ML 기반의 음성의 유/무성음 성분 분리

강명구, 유창동
한국통신 멀티미디어연구소 음성언어연구실

A VOICED/UNVOICED DECOMPOSITION OF SPEECH BASED ON MAXIMUM LIKELIHOOD METHOD

Kang Myung koo, Chang Dong Yoo
Spoken Language Research Team. Multimedia Technology Research Laboratory. Korea Telecom

Email : mgkang@smm.kotel.co.kr, cdyoo@smm.kotel.co.kr

요약

본 논문에서는 음성에 공존하는 유/무성음 성분을 추정하는 알고리즘을 제안하였다. 유성음 성분은 주기성을 띤 사인곡선의 형태로 표현되며, 무성음 성분은 자동회기의 결과로 표현된다. 두 성분을 각각 차례대로 추정할 경우 한 성분에 대한 추정치의 정확도가 나머지 성분의 추정에도 영향을 주기 때문에 제안된 알고리즘은 두 성분을 공동으로 추정한다. 실제 ML(maximum likelihood) 추정치는 구하기 어려워 이에 근접하는 추정치를 선형 방정식들을 iterative 방법으로 풀어 구하였다. 예비 시험결과 제안한 알고리즘이 정확하고 효율적으로 두 성분을 추정함을 알 수 있었고, 합성된 데이터 뿐만 아니라 실제 음성 데이터를 이용한 실험에서도 좋은 결과를 보여 주었다

I. 서론

음성모델을 기반으로 하는 음성처리 시스템의 성능은 모델의 정확성과 모델 계수 값들의 정확성에 따라 제한된다. 기존의 많은 음성 모델들은 유/무성음 성분의 판단을 요구하고 있어 유/무성음으로 분류할 수 없는 음성을 표현하는데 유연성이 부족하다. 다수의 parametric 음성 모델들은 여기 (excitation)와 spectral envelop의 계수 값들을 추정해야 하는데 추정하는 과정에서 계수 값들이 상호 영향을 준다. 대부분의 경우에 여기 계수 값들은 스펙트럴 모델의 추정 치들을 이용하게 된다. 이러한 방법은 여기계수 값이 스펙트럴 모델 추정의 정확성에 크게 영향을 받기 때문에 계수의 수가 일정하게 고정되어 있는 경우 음성을 표현하는데 효과적이지 못할 수 있다. 본 논문에서는 유/무성음이 공존하는 것을 허용하여 음성의 성분을 분별할 필요가 없어 오판으로 생길 수 있는 문제점들이 없다. 또 이 알고리즘은 유/무성음 두 성분을 동시에 추정함으로써 음성을 보다 정확하게 표현 할 수 있다.

음성이 유/무성음 성분들의 합으로 모델링(modeling) 하면 여러 응용분야에서 유용하게 이용될 수 있다. 음질개선에 있어서는 적절하게 각 성분들의 고유한 특성을 살리면서 개선하기 때문에 잡음제거 과정에서 생길 수 있는 음질왜곡 현상과 같은 단점을 없앨 수 있다[1,2]. 음성 성분 분리를 이용한 또 다른 유용한 응용분야는 피치 검출 이다. 기본적인 시험결과 이 방법을 이용한 피치 검출은, Gold-Rabiner 피치 검출 방법에 비해 좋은 결과를 보인 다 대역 여기(multiband excitation) 모델을 이용한 방법에 비해서도 정확도가 높았다. 본 논문에서 제안된 알고리즘을 이용한 피치 검출 방법은, 특히 유성음 대 무성음의 에너지 비가 낮을 경우 다 대역 여기 모델을 이용한 방법에 비해 피치 추정의 정확도가 현저하게 개선되었다.

유/무성음 성분을 추정하는데 maximum likelihood 추정 방법을 이용하면 계산이 복잡한 비선형(non-linear) 방정식을 풀어야 하는 문제가 생긴다. 그러나 실제 maximum likelihood 방법의 근사법을 이용하면 단지 선형 방정식을 반복적으로 푸는 문제로 단순화 시킬 수 있다. 두 성분을 각각 차례대로 추정할 경우 한 성분에 대한 추정치의 정확도가 다른 성분의 추정에 영향을 주기 때문에 제안한 알고리즘은 두 성분을 동시에 추정하는 구조를 제공한다

whitening 적응필터로 처리된 음성 신호의 minimum mean square 오차에 대해서 유성음은 추정된다는 전제를 알고리즘을 가진다. 이 필터는 유성음 대 무성음의 에너지 비가 낮은 주파수 성분은 강조를 하고 그 반대의 경우에는 약화시킨다. 이 필터의 frequency response은 무성음 성분의 spectral envelope의 역이다.

먼저, 제 2절에서는 유/무성음 성분이 공존하는 새로운 음성 모델을 설명하였고, 제 3절에서는 각 성분의 계수 값들을 추정하는 알고리즘을 설명하였으며 알고리즘을 합성된 데이터와 실제 음성 데이터에 적용해 보았다. 마지막으로 제 4절에서는 본 논문의 내용을 종합 정리한다.

II. 새로운 음성 모델

음성은 harmonic성분 **v**[n] 과 non-harmonic성분 **u**[n]의 합성으로 모델링 할 수 있다. 음성 신호는 시간에 따라 변화하는 본질 때문에 윈도우 함수 w[n]으로 고정된 구간으로 나누어 진다. 윈도우 함수에 의해 나누어진 음성 신호의 한 구간을 $s_w[n]$으로 다음과 같이

나타낸다.

$$s_w[n] = s[n] \bullet w[n] \quad (1)$$

$$s_w[n] = v_w[n] + u_w[n] \quad (2)$$

첨자 w는 윈도우 함수 w[n]에 의해서 구해지는 한 구간을 나타낸다. 유성음 성분은 주기성을 띤 성분으로 조정된 사인곡선들(harmonically modulated sinusoids)의 합으로 표현된다. $v_w[n]$은 수학적으로 다음과 같이 나타낼 수 있다.

$$v_w[n] = \sum_{m=-M}^{M} A_m e^{j\omega_0 mn} w[n], \qquad M = \left\lfloor \frac{\pi}{\omega_0} \right\rfloor \quad (3)$$

A_m과 ω_0은 각각 m번째 harmonic의 진폭과 기본 주파수를 나타낸다. 무성음 성분 $u_w[n]$은 차수가 p인 자동회기(autoregressive) 과정의 결과로 모델링 되며 수학적으로는 다음과 같이 표현된다.

$$\sum_{i=0}^{p} b_i u_w[n-i] + G \bullet d[n] = 0, \qquad b_0 = -1 \quad (4)$$

G와d[n]은 각각 gain과 평균값이 0이고 분산이 1인 백색 Gaussian 잡음을 나타낸다. 위의 공식들을 보면 음성신호 $s_w[n]$을 구하는데 2M+2p+3개의 계수들이 필요하다. 2M+1개의 진폭 값 $\{A_{-M},\ldots,A_M\}$, p개의 선형 예측 계수 값 $\{b_1,\ldots,b_p\}$, 기본 주파수 ω_0, gain G, p개의 초기 상태 값 $\{s_w[-1],\ldots,s_w[-p]\}$이 필요하다.

III. 유/무성음 성분의 추정 알고리즘

결정된 계수 값들은 다음과 같이 표시한다.

$$\theta = [\omega_0 \quad \mathbf{a} \quad G \quad \mathbf{b}]^t$$

$$\mathbf{a} = [A_{-M},\ldots,A_M], \quad \mathbf{b} = [b_1,\ldots,b_p]^t$$

p개의 초기상태 값은 알려진 것으로 가정하고 $\mathbf{s}_I$로 표시한다. 계수 값들의 공간 Ω는 (2M+p+3)차원의 Euclidean 공간이다.

$\mathbf{s}_O$는 조건부 확률 밀도 함수 $\mathbf{f}_s(\mathbf{s}_O \mid \theta;\mathbf{s}_I)$를 가지는 observed vector $[s_w[N-1],\ldots,s_w[0]]^t$를 나타낸다. maximum likelihood 추정치 θ_{ML}은 $\mathbf{f}_s(\mathbf{s}_O \mid \theta;\mathbf{s}_I)$ 또는 $\log \mathbf{f}_s(\mathbf{s}_O \mid \theta;\mathbf{s}_I)$ 값이 최대가 되는 θ의 값이다. 식은 다음과 같다.

$$\theta_{ML} = \arg\max_{\theta\in\Omega} \log \mathbf{f}_s(\mathbf{s}_O \mid \theta;\mathbf{s}_I) \quad (5)$$

조건부 확률 밀도 함수 $\mathbf{f}_s(\mathbf{s}_O \mid \theta;\mathbf{s}_I)$는 간단하게 다음과 같이 구할 수 있다.

$$\mathbf{f}_s(\mathbf{s}_O \mid \theta;\mathbf{s}_I) = \prod_{n=0}^{N-1} f(s[n] \mid \theta, s_w[n-1],\ldots,s_w[-p]) = \prod_{n=0}^{N-1} f(s[n] \mid \theta, s_w[n-1],\ldots,s_w[n-p])$$

식 4로부터 log-likelihood 함수 $\log \mathbf{f}_s(\mathbf{s}_O \mid \theta;\mathbf{s}_I)$를 다음과 같이 구할 수 있다.

$$\log \mathbf{f}_s(\mathbf{s}_O \mid \theta;\mathbf{s}_I) = -\frac{N}{2}\log 2\pi - N\log G - \frac{1}{2G^2}\sum_{n=0}^{N-1}\left(\sum_{k=0}^{p} b_k u_w[n-k]\right)^2 \quad (6)$$

Estimation of Harmonic Amplitude: 식 2를 식 6에 적용하고 A^*에 대해서 함수를 최대화 한 후 적절한 가정을 하면 다음과 같은 식을 구할 수 있다.

$$A_m = \frac{\int_{-\pi}^{\pi} |B(\omega)|^2 S_w(\omega) W^*(\omega - m\omega_0)\, d\omega}{\int_{-\pi}^{\pi} |B(\omega) W(\omega - m\omega_0)|^2 \, d\omega} \quad (7)$$

$B(\omega), S_w(\omega), W(\omega)$는 각각 $\{b_i\}_{i=0}^{p}, s_w[n], w[n]$의 Fourier 변환이다.

Estimation of Fundamental Frequency: 피치 ω_0의 추정치는 다음과 같이 $Q(\omega_0; B(\omega))$의 최대값을 구함으로써 얻을 수 있다

$$\omega_0 = \arg\max Q(\omega_0; B(\omega)) \quad (8)$$

$$Q(\omega_0; B(\omega)) = \sum_{m=-M}^{M} \frac{\int_{-\pi}^{\pi} |B(\omega)|^2 S_w(\omega) W^*(\omega - m\omega_0) d}{\int_{-\pi}^{\pi} |B(\omega) W(\omega - m\omega_0)|^2 \, d\omega}$$

Estimation of Linear Prediction Coefficient: b_k에 대한 log-likelihood 함수의 최대값은 선형예측에 있어서 일반적인 공분산을 따르면[5,6] 다음과 같이 표현할 수 있다.

$$\begin{bmatrix} R_{(1,1)} & \cdots & R_{(1,p)} \\ \vdots & \ddots & \vdots \\ R_{(p,1)} & \cdots & R_{(p,p)} \end{bmatrix} \bullet \begin{bmatrix} b_1 \\ \vdots \\ b_p \end{bmatrix} = \begin{bmatrix} R_{(1,0)} \\ \vdots \\ R_{(1,p)} \end{bmatrix} \quad (9)$$

$$R_{(i,k)} = \sum_{n=0}^{N-1} u_w[n-i]u_w[n-k]$$

i,k = 1,.....p

Estimation of Gain: G의 추정치는 log-likelihood 함수를 G에 대해서 최대화 함으로서 구할 수 있다.

$$G = \frac{1}{N^{\frac{1}{2}}}\left(R_{(0,0)} - \sum_{j=1}^{p} b_j R_{(j,0)} \right)^{1/2} \quad (10)$$

3.1 반복 알고리즘(Iterative Algorithm)

maximum likelihood에서 θ 값을 구하기 위해서는 비선형 방정식 (7),(8),(9),(10)을 풀어야 한다. 실제 maximum likelihood 추정 과정은 선형 방정식을 푸는 반복적인 방법으로 근사화 할 수 있다. 선형 예측 계수값 $\{b_k\}_{k=1}^{p}$을 알고 있을 때 기본 주파수 ω_0는 식 (8)에 의해서 구해진다. $\{b_k\}_{k=1}^{p}$와 ω_0를 알고 있을 때 harmonic 진폭 $\{A_m\}_{m=-M}^{M}$은 식 (7)을 이용해서 계산할 수 있다. $\mathbf{b}, \omega_0, \mathbf{a}$를 알고 있을 때 gain G는 식 (10)을 이용해서 구할 수 있다. 유사하게 A_m, ω_0를 알고 있으면 b_k는 식 (9)를 이용해서 구할 수 있다. 이러한 관찰은 자연스럽게 반복적인 과정을 통해서 최적의 계수 값들을 계산할 수 있다는 것을 제시한다. b_k의 초기상태(예: b_k = 0 for{k|1, ..p}는 쉽게 가정할 수 있고 나머지 계수 값들($\omega_0, \mathbf{a}, G$)은 주어진 차수에서 반복적인 계산을 통해서 구할 수 있다.계산은 원하는 수렴수치가 만족될 때까지 반복적으로 수행한다. 알고리즘의 반복 수행 시 마다 likelihood 함수의 값은 증가한다고 볼 수 있다.

알고리즘을 관찰해 보면 유성음은 frequency response $B(\omega)$를 가지는 whitening 적응필터로 처리된 음성 신호를 MMSE (minimum mean square error)에 대해서 추정된다. 다른 관점에서 보면 다음과 같은 유성음 추정 문제로 볼 수 있다:

$$\{A_m\}_{m=-M}^{M}, \omega_0 = \arg\min E,$$

$$E = \frac{1}{2\pi}\int_{-\pi}^{\pi} |X_w(\omega) - V_w(\omega)|^2 \, d\omega,$$

$$X_w(\omega) = B(\omega)S_w(\omega),$$

$$V_w(\omega) = B(\omega)\sum_{m=-M}^{M} A_m W(\omega - m\omega_0)$$

여기서 gain 함수는 유성음 대 무성음의 에너지 비가 높은 주파수 성분을 강조하게 된다. 이 gain 함수는 무성음 성분의 spectral envelope의 역함수 이다. 무성음 성분의 spectral envelope은 $s_w[n]$과 유성음 추정치 $\hat{v}_w[n]$의 차, $e_w[n] = s_w[n] - \hat{v}_w[n]$ 의 LPC 분석에 의해서 결정된다.

3.2 알고리즘의 응용

제 3절에서 설명한 반복 알고리즘의 실례를 보이기 위해, 알고리즘에 합성 데이터와 실제 음성 데이터를 적용해 보았다. 합성 데이터는 자동회기 과정(autoregressive process)의 출력(무성음)에 harmonically modulated 신호(유성음)를 인가하여 구한다. 윈도우 처리된 유성음 신호의 스펙트랄(spectral) 크기 $V_w(\omega)$는 그림 1(a)에 보여준다. 그림에서 보면 동일한 크기의 진폭을 가지며 주파수는 $\frac{2\pi}{40.3}$ 이다. 그림 1(b)는 자동회기(autoregressive) 신호의 스펙트럼으로, $0.8e^{j0}$에서 double pole을 가지는 2차 all-pole 필터의 자극으로 생성한다. 그림 1(c)는 합성된 유성음과 무성음 성분의 합의 스펙트랄 크기를 보여준다. 이 예에서는 $\beta = 7$인 Kaiser 윈도우를 사용하였다.

초기값으로 $B(\omega)$=1로 가정하였고, 폴(pole)의 수는 10으로 가정하였다. 첫 번째 수행 후 추정된 유/무성음 성분의 스펙트럼이 2(a)와 (b)이다. 추정된 기본 주파수는 $\frac{2\pi}{37.1}$이다. 두 번째 반복 수행 후 알고리즘은 수렴하여 유/무성음 성분을 그림 2(c)와 (d)로 분리하였다. 추정된 기본 주파수는 $\frac{2\pi}{40.3}$이다. 추정된 유성음 성분은 점선으로 표시된 원 유성음 성분을, 유성음 대 무성음의 에너지 비가 높은 고주파 대에서는 잘 모델링하고 에너지 비가 낮은 저주파 대에서는 제대로 모델링하지 못하는 것을 알 수 있다. 유성음 성분을 추정할 때 알고리즘은 유성음 대 무성음의 에너지 비가 높은 스펙트럼 영역을 찾는다. 알고리즘의 기본 주파수 추정치가 정확하다는 사실을 주목해야 한다.

다음 실험은 유성음 대 무성음의 에너지 비가 전체적으로 작은 실제 음성 데이터를 알고리즘에 적용하였다. 그림 3(a)는 실제 음성 데이터 "leisure"의 /zh/발음 스펙트럼이다. 이러한 데이터는 제한적인 유성음 성분을 가지고 있기 때문에 전통적인 방법으로는 피치(pitch)를 추정할 수 없다[6,8]. 그러나 제안된 알고리즘의 adaptive gain 함수를 이용하면 피치를 추정하는 게 가능하다. 초기값으로 $B(\omega)$=1로 설정하고 단지 두 번의 반복 수행 후 알고리즘은 수렴하였고, 분리된 유성음과 무성음 성분은 그림 3(b)와 (c)이다. 그림 3(b) 와 (c) 를

비교하면 알고리즘은 유성음 대 무성음의 에너지 비가 전체적으로 작은 경우에도 스펙트럼의 유성음 성분을 잘 모델링 한다.

4. 결론

알고리즘은 공존하는 유성음과 무성음 성분의 추정을 보여주었다. 실제 maximum likelihood 추정 과정의 근사화 하는 시도는 일련의 선형 방적식을 푸는 iterative 방법으로 귀결되었다. 한 성분의 추정치가 다른 성분의 추정에 영향을 미치게 되는 차례대로 각 성분을 추정하는 방법 대신 알고리즘은 두 성분을 동시에 추정하는 구조를 제공하였다. 알고리즘을 자세히 관찰해 보면 유성음은 whitening 적응필터로 처리된 음성 신호를 MMSE에 대해서 추정된다. 알고리즘에서 사용되고 있는 gain 함수는 유성음 대 무성음의 에너지 비가 높은 주파수 성분들을 강조하게 된다. whitening 필터의 frequency response은 무성음 성분의 spectral envelope의 역함수 이다.

6. 참조문헌

[1] J.Hardwick, C.D.Yoo, and J.S. Lim, "Speech enhancement using the dual excitaion model", ICASSP, pp.367-370, April 1993.

[2] C.D.Yoo and J.S.Lim, "Speech enhancement based on the generalized dual excitation model with adaptive analysis window", ICASSP,pp.832-835, May 1995.

[3] D.W. Griffin and J. Lim, " A new pitch estimation algorithm," Proc. of the Int. Conf. on Acoustics, Speech and Signal Proc., vol. 67, pp. 592-601, March 1984

[4] D.W. Griffin, Multi-Band Excitation Vocoder. PhD thesis, MIT. E.E.C.S. Department, 1987.

[5] J. Markel and J. A.H Gray, Linear Prediction of Speech. Berlin: Springer-Verlag., 1976.

[6] L. Rabiner and R. Schafer, Digital Processing of Speech Signals. New Jersey: Prentice-Hall., 1978.

[7] J. Hardwick, The Dual Excitation Speech Model. PhD thesis. MIT. E.E.C.S. Department, June 1992.

[8] J.John R. Deller, J.G. Proakis, and J.H.L. Hansen, Discrete Time Processing of Speech Signals. New York: McMillan, 1993.

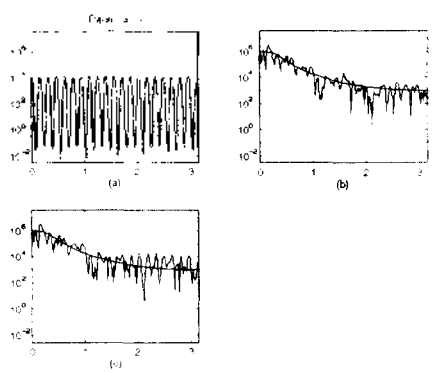

그림1. 합성데이타

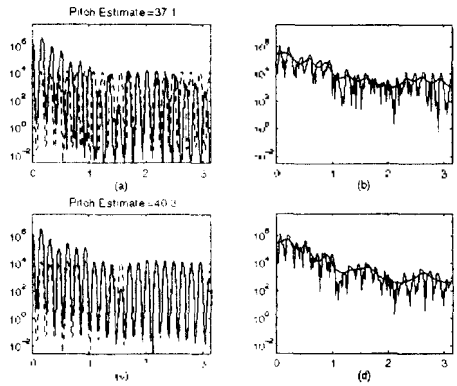

그림 2. a) 한번 수행 후 유성음 스펙트럼 b) 한번 수행 후 무성음 스펙트럼 c) 두번 반복 후 유성음 스펙트럼 d) 두번 반복 후 무성음 스펙트럼. 점선은 그림 1. a) 원 유성음 신호 스펙트럼

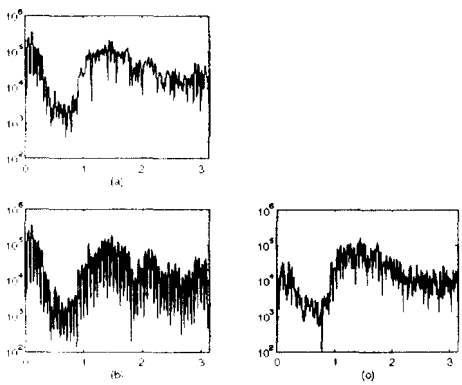

그림 3. ML 합성의 예 a) 원 음성의 스펙트럼 b) 두번 반복 후 추정된 유성음 스펙트럼 c) 두번 반복 후 추정된 잔존신호 스펙트럼