일반화된 음원 모델로서의 기저함수합계 모델

홍 준 모°*, 홍 성 훈*, 안 수 길*
서울대학교 전자공학과*

Sum-of-Basis-Functions Model As a Generalized Voice Source Model

JoonMo HONG°*, SungHoon HONG*, SouGuil ANN*
Dept. of Electronics Eng., Seoul Nat'l Univ.*

요 약

본 논문에서는 음원을 모델링하기 위한 새로운 음원 모델로서 기저함수합계 모델을 제안하고 그 모델의 변수를 추정하는 방법에 관하여 설명한다. 기존 모델들이 다양한 음원신호를 표현하는데 부족함이 많았던데 비해 기저함수합계 모델은 다양한 음원신호를 표현하기에 적합하며 ML (Maximum Likelihood)이라는 통일된 추정 방법을 통해 모델의 변수들을 구할 수 있다. 또한 기저함수합계 모델은 기존의 모델들을 포함하는 일반화된 음원 모델이 됨을 보인다.

I. 서 론

음성분석, 음성합성 및 음성코딩 등에 사용되는 음성발생 모델에서 음성은 음원(voice source)이 성도(vocal tract)라는 필터를 통과해 생성되는 출력으로 볼 수 있다 [1]. 따라서 입력으로 사용되는 음원과 필터로 표현되는 성도를 정확하게 모델링하고 그 모델의 변수들을 추정하는 것이 중요하다고 할 수 있다. 지금까지의 연구 결과 성도는 전극점(all pole) 모델 또는 영점-극점(pole-zero) 모델로 상당히 정확하게 표현될 수 있다는 것이 알려졌다. 하지만 음원은 소리의 본질적인 특징을 결정하는 중요한 요소임에도 불구하고 매우 복잡하고 다양한 모양을 띠고 있기 때문에 모델링하기가 수월하지 않았다. 최근에 보다 정확한 음원의 모델링이 합성음의 자연성을 증가시키는 데 중요한 역할을 하고 다양한 음성특징을 구별하는 데 도움이 된다는 사실이 알려졌다.

추정된 음원신호를 바탕으로 음원을 모델링하려는 시도는 오래전부터 행해져 왔다. Rosenberg는 다항식으로 음원신호를 모델링하였고 Liljencrants, Fant 등은 지수함수와 사인함수의 조합을 통해 음원신호를 표현하였다 [2][3]. 이외에도 Fujisaki, Thomson 등에 의해 음원 모델이 제안되었다 [4][5]. 이같이 많은 모델들이 제안된 것은 음원신호를 모델링하기가 그만큼 까다롭다는 것을 의미한다.

그러나 많은 모델들이 제안되었음에도 불구하고 기존 모델들은 여전히 여러 가지 문제점들을 안고 있다. 우선 단순한 함수로 음원신호를 표현하려고 하여 음원신호의 다양하고 복잡한 특성을 충분히 표현해 내지 못했다. 그리고 변수의 수를 증가시켜 좀더 다양하게 음원을 표현하려한 경우에는 추정방법상 오류가 생길 여지가 많이 있었다. 또한 추정기법들이 모델마다 거의 달랐다.

따라서 본 논문에서는 위의 문제점들을 해결하기 위해 다음과 같은 방법을 제시한다. 우선 새로운 음원 모델로 기저함수합계 모델을 제안하고, 이 모델이 다양한 신호를 생성하는데 우수함을 보인다. 다음으로 ML 추정방법을 통해 이 모델의 변수가 추정될 수 있음을 보인다. 그리고 기존의 모델들이 대부분 기저함수합계 모델로 표현됨을 보임으로써 기저함수합계 모델이 일반화된 음원 모델임을 보인다.

논문의 구성은 다음과 같다. 제 II 장에서는 음원과 음원 모델링의 정의, 기존 모델의 단점에 대해 설명한다. 제 III 장에서는 기저함수합계 모델을 새로운 음원 모델로서 제시한 후 이 모델의 변수가 ML 추정문제를 해결하는 일반적인 방법인 EM (Estimate Maximize) 알고리듬에 의해 추정됨을 보인다. 제 IV장에서는 기저함수합계 모델이 일반화된 음원 모델로서 기존의 모델들을 포함하며 다양한 음원을 구현하는데 우수함을 보인다.

II. 음원 모델링

음성은 폐에 의해 여기된 공기의 흐름이 기도, 성대, 구강, 입술 등을 통과해 밖으로 분출된 음향 신호이다. 이것을 그림으로 나타내면 그림 1.(a)와 같은 모델이 된다. 즉 음원이 성도를 통과하고 입술을 지나 음성이 생성되는 선형생성 모델로 표현된다. 이때 음원은 폐에서 분출된 공기가 성대를 거쳐 나온 체적속도(volume velocity)인 성문파(glottal flow)를 지칭한다. 만일 입술이 가지는 미분 특성을 음원에 미리 고려해 준다면 그림 1.(b)와 같이 효과음원(effective voice source)과 음성이 성도에 의해 연결된 것으로 볼 수 있고 효과음원은 성문파의 미분파형(glottal flow derivative)을 뜻하게 된다.

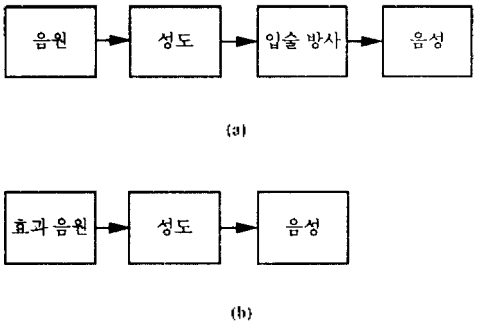

그림 1 음성생성 모델
Fig. 1. Speech production model.

음원을 모델링한다는 것은 성문파나 성문파의 미분파형을 일정한 함수 형태로 표시하고 그 변수들을 정의, 추정하는 것을 말한다. 흔히 마찰음이나 파찰음의 경우는 가산잡음(additive noise) 형태의 음원을 사용하므로 보통 음원 모델링이라고 하면 성대의 떨림이 생기는 유성음의 경우에만 적용되는 개념으로 생각한다. 그리고 분석 구간은 성문파의 한 주기 동안에만 국한시키는 것이 일반적이다.

음원을 표현하기 위해 음원신호를 수학적 함수로 표시하려는 노력들이 많이 이루어져 왔다. 간단한 다항식에서부터 지수함수, 삼각함수의 조합들을 이용한 모델들이 형태를 조금씩 달리해 가면서 제안되어 왔다. 그런데 음원 신호는 매우 복잡하고 다양한 특성들을 포함하고 있어서 개인마다 많은 변화가 있고 또 동일 화자의 경우라도 상황에 따른 차이가 있게 마련이다. 기존의 음원 모델은 단순화된 함수로 음원을 표시한 까닭에 이러한 음원의 다양한 성질들을 표현하기에는 부족한 점이 많았다. 대개의 경우 변수들을 추정하기가 쉽지 않고 추정방법상 오류가 발생할 여지가 있었다. 예를 들면 Thomson의 경우 높은 차수에서 문제가 발생할 가능성이 있어서 차수를 낮추면 너무 단조로운 다항식으로 환원되는 경향이 있었다. 그리고 모델마다 일관된 추정방법이 없었다. 본 논문에서는 이러한 문제점들을 해결하기 위해 새로운 음원 모델로서 기저함수합계 모델을 제안한다.

III. 기저함수합계 음원 모델

식 (3.1)과 같은 AR 모델을 이용하여 음성신호 $s(n)$의 표현이 가능하다.

$$s(n) = \sum_{i=1}^{M} a_i s(n-i) + r(n) \tag{3.1}$$

여기서 a_i는 필터 계수를 나타내고 $r(n)$은 오차신호를 나타낸다. 즉, 성도전달함수를 나타내는 필터계수들을 정확히 구하면 음성신호로부터 음원 부분이 추출될 수 있다는 것이다. 성도전달함수를 구현하는 필터 계수들의 정확한 추정은 잡음과 음원의 영향으로 인해 어렵다고 알려져 있다. 하지만 성도전달함수를 표현하는 비교적 정확한 방법들이 여러가지 소개되어 있으므로 성도의 특성이 제거된 음원이 $r(n)$의 형태로 추출된다. 이때 $r(n)$은 성대와 성도와의 상호 작용, 모델의 오차등이 포함된 음원신호이다. 많은 특성들이 포함된 만큼 복잡한 형태를 띠게 되므로 적당한 함수로 간략화시켜 음원을 표현하고 표현된 함수에서 변수를 추정하고자 하는 노력들이 많이 진행되어 왔다. 그러나 기존의 음원모델들은 II 장에서 언급한 바와 같이 음원과형의 다양한 특성들을 표현하는 데 많은 한계를 지니고 있었다. 따라서 이 장에서는 다양한 음원신호를 표현할 수 있는 새로운 음원 모델로 기저함수합계 형태의 모델을 제안하고 그 특성들을 살펴보도록 한다.

3.1 기저함수합계 모델의 정의

기저함수합계(sum of basis functions) 모델이란 식 (3.2)와 같이 기저함수들의 가중합계(weighted sum)로 신호를 표현한 모델을 뜻한다.

$$r(n) = \sum_{i=1}^{M} \alpha_i b_i(n-\tau_i) \qquad (3.2)$$

- $b_i(n)$: i 번째 기저함수
- τ_i : i 번째 기저함수의 시간지연
- α_i : i 번째 기저함수의 가중치
- M : 기저함수의 개수

식 (3.2)에 나타나 있는 것처럼 기저함수합계 모델은 여러 개의 기저함수 $b_i(n)$들과 그것들의 가중치 변수, 시간지연 변수로 구성되어 있다. 각 기저함수들을 적당한 시간만큼 옮기고 여기에 적절한 가중치를 두면 이것들의 합으로 음원신호 $r(n)$을 생성할 수 있다는 것이다. 기저함수의 형태와 개수에 따라 여러 가지 복잡한 모양을 만들 수 있음을 알 수 있다. 일단 기저함수 $b_i(n)$들은 어떤 형태로 명확히 정의되어야 한다. 만일 정의된 기저함수가 여러 변수들을 포함하고 있는 경우라면 적절한 방법으로 그 함수의 변수들을 추출한 것으로 가정한다. 그러면 기저함수합계 모델에서는 i 번째 기저함수의 가중치와 시간지연이 구해야 할 변수가 된다. 그런데 기저함수합계 모델과 같은 수식의 해는 ML을 해결하는 일반적인 방법인 EM 알고리듬으로 쉽게 구할 수 있다.

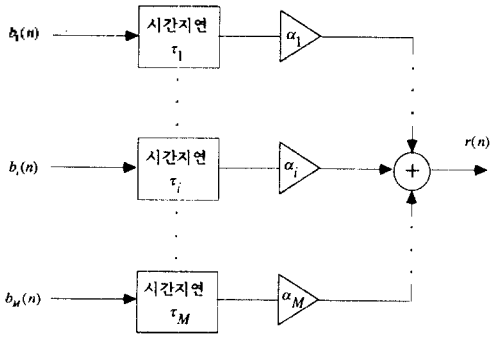

그림 2 기저함수합계 모델
Fig. 2. Sum of basis functions model.

3.2 기저함수합계 모델 변수의 추정

식 (3.2)에 나타난 기저함수합계 모델의 변수를 구하기 위해서 다음과 같은 가정을 하였다.

1) 관측된 음원신호 $y(n)$은 식 (3.3)과 같이 표현된다. 실제 음원은 기저함수합계로 표현되는데 여기에 잡음이 섞인 것으로 가정한다.

$$y(n) = \sum_{i=1}^{M} \alpha_i b_i(n-\tau_i) + e(n) \qquad (3.3)$$

2) 오차 성분 $e(n)$는 백색 가우시안 잡음(white Gaussian noise)으로 이의 공분산행렬(covariance matrix)은 식 (3.4)과 같이 표현된다.

$$E[\ e(n)\, e^*(k)\ ] = Q\delta(n-k) \qquad (3.4)$$

실제의 완전한 데이타(complete data)는 잡음에 의해 교란되어 확실히 알 수 없고 관측된 데이타(observed data)만 존재하는 경우, 일정한 모델을 가정하여 그 모델로 관측된 데이타를 표현하려고 한다. 즉 모델의 변수가 어떤 값을 가질 때 관측된 신호를 가장 잘 표현하게 되는가를 추정하는 문제이다. 이것은 ML 추정문제이다.

식 (3.3)과 같은 문제의 log-likelihood 함수는 식 (3.5)와 같이 나타나고 잡음으로 인한 오차를 최소화시키기 위한 사용된 ML 추정식의 표현은 식 (3.6)과 같다 [6]. 식 (3.5)에서 θ는 모델 변수를 나타낸다.

$$L(\theta) = c - \frac{\lambda}{2} \int_T [\ y(n) - \sum_{i=1}^{M} \alpha_i b_i(n-\tau_i)\ ]^* \cdot Q^{-1} [\ y(n) - \sum_{i=1}^{M} \alpha_i b_i(n-\tau_i)\ ]\, dn \qquad (3.5)$$

$$\min_{\substack{\tau_1, \tau_2, \cdots, \tau_M \\ \alpha_1, \alpha_2, \cdots, \alpha_M}} \left[\int_T |y(n) - \sum_{i=1}^{M} \alpha_i b_i(n-\tau_i)|^2 dn \right] \qquad (3.6)$$

그런데 식 (3.6)과 같은 문제를 직접 풀기는 매우 어려우므로 ML 문제를 푸는 일반적인 방법인 EM 알고리듬을 사용하도록 한다. EM 알고리듬은 식 (3.6)과 같은 다중변수최적화(multi-parameter optimization) 과정을 독립된 M 개의 최적화 과정으로 변환하여 변수를 추정할 수 있도록 해주는 방법이다 [7].

식 (3.7)과 (3.8)에서 보는 바와 같이 EM 알고리듬은 추정치를 구하고, 그 추정치를 바탕으로 오차를 최소화하는 (즉 유사도를 최대로 하는) 2 단계의 알고리듬이다. 일단 모델 변수들의 초기값들을 임의로 잡아 미리 규정된 완전한 데이타의 추정치를 구한 후 이 추정치를 바탕으로 오차가 최소가 되는 모델 변수들을 구한다. 이 모델 변수들

을 바탕으로 새로운 추정치를 구하고 다시 같은 과정을 반복한다.

$i = 1, 2, \ldots, M$에 대해

$$\hat{x}_i^{(l)}(n) = \hat{a}_i^{(l)} b_i(n - \hat{\tau}_i^{(l)}) + \beta[\, y(n) - \sum_{i=1}^{M} \hat{a}_i^{(l)} b_i(n - \hat{\tau}_i^{(l)})] \tag{3.7}$$

$i = 1, 2, \ldots, M$에 대해

$$\min_{a,\tau} \int_T | \hat{x}_i^{(l)} - a b_i(n-\tau)|^2 dn \quad \rightarrow \quad \hat{a}_i^{(l+1)}, \ \hat{\tau}_i^{(l+1)} \tag{3.8}$$

윗 식에서 l은 반복 횟수이고 β는 상수이다.

식 (3.7)에서 a와 τ는 다음 식 (3.9)로부터 구할 수 있다. g^*는 g의 정합필터(matched filter)를 나타낸다.

$$\max_{\tau} |f_k(\tau)| \rightarrow \tau_k, \qquad a_k = \frac{f_k(\tau_k)}{E} \tag{3.9}$$

여기서 $E = \int_T |b_k(t)|^2 dt$이고 $f_k(\tau) = \int_T \hat{x}_k(t) b_k^*(t-\tau) dt$

이다.

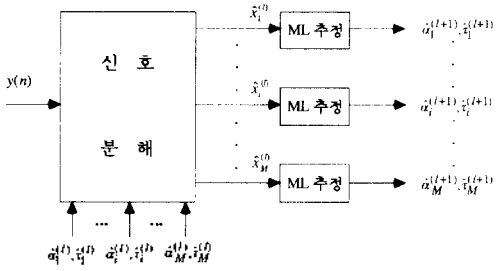

그림 3 EM 알고리듬을 이용한 음원 모델 변수의 추정
Fig. 3. Estimation of voice souce model parameters using EM algorithm.

IV. 기존 모델의 기저함수합계 표현

기존 모델들 대부분이 기저함수합계 모델로 표현될 수 있다. 다음에 기존의 모델들을 이산신호 형태로 정리하여 기저함수합계 모델로 표현해 보자.

4.1 Rosenberg 모델

Rosenberg는 성문파를 식 (4.1)과 같이 다항식의 형태로 모델링하였다.

$$u(n) = a_1 n + a_2 n^2 + a_3 n^3 \tag{4.1}$$

기저함수합계 표현은 $u(n) = \sum_k a_k b_k(n - \tau_k)$으로 기저함수와 가중치는 다음과 같다.

b_k : $b_1(n) = n$, $b_2(n) = n^2$, $b_3(n) = n^3$

a_k : a_1, a_2, a_3는 원래 모델에서의 a 값들이다.

τ_k : 모두 0

기저함수가 시간 n의 다항식으로 구성된 함수이고 a_i, $i = 1, 2, 3$ 가 가중치인 모델로 표현할 수 있다. 식 (3.7)과 (3.8)을 이용하여 모델 변수들을 구할 수 있다.

4.2 LF 모델

Krishnamurthy에 의해 정리된 것처럼 LF 모델은 지수함수합계(sum of exponential functions) 모델로 표시된다. 먼저 성문이 열렸을 때의 모델은 식 (4.2)과 같다.

$$\begin{aligned} g_1(n) &= E_0 e^{\alpha n} \sin \omega_g n \\ &= \frac{E_0}{2} e^{-j\frac{\pi}{2}} e^{(\alpha + j\omega_g)n} - \frac{E_0}{2} e^{j\frac{\pi}{2}} e^{(\alpha - j\omega_g)n} \\ &= c_0 z_g^n + c_0^* z_g^*, \quad T_0 < n < T_e \end{aligned} \tag{4.2}$$

여기서 $c_0 = \frac{E_0}{2} e^{-j\frac{\pi}{2}}$이고 $z_g = e^{(\alpha + j\omega_g)n}$이다.

기저함수합계 표현은 $g_1(n) = \sum_k a_k b_k(n - \tau_k)$으로 이때 기저함수와 가중치는 다음과 같이 표현된다.

b_k : $b_1(n) = z_g^n$, $b_2(n) = (z_g^*)^n$

a_k : $a_1 = c_0$, $a_2 = c_0^*$

모델 변수를 구하기 위해 우선 음원신호 $y(n)$이 주어지면 선형예측계수들을 구하여 다음과 같은 식을 얻는다.

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_p z^{-p}$$

$A(z) = 0$ 의 해를 구하여 z_i, $i = 1, 2, \ldots, p$ 들과 F_i들을 구한다.

$$F_i = \frac{F_s}{2\pi} \arg[\frac{Im(z_i)}{Re(z_i)}]$$

이 중에서 가장 적합한 F_i와 그에 해당하는 z_i를 구하여 그때의 z_i를 z_g로 놓는다. 식 (3.7)과 (3.8)을 이용하여 가중치들을 구한다.

한편 성문이 닫혔을 때의 모델은 식 (4.3)와 같다.

$$g_2(n) = \frac{-E_0}{\varepsilon T_a}\left[e^{-\varepsilon(n-T_a)} - e^{-\varepsilon(T_c-T_a)} \right]$$

$$= \frac{-E_0}{1-e^{-\varepsilon(T_c-T_a)}}\left[e^{-\varepsilon(n-T_a)} - e^{-\varepsilon(T_c-T_a)} \right]$$

$$, \quad T_e < n < T_c \tag{4.3}$$

$g_2(T_a) = -E_e$, $g_2(T_c) = 0$임을 이용하여 위의 식으로부터 ε을 구한다. 그림 4에 LF 모델로 표현된 성문파의 미분파형이 표현되어 있다.

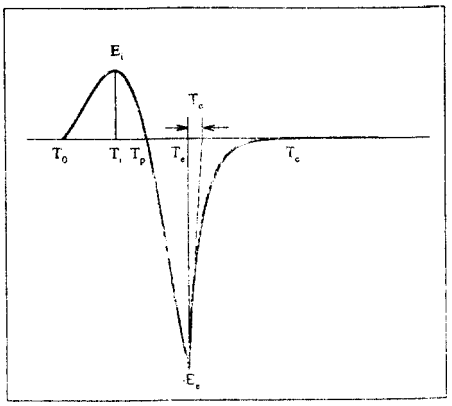

그림 4 LF 모델로 표현된 성문파의 미분파형
Fig. 4. Glottal flow derivative represented by LF model.

4.3 Fujisaki 모델

성문파의 미분파형을 4 구간으로 나누어 각 구간을 다항식으로 표시하였다. 즉, 개구간의 길이(W), 성문이 열리는 구간과 닫히는 구간의 상대적 비(R,F), 성문이 닫히는데 걸리는 시간(D) 등의 시간 변수들과 성문이 열릴 때 성문파의 기울기(A), 성문이 닫히기 전후의 성문파의 기울기(B,C) 등의 크기 변수들을 모두 포함하고 있다.

$$g(n) = A - \frac{2A+R\alpha}{R} n + \frac{A+R\alpha}{R} n^2, \quad 0 < n < R$$

$$= \alpha(n-R) + \frac{3B-2F\alpha}{F^2}(n-R)^2 - \frac{2B-F\alpha}{F^3}(n-R)^3$$

$$, \quad R < n < W$$

$$= C + \frac{2(C-\beta)}{D}(n-W) - \frac{C-\beta}{D^2}(n-W)^2, \quad W < n < W+D$$

$$= \beta, \quad W+D < n < T \tag{4.4}$$

여기서 $\alpha = \frac{4AR-6FB}{F^2-2R^2}$ 이고 $\beta = -\frac{CD}{D-(T-W)}$ 이다.

기저함수합계로 표현하면 다음과 같이 된다.

$$g(n) = \sum_k a_{1k} b_{1k}(n-\tau_{1k}), \quad 0 < n < R$$

$$= \sum_k a_{2k} b_{2k}(n-\tau_{2k}), \quad R < n < W$$

$$= \sum_k a_{3k} b_{3k}(n-\tau_{3k}), \quad W < n < W+D$$

$$= c, \quad W+D < n < T$$

기저함수들은 다음과 같이 표시되고 가중치와 시간지연은 각 구간별로 식 (3.7)과 (3.8)을 통해 얻어진다. 그림 5에 Fujisaki 모델로 표현된 성문파의 미분파형이 표현되어 있다.

$$b_{11}(n) = 1, \quad b_{12}(n) = n, \quad b_{13}(n) = n^2$$

$$b_{21}(n) = n, \quad b_{22}(n) = n^2, \quad b_{23}(n) = n^3$$

$$b_{31}(n) = 1, \quad b_{32}(n) = n, \quad b_{33}(n) = n^2$$

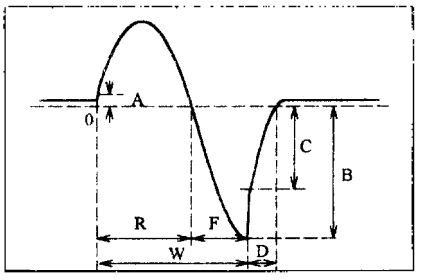

그림 5 Fujisaki 모델로 표현된 성문파의 미분파형
Fig. 5. Glottal flow derivative represented by Fujisaki model.

V. 결 론

본 논문에서는 음원파형을 표현하기 위한 새로운 음원 모델로서 기저함수합계 모델을 제안하였다. 기저함수합계 모델은 주어진 기저함수들의 가중합으로 음원을 표현한 모델로서 가중치와 시간지연들을 모델 변수로 갖는다. 그러한 모델 변수들은 ML 추정방법을 해결하는 일반적인 방법인 EM 알고리듬에 의해 구해진다.

제안한 모델은 다양한 음원을 표현할 수 없었던 기존 모델들과는 달리 기저함수를 적절하게 선정하면 여러 가지 모양들을 생성해 낼 수 있으므로 다양한 음원을 표현하기에 적합한 모델이다. 또한 모델마다 달랐던 기존의 추정방법 대신 ML 이라는 통일된 추정방법에 의해 모델 변수들을 구할 수 있는 장점이 있다. 다만 기저함수 자체의 변수들은 미리 구해야 한다. Rosenberg, Liljencrants, Fant, Fujisaki 등에 의해 제안된 기존 모델들은 대부분 기저함수합계 모델로 표현되어 기저함수합계 모델의 통일된 접근방법을 통해 변수추정이 가능하다. 따라서 기저함수합계 모델은 기존 모델들을 포함하는 일반화된 음원 모델이다.

이러한 기저함수합계 모델에서 가장 중요한 것은 음원파형을 가장 잘 표현할 수 있는 기저함수의 선정이다. 이 기저함수의 적절한 선택은 얼마나 정확하게 음원을 표현할 수 있을 것인가를 결정하는 중요한 문제이나 차후의 과제로 남겨 둔다.

참고 문헌

[1] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*. New York: Springer-Verlag, 1976.

[2] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Am.*, vol. 49, no. 2, pp. 583-590, Feb., 1971.

[3] G. Fant, J. Liljencrants, and Q. Lin, "A four parameter model of glottal flow," STL-QPSR 4/1985, 1985, pp. 1-13; also presented at the French-Swedish Symp., Apr. 22-24, 1985.

[4] A. K. Krishnamurthy, "Glottal source estimation using a sum of exponentials model," *IEEE Trans. SP*, vol. 40, no. 3, pp. 682-686, Mar., 1988.

[5] H. Fujisaki and M. Ljungqvist, "Proposal and evaluation of models for the glottal source waveform," *Proc. ICASSP-86*, pp. 1605-1608, 1986.

[6] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*. New York: Willey, 1968.

[7] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the EM algorithm," *IEEE Trans. ASSP*, vol. 36, no. 4, pp. 477-487, Apr., 1988.