말소리와 음성과학 제4권 제4호 (2012. 12. 31)
ISSN 2005-8063, pp. 79~86

# TBE 모델을 사용하는 HMM 기반 음성합성기 성능 향상을 위한 하모닉 선택에 기반한 MVF 예측 방법

## Harmonic Peak Picking-based MVF Estimation for Improvement of HMM-based Speech Synthesis System Using TBE Model

박 지 훈[1] · 한 민 수[2]
Park, Jihoon · Hahn, Minsoo

**ABSTRACT**

In the two-band excitation (TBE) model, maximum voiced frequency (MVF) is the most important feature of the excitation parameter because the synthetic speech quality depends on MVF. Thus, this paper proposes an enhanced MVF estimation scheme based on the peak picking method. In the proposed scheme, the local peak and the peak lobe are picked from the spectrum of a linear predictive residual signal. The normalized distance between neighboring peak lobes is calculated and utilized as a feature to estimate MVF. Experimental results of both objective and subjective tests show that the proposed scheme improves synthetic speech quality compared with that of the conventional one.

**Keywords: speakers with cleft palate, hypernasality, continuous positive airway pressure, efficacy**

## 1. 서론

음성은 사람이 다른 도구 없이 정보를 전달하는 매체로서 가장 많이 이용되고 있으며 가장 간편한 전달 방법이다. 그리고 사람이 기계 및 장치와 의사소통을 가능하게 하는 중요한 기술이 음성인식과 음성합성 기술이다. 음성인식기를 기계의 귀라 표현하면, 음성합성기는 기계의 입이라 할 수 있다. 음성합성기는 기계가 사용자에게 정보(information)을 텍스트나 그림이 아닌 음성 신호로 전달함으로서 운전 중 이거나, 맹인인 경우처럼 사용자가 작동하는 기계의 화면을 볼 수 없는 경우 음성합성 기술은 매우 유용하다. 근래에 들어, 스마트폰, 전자책 리더, 차량 네비게이션 등 개인 휴대용 장치의 개발과 보급이 활발하게 이루어짐으로써 사용자의 수도 급속도로 증가하였다. 따라서 개인 휴대용 장치에서 사용 가능한 임베디드 (embedded) 음성합성기가 요구되었다. 임베디드 음성합성기는 문자, 전자우편, 전자신문 등 무작위로 시변하는 글을 읽어주는 것과 같이 다양한 개인 휴대용 장치에서 사용 가능하다.

음성합성 기술은 합성 방식에 따라서 음편접합 기반 음성합성 방식과 HMM(Hidden Markov Model) 기반 음성합성 방식 크게 두 가지로 구분할 수 있다. 음편접합 기반 음성합성 방식은 음성합성 방식 중 가장 널리 사용되고 개발되어온 음성합성 방식으로, 음편 별로 파형 자체를 보유한 대용량의 코퍼스 데이터로부터 적당한 음편을 선택하여 음성을 합성하는 방식이다 [1]. 음편접합 음성합성 방식은 합성음의 음질이 우수하다는 장점이 있지만, 합성기의 용량이 크고, 전력소모가 크다는 단점이 있다. HMM 기반 음성합성 방식은 Tokuda에 의해 개발된 이후, 많은 연구와 개발이 진행되어 왔다[2-6]. HMM 기반 음성합성 방식은 그 방식 때문에 통계적 모델링을 이용한 파라미터릭 음성합성 방식이라고도 불린다. 이 방식은 이름에서도 알 수 있듯이 먼저 음성신호의 음편을 특징 파리미터로 변환하고 그 특징 파라미터들을 통계적으로 모델링하고 이를 보코더(vocoder)에 통과시켜 합성음을 생성하는 방식이다. HMM 기반 음성합성기는 합성기의 용량과 전력소모가 적고 합성기의 용량에 비해 만족스러운 음질을 제공하는 것으

1) 한국과학기술원, bato2n@kaist.ac.kr
2) 한국과학기술원, mshahn@ee.kaist.ac.kr

접수일자: 2012년 7월 26일
수정일자: 2012년 12월 4일
게재확정: 2012년 12월 12일

로 알려져 있다 [2,3]. 개인 휴대용 장치는 메모리 용량과 장치의 전력 소모에 제한이 있기 때문에, 음편접합 기반 음성합성 방식은 임베디드 음성합성 방식으로 적합하지 않다. 반면에 HMM 기반 음성 합성방식은 개인 휴대용 장치에서 임베디드 음성합성 방식으로 사용하기에 적합하다.

HMM 기반 음성합성기는 <그림 1>과 같이 훈련 과정과 합성 과정으로 구성되어 있다. 훈련 과정에서는 음성 신호를 스펙트럼 (spectrum), 여기 신호, 지속시간 정보를 문맥 정보에 기반하여 각각 독립된 가우시안(Gaussian) 확률분포를 가지는 HMM 모델로 훈련한다. 합성 과정에서는 입력 문장을 사용하여 훈련된 HMM 모델로부터 합성할 음성 파라미터들이 결정되면 MLSA(mel-log spectrum approximation) 필터링을 통해 최종 합성음이 생성된다 [5].

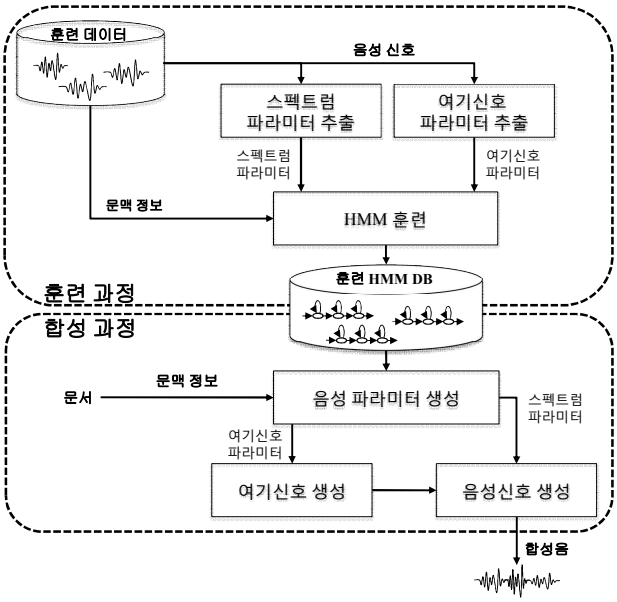


그림 1. HMM 기반 음성합성기 구조도
Figure 1. Structure of HMM-based speech synthesis

음편접합 기반 음성 합성기에 비해 초기 HMM 기반 음성 합성기는 합성음의 음질측면에서 만족할만한 합성음을 제공하지 못했다. 만족할만한 음질의 합성음을 얻기 위해 다양한 연구가 진행 되었고 많은 연구 결과에서 여기신호 파라미터 즉 여기신호 모델이 HMM 기반 음성기의 음질에 영향을 주는 것으로 알려져 왔다 [7-12]. 초기 HMM 기반 음성합성기의 여기신호 모델은 CE(conventional excitation) 모델을 사용하였다. CE 모델은 음성생성모델에 기반한 여기신호 모델로써 무성음 구간은 랜덤(random) 잡음을 사용하고, 유성음 구간에서는 피치(pitch) 길이를 가지는 주기적인 펄스 트레인 (pulse train)을 사용하였다. 이러한 CE 모델을 사용한 합성음은 버지(buzzy)한 소리를 들려줌으로써 합성음의 자연성과 음질의 저하를 가져왔다. 버지 효과를 줄이기 위해 Yoshimura는 MELP(multi excitation linear predictive) 보코더의 여기신호 모델을 사용하여 ME (mixed excitation) 모델을 제안하였다. ME 모델에서는 여기신호를 주파수 영역에서 다섯 개의 밴드로 나누고 각 밴드의 주기성에 따라 주기적인 밴드는 피치 간격의 주기적인 펄스 트레인을, 비주기적인 밴드는 랜덤 잡음을 사용하여 여기신호를 생성한다. ME 모델이 버지한 소리를 줄이더라도 MELP 보코더가 8 kHz 협대역 음성 신호에서 사용된 보코더이기 때문에 밴드의 대역폭이 16 kHz 광대역 음성 신호에는 최적화 되어 있지 않다. 또한 광대역 신호에서의 ME 모델을 위해 밴드의 수를 늘리게 되면, ME 모델을 사용하는 HMM 기반 음성합성기의 용량과 전력소모 또한 늘어나게 된다. 이러한 문제를 해결하고자 Kim은 TBE(two-band excitation) 모델을 사용하는 HMM 기반 음성합성기를 제안하였다 [8-10]. TBE 모델은 유성음을 주파수 영역에서 MVF(maximum voiced frequency)라는 파라미터에 의해 주기적인 저주파 부분과 비주기적인 고주파 부분으로 나눈다. TBE 모델을 사용한 여기신호 생성은 ME 모델과 같이 주기적인 저주파 대역은 피치 길이를 가지는 펄스 트레인을 사용하고, 비주기적인 고주파 대역은 랜덤 잡음을 사용하여 여기신호를 생성한다. 다시 말해 MVF는 TBE 모델에서 가장 중요한 파라미터로서 MVF의 정확도가 TBE 모델의 성능뿐만 아니라 전체 HMM 기반 음성합성기의 성능까지 영향을 미친다.

정확한 MVF를 계산하기 위해 Kim은 고주파 통과 필터 기반의 MVF 예측 방법을 사용하였다. 이 방법은 고주파 통과 필터링을 거친 음성 신호와 그 신호의 피치 주기 지연을 갖는 신호의 정규화 자기 상관도를 사용한다 [8]. 고주파 통과 필터는 스펙트럼 포락선 정보가 포함된 음성 신호를 사용하여 MVF를 예측하기 때문에 MVF를 정확하게 예측할 수 없다. HMM 기반 음성 합성기의 훈련은 오프라인(off line)으로 이루어지기 때문에 Han은 보다 정확한 MVF 예측을 위해 고주파 통과 필터 기반의 MVF 예측 방법의 후처리 개념으로 ABS(analysis-by-synthesis) 기반의 MVF 최적화 방법을 제안하였다 [11]. MVF 최적화 방법은 초기 MVF 값이 예측되면 예측된 MVF의 주위 값들을 사용하여 합성음을 생성하여 훈련 과정에서의 원 음성과의 왜곡 측정을 통해 그 왜곡이 적은 MVF를 찾는 방법이다. 하지만 MVF 최적화 방법은 MVF 초기값이 본래 MVF 값과 차이가 많으면 정확한 MVF를 예측할 수 없다는 문제점이 있다. 그래서 본 논문에서는 초기 MVF의 정확도를 향상시키기 위해서 주파수 영역에서의 스펙트럼 하모닉(harmonic) 선택 방법에 기반한 MVF 예측 방법을 제안한다. 제안하는 방법에서는 입력신호의 선형 예측 잔차 신호를 사용하여 MVF를 예측함으로써 기존 고주파 통과 필터 기반의 MVF 예측 방법의 문제점을 해결하였다.

본 논문의 구성은 다음과 같다. 2장에서는 기존 MVF 예측 방법인 고주파 통과 필터 기반의 MVF 예측 방법과 MVF 최

적화 방법에 대해서 설명한다. 그리고 3장에서는 본 논문에서 제안하는 MVF 예측 방법에 대해 설명한다. 4장에서는 실험과 결과로 제안하는 MVF 예측 방법의 성능을 확인하고 5장에서 결론을 맺는다.

## 2. 기존의 MVF 예측 방법과 문제점

### 2.1 고주파 통과 필터 기반의 MVF 예측 방법

음성 생성 모델 이론에 따라 유성음성을 주파수상에서 분석하면, 유성음성의 여기신호가 펄스 트레인만으로 이루어져 있지 않다. 이러한 분석에서 접근한 방법이 TBE 모델이다. TBE 모델은 유성음성의 스펙트럼이 주기적인 하모닉 성분이 있는 저주파 대역과 비주기적인 성분이 있는 고주파 대역으로 나누는 모델이다. TBE 모델에서 MVF는 주기-비주기 대역을 구분하는 경계선으로 정의할 수 있다. <그림 2>에 나타나듯이, 유성음성의 스펙트럼은 MVF에 의해 주기적인 저주파 대역과 비주기적인 고주파 대역으로 나눌 수 있다. 그래서 TBE 모델을 사용할 때, 주기적인 부분과 비주기적인 부분으로 나누는 MVF가 가장 중요한 파라미터이고 또한 MVF를 정확하게 예측하는 것이 중요하다.

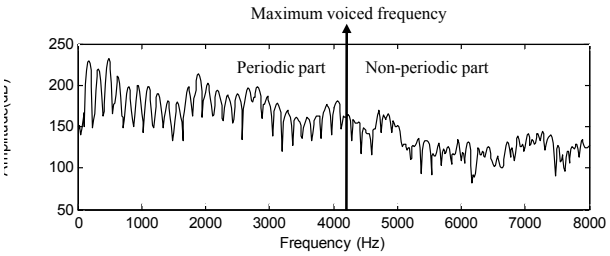


그림 2. TBE 모델과 MVF의 예
Figure 2. Example of TBE model and MVF

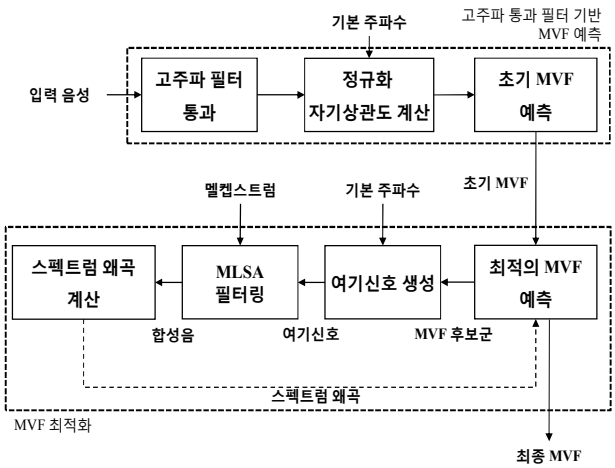


그림 3. 고주파 통과 필터 기반의 MVF 최적화 구조도
Figure 3. High-pass filter based MVF optimization method

MVF 예측을 위해 Han은 기존의 고주파 통과 필터 기반의 MVF 예측 방법의 후처리 개념으로 ABS(analysis-by-synthesis) 기반의 MVF 최적화 방법을 제안하였다 [11]. 고주파 통과 필터 기반의 MVF 최적화 방법의 전체적인 과정은 다음 <그림 3>과 같이 나타난다.

이 MVF 예측 방법은 초기 MVF를 예측하는 과정과, MVF 최적화를 수행하는 과정으로 나누어져 있다. 초기 MVF 예측 과정으로써, 고주파 통과 필터 기반의 MVF 예측 방법이 사용된다 [8]. 이 초기 MVF 예측 방법은 고주파 통과 필터를 통과한 신호의 피치 주기 지연을 갖는 신호의 정규화 자기 상관도를 사용한다. $f$ 의 차단주파수를 갖는 고주파 통과 필터를 $h_{HPF}^{f}(n)$ 라 하면, 이 필터를 통과한 음성 신호 $s_{HB}^{f}(n)$ 은 다음 수식 (1)과 같이 정의된다.

$$s_{HB}^{f}(n) = s(n) * h_{HPF}^{f} \tag{1}$$

여기서 $s(n)$ 은 입력 음성 신호이다. 수식 (1)에서 구한 고주파 대역 음성 신호를 사용하여 피치 주기 $\tau$ 만큼 지연을 갖는 신호와의 정규화 자기 상관도 $R_f(\tau)$ 를 다음 수식 (2)와 같이 계산한다.

$$R_f(\tau) = \frac{\sum_{n=0}^{N-1} s_{HB}^{f}(n)\, s_{HB}^{f}(n+\tau)}{\sqrt{\sum_{n=0}^{N-1}\left[s_{HB}^{f}(n)\right]^2 \sum_{n=0}^{N-1}\left[s_{HB}^{f}(n+\tau)\right]^2}} \tag{2}$$

여기서 $N$ 은 분석 프레임의 사이즈이다. 이 정규화 자기 상관도 값을 차단 주파수에 따라 계산해보면, 이 값은 -1에서 1 안에 존재한다. 만약 차단 주파수가 MVF보다 작다면, 고주파 통과 음성 신호에 주기적인 성분이 존재할 것이고 정규화 자기 상관도 값은 1에 가까울 것이다. 반대로 차단 주파수가 MVF보다 크다면, 고주파 통과 음성 신호에 주기적인 성분이 존재하지 않을 것이고 정규화 자기 상관도 값은 0에 가까울 것이다. 다음 과정인 MVF 최적화 과정에서는 이전 과정에서 나온 초기 MVF를 사용해서 최종 MVF를 예측하게 된다. 첫 번째로 초기 MVF와 피치 주기의 역수인 기본 주파수를 사용하여 여기 신호를 생성한다. 다음으로 생성된 여기 신호와 미리 계산된 멜켑스트럼(Mel-cepstrum)를 MLSA(mel-log spectrum approximation) 필터에 통과시켜 합성음을 생성한다. 합성음은 MVF의 위치를 이동해가며 여러 번 생성하게 되며 원음과의 객관적 음질 평가를 통해 음질 열화가 가장 적은 합성음의 MVF를 최적화된 MVF로 예측하게 된다. MVF 최적화는 모든

MVF에 대하여 객관적 음질 평가를 수행하지 않고 초기 MVF 예측 값의 근처 MVF 값들만을 가지고 음질 평가를 하여 MVF 최적화를 수행하기 때문에, 초기 MVF 예측이 중요하다.

### 2.2 기존 MVF 예측 방법의 문제점

기존 고주파 통과 필터 기반의 MVF 예측 방법은 크게 두 가지 문제점이 있다. 첫 번째 문제점은 음성신호의 스펙트럼 포락선 즉 포만트 정보이다. 여기신호는 포만트에 의해 음성신호로 변환되는데, 그 포만트 정보에 의해 스펙트럼의 주기성 또한 변환되어 정확한 MVF의 위치를 찾기가 어렵다. 하지만 포만트 정보가 제거된 여기 신호를 사용하여 고주파 통과 필터를 통과시키고 피치 주기만큼 지연된 정규화 자기 상관도를 구해보면 그 값이 너무 작아 파라미터로서 사용할 수 없다. 두 번째 문제점은 고주파 통과 필터의 개수 와 MVF 해상도가 트레이드 오프(trade-off) 관계에 있다는 것이다. MVF의 해상도는 고주파 통과 필터의 차단주파수와 관계가 있는데, MVF의 해상도를 위해 너무 많은 고주파 통과필터를 수행하게 되면 시스템의 계산 량이 너무 많아지게 된다. 반면에 너무 적은 고주파 통과 필터를 사용하게 되면 MVF의 해상도는 떨어지게 되어 잘못된 MVF를 예측하게 된다. 이처럼 MVF가 잘못 예측되어 너무 높게 MVF가 예측되면 합성음이 기존의 CE 모델로 만들어진 합성음과 같이 버지한 소리가 난다. 반면에 MVF가 너무 낮게 잘못 예측되면 허스키한 합성음을 제공하여 전체적인 합성음의 음질에 문제가 생기게 된다. 이처럼 TBE 모델에서는 정확한 MVF를 예측하는 것이 가장 중요하다. 그래서 기존 방법의 문제점을 해결하고 좀더 정확한 MVF를 예측 하기 위한 MVF 예측 방법이 제안되어야 한다.

## 3. 제안하는 MVF 예측 방법

본 논문에서는 주파수 도메인에서 여기 신호의 주기적인 하모닉 성분을 검출하는 방법을 제안한다. 제안하는 방법은 입력 신호의 여기 신호를 사용함으로써 스펙트럼 포락선을 제거하여 보다 정확한 MVF를 예측 가능하며, 추가적으로 주파수 도메인에서 MVF의 위치를 찾기 때문에 디지털 푸리에 변환의 크기만큼 그 해상도가 증가하게 된다. 제안하는 방법의 전체적인 과정은 <그림 4>와 같이 나타난다. 제안하는 방법에서도 2장에서 설명한 MVF 최적화 과정을 후처리 과정으로서 사용한다.

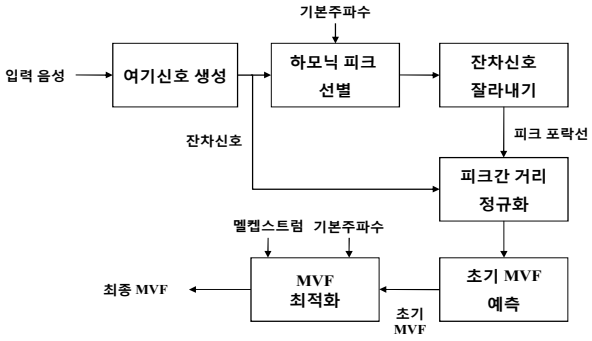


그림 4. 제안하는 MVF예측 방법의 구조도
Figure 4. Proposed MVF estimation method

### 3.1 여기신호 생성

제안하는 방법의 첫 번째 과정으로 입력 음성으로부터 여기신호를 생성한다. 여기신호는 선형 분석법을 통해 생성된 잔차신호를 사용한다. 선형 분석법을 통한 잔차신호는 일반적으로 신호의 여기신호를 나타내는 것으로 널리 알려져 있다. 먼저 선형 분석법을 통해 선형 예측 계수를 계산한다. 선형 예측 계수는 레빈슨-더빈 (Levinson-Durbin) 알고리즘을 사용하여 계산한다. 다음으로 선형 분석법에 의한 잔차신호(LP 잔차신호)는 원 신호와 선형 예측 계수로 예측된 신호와의 차이로 생성되는데 그 식은 다음 수식 (3)과 같이 정의된다.

$$r(n) = x(n) - \sum_{i=1}^{p} a(i)\,x(n-i) \tag{3}$$

여기서 $r(n)$와 $x(n)$ 각각 잔차신호와 입력 음성신호를 나타낸다. 그리고 $a(i)$는 선형 예측 계수이고 $p$는 선형 예측 차수이다. 다음 <그림 5> 는 입력 음성과 선형 예측 분석을 통한 잔차신호의 스펙트럼 그림이다. <그림 5-b>에 보여지듯이 스펙트럼 포락선이 제거된 잔차신호에서 주기적인 부분과 비주기적인 부분의 경계를 보다 정확하게 확인 가능하다.

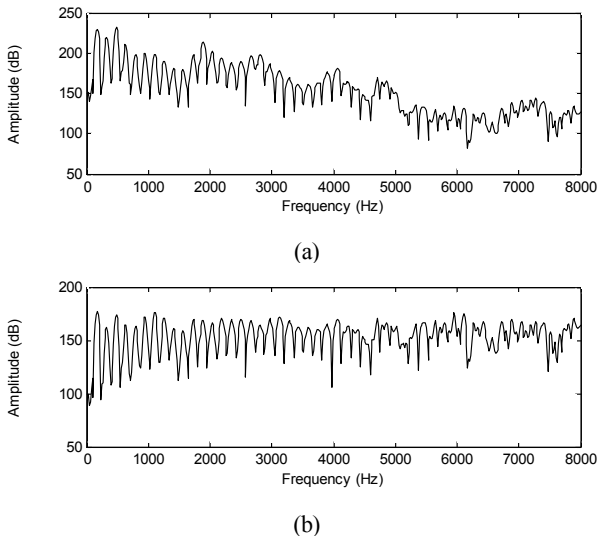


그림 5. (a): 입력 음성의 스펙트럼,
(b): LP 잔차신호의 스펙트럼
Figure 5. (a): Spectrum of input speech signal
(b): Spectrum of LP residual signal

### 3.2 하모닉 피크 선별

하모닉의 주기성을 파악하기 위해 하모닉 피크간 간격을 사용하였다. 하모닉은 기본주파수의 정수 배 간격으로 나타나기 때문에, 하모닉 피크간의 간격은 신호의 주기성과 연관이 깊다. 하모닉은 주파수 상에서 확인이 가능하므로, 푸리에 변환을 통해 여기신호를 주파수 신호로 바꾸어준다. 다음 수식 (4)와 같이 파워 스펙트럼을 구해준다.

$$P_R(k) = 20\log_{10}\left(\left|R(k)\right|\right) \tag{4}$$

여기서 $R(k)$ 는 3.1장에서 생성된 잔차신호의 주파수 신호이다. 하모닉 피크들을 찾기 위해 먼저 잔차신호의 파워스펙트럼에서 다음 수식 (5)를 사용하여 모든 피크들을 찾아준다.

$$h(k) = \begin{cases} P_R(k), & \left(\Delta P_R(k) > 0\right) and \left(\Delta P_R(k+1) < 0\right) \\ 0, & otherwise \end{cases} \tag{5}$$

여기서 $\Delta P_R(k)$ 는 다음 수식(6)과 같이 정의된다.

$$\Delta P_R(k) = P_R(k) - P_R(k-1) \tag{6}$$

다음으로 수식 (5)를 통해 찾은 전체 피크에서 기본 주파수 $F_0$ 값을 기준으로 하여 하모닉 피크의 위치들을 찾는다. 하모닉 피크들은 기본 주파수의 정수 배 간격으로 위치하기 때문에, 기본 주파수를 기준으로 하여 다음 수식 (7)을 사용하여 지역 하모닉의 피크 위치들을 찾아준다.

$$h_l(m) = \arg\max_i h\left(m\frac{F_0}{2} + i\right), \qquad m = 1,2,\cdots,\left\lfloor \frac{F_S}{F_0} \right\rfloor \tag{7}$$

여기서 $F_S$ 는 표본화 주파수이고, $i$ 의 탐색 범위는 $-F_0/2 < i < F_0/2$ 이며 $\lfloor \bullet \rfloor$ 은 버림 연산자이다. 지역 하모닉 피크는 기본주파수의 정수 배 근처에 있으므로 앞의 수식 (7)과 같이 표현 가능하다. 수식 (5)와 수식 (7)을 정리하면 다음 수식 (8)과 같이 지역 하모닉의 위치와 값들을 찾을 수 있다.

$$h_l(k) = \begin{cases} h\left(k\frac{F_0}{2} + h_l(m)\right), & m = 1,2,\cdots,\left\lfloor \frac{F_S}{F_0} \right\rfloor \\ 0, & otherwise \end{cases} \tag{8}$$

<그림 6>은 <그림 5-b>의 신호에 수식 (8)을 사용하여 찾은 지역 하모닉 피크들과 그 위치를 보여준다.

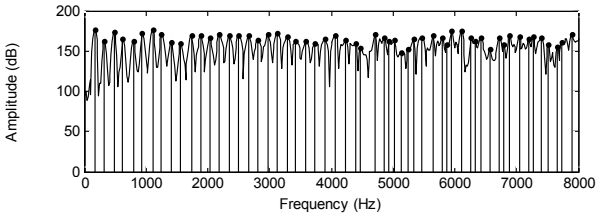


그림 6. 지역 피크 선별 결과
Figure 6. Results of local peak selection

### 3.3 잔차신호 잘라내기

하모닉들의 피크를 정확히 찾았다 할지라도 찾은 하모닉 피크들의 위치만으로 MVF를 예측하는 것은 어려운 문제다. 그렇기 때문에 선별한 지역 하모닉 피크들의 포락선 정보를 사용하여 주기성판단을 위해 잔차신호를 잘라낸다. 피크 포락선 $h_e(k)$ 은 지역 하모닉 피크들 사이를 선형보간법을 사용하여 만들고, 이 피크 포락선 보다 3 dB 낮은 위치에서 다음 수식 (9)와 같이 잔차신호를 절단한다.

$$P_T(k) = P_R(k) - \left(h_e(k) - 3\right) \tag{9}$$

수식 (9)를 사용하여 절단된 잔차신호는 <그림 7>에 나타나 있다. <그림 7>에 표시된 원형부분을 보면 하모닉이 어디까지 주기성을 갖고 나타나는지 <그림 6> 보다 정확하게 확인이 가능하다.

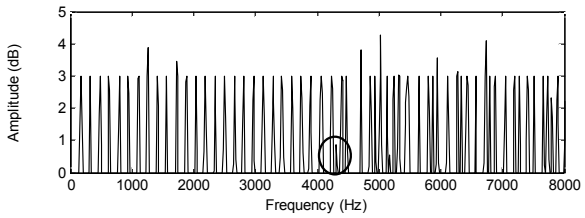


그림 7. 절단된 LP 잔차신호 결과
Figure 7. Results of truncated LP residual signal

### 3.4 피크간 거리 정규화 및 초기 MVF 예측

마지막 과정으로 절단된 잔차신호의 피크간 거리를 기본주파수로 정규화 작업을 수행 후 초기 MVF를 예측한다. 화자 또는 음소 별로 피크간 거리가 일정하지 않기 때문에, 정규화 작업이 필요하다. 하모닉은 기본 주파수의 정수 배 위치에서 생기기 때문에, 기본주파수로 피크간 거리를 정규화 수행한다. 정규화를 수행하고 나면 다음 <그림 8)>과 같은 그림이 나타난다.

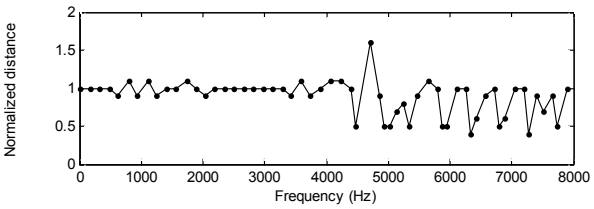


그림 8. 피크간 거리 정규화 결과
Figure 8. Results of peak distance normalization

정규화가 수행된 그림에서 보여지듯이 하모닉 피크들이 주기성을 갖는 곳에서는 정규화 값이 1에 가깝고, 주기성을 갖지 않는 곳에서는 1에서 그 값이 멀어진다. 정규화가 수행된 피크간 거리는 다음 수식에 의해 MVF 위치를 찾게 된다. 다음 수식 (10)과 같이 MVF 예측을 위한 문턱치로 0.5와 1.5를 사용하였고 정규화 값이 가장 먼저 문턱치를 벗어나는 주파수의 지점을 MVF위치로 예측한다.

$$f_{mvf} = \arg\left(\left(0.5 > D_n(k)\right) \text{ or } \left(D_n(k) > 1.5\right)\right) \tag{10}$$

여기서 $D_n(k)$ 는 정규화된 피크간 거리이다. 초기 MVF 위치 예측이 끝난 후, MVF 최적화 과정을 통해 최종 MVF를 예측한다.

## 4. 실험 환경 및 결과

### 4.1 실험 환경

실험을 위해서 단일 여성 화자가 발성한 4000 문장이 사용되었다. 이중 3000 문장은 훈련과정에 사용되었고 나머지 1000 문장은 합성 과정에서 성능평가를 위해 사용되었다. 모든 데이터의 평균 발성 길이는 2~3초 정도이며 16 kHz로 표본화 되었고 16 bit로 양자화 되었다. HTS 시스템은 2.1 버전을 사용하였고, 문맥 정보는 HTS에서 기본적으로 제공하는 문맥정보를 사용하였다. 0차를 포함한 13차 MFCC (mel-frequency cepstrum coefficient)가 사용되었으며, 분석을 위해 25 ms 해닝윈도우(hanning window)를 사용하였고 주파수 파라미터를 위해 5 ms 오버랩(overlap)을 사용하였다. 기본 주파수는 STRAIGHT(speech transformation and representation based on adaptive interpolation of weighted spectrogram)을 사용하여 계산하였다. 고주파 통과 필터 기반의 MVF 예측을 위해 차단 주파수를 500 Hz 해상도를 갖는 고주파 통과 필터들을 사용하였고, 초기 MVF 예측을 위해 정규화 자기 상관도 값은 0.5를 MVF 결정 기준으로 하였다. 제안 하는 방법에서 잔차신호를 계산하기 위해 16차 LPC가 사용되었고, 신호의 주파수 변화를 위해 512 포인트의 FFT (fast Fourier transform)이 사용되었다.

평가를 위해 CE 방법, ME 방법, 고주파 통과 필터 기반의 MVF 예측 방법(O-FTBE: optimization of filtering-based two-band excitation):, 그리고 제안하는 방법(O-FTBE: optimization of proposed two-band excitation)으로 훈련 및 합성하였다. 평가는 객관적 평가와 주관적 평가 모두 수행하였다. 먼저 객관적 평가는 훈련에 사용된 모든 데이터의 원 음성과 합성음의 왜곡 정도를 측정하였다. 왜곡 측정 방법으로 SKLD (symmetric Kullback-Leibler distortion)과 LSD(log spectral distortion)을 다음 수식 (11), (12)과 같이 측정하였다 [13], [14].

$$D_{SKLD} = \sum_k \left( \left(P(k) - Q(k)\right) \log \frac{P(k)}{Q(k)} \right) \tag{11}$$

$$D_{LSD} = \sqrt{\sum_k \left[ 10 \log \frac{P(k)}{Q(k)} \right]^2} \tag{12}$$

여기서 $P(k)$ 는 원 음성신호이고, $Q(k)$ 는 HTS를 통해 합성된 합성음이다. 또한 원음성과 비교 음성의 발성 길이를 맞추기 위해 DTW(dynamic time wrapping)을 사용하여 길이를 맞춘 후 객관적 평가를 수행하였다. 주관적 수행 평가로는 MOS(mean opinion score)와 선호도 평가를 수행하였다 [14]. 주관적 수행 평가를 위해 10명의 청취자가 평가를 하였고, 훈련 과정과 합성 과정에 사용된 각각의 문장들 중 임의로 5 문장씩 총 10 문장이 사용되었다. MOS에는 CE, ME, O-FTBE, O-PTBE 방법을 모두 평가하였고, 선호도 평가는 O-FTBE와 O-PTBE 두 가지 방법으로 합성된 합성음에 대하여 수행하였다.

### 4.2 실험 결과

객관적 평가 결과는 <표 1>에 정리되어 있다. <표 1>에서 보듯이, 훈련과정에서의 SKLD와 LSD는 ME 방법이 201.37, 68.10 dB로 모든 방법 중에 가장 좋은 성능을 보여준다. 하지만 합성과정의 SKLD와 LSD는 제안하는 방법이 317.71, 82.40 dB로 비교 방법들 중 가장 좋은 성능을 보여주었다.

표 1. 객관적 평가 결과 (단위: dB)
Table 1. Results of objective test (Unit: dB)

| 방법 | 훈련 과정 | | 합성 과정 | |
|---|---|---|---|---|
| | SKLD | LSD | SKLD | LSD |
| CE | 207.86 | 69.83 | 328.37 | 83.24 |
| ME | 201.37 | 68.10 | 322.29 | 82.87 |
| O-FTBE | 203.35 | 68.32 | 320.32 | 82.53 |
| O-PTBE | 201.79 | 68.14 | 317.71 | 82.40 |

주관적 평가 결과인 MOS는 <그림 9>에 선호도 테스트는 <그림 10>에 나타나 있다. <그림 9>에 보여지듯이 제안하는 방법이 기존 O-FTBE 방법에 비해 0.33 점의 MOS 점수가 더 높고, 모든 비교 방법들 중에서 가장 높은 점수를 얻은 것을 확인할 수 있다.

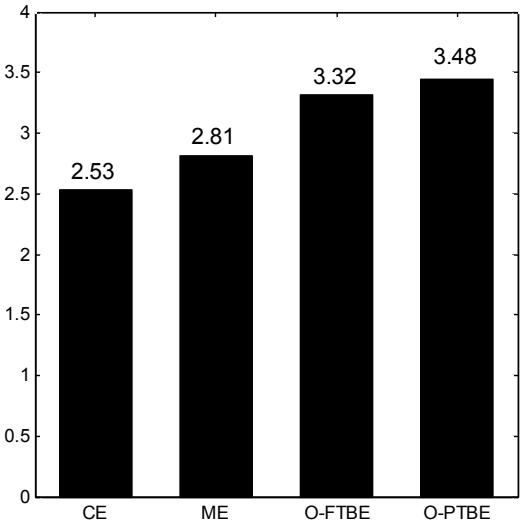


그림 9. MOS 평가 결과
Figure 9. Results of MOS test

또한 <그림 10>에서 보여지듯이 선호도 테스트 결과를 비율로 나타냈는데, 65%의 청취자가 제안하는 방법을 O-FTBE 보다 낫다고 선호하였다. 객관적 평가 결과와 주관적 평가 결과 모두에서 나타나듯이 제안하는 방법으로 MVF를 예측하여 훈련, 합성하는 것이 기존의 O-FTBE 방법보다 더 나은 성능을 보여주는 것을 확인할 수 있다.

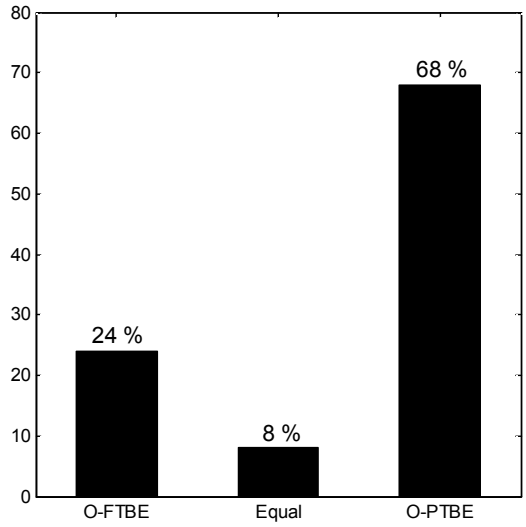


그림 10. 선호도 평가 결과
Figure 10. Results of preference test

<그림 11>은 훈련 과정에서 O-FTBE방법과 제안하는 방법으로 계산된 MVF의 결과이다. 그림의 배경은 스펙트로그램이고 실선은 계산된 MVF이다. 샘플 결과에서도 보여지듯이 기존 방법의 MVF는 고주파 통과 필터의 차단 주파수에 의해 해상도가 결정되므로 해상도가 낮은 것을 알 수 있다. 0.5 ~ 1 초 사이의 MVF 결과를 보면 기존 O-FTBE 방법은 너무 높게 MVF가 예측된 반면, 제안하는 방법은 하모닉이 존재하는 곳까지 적절하게 MVF가 예측된 것을 확인할 수 있다.

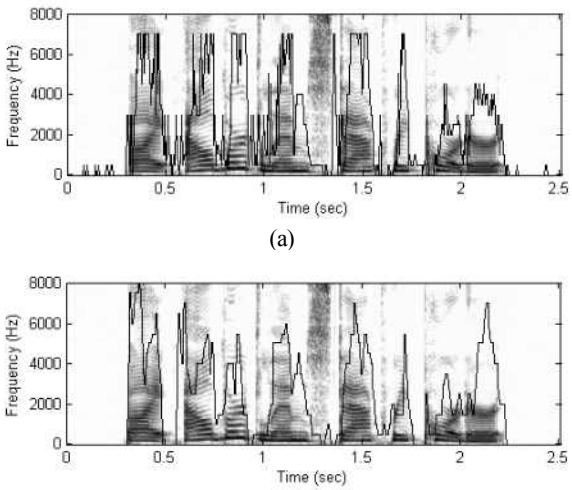


그림 11. 스펙트로그램과 MVF 예측 결과
((a): O-FTBE, (b): O-PTBE)
Figure 11. Spectrograms and MVF estimation results
((a): O-FTBE, (b): O-PTBE)

## 5. 결론

TBE 모델은 HTS 합성기에서 우수한 음질을 제공하는 여기신호 모델이다. 그리고 이 여기신호 모델에서는 MVF가 가장 중요한 파라미터이다. 하지만 기존의 고주파 통과 필터 기반의 MVF 예측 방법은 분석 음성의 스펙트럼 포락선에 의해 정확한 MVF 예측이 어렵다는 단점이 있다. 그래서 본 논문에서는 스펙트럼 포락선이 제거된 LP 잔차신호와 주파수상에서 주기성을 나타내는 하모닉 정보를 이용하여 MVF를 예측하는 방법을 제안하였다. 제안된 방법은 객관적 성능평가와 주관적 성능평가 모두에서 제안하는 MVF예측 방법이 기존의 방법보다 우수한 성능을 보여주었다. 합성 시스템의 메모리와 비교하여 음질이 상대적으로 우수하지만 아직 실생활에서 사용하기에는 합성음 음질 향상이 더 필요하다. 앞으로의 일로 여기신호 새로운 생성 모듈의 연구가 필요하다.

## 감사의 글

이 논문은 2012년도 한국연구재단의 지원을 받아 수행되었습니다(과제명: 지능형 로봇과 휴대용 장치를 위한 초소용량 고음질 내장형 음성 합성 시스템 개발, 과제번호: 2012009563).

## 참고문헌

[1] Hunt, A. & Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database, *Proc. IEEE ICASSP*. Vol. 1, 959-962.

[2] Tokuda, K., Kobayasho, T. & Imai, S. (1995). Speech parameter generation form HMM using dynamic features, *Proc. IEEE ICASSP*. Vol. 1, 660-663.

[3] Tokuda, K., Masuko, T., Yamada, T., Kobayashi, T. & Imai, S. (1995). An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features, *Proc. Eurospeech*. Vol. 1, 757-760.

[4] Tokuda, K., Zen, H. & Black, A. W. (2002). An HMM-based speech synthesis system applied to English, *Proc. IEEE Workshop on Speech Synthesis*..227-230.

[5] Fukada, T., Tokuda, K., Kobayashi, T. & Imai, S.. (1992). An adaptive algorithm for mel-cepstral analysis of speech, *Proc. ICASSP*. Vol. 1, 137-140.

[6] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. & Kitamura, T. (2000). Speech parameter generation algorithm for HMM-based speech synthesis, *Proc. ICASSP*. Vol. 1, 1315-1318.

[7] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. & Kitamura, T. (2001). Mixed excitation for HMM-based speech synthesis, *Proc. Eurospeech*. Vol. 3, 2263-2266.

[8] Kim, S., Kim, J. & Hahn, M. (2006). HMM-based Korean speech synthesis system for hand-held devices, *IEEE Trans. Consumer Electronics*. Vol. 52, No. 4, 1384-1390.

[9] Kim, S., Kim, J. & Hahn, M. (2006). Implementation and evaluation of an HMM-based Korean speech synthesis system, *IEICE Transactions on Information and Systems*. Vol. E89-D, No.3, 1116-1119.

[10] Kim, S., Kim, J. & Hahn, M. (2007). Two-band excitation for HMM-based speech synthesis, *IEICE Trans. Information and Systems*. Vol. E90-D, No 1, 378-381.

[11] Han, S., Jeong, S. & Hahn, M. (2009) Optimum MVF estimation-based two-band excitation for HMM-based speech synthesis, *ETRI Journal,* Vol. 31, No. 4, 457-459.

[12] Zen, H., Toda, T., Nakamura, M. & Tokuda, K. (2007) Details of Nitech. HMM-based speech synthesis system for the Blizzard Challenge. 2005, *IEICE Trans. Information and Systems,* Vol. E90-D, 325-333.

[13] Klabbers, E. & Veldhuis, R. (2001). Reducing audible spectral discontinuities, *IEEE Trans. Speech and Audio Proc.,* Vol. 9, No. 1, 39-51.

[14] Huang, X., Acero, A. & Hon, H.-W. (2001). *Spoken language processing: a guide to theory, algorithm, and system development,* NY: Prentice Hall.

• **박지훈(Park, Jihoon), 교신저자**
한국과학기술원 전기 및 전자 공학과
대전시 유성구 구성동 대학로 291
Tel: 042-350-5474
Email: batho2n@kaist.ac.kr
관심분야: 음성합성, 오디오 코딩
현재 전기 및 전자 공학과 대학원 박사과정 재학 중

• **한민수(Hahn, Minsoo)**
한국과학기술원 전기 및 전자공학과
대전시 유성구 구성동 대학로 291
Tel: 042-350-8074
Email: mshahn@ee.kaist.ac.kr
관심분야: 음성합성, 잡음제거, 음성 코딩, 오디오 코딩