성문 닫힘 구간 가변에 의한 피치변경

강동규, 김상훈, 이정철
한국전자통신연구원

Glottal Closure Interval Extrapolation Technique Based Pitch modification

Dong-Gyu Kang, Sanghun Kim, Jungchul Lee
ELECTRONICS AND TELECOMMUNICATIONS RESEARCH INSTITUTE
E-mail: dgkang.etri.re.kr

요약문

시간영역에서 유성음의 피치를 조절하기 위해 한 피치구간의 신호 중에서 성문이 닫힌 구간의 특성을 추정한 파라미터로 성문 닫힌 구간의 신호에 연속하여 선형적으로 연장 또는 축소하므로서 고 음질을 유지하면서도 자유롭게 피치를 조절할 수 있는 방법을 제안하였다. 제안된 방법은 PSOLA 기법에서와 같은 window의 적용이나 신호의 겹침에 의한 영향이 최소화되므로 보다 명료한 합성음을 얻을 수 있다.

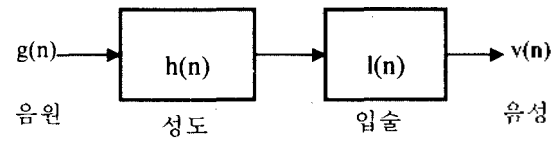
1. 서론

음성 합성 방법에는 합성 가능한 어휘의 범위에 따라 제한 어휘 합성과 무제한 어휘 합성 방식으로 분류할 수 있다. 제한 어휘 합성방식은 합성하고자 하는 어휘들을 미리 분석하여 데이타 베이스로 저장하였다가 필요에 따라 이들을 조합하여 음성을 합성하는 방법으로 현재 지하철 안내 방송이나 전화의 114 안내 방송 등에서 사용되고 있다. 이 방식은 비교적 양질의 합성음을 얻을 수 있으나 제한된 어휘만을 합성할 수 있는 단점이 있다. 무제한 어휘 합성방식은 음소, 반음소, 반음절, 음절 등을 기본 단위로 하여 합성 단위의 데이타 베이스를 구성하였다가 규칙에 의해 기본단위를 조합하여 음성을 합성한다. 이 방식은 제한 어휘 합성 방식보다 음질은 좋지 않으나 임의의 문장을 합성할 수 있으므로 문자/음성(TTS : Text-To-Speech) 변환 시스템에 적용되고 있다.

무제한 어휘 합성 방식에는 파라미터 방식인 포먼트, LPC, LSP 합성 방법 등이 연구되었으며 이 방법들은 음질은 다소 떨어지지만 음원과 성도 파라미터 등을 조절하여 다양한 합성음을 만들 수 있는 장점이 있다. 고품질의 합성음을 얻기 위해 자연 음성신호를 집속하여 시간영역에서 피치를 가변할 수 있는 PSOLA 기법이 제안되었다. PSOLA 방법은 피치의 변경율이 클수록 피치 단위별로 적용하는 window의 영향과 피치구간 전체가 중첩되면서 발생하는 스펙트럼 왜곡이 커져 합성음의 명료도가 저하되는 것이 단점으로 알려져 있다. 이와 같은 PSOLA 기법의 단점을 극복하기 위해 본 연구에서는 한 피치 구간에서 창 함수 영향을 최소화할 수 있도록 성문 닫힘 구간에 연속적인 신호를 임의의 길이까지 연장 혹은 축소하여 피치를 변경할 수 있는 방법(Glottal closure interval Extrapolation Technique: GET)을 제안하였다.

2. 음성 발성 모델

음성 발성은 음원신호를 $g(n)$, 성도 함수를 $h(n)$, 발성된 음성신호를 $v(n)$이라 할 때 [그림 1]과 같이 음원이 성도 필터를 통과해 입술에서 방사되어 발생하는 선형 시스템으로 모델링될 수 있다.



[그림 1] 음성 생성 모델

비음을 제외한 유성음의 주파수 응답 $V(z)$는 식 (1)과 같이 표현할 수 있다.

$$V(z) = G(z) \times H(z) \times L(z)$$
$$= \frac{G'(z)}{A(z)} = \frac{G'(z)}{1 + \sum_{k=1}^{p} a_k z^{-k}} \quad \text{--- (1)}$$

여기서 a_k는 선형 예측 계수이고 $G'(Z)=G(Z)\cdot L(Z)$이다.

음성 발생은 유성음의 경우 성대의 진동에 의한 여기 신호가 성도를 통과하면서 공명을 일으켜 발생된다. 성대는 베루누이 효과(Bernoulli Effect)에 의해 설명되는 진동을 일으키며 급격히 닫히고 서서히 열리는 특성을 나타낸다. 유성음 신호는 성대가 급격히 닫히는 시점에서 최대의 에너지로 여기되고 성문이 닫혀 있는 동안에는 여기원이 없으므로 조음구조와 성도의 물리적 특성에 따른 자연스런 감쇠진동을 일으킨다. 성문이 서서히 열리면서부터는 열린 성문과 음원 신호에 의해 자연스런 감쇠진동을 방해를 받으므로 공명 주파수가 변화하고 더욱 급격한 감쇠 진동을 하다가 다시 성문이 급격히 닫히면서 위와 같은 과정을 반복한다.

식 (1)을 다른 형태로 나타내면 식 (2) 와 같이 나타낼 수 있다.

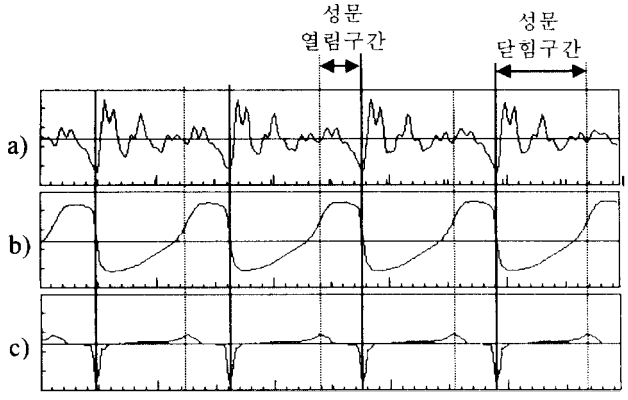
$$v(n) = g(n) + \sum_{k=1}^{p} a_k v_{n-k} \quad \text{--- (2)}$$

성문 닫힘 구간에서는 음원 특성인 식 (2)의 $g(n)$이 영(zero) 혹은 상수가 되므로 이 구간의 신호는 zero-input 응답으로 모델링될 수 있고 이 구간 내의 음성신호는 한 피치 구간 내에서 대부분의 에너지와 포먼트 정보를 포함하고 있다. 성문 닫힘 구간에서는 성도 특성이 선형적이고 출력 신호가 zero-input 응답이어서 보다 정확한 분석이 가능하므로 이 구간의 신호를 분석하여 구한 성도 특성으로 성문 닫힘 구간의 신호를 역필터링하면 보다 정확하게 음원 특성인 성문파를 추정할 수 있다[4][5]. 유성음에서 성문 닫힘과 열림 구간에 대한 정보를 알면 시간영역에서 한 피치 구간의 신호를 음원과 성도에 대한 특성으로 분리할 수 있으므로 식 (2)에 의해 성문 닫힘 구간의 신호를 성도 특성에 따라 시간영역에서 선형적으로 연장하거나 줄여서 유성음의 피치를 임의로 조절할 수 있다.

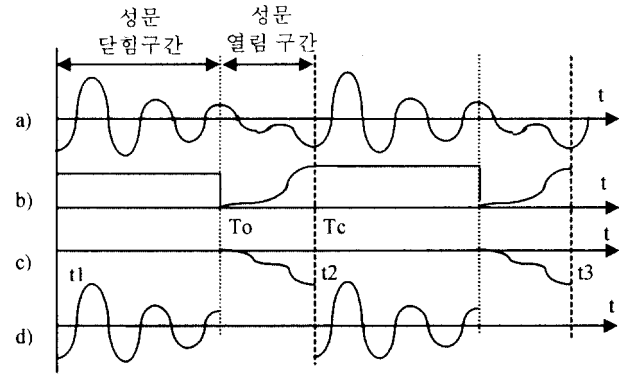
3. 유성음에서 성문과 성도 특성의 분리

성문 닫힘 구간 분석에 의해 한 피치 구간에서 성도와 음원 특성의 신호로 분리하기 위해서는 성문 닫힘 시점의 검출이 선행되어야 한다. 성문 닫힘 시점은 성문 진동을 관측할 수 있는 EGG(ElectroGlottoGraph)신호를 음성과 동시에 녹음하여 검출하거나 음성신호를 처리하여 epoch를 검출함으로서 구할 수 있다. 후자의 방법은 임의의 음성을 사용할 수 있는 반면 성문 열림 구간을 결정하기 어렵고 정확도가 전자에 비해 낮으므로 수작업으로 후 처리해야 할 필요가 있다. 전자의 경우에는 [그림 2]에서와 같이 검출이 용이하고 정확도가 높으며 성문 열림 정보도 비교적 정확하게 구할 수 있는 반면 부가적인 장비(EGG 측정기)를 착용하여 음성신호와 동시에 녹음해야 하며 기존의 녹음된 음성신호를 활용할 수 없는 단점이 있다.

시간 영역에서 성도 특성의 신호는 성문 닫힘 구간의 신호를 분리하여 쉽게 얻을 수 있지만 성문 특성에 의한 신호는 성문 열림 구간의 신호에서 성도 특성을 제거해야 하므로 복잡하고 정밀한 처리가 필요하다.



[그림 2] EGG 신호에 의한 성문 닫힘과 열림 구간의 검출, a) 음성파형, b) EGG 신호, c) 1 차 미분된 EGG 신호



[그림 3] 성도와 성문 특성신호의 근사적 분리, a) 음성신호, b) 가중함수, c) 음원 특성신호, d) 성도 특성신호

그러나 성문 열림 구간에서 성문과 성도 특성의 에너지 비율이 상대적으로 성문 특성 쪽이 현저히 크므로 [그림 3] b)와 같이 성문 열림 구간의 신호에서 성문 특성이 많은 쪽에 큰 가중치를 주면 근사적으로 음원 신호를 분리할 수 있다. 이와 같이 분리된 음원 신호는 두 피치를 접속할 때 신호의 자연스런 연속성을 유지시킬 수 있다.

성문 열림 구간은 일반적으로 한 피치의 30~70% 정도를 차지하므로 가중함수의 길이는 이에 따라 결정할 수 있다[3]. 실험에 의하면 가중 함수의 길이를 피치의 40~60% 정도로 하였을 경우 음질의 저하가 가장 적게 나타났다. 본 연구에서 사용한 성문 닫힘 구간의 검출 방법은 EGG 신호를 이용할 경우에는 미분된 EGG 신호에서 피크 피킹(peak picking)하여 마이너스(-) 측의 큰 피크는 성문 닫힘 시점으로, 플러스(+) 측의 작은 피크는 성문 열림 시점으로 사용하며[4][5] 신호처리 기법에 의한 epoch 검출기를 이용할 경우에는 epoch 시점으로부터 한 피치 구간의 40~50%로 근사화 하였다. 성문 열림 구간은 EGG 신호를 이용하는 경우에는 검출된 성문 열림 구간을 적용하고 epoch 검출기를 이용하는 경우에는 성문 닫힘 시점의 직전에 위치하는 피치의 40~60%를 적용하였다.

4. 성문 닫힘 구간 의 가변에 의한 피치 변경

성문 닫힘 구간 신호의 가변에 의한 피치 수정 방법은 다음과 같이 4 단계에 걸쳐 수행된다.

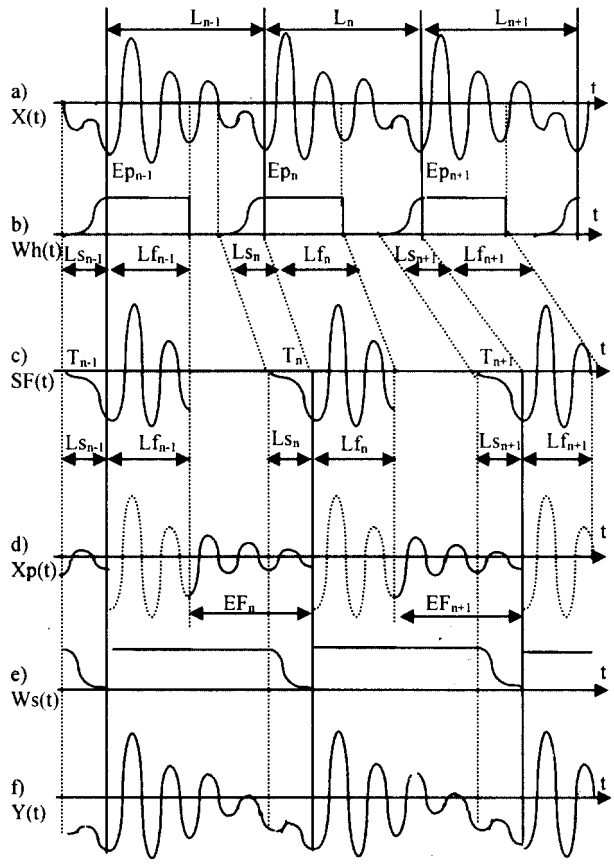
1 단계 : 성문 닫힘 구간 검출 및 성도 파라미터 추정

2 단계 : 성문 닫힘 구간에서의 음성신호와 성문 열림 구간의 신호 분리

3 단계 : 성문 닫힘 구간 신호의 연장 또는 축소

4 단계 : 성문 닫힘 구간이 변경된 신호에 성문 열림 구간의 신호 접합

먼저 1 단계에서 성문 닫힘 구간은 EGG 보다는 정확도가 낮지만 일반적인 경우를 고려하여 epoch 검출기를 이용하여 검출한다[1]. 성문 닫힘 구간 연장에 필요한 성도 파라미터의 정밀도는 합성음의 품질에 영향을 주므로 가능한 안정되고 정밀한 분석 기법이 요구된다. 실험에 의하면 일반적으로 프레임 동기식 분석기법으로도 높은 음질을 유지할 수 있으나 피치가 매우 짧거나 성문의 특성이 불안정한 경우에는 추출된 성도 파라미터의 정밀도가 낮아서 음질이 저하될 수 있다.



[그림 4] 성문 닫힘 구간 연장에 의한 피치 가변, a) 음성 신호, b) 성도 및 음원 특성 분리용 가중함수, c) 분리된 성도 및 음원 특성 신호, d) 성문 닫힘 구간이 연장된 신호, e) 중첩용 가중함수

본 연구에서는 성문 닫힘 구간 개선(Glottal Open Phase Enhancement: GOPE)에 의한 피치 동기식 분석 기법을 사용하였다[2].

2 단계에서는 [그림 4] b)의 가중함수 *Wh(t)*을 이용하여 성문 닫힘 구간에서의 음성신호와 성문 열림 구간에서의 음원을 근사적으로 분리한다. 성문 닫힘 구간, *Wh(t)*의 *Lf*는 해당 피치의 40~50% 정도로 하고 성문 열림 구간인 *Wh(t)*의 *Ls*를 해당 피치의 40~60% 정도로 하면 근사적으로 음원신호를 분리할 수 있다.

$$Wh(t) = 0.5 - 0.5 \times COS(2\pi \frac{1}{2Ls} t), \quad t_n - Ls \le t < t_n$$

$$Wh(t) = 1, \quad t_n \le t < t_n + Lf$$

$$Wh(t) = 0, \quad t_n + Lf \le t < t_{n+1} - Ls \quad \text{---(3)}$$

식 (3)과 같은 가중 함수를 음성신호에 곱하여 구한 신호를 각각의 변경하고자 하는 피치 길이로 이동하여 위치시키면 [그림 4] c)의 SF(t)와 같은 신호를 얻을 수 있다.

3 단계에서는 1단계에서 구한 성도 파라미터를 이용하여 성문 닫힘 구간의 음성신호에 연속해서 원하는 피치 길이까지 선형적인 신호([그림 4] d)의 *Xp(t)*에서 실선)를 연장 혹은 축소한다.

4 단계에서는 식 (4)와 같이 3 단계에서 얻어진 [그림 4] d)의 Xp(t)에 가중함수 Ws(t)를 곱하여 2 단계에서 구한 성문 열림 구간에서의 음원 신호인 *SF(t)*를 중첩하여 더하는 과정으로서 인접 피치간에 신호의 연속성을 유지시켜 [그림 4] f)의 자연스런 합성음 *Y(t)*를 얻을 수 있다.

$$Y(t) = Xp(t) \times Ws(t) + SF(t) \quad \text{--- (4)}$$

여기서 *Ws(t)*는 음원 특성신호를 구할 때 사용된 가중함수와 상호 보완되는 함수이다.

5. 실험 및 결과

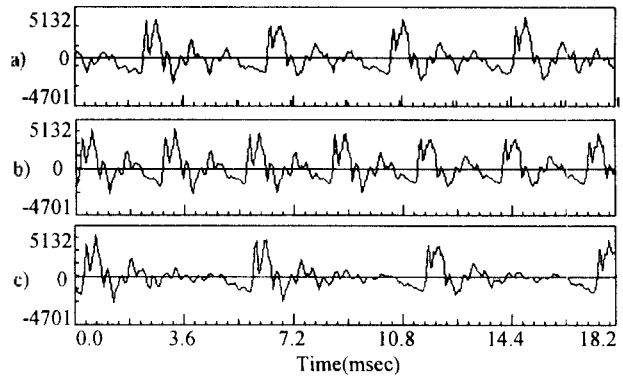
성문 닫힘구간 연장에 의한 피치 변경 기법을 컴퓨터 시뮬레이션하기 위해 16-bit, 10kHz로 표본화된 남자의 음성신호에서 epoch 검출기에 의해 epoch를 검출한 후 수동으로 에러를 수정한 다음 제안된 [그림 4]와 같은 방법으로 처리하였다.

실험에서 *Wh(t)*의 *Ls*는 해당 피치의 40~60%로 하고, *Lf*는 40~50%로 적용하였다. GOPE 기법에 의한 피치 동기식 분석으로 정밀한 성도 파라미터를 추출하고[2] LP 합성에 의해 성문 닫힘 구간의 신호를 연장한 후 음원 신호를 중첩시켜 피치를 변경하였다.

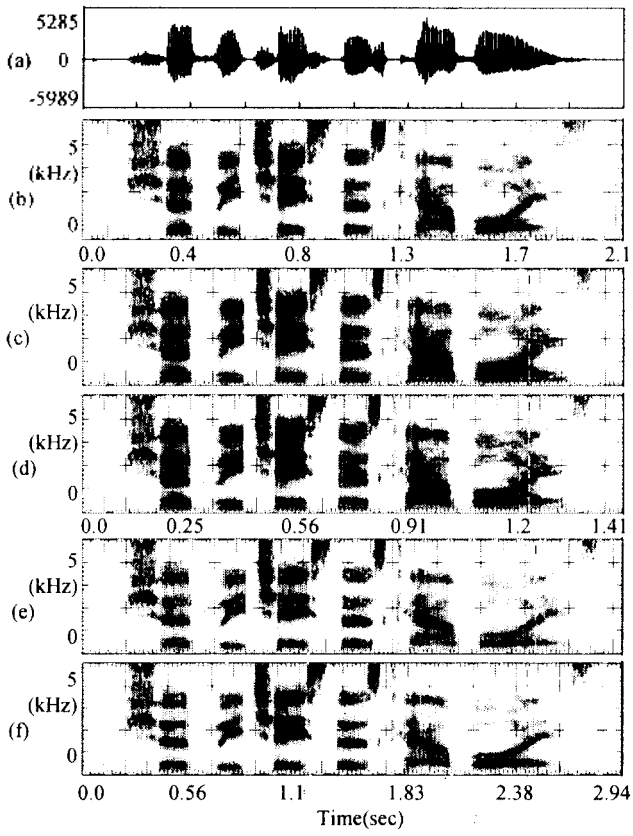
유성음에 대하여 부분적으로 처리된 결과인 [그림 5]에서 볼 수 있듯이 피치가 시작되는 에너지가 큰 구간에서는 원래의 파형을 그대로 유지하고 피치가 연장된 경우에는 자연스럽게 공명과 감쇠가 유지되는 것을 확인할 수 있으며 피치의 시작점 앞에 삽입된 음원 신호도 원래의 모양과 유사한 파형을 보이고 있다.

LPC cepstrum 30차에 의한 원음과의 유사도 측정에서 남성화자의 경우 제안된 GET에 의한 합성음이 PSOLA에 의한 합성음 보다 40%이상 개선되었으나 여성화자의 경우 20% 정도의 개선을 보였다. 피치를 늘리는 경우보다 줄이는 경우에 좀더 많은 개선을 나타내었다.

남녀 10명에 대한 PSOLA에 의한 합성음과의 주관적 비교 평가에서는 대체적으로 PSOLA에 의한 방법에서 부분적으로 나타나고 있는 고음 약화, 저음 강조, 복원 현상 및 buzz 현상이 최소화되어 자연성 및 명료도가 향상된 것으로 평가되었다.



[그림 5] 제안된 방법에 의해 피치가 변경된 음성파형, a) 원래의 음성파형, b) 제안된 방법에 의해 70% 줄인 음성파형, c) 제안된 방법에 의해 140% 늘인 음성파형



[그림 6] 남성화자가 발성한 "Should we chase those cowboys?"에 대한 PSOLA 기법과 제안된 방법과의 처리 결과, a) 원래의 음성파형, b) 원래의 spectrogram, c) PSOLA에 의해 70% 줄인 spectrogram, d) 제안된 방법에 의해 70% 줄인 spectrogram, e) PSOLA에 의해 140% 늘인 spectrogram, f) 제안된 방법에 의해 140% 늘인 spectrogram

청취시험 결과 두 가지 방법에 의한 합성음에 대하여 57.68%가 "비슷하다"고 응답하였으며 PSOLA에 의한 합성음이 좋다고 판정한 경우는 3.6%, 제안된 GET에 의한 합성음이 좋다고 응답한 경우는 38.72%인 것으로 나타났다.

[그림 6]에는 원래 피치의 70%, 140%로 변경한 PSOLA 방법과 제안된 방법에 의한 처리 결과의 spectrogram을 각각 도시하였다. Spectrogram에서 볼 수 있듯이 제안된 GET 방법이 PSOLA 방법보다 원래의 형태에 가까운 spectrogram를 나타내고 있음을 관찰할 수 있다.

6. 결론

음성합성 분야에서 원래의 자연 음성신호를 접속하여 고 품질의 합성음의 얻기 위해서는 자연 음성에 대한 피치의 변경이 필요하다. 고품질을 유지하면서 손쉽게 피치를 가변하기 위해서는 한 피치 구간에서 음성 대부분의 에너지와 성도 특성을 포함하고 있는 성문 닫힘 구간의 신호를 연장하거나 축소할 수 있어야 한다.

본 연구에서는 PSOLA 기법이 창함수의 적용과 두 신호의 중첩으로 발생되는 스펙트럼 왜곡을 극복하기 위해 먼저 유성음 구간에서 성도 파라미터를 추출한 다음 한 피치구간에서 epoch를 참조하여 국사적으로 성문 닫힘 구간을 결정하고 이 구간에 선형적으로 연속되는 신호를 연장하거나 축소하므로서 피치를 가변할 수 있는 방법을 제안하였다.

제안된 방법은 한 피치 구간에서 성문 닫힘 구간을 성도 파라미터에 의해 연상하고 성문 열림 구간에서의 음원 신호에 대해서만 창 함수를 적용하므로서 창 함수에 의한 영향과 신호의 중첩에 의한 영향이 최소화되어 보다 명료한 합성음을 얻을 수 있었다. 제안된 방법에 의한 합성음의 품질은 정확한 epoch, 정밀한 성도 특성 분석 방법에 따라 영향을 받으므로 이들에 대한 정밀한 정보가 필요하다.

감사의 글

본 연구는 정보통신부 출연 "HCI를 위한 음성 입출력 처리기술 개발"과제의 연구결과 입니다.

참고 문헌

[1] Minsoo Hahn, Dong-Gyu Kang, "Precise Glottal Closure Instant Detector for Voiced Speech," IEE Electronics Letters, Vol. 32, No. 23, pp. 2117~2118, November 1996.

[2] 강동규, 한민수, "유성음 구간에서 피치 동기식 포먼트 추출", 한국음향학회 학술발표대회 논문집 제 15권 제 1(s)호, 1996

[3] Wolfgang J. Hess, *Pitch Determination of Speech Signals*, pp. 52, (zero-crossing and Excursion Cycles), Springer-Verlag, 1983

[4] A. K. Krishnamurthy, D. G Childers, "Two-Channel Speech Analysis," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-34, no. 4, pp. 730~743, August 1986.

[5] D. E. Veeneman, S. L. Bement, "Automatic Glottal Inverse Filtering from Speech and Electroglottographic Signals," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-33, no. 2, pp. 369~377, April 1985.