음성 합성을 위한 음성 파라미터 분석법의 개선에 관한 연구

방호균°, 조철우
창원대학교 제어계측공학과 신호 및 시스템 연구실

A Study on Improvements of Speech Analysis Methods for Speech Synthesis

Hogyun Bang, Cheol-Woo Jo
Acoustic and Speech Group, Dept.of Control and Instrumentation Eng.
ChangWon National University

요 약

본 논문에서는 포만트 합성에 필요한 음성 파라미터를 분석하는 방법의 개선에 관하여 논한다. 내용은 주로 피치 동기 분석을 위한 피치 위치 추정법의 개선과 포만트 분석서 발생하는 스펙트럼의 왜곡 현상을 기존의 포만트 분석법 및 선형예측분방법과 비교한다.

1. 서 론

현재 음성 합성을 위한 많은 방법들이 제안되고, 실용화 되고 있지만 아직은 인간의 음성에 비해 명료도와 자연성이 떨어진다. 이러한 특성은 음성의 비선형성을 선형적 모델로 근사화 과정에서 야기된 것이다. 근사화에 따른 오차는 시간축 해석 방식 보다 계수에 의한 합성법에서 두드러지게 나타난다[1]. 시간 파형에서의 합성법은 우선적으로 자연성을 쉽게 얻을 수 있다는 장점 때문에 선호되고 있으나[2], 기억용량이 많고, 대량의 데이타베이스가 필요한 관계로 다양한 음질을 합성할 수 있는 합성기에는 부족한 점이 있다. 현재 계수 합성법으로 가장 널리 알려진 포만트 합성기의 경우 포만트 주파수와 피치의 가변이 가능하고 음원 제어의 편의성을 가지고 있다는 장점이 있으나, 유성음과 유사 유성음 이외의 음성에서는 적지 않은 문제점을 내포하고 있다. 이러한 문제점을 해결하기 위한 방안으로 시간 영역과 주파수 영역에 혼합 방식과 포만트 주파수를 추정하는 과정에서 발생한 오차 보상 등에 혼합 합성기 개발에 관한 연구가 보도되고 있다[1][3].

본 연구에서는 고차 선형 예측법을 이용한 포만트 추정과 포만트 추정시 발생하는 오차를 보상할 수 있는 방법에 대해 제안한다. 포만트 분석은 피치 동기식에 의한 접근을 한다. 피치 검출은 필자들이 제안한 개선된 SGCID(Sequential Glottal Closure Instant Detect)법을 이용한다[3].

2. 피치위치 추출법의 개선

우선 피치동기방식의 분석을 행하기 위해 필자들이 지난해 제안했던 SGCID(Sequential Glottal Closure Instant Detect)법을 보완하였다.

SGCID법은 Hilbert 변환을 이용한 EFLPR(Epoch Filtering of Linear Prediction Residual)법을 개선한 것으로 시간 및 주파수 영역에서 피치를 검출하는 기법이다[4] [5] [6].

SGCID는 미리 검출된 기준 피치에 의해 유성음의 피치를 순차적으로 찾아가는 기법이다. 이러한 특성으로 인해 불특정 음성에 대하여 초기 피치를 정확히 추정하지 못하는 결점을 가지고 있다. 초기 피치들의 부정확성을 보상하기 위하여 후향 탐색(Backward Scan)을 시도했다.

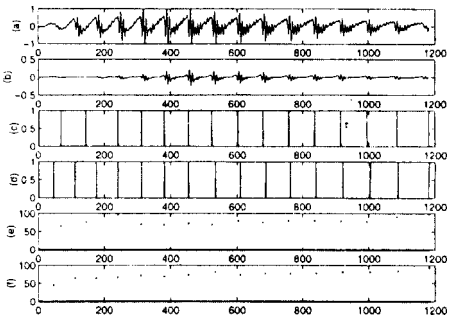

그림 1. SGCID법의 단모음 / ㅏ/ 적용결과
(a) 남성음 /파/의 / ㅏ/, (b) 잔차신호, (c) SGCID에 의한 피치, (d) 개선된 SGCID에 의한 피치, (e) SGCID에 의한 피치 궤적, (f) 개선된 SGCID에 의한 피치 궤적

전향 탐색(Forward Scan)법은 기존의 SGCID법과 동일하다. 다만, 잔차 신호 추정을 Modified Covariance Method로 하고, 12차의 선형예측 차수를 8차로 하여 계산량의 감소를 시도했다.

그림 1 (a)는 한국음 /파/에서 영교율과 에너지값에 의해 유성음 / ㅏ/를 검출했다. 화자는 20대 중반의 한국인 남성이며, SUN SPARC 10 workstation에서 8 kHz로 표본화했다.

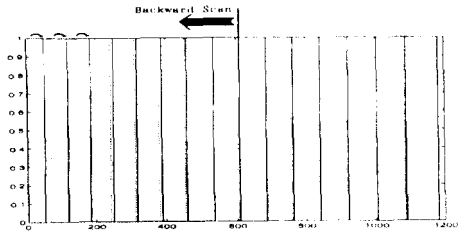

그림 2. 단음절 /파/의 / ㅏ/부분의 SGCID와 개선된 SGCID에 의한 피치

후향 탐색은 전향 탐색에서 설정된 피치 정보를 기준으로 행해진다. 후향 탐색의 시작점은 유성음 구간의 중간 부분으로 설정했다. 사용된 음성의 경우 시작점과 끝점을 포함하여 총 16 개의 피치가 존재고, 후향 탐색을 위한 시작점은 8 번째 피치가 된다. 그림 2는 그림 1의 (c)와 (d)를 중복시켜 나타낸 것이다. 점선은 SGCID, 실선은 개선된 SGCID에 의한 결과이다. 가로축 200점 이후 나타난 지연 현상은 방사 효과를 제거하고, Modified Covariance Method에 의한 선형예측을 사용한데 따른 결과이다. 성문파 추정을 위한 실험 결과 개선된 SGCID에 의한 피치 검출법이 더욱 유효한 것으로 생각된다.

개선된 피치 검출법은 그림 1과 2에서 보여주는 것과 같이 초기 1~3개의 부정확 피치에 대한 보정효과를 얻을 수 있다. 연속음 적용시, 영교율과 에너지 값에 의한 유성음 검출로 변이음에 대한 처리가 부족한 관계로 변이 부분에서 약간의 오차를 가지고 있다.

3. 음성 계수 분석법

일반적으로 음성 계수의 추정은 LPCP에 의한 방법이 주류를 이루고 있다. 이러한 계수 추정법의 경우 계수화에 따른 오차로 자연음에 충실도를 유지하고 힘들다. 특히 포만트 계수 추정의 의한 경우 오차율의 누적으로 오차에 대한 보정법이 요구된다.

본 논문에서는 보편적으로 사용되어 온 계수화 분석법를 고찰하고, 제안된 혼합 합성기에 의한 결과와 비교한다.

3-1. LPC Analysis

선형예측분석법은 음성 코딩과 계수 분석을 하기 위해 사용되는 가장 일반적인 방법 중 하나이다[1][7]. 음성 S(z)를 p차 LPC로 표현하면,

$$S(z) = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-1}} = \frac{1}{A(z)} \tag{1}$$

여기서 a_i는 음성 S(z)의 i번째 polynomial 계수다. 그리고 A(z)는 S(z)의 Polynomial로,

$$A(z) = A_s(z) A_u(z) = \prod_{i=1}^{q/2} (1 - b_i z^{-1})(1 - b_i^* z^{-1}) \prod_{i=q+1}^{p} (1 - c_i z^{-1}) \tag{2}$$

여기서 $A_u(z)$는 (p-q)차가 되고, 허수 짝을 이루지 못한 불안정근과 실수근으로 이루어 진다.

LPC분석에 의한 S(z)의 Polynomial은 B(z)가 된고, 최대 q/2개의 포만트를 가진다. LPC분석에서는 불안정한 근의 소거에 의한 차수의 감소로 충분히 높은 차수로 분석할 때 충실도를 갖는 분석이 가능하다.

3-2. Formant Analysis

성도의 공진 특성을 모델링 한 것으로, 공진주파수, 공신주파수의 크기와 대역폭의 추정이 필요하다. 필요한 계수의 추정을 위해서 LPC 분석이 선행되어야한다.

포만트 주파수 F는 (2)식의 안정화된 근 B(z)에서 실근과 허근을 분리하여,

$$F_i = \frac{1}{2\pi T} \cdot \tan^{-1}\left[\frac{\mathrm{Im}(z_i)}{\mathrm{Re}(z_i)}\right] \tag{3}$$

$$B_i = -\frac{1}{2\pi T} \log\left\{\mathrm{Re}(z_i)^2 + \mathrm{Im}(z_i)^2\right\} \tag{4}$$

구한다. 여기서 T는 표본화 시간을 나타낸다. i는 1, . . . , q로 주어진다. 구해진 포만트 주파수와 대역폭에 개수를 제한하여 4~5개를 선택하여 포만트 주파수로 설정한다.

이러한 방식은 근사화에 의한 오차가 누적되어, 음성의 충실도를 잃을 수 있다. 현재 사용되는 ABS(Analysis-by-Synthesis)법은 충실도는 가지나, 엄청난 연산량과 불필요하게 많은

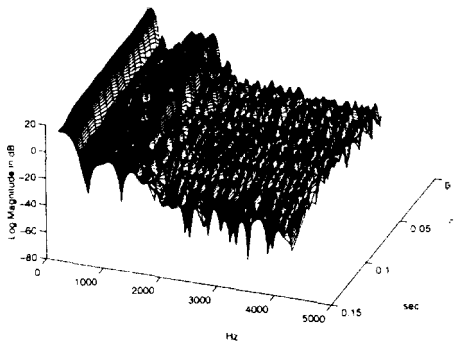

그림 3. (a) FFT Contour

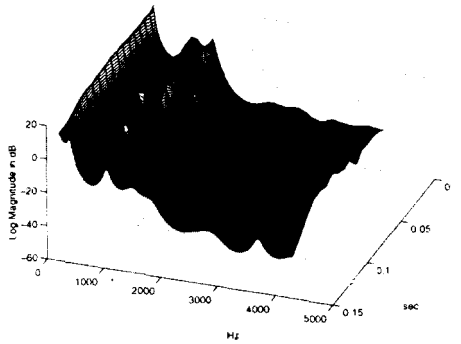

그림 3 (b) LPC Contour

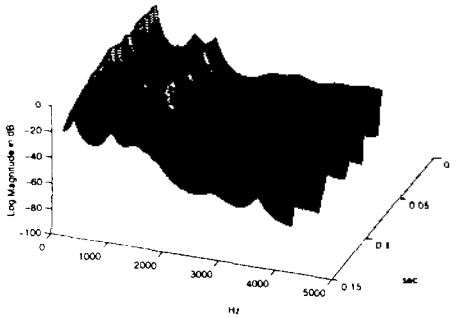

그림 3. (c) Formant Coutour

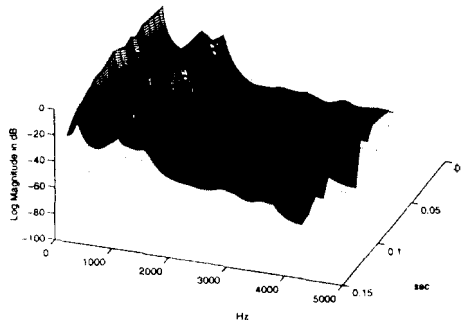

그림 3. (d) Hybrid Formant Contour

반복 검증에 대한 문제점이 있다.

많은 문제점에도 불구하고, 포만트 합성기는 여러가지 음원에 대한 적용 가능성 때문에 계수 분석법과 합성기의 개선에 관한 연구가 계속되고 있다.

3-3. Hybrid Formant Analysis

혼합형 합성기는 포만트 합성기의 음원 제어 및 변경에 대한 잇점을 살리고, 근사화 과정에서 발생한 오차를 보상하기 위한 보상기 설계에 목적이 있다. 더불어 유성음과 유사 유성음을 제외한 음성에서 취약한 구조를 보완하기 위한데 부차적인 목적을 가지고 있다.

포만트 합성기에서 발생되는 오차는 근사화 과정에 의한 것으로, (2)식의 $A_u(z)$와 (3)식과 (4)식에서 부적합한 포만트의 제외에 따른 결과이다.

$$A_s(z) = A_{s1}(z)A_{s2}(z) \tag{5. a}$$

$$A_u(z) = A_{u1}(z)A_{u2}(z) \tag{5. b}$$

$$A_{s1}(z) = \prod_{i=1}^{r}\left(1-a_i z^{-1}\right)\left(1-a_i^* z^{-1}\right),$$
$$A_{s2}(z) = \prod_{i=r+1}^{q/2}\left(1-a_i z^{-1}\right)\left(1-a_i^* z^{-1}\right) \tag{5. c}$$

식 (5. a)에서 $A_{s1}(z)$는 포만트로 설정된 근들이다. $A_{s2}(z)$는 대역폭과 제한된 공진 대역 조건에 의해 부적합한 포만트로 설정된 근들이다. r은 포먼트의 갯수로 고정형 합성기인 경우 3~4정도의 값이 적당하다.

$$C(z) = A_{s1}(z)S(z) = \frac{1}{A_{s2}(z)A_{u1}(z)} \tag{6}$$

C(z)는 근사화 방법에서 발생하는 오차로 보상기를 통해 고려되어야 할 부분이다. 보상기에서 $A_{u1}(z)$ 실수로 이루어진 근만을 고려한다. 그리고 $A_{u2}(z)$는 허수쌍을 이루지 못하는 불안정한 근이다. 실수근은 공진기 구성에 직접적인 참여를 하지 않는 관계로 공진기 설계시 불필요한 근으로 간주되어왔다. 하지만 포

만트 합성기에서 발생하는 포락선의 감쇄 현상을 보안하기 위한 중요한 요소가 될 수 있다.

4. 실험 결과

그림 3에서 보여준 음성은 SGCID에서 적용한 한국음 /파/중에서 / ㅏ/를 분석한 것이다. 분석 방법은 피치 동기식을 주로하여 시행되었다. 피치 검출 방식은 필자들이 제안한 SGCID법을 이용하였다. 그림 3 (a)에서 고속퓨리에 변환은 1024점으로 행해졌다. 피치는 85~105정도로 퓨리에 해석시 zero-padding 효과를 이용하여 해상도를 높였다. 그림 3 (b)는 선형예측 분석법에 의한 3차원 Contour를 보여준것이다. 분석 차수는 16 차로 전체적인 Contour가 매우 매끄러운 것을 확인할 수 있다. 200 Hz 미만에 존재하는 최고점은 피치 성분이 반영된 것으로 포만트는 아니다. 그림 3 (c)는 16차 LPC분석된 계수에서 공진주파수와 대역폭을 구하여 병렬공진필터를 통과 시킨 결과이다. 첫번째와 두번째 포먼트 주파수의 추정은 가능하나 대역폭과 크기의 부정확한 추정으로 불연속적인 영역을 보인다. 그림 3 (d)는 혼합형 포만트 필터를 통과 시킨것으로 첫번째와 두번째 포만트의 궤적이 비교적 선형적임을 확인할 수 있다.

5. 결 론

본 논문에서는 피치동기 음성분석을 위한 피치 추정방식의 개선점을 먼저 논한 뒤, 계수 합성 방식 문제점을 고찰하고, 문제점 해결을 위한 혼합형 합성 방식을 제안하고 실험하였다. 혼합형 합성기는 포만트 합성기와 LPC합성기를 조합한 것으로, 한국어 단모음에의 적용 결과 포만트 합성방식에 비해 궤적의 추정이 우수한 등의 특성을 보이고 있다. 앞으로 이러한 방식을 규칙합성에 적용하고 분석방법을 개선하기 위한 연구가 계속될 예정이다.

6. 참고 문헌

[1] Georg Fries, "Hybrid Time and Frequency-Domain Speech Synthesis with Extended Glottal Source Generation", IEEE Proc. ICASSP, no. 1, pp581-584, 1995.

[2] Sang-Hun Kim, Minjo Zhi, Un-Cheon Choi, "Application of TD-PSOLA to Korean Text-to-Speech Conversion", Proc. SCAS, vol. 10, no. 1, pp. 291-294, 1993.

[3] Joseph P. OLive, "Mixed spectral representation-Formant and linear predictive coding(LPC)", JASA., vol. 92., pp1837-1840, Oct. 1992.

[4] 방호균, 조철우, "연속선형예측을 이용한 성문폐쇄시점 검출법의 개선에 관한 연구", Proc. SCAS, vol. 7, no. 1 pp. 762-765, 1994.

[5] T. V. Ananthapadmanbha and B. Yengnanarayana, "Epoch extraction of voiced speech," IEE Trans. Acoust., Speech, Signal Processing, vol. 23, no. 6, pp. 562-570, Oct. 1975.

[6] T. V. Ananthapadmanbha and B. Yengnanarayana, "Epoch extraction from linear prediction rdsidual for identification and close glottis interval," IEEE Trans. Acoust., Speech, Signal Processing, vol. 27, no. 4, pp. 309-319, Aug. 1979.

[7] M. Sandler, "Algorithm for high precision root finding from high order LPC models," IEE Proceedings-I, vol. 138, no. 6, Dec. 1991