GMM을 이용한 MFCC로부터 복원된 음성의 개선

최원영(부산대), 최무열(부산대), 김형순(부산대)

<차 례>

<Abstract>

Improvement of Speech Reconstructed from MFCC Using GMM

Won Young Choi, Mu Yeol Choi, Hyung Soon Kim

The goal of this research is to improve the quality of reconstructed speech in the Distributed Speech Recognition (DSR) system. For the extended DSR, we estimate the variable Maximum Voiced Frequency (MVF) from Mel-Frequency Cepstral Coefficient (MFCC) based on Gaussian Mixture Model (GMM), to implement realistic harmonic plus noise model for the excitation signal. For the standard DSR, we also make the voiced/unvoiced decision from MFCC based on GMM because the pitch information is not available in that case. The perceptual test reveals that speech reconstructed by the proposed method is preferred to the one by the conventional methods.

*Keywords : Speech reconstruction, Gaussian mixture model

1. 서 론

휴대 전화나 PDA와 같은 휴대용 단말기는 이미 현대 생활의 필수품이 되었으며, 단말기를 통한 음성인식도 관심의 대상이 되어왔다. 그러나 현재 사용되고 있는 휴대용 단말기들은 대어휘 음성인식에 충분한 하드웨어 처리능력을 가지고 있지 못한 경우가 대부분이다. 그 대안으로 단말기 대신 서버에 인식 시스템을 두는 방식을 사용할 수 있으나, 이 경우에는 음성코덱과 채널 오류로 인해 음성인식 성능이 저하되는 문제가 있다. 이러한 문제의 해결을 위해 분산음성인식(Distributed Speech Recognition(DSR)) 방식이 제안되었다[1]. 이 방식에서 단말기는 음성 특징벡터를 추출하고 데이터 채널을 통해서 그 특징 벡터를 서버로 보낸다. 서버의 음성인식기에서는 전송 받은 특징 벡터를 사용하여 인식과정을 수행함으로써, 인식성능의 향상을 가져올 수 있다. 이러한 분산음성인식 방식에는 일반적인 음성인식 특징벡터인 Mel-Frequency Cepstral Coefficient(MFCC)만을 전송하는 standard DSR 방식과 Mandarin과 같은 tonal language를 인식하기 위해 특징벡터와 피치 주파수를 함께 전송하는 extended DSR 방식이 있다.

서버에서의 음성인식기능과 더불어 금융거래에 있어서 비밀번호와 같은 법률상 민감한 정보들의 저장 및 확인을 위해 서버 측에서 음성신호 자체를 복원할 필요성이 제기되고 있다. 그러나, DSR에서의 특징벡터의 전송방식은 음성신호를 직접 전송하는 방식에 비해 음성을 복원하기가 쉽지 않은 문제점이 있다. 이 문제를 해결하기 위하여 최근 들어 서버 측에서 MFCC와 피치 주파수를 이용하여 음성을 복원하는 연구들이 이루어지고 있다[2][3][4][5].

정현파 모델을 사용하여 음성을 복원하는 기존 연구들은 extended DSR 방식에서 전송되는 MFCC와 피치 정보를 모두 사용한다. 즉, MFCC로부터 정현파의 진폭을, 피치 주파수로부터 정현파의 주파수 및 위상을 추정한다. 정현파의 파라미터를 추정하는 과정에는, 유성음 프레임의 고주파 영역에서 나타나는 잡음특성을 고려하여 Harmonic plus Noise Model(HNM)을 이용하는 방법[3]과 잡음특성을 고려하지 않고 하모닉 성분만을 이용하는 방법[4]이 있다. 전자의 경우, 미리 정해진 주파수보다 높은 주파수 영역에서 잡음 특성이 나타나도록 한다. 그러나, 이들 두 가지 방법 모두, 실제음성에서 주파수영역의 프레임에 따라 보이는 잡음 특성의 변화를 제대로 표현하지 못한다. 이 문제의 해결을 위해 본 논문에서는 Gaussian Mixture Model(GMM)을 기반으로 매 프레임 별로 잡음특성을 나타내는 주파수영역을 MFCC로부터 추정하는 방법을 도입한다. 그리고 피치정보 없이 단지 MFCC만이 전송되는 standard DSR 방식에서, MFCC로부터 GMM을 기반으로 하여 유/무성음을 검출하여 음성을 복원하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 서론에 이어 2장에서는 MFCC로부터 음성을 복원하는 기존의 방법에 대하여 설명하고, 3장에서는 통계적 모델링 방법에 의해

영역 분리 주파수의 추정을 통한 음성복원 방법을 제안한다. 4장에서는 MFCC로부터 유/무성음 검출을 통한 음성복원 방법을 제안하며, 5장에서는 실험 및 결과를 기술하고, 마지막으로 6장에서 이 논문의 결론을 맺는다.

2. MFCC로부터 음성 복원

본 논문에서 사용되는 음성복원 방법으로는 M개의 정현파의 합으로 음성신호 $x(n)$을 합성하는 정현파 모델을 이용하였다[6].

$$x(n) = \sum_{m=1}^{M} A_m \cos(\omega_m n + \Theta_m) \tag{1}$$

여기서 A_m, ω_m 및 Θ_m는 각각 m번째 정현파의 진폭, 주파수 및 위상이다.

2.1 스펙트럼 포락선 추정

정현파 모델에 사용되는 파라미터들 중에서 진폭은 MFCC 벡터로부터 추정한 스펙트럼 포락선을 이용하여 얻게 된다[2]. 먼저 멜-주파수 필터에 대한 효과와 pre-emphasis에 대한 효과를 제거하기 위해서 MFCC를 수정하게 된다. 멜-주파수 필터에 대한 효과는 각 멜-주파수 필터들의 면적 값들을 그에 대응하는 MFCC값에서 빼주는 형태로 제거하게 되며, pre-emphasis 필터 효과는 멜-주파수 필터뱅크의 중심주파수에 해당하는 필터값들을 켑스트럼 계수로 변환하여 MFCC값에서 빼줌으로써 제거한다. 이렇게 수정된 MFCC로부터 spectral bin만큼의 높은 분해도로 IDCT를 하여 smoothed log magnitude spectrum을 얻게 된다.

2.2 정현파 모델 파라미터 추정

정현파 모델에 사용되는 파라미터들 중 정현파의 하모닉 주파수(ω_m)들은 피치 주파수(ω_0)의 배수($\omega_m = m \cdot \omega_0$)로써 추정하게 된다.

정현파의 진폭(A_m)은 앞서 구한 smoothed magnitude spectrum으로부터 하모닉 주파수에 해당하는 magnitude값 ($A_m = |X(m \cdot \omega_0)|$) 으로 추정할 수 있다.

위상(Θ_m)은 유성음 프레임과 무성음 프레임을 나누어 추정하게 된다. 무성음 프레임에서는 $(-\pi, \pi)$사이에서 균일한 분포를 가지는 랜덤 위상을 사용한다. 유성음 프레임에서 유성음 주파수 구간은 식 (2)와 같은 선형위상모델[6]을, 무성음

주파수 구간은 무성음 프레임에서 사용한 것과 같은 랜덤 위상을 사용한다.

$$\begin{aligned}\Theta_m(iT) &= \int_0^{iT} \omega_m(\sigma)d\sigma \\ &= \Theta_m[(i-1)T] + (\omega_m^{i-1} + \omega_m^{i})\frac{T}{2}\end{aligned} \tag{2}$$

여기서 i는 프레임 번호이며 T는 한 프레임의 길이, ω_m^i은 i번째 프레임의 m번째 하모닉 주파수, 그리고 $\Theta_m(iT)$는 i번째 프레임의 위상이다.

3. Maximum voiced frequency(MVF)를 이용한 복원알고리즘의 개선

일반적으로 유성음 프레임에서 고주파수 영역은 정현파의 하모닉 성분이 저주파수 영역에 비해 뚜렷하지 않아 잡음 특성이 나타난다. 이러한 특성을 표현하기 위해 Harmonic plus Noise Model(HNM)이 제안되었다[7]. HNM에서는 유성음 스펙트럼을 순수 하모닉 영역과 잡음특성이 나타나는 영역으로 구분하며, 이들을 구분하는 주파수 값을 Maximum Voiced Frequency(MVF)라고 부른다. MVF는 음성스펙트럼의 각 하모닉 주파수 대역폭내의 가장 높은 봉우리에서의 피크값과 나머지 봉우리들에서의 피크값 사이의 비율을 이용해서 구한다[7].

분산음성인식 특징 파라미터로부터 음성을 복원할 경우 MVF값이 제공되지 않으므로, 기존 방식에서는 HNM을 사용하지 않거나[4] 고정된 MVF값을 이용한 HNM을 사용하였다[3]. 본 논문에서는 MFCC와 MVF사이의 연관관계를 GMM으로 모델링하여 MVF를 추정하는 방식을 제안한다. 제안된 방식에서는 우선 데이터의 유성음 프레임에서 구한 MVF와 MFCC 벡터를 이용하여 새로운 벡터 $\boldsymbol{y}$를 만든다.

$$\boldsymbol{y} = [\,\boldsymbol{x}^T, f\,]^T \tag{3}$$

여기서 $\boldsymbol{x}$는 MFCC($c_0 \sim c_{12}$)이고 f는 MVF이다.

벡터 $\boldsymbol{y}$로부터 식(4)와 같이 K개의 Gaussian mixture를 분포로 모델링한다.

$$P(y\,|\lambda) = \sum_{k=0}^{K} \alpha_k N(y;\,\mu_k \Sigma_k) \tag{4}$$

$$\sum_{k=0}^{K} \alpha_k = 1,\ \alpha_k \geq 0\ \lambda = \{\alpha_k,\ \mu_k,\ \Sigma_k \text{ for } k = 1, 2, \cdots, K\} \tag{5}$$

여기서 μ_k와 Σ_k는 k번째 Gaussian 분포의 평균과 공분산이며, λ는 GMM의 파라미터모델이다. 그리고 $N(\boldsymbol{y}, \mu_k, \Sigma_k)$는 다음 식과 같은 Gaussian 분포이다.

$$N(\boldsymbol{y}, \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{D/2} |\Sigma_k|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{y}-\mu_k)^T \Sigma_k^{-1}(\boldsymbol{y}-\mu_k)\right] \quad (6)$$

식 (4), (5)에서의 α_k는 전체 벡터의 분포에서 k번째 Gaussian 분포가 차지하는 가중치이고, 식 (6)에서의 D는 벡터 $\boldsymbol{y}$의 차원이다.

이렇게 GMM모델이 구성되고 나면, 입력된 특징 벡터 $\boldsymbol{x}$로부터 MVF f를 추정하기 위한 변환 함수 F는 식 (7)과 같이 자승오차를 최소화하는 함수가 된다.

$$\varepsilon_{mse} = E[\,||f - F(\boldsymbol{x})||^2] \quad (7)$$

여기서 E는 기대값이다. 변환 함수는 식 (8)와 같이 회귀식으로 표현 된다[9].

$$\begin{aligned} F(\boldsymbol{x}) &= E[f|\boldsymbol{x}] = \int f \cdot p(f|\boldsymbol{x}) df \\ &= \sum_{k=1}^{K} h_k(\boldsymbol{x}_i)\left(\mu_k^f + \Sigma_k^{fx}(\Sigma_k^{x})^{-1}(\boldsymbol{x}_i - \mu_k^x)^T\right) \end{aligned} \quad (8)$$

여기서 i는 프레임 번호이며, $h_k(\boldsymbol{x}_i)$, μ_k^y 및 Σ_k^y는 다음과 같다.

$$h_k(\boldsymbol{x}_i) = \frac{\alpha_k p(\boldsymbol{x}_i|c_k^x)}{\sum_{k=1}^{K} \alpha_k p(\boldsymbol{x}_i|c_k^x)} \quad (9)$$

$$\mu_k^y = \begin{bmatrix} \mu_k^x \\ \mu_k^f \end{bmatrix} \quad \text{and} \quad \Sigma_k^y = \begin{bmatrix} \Sigma_k^{xx} & \Sigma_k^{fx} \\ \Sigma_k^{xf} & \Sigma_k^{ff} \end{bmatrix} \quad (10)$$

이때, $p(\boldsymbol{x}_i|c_k^x)$는 k번째 mixture의 MFCC 벡터의 marginal distribution, 그리고 α_k는 k번째 mixture의 가중치이며, μ_k^y와 Σ_k^y는 각각 벡터 $\boldsymbol{y}$의 k번째 mixture의 평균과 공분산이다.

위와 같은 방법으로 MFCC로부터 MVF를 추정하게 되면, 피치 주파수로부터 추정한 각 정현파의 주파수와 MVF를 비교한다. 이때 정현파의 주파수가 MVF보다 낮으면 유성음 주파수의 정현파로 판단하여 선형위상모델을 이용한 정현파의

위상을 추정한다. 반면 정현파의 주파수가 MVF보다 높으면 무성음 주파수의 정현파로 판단하여 $(-\pi, \pi)$사이의 랜덤위상으로 정현파의 위상을 추정한다.

4. 피치정보가 없을 경우의 MFCC로부터 음성복원

2장에서 설명한 MFCC로 음성을 복원하는 방식은 MFCC와 피치 주파수를 함께 전송하는 extended DSR시스템에서 사용 가능한 방식이다. 그러나 standard DSR 시스템은 피치 정보는 전송하지 않고 단지 MFCC만을 전송한다. 이 경우 피치 주파수를 알 수 없으므로 앞서 설명한 MFCC로부터 음성을 복원하는 방법을 그대로 사용할 수가 없다. 이때 피치 주파수를 고정된 상수로 사용한다고 하더라도 각 프레임의 유/무성음 정보는 필요하다. 따라서 본 장에서는 피치정보가 없을 때 MFCC만을 가지고 유/무성음 검출을 통해 음성을 복원하는 방법을 제안하고자 한다

4.1 MFCC로부터 유/무성음 검출

훈련 데이터로부터 유성음 프레임에 해당하는 MFCC 벡터들로 GMM을 만들고, 이와 더불어 무성음 프레임에 해당하는 MFCC 벡터들로 구성된 GMM을 만든다. 결과적으로 무성음과 유성음의 경우에 해당하는 각각의 GMM 확률분포 $P_u(\boldsymbol{x})$와 $P_v(\boldsymbol{x})$를 식 (11) 및 식 (12)와 같이 표현할 수 있다.

$$P_u(\boldsymbol{x}) = \sum_{i=0}^{M} \alpha_i^U N(X; \mu_i^U \Sigma_i^U), \qquad \sum_{i=0}^{M} \alpha_i^U = 1 \tag{11}$$

$$P_v(\boldsymbol{x}) = \sum_{i=0}^{M} \alpha_i^V N(X; \mu_i^V \Sigma_i^V), \qquad \sum_{i=0}^{M} \alpha_i^V = 1 \tag{12}$$

여기서 α_i^U, μ_i^U 및 Σ_i^U는 무성음 GMM의 i번째 mixture의 가중치, 평균벡터 및 공분산 행렬이며 α_i^V, μ_i^V 및 Σ_i^V는 유성음 GMM의 i번째 mixture의 가중치, 평균벡터 및 공분산 행렬이다.

유무성음 판별을 위한 테스트 음성 프레임의 MFCC 벡터 $\boldsymbol{x}$에 대하여 $P_u(\boldsymbol{x}) > P_v(\boldsymbol{x})$이면 무성음이라 판단하고, 그렇지 않을 경우 유성음으로 판단한다.

4.2 정현파 모델 파라미터 추정

정현파의 파라미터 중 각 정현파의 주파수들은 피치 정보가 제공되지 않기 때문에 직접적으로 구할 수 없다. 그래서 유성음이라고 판단된 프레임의 경우는 고정된 주파수, ω_f를 이용하여 기본 주파수 ω_0대신 사용한다. 그리고 고정된 주파수의 배수를 사용하여 정현파의 주파수를 얻게 된다.

$$\omega_m = m \cdot \omega_f \tag{13}$$

이때 음성을 복원하기 위해서 피치 주파수를 고정된 상수로 사용하더라도 각 프레임의 유/무성음 정보는 필요하다. 그래서 4.1절에서 설명한 것처럼 MFCC를 이용하여 유/무성음을 추정한다. 무성음이라고 판단한 프레임에서는 2장에서 설명한 것과 마찬가지로 각 FFT bin에 해당하는 주파수들을 정현파의 주파수로 사용한다. 나머지 파라미터들은 2.1절과 2.2절에서 설명한 것과 동일하게 얻을 수 있다.

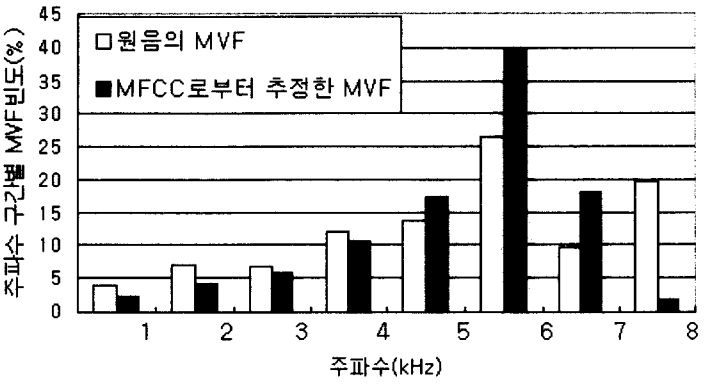
5. 실험 및 결과

5.1 실험 환경

훈련과 테스트를 위하여 음성정보기술산업지원센터(SITEC)에서 구축한 Phonetically Balanced Sentence(PBS) 음성 데이터 베이스 중 일부를 사용하였으며, 이 데이터베이스는 16kHz로 샘플링된 것이다. 훈련을 위해서 남녀 각 10명씩, 총 4173개의 문장을 사용하였고. 테스트를 위해서 훈련에 사용되지 않은 남녀 각 2명씩 473개의 문장을 사용하였다.

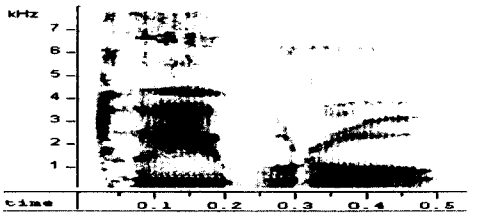
5.2 MFCC를 이용한 MVF추정 실험

<그림 1>은 테스트 데이터로부터 직접 구한 MVF와 MFCC로부터 추정된 MVF에 대한 주파수 구간대별 분포를 보여준다. 테스트 데이터로부터 구한 MVF와 MFCC로부터 추정한 MVF값의 평균오차는 1162Hz로 비교적 크게 나타났다. 이는 <그림 1>에서 보는 바와 같이 많은 프레임에서 원음의 MVF가 8000Hz 가까이에 분포하고 있으나, 이 프레임에서의 MFCC로 추정된 MVF값에서는 원음의 MVF 평균인 5241Hz에 가까운 주파수 대역으로 추정되어 평균오차가 크게 난 것으로 분석된다.

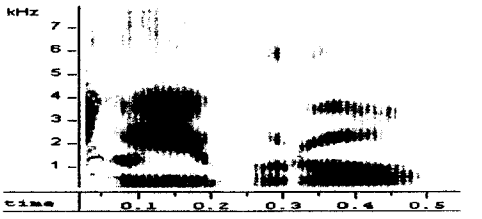


<그림 1> 주파수 구간에 따른 MVF 분포

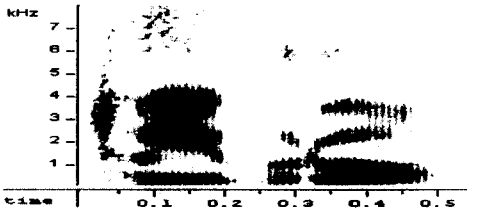
<그림 2>에서 기존 방식 및 제안된 방식에 의해 복원된 음성의 스펙트로그램을 비교하여 나타내었다. <그림 2>에서 (a)는 원음성이고, (b)는 HNM을 이용하지 않고 유성음의 모든 주파수대역이 하모닉 특성을 가지도록 복원한 것이며, (c)와 (d)는 고정된 MVF(5300Hz)와 MFCC로부터 추정한 MVF를 적용하여 스펙트럼을 유성음 영역과 무성음 영역으로 나누어 복원한 음성이다. <그림 2>의 (a) 와 (d)를 보면 0.1~0.2초 사이의 4kHz 근처에서 잡음특성이 제안된 방식과 원음이 비슷한 잡음특성을 나타내는 것을 알 수 있다.



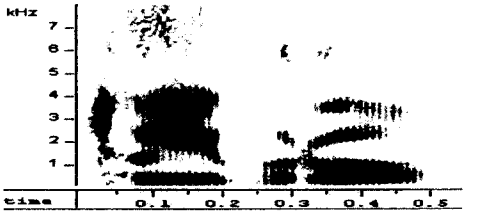
(a) 원음



(b) HNM을 이용하지 않은 복원음성



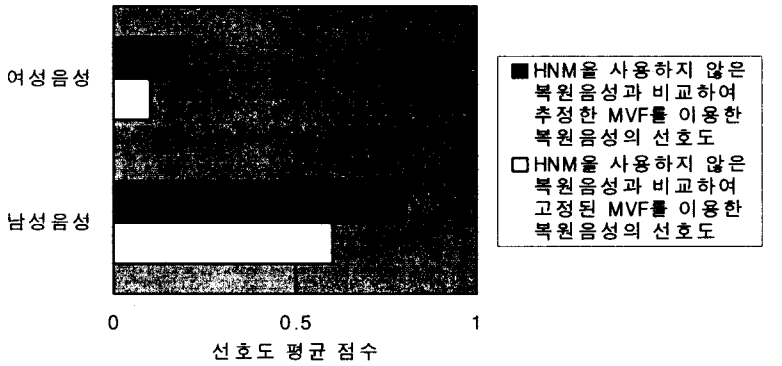
(c) 고정 MVF 기반의 HNM을 이용한 복원음성



(d) 추정된 MVF 기반의 HNM을 이용한 복원음성

<그림 2> 원음 및 복원한 음성들의 스펙트로그램

그 다음으로 20대 성인 남자 11명을 대상으로, HNM을 이용하지 않은 방식과 비교하여 고정된 MVF 및 추정된 MVF를 기반으로 HNM을 적용한 방식의 선호도에 대한 청취평가를 수행하였으며, 그 결과를 <그림 3>에 나타내었다. 선호도는 HNM을 사용하지 않는 경우와 비교하여 상대적으로 아주 좋음(2), 좋음(1), 같음(0), 나쁨(-1), 아주 나쁨(-2)의 다섯 단계로 평가하였다.



<그림 3> HNM을 사용하지 않은 복원음성에 비해 HNM을 사용한 복원음성의 선호도

평가 결과는 MVF를 이용한 복원한 음성을 그렇지 않은 것보다 선호하는 것으로 나타났으며, MVF를 이용한 복원음성의 경우, 고정된 MVF를 이용한 복원음성보다 추정한 MVF를 이용한 복원음성을 약간 더 선호하는 것으로 나타났다. 그림에서 여성 음성의 경우, 고정된 MVF와 추정된 MVF를 이용한 복원음성이 MVF를 이용하지 않고 복원한 음성과 큰 차이가 없는 것으로 나타났다. 이는 여성 음성의 경우에는 잡음특성이 나타나는 주파수 영역이 남성 음성보다 높은 주파수에 있기 때문에, MVF 적용 여부가 복원 음질에 큰 차이를 나타내지 않기 때문인 것으로 분석되며, 향후 추가적인 검토가 필요하다.

5.3 MFCC를 이용한 유/무성음 검출 실험

MFCC 벡터를 이용하여 유/무성음 검출 실험결과로서 식 (14)과 같이 분류 오류율 (classification error rate)을 측정하였다.

$$E_c = \frac{N_{U\to V} + N_{V\to U}}{N_{Total}} \times 100\% \tag{14}$$

여기서 $N_{U\to V}$는 무성음 프레임을 유성음 프레임으로 판단한 프레임의 수이고,

$N_{V\to U}$는 유성음 프레임을 무성음 프레임으로 판단한 프레임 수, N_{Total}음 전체 프레임의 수이다.

분류오류율의 측정결과 6.3%의 오류율을 보였다. 그리고 보다 세부적인 유/무성음 검출 결과를 <표 1>에 나타내었다. <표 1>에서 보는 바와 같이 무성음을 무성음으로 추정한 비율이 93.9%로 나타났고, 유성음을 유성음으로 판단한 경우는 93.6%로 나타났다. 잘못 검출한 경우에 대해 살펴보면 무성음을 유성음으로 검출한 경우가 38.5%이고, 유성음을 무성음으로 잘못 판단하는 경우가 61.5%로서 후자의 경우가 더 많았다.

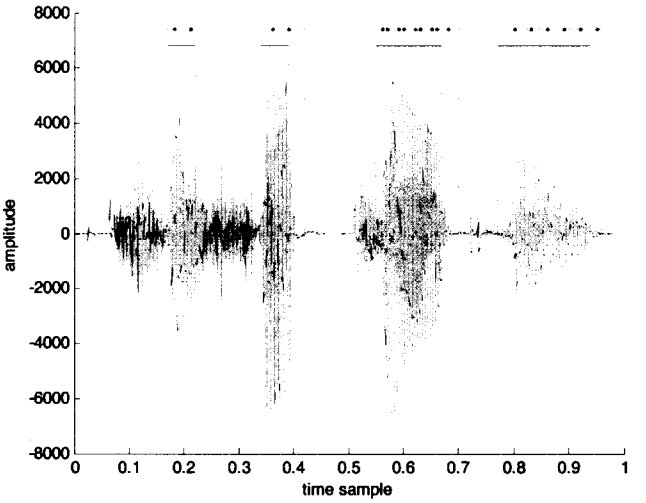
<그림 4>는 음성신호와 함께 ESPS에서 사용한 방법[6]으로 검출한 유/무성음 결과(점선)와 MFCC로 추정한 유/무성음을 검출한 결과(실선)를 나타낸 것이다. 전체 96프레임 중 87프레임이 일치하며, 이 그림으로부터 MFCC를 이용하여 유/무성음 검출한 결과가 ESPS을 이용한 유/무성음 검출의 결과와 유사한 결과를 보임을 확인할 수 있다. <그림 5>는 원음성과 피치정보가 없이 MFCC만으로 음성을 복원하였을 경우의 스펙트로그램이다. 여기서 유성음 프레임에 대한 고정 피치값은 210Hz를 사용하였다. 복원된 음성은 고정된 피치의 사용으로 단조로운 음색의 tonal sound로 복원되었다. ESPS를 이용하여 유/무성음을 검출한 결과를 이용하여 복원한 음성인 (c)와 MFCC로 추정한 유/무성음 검출을 이용하여 복원한 음성인 (d) 사이의 차이는 거의 나지 않았다.

<표 1> MFCC를 이용한 유/무성음 검출 성능

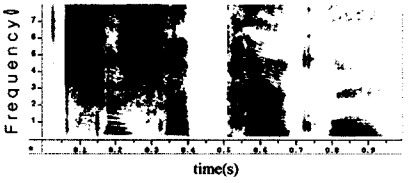
올바로 판단한 경우	유성음을 유성음으로 판단한 경우	93.6%
	무성음을 무성음으로 판단한 경우	93.9%
잘못 판단한 경우	무성음을 유성음으로 판단한 경우	38.5%
	유성음을 무성음으로 판단한 경우	61.5%

6. 결 론

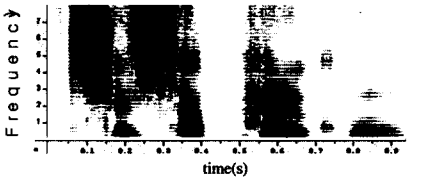
본 논문에서는 분산음성인식에서 사용되는 음성특징 파라미터인 MFCC를 이용하여 음성신호를 복원하는 방법의 개선을 위하여 MFCC로부터 통계적 모델링에 의해 MVF를 추정하는 방법을 제안하였다. 또한 피치정보가 제공되지 않는



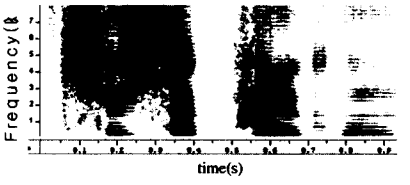
<그림 4> 음성신호와 유/무성음 검출결과
(점선 : ESPS를 이용한 유/무성음 검출 결과,
실선 : MFCC를 이용하여 추정한 유/무성음 검출 결과)



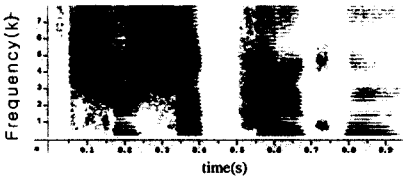
(a) 원음성



(b) 모든 프레임을 유성음으로 가정하여 MFCC만 이용한 복원 음성



(c) 원음성의 유/무성음 정보와 MFCC를 이용한 복원 음성



(d) MFCC 및 이로부터 추정한 유/무성음 정보를 이용한 복원 음성

<그림 5> 원음성과 MFCC만으로 복원한 음성의 스펙트로그램

Standard DSR 방식에서 MFCC만을 이용하여 음성을 복원하기 위한 방법으로 유/무성음 검출을 이용한 음성복원 방법을 제안하였다. 이 경우 피치 주파수를 고정된 값으로 사용하더라도 각 프레임의 유/무성음 정보가 필요하므로, 이에 입력 받은 MFCC로부터 각 프레임의 유/무성음 정보를 검출하는 방법을 제안하였다.

제안된 방법에서 추정된 MVF의 값을 이용하여 MFCC로 음성을 복원한 결과, 선호도 테스트에서 남성음성에 대한 음질 향상이 있음을 확인할 수 있었다. 그리고 유/무성음 검출방법을 이용하여 음성을 복원한 결과, 전구간을 유성음으로 가정하여 음성을 복원한 것에 비해 음질이 향상되었다.

참고문헌

[1] European Telecommunications Standards Institute, "Speech Processing, Transmission and Quality aspects; Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms", ETSI Standard ES 201 108 v1.1.2, Apr. 2000.

[2] B. Pilner and X. Shao, "Speech reconstruction from MFCCs using a source-filter model", *in Proc. of ICSLP*, pp. 2421-2424, Sep. 2002.

[3] D. Chasan et al, "Speech reconstruction from mel frequency cepstral coefficients and pitch", *in Proc. of ICASSP*, pp. 1299-1302, June, 2000.

[4] X. Shao and B. Milner, "Clean speech reconstruction from noisy mel-frequency cepstral coefficients using a sinusoidal model", *in Proc. of ICASSP*, pp. I.704-I.707, Apr. 2003.

[5] H. G. Kang and H. K. Kim, "A phase generation method for speech reconstruction from spectral envelope and pitch intervals", *in Proc. of ICASSP*, pp. 2645-2648, May, 2002.

[6] R. J. McAulay and T. F. Quatieri, "Sinusoidal Coding", *Speech coding and Synthesis* (W.B. Kleijn and K.K. Paliwal, eds.), Ch. 4, pp. 121-170, Elsevier, 1995.

[8] Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification", Ph.D. diss., Ecole Nationale Supèrieure des Télécommunications, Paris, France, Jan. 1996.

[9] A. Kain and Y. Stylianou, "Stochastic modeling of spectral adjustment for high quality pitch modification", *in Proc. of ICASSP*, pp. II.949-II.952, June, 2000.

접수일자 : 2005년 2월 10일
게재결정 : 2005년 3월 15일

▶ 최원영 (Won Young Choi)
주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 공과대학 전자공학과
소속: 부산대학교 전자공학과 음성통신연구실
전화: 051) 516-4279
E-mail: nan002@pusan.ac.kr

▶ 최무열 (Mu Yeol Choi)
주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 공과대학 전자공학과
소속: 부산대학교 전자공학과 음성통신연구실
전화: 051) 516-4279
E-mail: mychois@pusan.ac.kr

▶ 김형순 (Hyung Soon Kim)
주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 공과대학 전자공학과
소속: 부산대학교 전자공학과 음성통신연구실
전화: 051) 510-2452
E-mail: kimhs@pusan.ac.kr