1998년도 한국음향학회 학술발표대회 논문집 제 17 권 제 1(s) 호

기저함수의 가중합을 이용한 음원의 모델링

강상기, 양기혁, 성굉모
서울대학교 전기공학부

Voice Source Modeling Using Weighted Sum-of-Basis-Functions Model

Sang-Ki Kang, Ki-Hyuk Yang, and Koeng-Mo Sung
School of Electrical Engineering, Seoul National University, Seoul 151-742, KOREA
TEL : 02-880-7263, FAX : 02-886-0791

요약

본 논문에서는 음성합성(speech synthesis) 및 부호화(coding) 시스템에 있어서 음원(voice source) 모델링에 관한 문제를 살펴보고자 한다. 기존의 음원 모델링 시스템이 가지고 있는 여러 문제들을 극복하고자 기저함수(basis function)의 가중 합(weighted -sum)으로 음원을 모델링 하는 새로운 기법을 제안하고자 한다. 제안한 방법에서는 음원 파형(voice source waveform)을 적절히 표현하기 위해서 필터 뱅크(filter bank)에 기초한 기저함수의 가중 합으로 나타낸다. 다양한 음원 특성을 효과적으로 나타내는 음원 파라미터를 구하기 위하여 EM(estimate maximize)에 기초한 구조에 관해 조사한다. 제안한 방법을 이용하여 다양한 유성음에 대해 실험을 수행하였다. 실험결과 제안한 추정(estimation) 방법 및 모델링 방법을 이용하면 기존의 방법에 비해 더 정확한 음원 파형을 추정할 수 있고, 다양한 음원 특성을 나타낼 수 있다. 또한 음성합성 및 부호화에서도 음성 품질(voice quality)를 개선시킬 수 있으리라 기대된다.

1. 서론

현재 대부분의 음성분석 및 합성은 그림 1과 같이 음성이라는 입력이 성도(vocal tract)라는 필터를 통과한 출력으로 음성 파형을 간주하는 선형생성이론

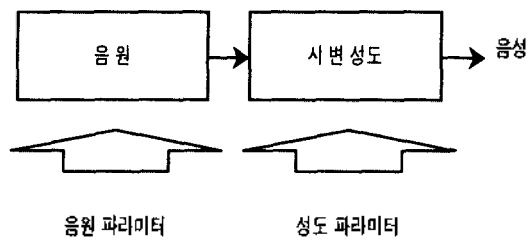


그림1. 음성 생성 모델

에 기초를 두고 있다. 따라서 음원 과 성도를 모델링하고 정확한 모델 파라미터를 추정하는 것이 중요하다. 선형 예측 부호화 방법(linear predictive coding)이 음성 분석이나 합성에 주로 사용된다[1]. 전극 필터(all-pole filter)와 관계된 선형 예측 이론이 우수함에도 불구하고, 고려되어야 할 몇몇 여지가 있다. 1)시변음원(time-varying voice)의 특성을 적절히 모델링 할 수 없다. 2) 음원의 특성을 효과적으로 모델링 할 수 없다. 3) 음원 파라미터를 추정하기 어렵다. 최근에 음원 모델링에 관한 많은 방법이 연구되었다[2]. 그러나 기존의 방법으로 음원을 모델링 할 경우 부정확한 또는 의미 없는 결과를 얻기 쉽고, 잡음에 관한 영향도 제거할 수 없다[3].

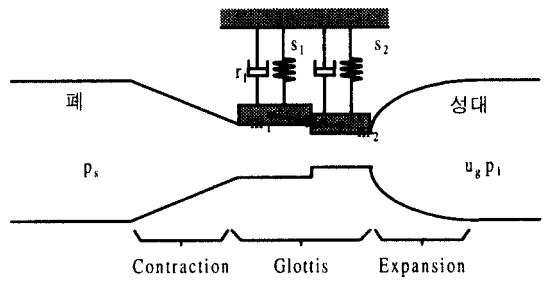
비교적 적은 파라미터로 다양한 시간 또는 주파수 특성을 나타내는 것이 바람직한 음성 분석 및 합성에 있어서 음원 파형의 계수화는 중요한 역할을 한다. 음성신호를 계수화 하는 많은 일반적인 방법은

신호처리의 한 부분이다. 음원 파형을 정적(stationary) 이거나 시변(time-varying)인 파라미터로 나타낼 수 있는 지수 합과 다항식을 이용한 몇몇 모델들이 있다. 이 파라미터는 수학적으로 표기한 함수의 계수들이며, 이들 함수들의 선형결합(linear combination)을 통해 음원 파형을 근사 화 할 수 있다. 기존의 이런 음원 모델은 음원 파형을 시간 영역에서 지수 합이나 다항식으로 나타낼 수 있다. 이와 같은 간략화 된 모델로는 주파수 영역에서 다양한 음원 파형과 복잡한 특성을 나타낼 수 없다. 이러한 문제들을 해결하고자 많은 음원 모델들이 최근에 제안되었다[5]. 이 모델들은 음원 파형을 기저함수(basis function)의 가중 합(weighted sum)으로 나타낸다. 그러나 이 모델은 몇몇 중요한 문제들을 가지고 있다. 1)실제 음원 파형은 추정된 계수로부터 효과적으로 모델링 할 수 없다. 2)이런 문제들을 해결하기 위한 저차(reduced order)모델은 과도하게 간략화 된 모델이다. 3)모델 파라미터의 추정 기법이 완전하지 않다.

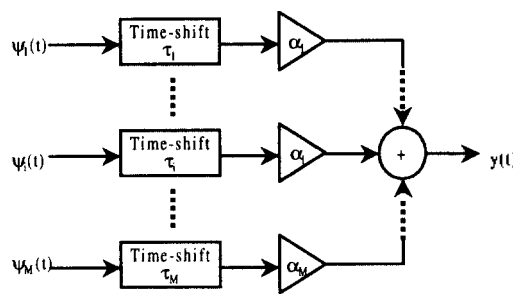
추출된 음원 파형으로부터 생리학적인 성문 파형(glottal waveform)을 예측하는 일반적인 수학적 표현이 없다. 이 분야에서의 연구성과 부족은 모델의 생리학적 파라미터의 측정이 어렵다는데 기인된다. 본 논문에서 기존의 음원 모델링 기법의 문제들을 해결하고자 기저함수의 가중 합을 이용하는 새로운 모델을 제안한다. 생리학적 모델 중에서 잘 알려진 이중 질량 모델(two-mass model)을 근이 화 하고 기저함수의 가중 합을 이용한 함수를 구한다. 이 모델은 음원 파형을 최적화 된 기저함수의 가중 합으로 나타낸다. 제안한 모델로는 다양한 음원 특성을 나타낼 수 있다. 마지막으로 제안한 음원 모델의 가중(weight)값 과 시간추이(time shift) 파라미터는 구역 분할 개념(region division concepts)으로 EM(estimate maximize) 알고리듬을 이용하여 추정된다. 본 논문의 구성은 다음과 같다. 2장에서는 기저함수의 가중 합 모델이 제안된다. 파라미터 추정 기법은 3장에서 검토된다. 성능평가를 위한 컴퓨터 실험결과는 4장에 제시된다.

2. 역 필터링된 파형의 모델링

유성음에 있어서 여기(excitation)는 성대(vocal cords)의 진동으로 인한 폐(lung)로부터의 공기의 유사 주기적(quasi-periodic)흐름에 기인한다. 그러나 이런 음원 생성 메카니즘은 단순한 다항식 모델로 나타내기에는 너무나 복잡하다. 음원 특성을 효과적으로 나타내기 위하여 많은 종류의 생리학적인 모델이 제안되었다. 생리학적인 모델에 있어서 음원은 음성 신호의 특성뿐만 아니라 음성신호를 만들어 내는 성도구조(vocal apparatus)에 따라 특징 지어진다. 본 논문에서 유성음 여기를 위해 사용하는 생리학적인 모델은 잘 알려진 이중 질량 모델이다[3]. 이 모델은 그림 2에 나타나 있다.



(a)



(b)

그림 2. 음원 모델링 기법 : (a) 이중 질량을 이용한 성도 모델; (b) 기저함수의 가중 합을 이용한 모델

성문 체적 속도 $u_g(t)$는 식(1)과 같은 미분 방정식을 만족한다.

$$R_{tot}\, u_g + L_{tot}\frac{du_g}{dt} = \Delta p \qquad (1)$$

여기서 Δp는 성문에서의 압력이다. 이중 질량 모델을 이용한 성도모델은 생리학적으로 어느 정도의 타당성이 있다. 하지만 음성신호로부터 모델의 파라미터를 추정할 수 없다는 문제점을 가지고 있다. 이런 문제들을 해결하기 위하여 그림 2(b)와 같이 기저함수의 가중 합으로 모델링 하는 기법이 유도된다. 시

간 단(time step)을 δt 라하고 이산시간을 도입함으로서, 또한 시간 $n\delta t$ 에서 $u_g{}^n = u_g(n\delta t)$을 최적해로 두면 식 (1)은 식 (2)과 같이 나타낼 수 있다.

$$L_{tot}\ u_g{}^{n+1} = (L_{tot} - R_{tot}\delta t)\ u_g{}^n + \delta\Delta p^n \quad (2)$$

기저함수 $\phi_{j,k}$에 대한 적성 가정을 통해 식 (3)을 구할 수 있다.

$$\theta_{j,k} = L_{tot}{}^{-1}(L_{tot} - R_{tot}\delta t)\ \phi_{j,k} \quad (3)$$

이로부터 식 (2)은 다음과 같은 식 (4)으로 표현할 수 있다.

$$\langle u_g{}^{n+1} \mid \phi_{j,k}\rangle = \langle u_g{}^n \mid \phi_{j,k}\rangle + \delta t\langle \Delta p^n \mid \theta_{j,k}\rangle \quad (4)$$

여기서 $\langle\ |\ \rangle$는 스칼라 적을 나타낸다. 따라서 식 (5)와 같은 이중 질량 모델에 대한 근이 해를 구할 수 있다.

$$u_g(t) = \sum_{i=0}^{M} \alpha_i(t)\phi(\frac{t-\tau_i}{c_i}) \quad (5)$$

여기서 ϕ는 기저함수, α_i, c_i, τ_i는 각각 진폭, 스케일, 그리고 위치 파라미터 값이다.

기저함수의 가중 합을 이용한 모델의 완전한 구현을 위해 음원을 근이 화하는 최적의 기저함수를 구한다. 음원을 최적화 하는 방법은 주파수 영역에서 L^p norm 의 근이 화 에러를 최소 화 함으로서 구할 수 있다. 최적의 기저함수를 찾는 방법은 식(6)과 같이 분산을 구하는 것이다.

$$\sigma_2 = \int_{\Omega} |Y(w)|^2 \Psi(\frac{w}{M})^2 dw \quad (6)$$

여기서 $Y(w)$ 는 한 피치주기 내에서 음원의 퓨리에 변환이고 $\Psi(w)$는 기저함수의 퓨리에 변환, 그리고 f는 정밀도(resolution) 이다.

기저함수의 가중 합 모델은 기존의 방법에 비해 몇몇 이점을 가지고 있다. 기존의 음원 모델은 음원 파형을 시간영역에서 간소화된 지수 합이나 다항식으로 나타내었다. 이런 기존의 간소화된 모델로는 다양한 음원 특성을 나타낼 수 없고 주파수영역에서 표현하는 데 어려움이 있다. 이러한 문제들을 해결하고자 지수함수의 가중 합을 이용하는 새로운 음원 모델이 제안되었다. 그러나 지수 합 모델은 몇몇 문제들을 가지고 있다. 1)실제 음원 파형을 추정된 계수로부터 효과적으로 모델링 할 수 없다. 2)1)의 문제를 해결하기 위해 저차의 모델은 너무 간략화 된 형태이다. 3)모델 계수의 추정 기법이 완전하지 않다. 4)음원의 주파수 특성을 잘 나타낼 수 없다.

3.기저함수의 가중합모델에관한 계수추정

음원 파형을 나타내는 기저함수의 가중 합에서 기저함수를 구하는 것이 중요하다. 실험결과 기저함수를 이용이 기존의 방법에 비해 더 좋은 결과를 얻을 수 있기 때문이다. 기저함수의 가중 합을 이용한 모델의 계수는 기저함수의 가중 값 과 시간 추이 값이다. 이 값들은 EM알고리듬을 이용하여 추정할 수 있다. 제안된 파라미터의 계수 값을 얻기 위하여 다음의 2가지 가정을 한다. (1)관측된 신호 $y(t)$는 어떤 기저함수의 가중 합으로도 구할 수 있다. (2)근이화 에러 $n(t)$는 평균이 영인 백색 가우시안 잡음이고, 분산 행렬은 $E(n(t)*n(\sigma)) = Q\delta(t-\sigma)$로 나타낼 수 있다. 위의 가정으로부터 로그와 유사한 식(7)과 같은 결과를 얻을 수 있다.

$$L(\theta) = d - \frac{\lambda}{2}\int_T [y(t) - \sum_{i=1}^{M}\alpha_i\phi(\frac{c_i}{t-\tau_i})]*$$
$$Q^{-1}[y(t) - \sum_{i=1}^{M}\alpha_i\phi(\frac{c_i}{t-\tau_i})]dt \quad (7)$$

따라서 ML방법에서 기저함수의 가중 값과 시간 변이 값을 얻기위한 식은 다음과 같다.

$$\min \int_T |y(t) - \sum_{i=1}^{M}\alpha_i\phi(\frac{c_i}{t-\tau_i})|^2 dt \quad (8)$$

식(8)은 복잡한 멀티파라미터 최적화 문제이다. 이 값 을 구하는 데는 복잡한 계산을 필요로 하고 시간 또 한 많이 소비된다. 하지만 EM알고리듬을 이용하면 복잡한 멀티파라미터 최적화 문제를 간략화 시킬 수 있다.

추정단계

$$For\ i = 1,2,..,M \qquad \hat{x}_i{}^{(n)} = \hat{\alpha}_i{}^{(n)}\phi(\frac{t-\hat{\tau}_i}{c_i{}^{(n)}})$$
$$+\ \beta_i[y(t) - \sum_{i=1}^{M}\hat{\alpha}_i{}^{(n)}\phi(\frac{t-\hat{\tau}_i}{c_i{}^{(n)}})] \quad (9)$$

최적화 단계

$$For\ i = 1,2,..,M$$
$$\min \int_T |\ \hat{x}_i{}^{(n)} - a\phi(\frac{t-\tau}{c})|^2 dt$$
$$\rightarrow \hat{\alpha}_i{}^{(n+1)},\ \hat{\tau}_i{}^{(n+1)}$$

여기서 n은 반복 횟수이다.

그러나 이 알고리듬을 기저함수의 가중 합 모델에서의 파라미터 추정에 직접 사용할 수 는 없다. 왜냐하면 추정된 파라미터 값들이 지역최적값(local optimum)에 수렴하기 때문이다. 본 논문에서는 이런 문제를 해결하기 위해 음원 파형을 구간 선택 개념을 도입하여 두 구간으로 나눈다. 즉 기울기가 양인 구간과 음인 구간으로 나눈다. 다음장의 실험결과에서 보듯이 구간 분할 개념이 기존의 EM기법에 있어서 지역 적인 최적수렴의 문제를 해결하고 있음을 볼 수 있다.

4. 실험 결과

본 논문에서는 위와 같은 분석 절차를 통해 유성음에 대해 실험해 보았다(남성 과 여성 각각 1명). 유성음은 마이크로폰을 통해 녹음하였고, A/D 변환 후 샘플링 주파수는 8kHz, 양자화는 8비트로 하였다. 성문 닫힘 구간(glottis open region) 과 성문 열림 구간(glottis close region)만 정확히 찾는다면 입은 실험결과는 여기에서 보여준 예와 일치한다. 닫힘과 열림 구간에 관한 알고리듬은 참고문헌 [9]에 제시된다. 본 논문에서는 참고문헌 [9]의 역 필터링 알고리듬을 사용했다. 이 실험에서 음원 모델의 차수는 6차로 하였다. 제안된 분석 방법에 관한 성능을 그림 3을 통해 제시한다. 그림 3에서 보듯이 제안한 모델이 음성의 복잡한 특성을 더 잘 모델링 하고 있음을 알 수 있다. 합성음에 있어서 기존의 방법에 비해 약 4dB 정도 개선되었다.

5. 결론

본 논문에서는 다양한 음원 특성을 나타낼 수 있는 새로운 음원 모델링 기법을 제안하였다. 이 기법은 기저함수의 가중 합을 이용한 모델링 기법이다. 이 모델은 추정된 음원 파형을 미리 정의된 기저함수의 가중 합으로 나타낼 수 있다. 마지막으로 제안한 음원 모델의 가중 값과 시간 변이파라미터들은 구간 개념을 적용한 EM알고리듬을 사용하여 추정할 수 있었다. 실험결과 제안한 추정 기법과 모델링 기법이 기존방법에 비해 더 정확한 음원 파형을 추정할 수 있었고, 또한 다양한 음원 특성도 모델링 할 수 있었다. 제안한 방법을 이용하여 음성 합성이나 부호화 기술에서 음성의 질을 향상시키는데 기여할 수 있으리라 생각된다.

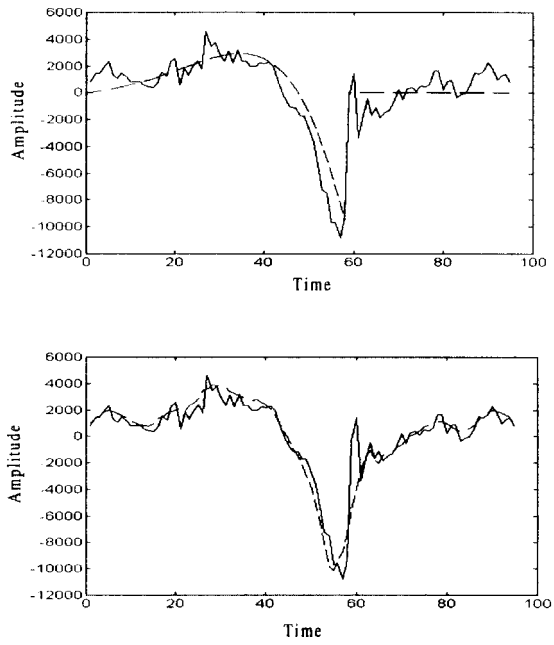


그림 3. 음원모델링 결과:(a)LF 모델 (b)제안한 방법

참고 문헌

[1] J. D. Markel and A. H. Gray, Jr.,, *"Linear Prediction of Speech,"* Springer-Velag, 1976.

[2] S. Crisafulli, J. D. Mills and R. R. Bitmead,"Kalman filtering techniques in speech coding", *Proceedings of ICASSP*, vol.1,pp.77 80, 1992.

[3] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," Bell Syst. Tech. J. 51(6) pp.1233-1268, 1972.

[4] G.Fant, J. Lilencrants, and Q. Lin, "A four parameter model of glottal flow," STL-QPSR 4/1985, pp. 1-13, 1985

[5] M.M.Thomson,"A New model for determining the vocal tract transfer function and its excitation from voiced speech,"*Proceedingsof ICASSP*, vol.2,pp37-40,1992

[6] R.Vaccaro,*SVD and Signal Processing II:Algorithms, Analysis and Applications*, Elsevier Publishers B.V., 1991.

[7] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using EM algorithm," *IEEE Trans. on ASSP*, vol 36, n0.4, Apr.,1988.

[8] Y. M. Cheng and D. O'Shaughnessy," Automatic and reliable estimation of glottal closure instant and period," *IEEE Trans. on ASSP*, vol. 37, n0.12, Dec.,1989.

[9] S. H. Hong, S. K. Kang and S. Ann,"Voice source estimation using sequential SVD and EM algorithm, " *Proceedings of ICSLP*, May 1994.