) - C^{-}_{N-1}(z). \tag{6}$$

3. DERIVATION OF FILTER STRCTURE FOR SPEECH SYNTHESIS APPLICATION

Two-pair shown in Fig. 2(a) represents the generalized vocal tract system in compliance with (6), where

$$\begin{aligned} A_{*}(z) \;=\; & \begin{bmatrix} \mu_{M+L} & 1 \end{bmatrix} \begin{bmatrix} 1 & \mu_{M+L-1} \\ \mu_{M+L-1}z^{-1} & z^{-1} \end{bmatrix} \cdots \begin{bmatrix} 1 & \mu_{M} \\ \mu_{M}z^{-1} & z^{-1} \end{bmatrix} \cdot \begin{bmatrix} P(z) & Q(z) \\ R(z) & S(z) \end{bmatrix} \\ & \begin{bmatrix} 1 & \mu_{M-1} \\ \mu_{M-1}z^{-1} & z^{-1} \end{bmatrix} \cdots \begin{bmatrix} 1 & \mu_{1} \\ \mu_{1}z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} 1 \\ z^{-1} \end{bmatrix}. \end{aligned} \tag{7}$$

In mathematical expressions,

$$\begin{bmatrix} 1 \\ A_{*}(z)/A(z) \end{bmatrix} = (1-\sigma)\Pi(z) \begin{bmatrix} H(z) \\ 0 \end{bmatrix} \tag{8}$$

for the chain matrix $\Pi(z)$. If we define the chain matrices $\Pi(\mu_i, z)$ and $\Pi_B(z)$ respectively by

$$\Pi(\mu_i, z) \equiv \begin{bmatrix} 1 & \mu_i \\ \mu_i z^{-1} & z^{-1} \end{bmatrix}, \tag{9}$$

$$\Pi_B(z) \equiv \frac{1}{B(z)} \begin{bmatrix} P(z) & Q(z) \\ R(z) & S(z) \end{bmatrix}, \tag{10}$$

we obtain, by (11) and (12), the expression

$$\Pi(z) = \Pi(\mu_{M+L}, 1)\Pi(\mu_{M+L-1}, z) \cdots \Pi(\mu_{M}, z)\Pi_B(z)\Pi(\mu_{M-1}, z) \cdots \Pi(\mu_1, z)\Pi(1, z). \tag{11}$$

[1] In [4], w and $-C^{-}_{N-1}(z)$ are used in place of σ and $C^{-}_{N-1}(z)$, respectively.

A two-pair characterized by the chain matrix $\Pi(\mu_i, z)$ has the signal flowgraph shown in Fig. 3. We combine (4), (8) and (14), to get

$$\Pi_B(z) = \frac{1}{1 - C(z)} \begin{bmatrix} 1 - (1-\sigma)C(z) & -\sigma \\ \sigma C(z) & (1-\sigma) - C(z) \end{bmatrix}, \tag{12}$$

where

$$C(z) \equiv \frac{C^-_{N-1}(z)}{C^+_{N-1}(z)}. \tag{13}$$

Then it is not difficult to show that the chain matrix $\Pi_B(z)$ in (16) can be realized as shown in Fig. 4(a), where the rectangular block corresponds to the *three-port series wave adaptor*[7] symbolized in Fig. 4(b). We also have

$$\sigma = R^c_{N-1}/(R_{M-1} + R^c_{N-1}). \tag{14}$$

Eq. (3) can be arranged into the expression

$$\begin{bmatrix} 1 \\ C(z) \end{bmatrix} = \Pi_C(z) \begin{bmatrix} 1/C^+_{N-1}(z) \\ 0 \end{bmatrix} \tag{15}$$

for the chain matrix $\Pi_C(z)$ defined by

$$\Pi_C(z) = \Pi(\mu^c_{N-1}, z) \cdots \Pi(\mu^c_1, z)\Pi(\mu^c_0, z) \tag{16}$$

with the value $\mu^c_0 = -1$. Hence $C(z)$ is realized as in Fig. 5. The parameter $\mu^c_0 = -1$ in the last two-pair in (20) indicates that the corresponding termination is completely closed. With $\mu^c_0 = -1$ we can easily deduce from (3) the relation $C^-_{N-1}(z) = -z^N C^+_{N-1}(z^{-1})$, and in view of (8) we can confirm the symmetry $B(z^{-1}) = B(z)$. This means that all zeros of $H(z)$ are located either in quad or on the unit circle in the z-plane. In general, $H(z)$ with symmetric $B(z)$ is not well-suited for modeling of speech signals, since the symmetry brings about too strong a constraint in the modeling. In fact, any value of μ^c_0 can serve as the objected parameter as long as it lies within the open interval $(-1,\ 1)$. This new parameter $\mu^c_0 \in (-1,\ 1)$ breaks the symmetry that used to reside in the numerator of the existing lossless model, thus allowing for more freedom in modeling of speech signals. If we combine the three building blocks in Figs. 3 to 5 in accordance with the expression (15), we obtain the signal flowgraph of the finalized lossy pole-zero model shown in Fig. 6.

Note that in case $\sigma = 0$ the series adaptor becomes a direct connection and $\Pi_B(z)$ in (16) becomes an identity matrix, thus making $\Pi(z)$ in (15) a chain matrix for an all-pole two-pair. Therefore the transfer function becomes pole-zero when $0 < \sigma < 1$, and becomes all-pole when $\sigma = 0$. Furthermore, it becomes lossy when $-1 < \mu^c_0 < 1$, and becomes lossless when $\mu^c_0 = -1$. The resulting signal flowgraph is, therefore, of a general form including most possible cases as well as the existing Gray-Markel all-pole filter. Hence proposed the signal flowgraph can best serve as a speech synthesis filter.

4. EVALUATION OF REFLECTION COEFFICIENTS

All equations for the lossless pole-zero model introduced in Section II are also valid for the lossy model except that equation (3) needs to be modified into

$$\begin{bmatrix} C^+_{N-1}(z) \\ C^-_{N-1}(z) \end{bmatrix} = \Pi(\mu^c_{N-1}, z) \cdots \Pi(\mu^c_1, z)\Pi(\mu^c_0, z) \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \tag{17}$$

for the parameter μ^c_0 in (-1, 1). Reflecting this modification to (8), we get

$$B(z) = [\ 1 \quad -1\ ]\, \Pi(\mu^c_{N-1}, z) \cdots \Pi(\mu^c_1, z)\Pi(\mu^c_0, z) \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{18}$$

Now we reverse the flow of the signal flowgraph representing this matrix equation. Then, since the signal flowgraph is for a one-input one-output linear system, the transfer function is preserved over this reversion process. We also spread the effect of the negative sign, finally obtaining the signal flowgraph whose matrix equation is

$$B(z) = [\ 1 \quad -\mu_0^c\ ]\,\Pi(-\mu_1^c, z)\cdots\Pi(-\mu_{N-1}^c, z)\begin{bmatrix} 1 \\ z^{-1} \end{bmatrix}. \qquad (19)$$

This represents the Gray-Markel lattice structure [5]. Therefore it is possible to apply the so-called "stepdown" algorithm [5] to $B(z)$ to determine the lattice coefficients $k_1, k_2, \cdots, k_N$, which correspond to $-\mu_{N-1}^c, -\mu_{N-2}^c, \cdots, -\mu_0^c$, respectively. This solves the problem of determining the parameters μ_i^c's, $i = 0, 1, \cdots, N-1$, for a given numerator function $B(z)$. But it is a difficult task to determine this moving average parameter $B(z)$ itself out of raw speech signals. In order to estimate $B(z)$ in an optimal way (i.e., in terms of maximum likelihood estimation) we have to estimate the minimum point of the nonlinear multivariate likelihood functions, which is very difficult in practice. To get around this difficulty, several suboptimal methods have been proposed, instead, among which the Durbin's inverse linear prediction method is the most popular and practical [6], [2]. If we employ this method, the closed tract reflection coefficients μ_i^c's can be determined from the given speech signal $s(n)$ through the following three steps: First, estimate $B(z)$ from $s(n)$ using Durbin's method; then evaluate k_1 through k_N by applying the step-down routine to $B(z)$; and finally put $\mu_0^c = -k_N$, $\mu_1^c = -k_{N-1}, \ldots, \mu_{N-1}^c = -k_1$.

$A(z)$ is a polynomial of order $L + M + N$ but with the degree of freedom of $L + M + 1$. We first estimate $\hat{A}(z)$ of order $L + M + N$ by applying $(L + M + N)$th order linear prediction (i.e., all-pole modeling) to $\tilde{s}(n)$, and then fit $A(z)$ to $\hat{A}(z)$. Then μ_1 through μ_{L+M} can be evaluated. As the performance measure we employ the error function

$$J = \sum_{n=0}^{L+M+N} (\hat{a}(n) - a(n))^2 \qquad (20)$$

for the coefficients $\hat{a}(n)$'s in $\hat{A}(z) = \sum_{n=0}^{L+M+N} \hat{a}(n)z^{-n}$, and $a(n)$'s in $A(z) = \sum_{n=0}^{L+M+N} a(n)z^{-n}$. We apply the steepest descent method to this error function.

5. EXAMPLE

We take the nasal sound /m/ as the speech signal $s(n)$ to be analyzed. In case sampling frequency f_s is 10kHz, we get the length of one section $l = 1.7cm$ via the relation $f_s = c/2l$ [4], with c roughly 340 m/s. The length of the vocal tract including the pharynx and the oral tract is about 16.5 to 19.5 *cm* for voiced sounds,[2] we assume the lengths of 7 *cm*, 10 *cm*, and 8 *cm* respectively for pharynx, the nasal tract, and the oral tract. Since the length of one section is 1.7 *cm*, we obtain the section numbers $L = 4$, $M = 6$, and $N = 5$. Therefore the orders of the polynomials $A(z)$ and $B(z)$ become 5 and 15, respectively. Table I and II display the evaluated coeeficients. Fig. 7 shows the plots of frequency characteristics of the conventional models, where (a) plots the magnitude of the 512-point speech signals $s(n)$ in overlap with that of the 30th order LPC all-pole model, (b) plots the magnitude of the conventional signal-based model of order 5/15 (i.e., numerator of order 5 and denominator of order 15), and (c) plots the magnitude of the proposed lossy pole-zero model with (L, M, N)=(4, 6, 5). We define spectral distance(SD) by

$$SD \equiv \sqrt{\frac{1}{N}\sum_{k=1}^{N} e^2(k) - \left\{\frac{1}{N}\sum_{k=1}^{N} e(k)\right\}^2}, \qquad (21)$$

where $e(k) = F_r(k) - F_t(k)$ for the reference and test frequency characteristics $F_r(k)$ and $F_t(k)$. Taking the 30th order all-pole characteristic in Fig. 7(a) as the reference $F_r(k)$, we obtain the

[2] According to [5], the typical length of vocal tract for the voiced sounds /a/, /e/, /i/, /o/, /u/ are respectively, 17.0, 16,5, 16.5,18.5, and 19.5 *cm*.

SD's for the 5/15th order existing pole-zero model and the proposed lossy pole-zero models for the choice of $(L, M, N) = (4, 6, 5)$ in Table III (a) and (b). The table shows the standard deviations are all comparable.

5. CONCLUDING REMARKS

In this paper we presented a lossy pole-zero modeling of speech signals in the form of a signal flowgraph as well as in terms of a pole-zero type transfer function. It becomes pole-zero or all-pole type and lossy or lossless, by choosing σ and μ_0^c properly. Hence the proposed signal flowgraph can best serve as a speech synthesis filter.

REFERENCES

[1] J. Makhoul, "Linear prediction : A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561-580, Apr. 1975.

[2] S. M. Kay, *Modern Spectral Estimation*, Prentice Hall, 1988.

[3] M. G. Kang and B. G. Lee, "A generalized vocal tract model for pole-zero type linear prediction," *Proceeding of International Conference on ASSP*, S14.10, 1988.

[4] I. -T. Lim and B. G. Lee, "Lossless pole-zero modeling of speech signals," *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 3, pp. 269-276 July 1993.

[5] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*, Spring-Verlag, New York, 1976.

[6] J. Durbin, "Efficient estimation of parameter in moving average models," *Biometrika*, vol. 46, pp. 306-316, 1959.

[7] A. Fettweis, "Wave digital filters: theory and practice," *Proc. IEEE*, pp. 270-327, February 1986.

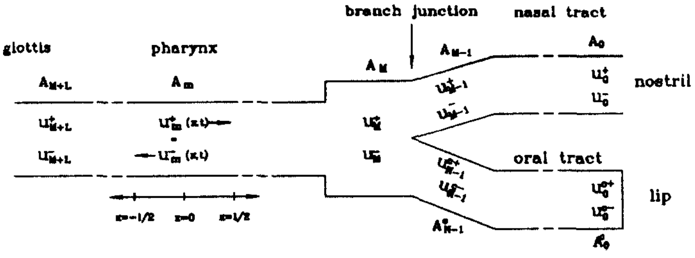

Fig. 1. Generalized vocal tract model.

TABLE I

REFLECTION COEFFICIENTS OF ORAL CAVITY

μ_0^c	-0.28604275
μ_1^c	-0.09158915
μ_2^c	-0.06728477
μ_3^c	-0.32191045
μ_4^c	-0.69770125

TABLE II

REFLECTION COEFFICIENTS OF PHARYNX AND NASAL CAVITY, AND AREA RATIO σ.

μ_1	-0.95516754
μ_2	0.11108155
μ_3	0.18113914
μ_4	0.01216234
μ_5	0.26472138
μ_6	0.54511112
μ_7	-0.19794235
μ_8	0.30157218
μ_9	-0.40789313
μ_{10}	-0.03154621
σ	0.13703220

TABLE III

VALUES OF SPECTRAL DISTANCE(SD).

	Model	SD
(a)	5/15 existing pole-zero model (degree of freedom: 20)	3.5
(b)	(4, 6, 5) lossy pole-zero model (degree of freedom: 16)	3.3

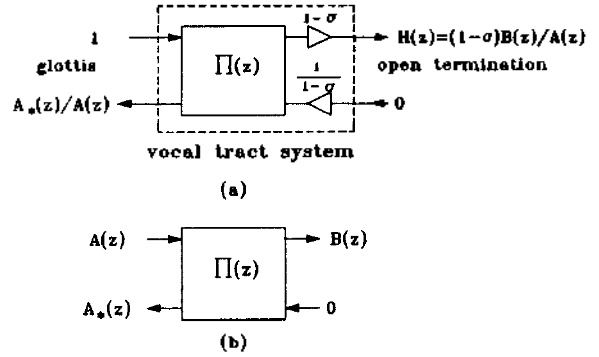

Fig. 2. A vocal tract system two-pair that generates $H(z)$ and $A_*(z)/A(z)$ (a), and the two-pair characterizing $\Pi(z)$ (b).

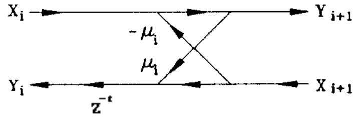

Fig. 3. Signal flowgraph of the two-pair characterized by the chain matrix $\Pi(\mu_i, z)$.

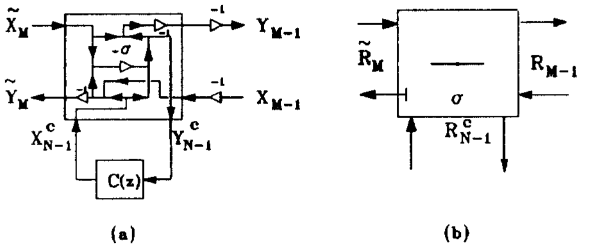

Fig. 4. (a) Signal flowgraph of the two-pair characterized by the chain matrix $\Pi_B(z)$; (b) series wave adaptor symbol representing the box in (a).

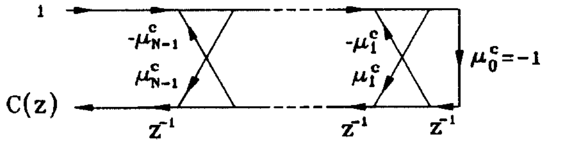

Fig. 5. Signal flowgraph for $C(z)$.

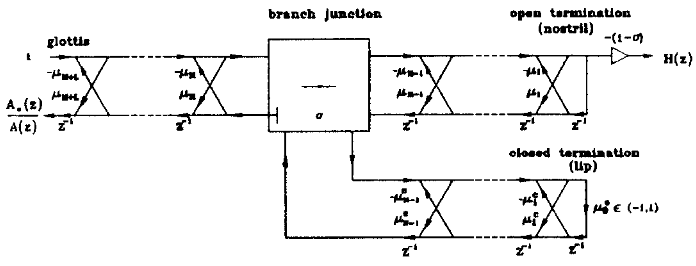

Fig. 6. Overall signal flowgraph for the lossy pole-zero model.

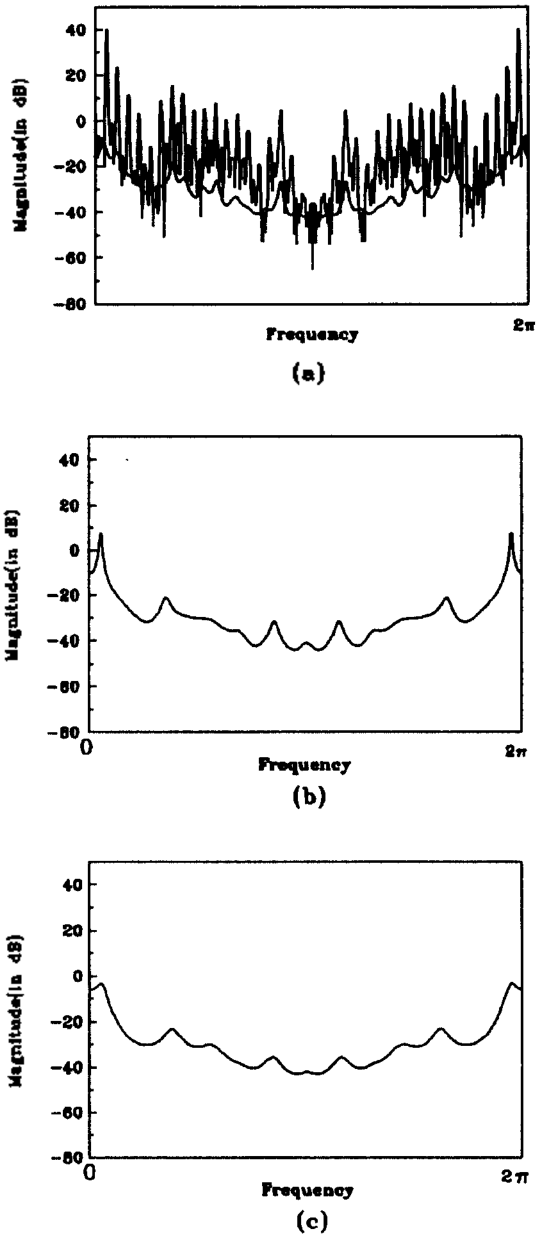

Fig. 7. Comparison of frequency characteristics of conventional models. (a) Magnitude characteristic of 30th-order LPC all-pole model in overlap with the DFT of the 512-point speech signal /m/; (b) magnitude characteristic of the conventional signal-based pole-zero model of order 5/15; (c) magnitude characteristic of the proposed lossy pole-zero model for the choice of (L, M, N) = (4, 5, 6).