한국음향학회지 제23권 제1호 pp. 61~74 (2004)

채널에 강인한 화자 인식을 위한 채널 정규화 피치 동기 켑스트럼에 관한 연구

A Study on the Channel Normalized Pitch Synchronous Cepstrum for Speaker Recognition

김 유 진*, 정 재 호*
(Yu-Jin Kim*, Jae-Ho Chung*)

*인하대학교 전자공학과 디지털 신호처리 연구실
(접수일자: 2003년 9월 29일; 채택일자: 2003년 11월 13일)

본 논문에서는 채널 환경에 강인한 화자 인식 시스템을 위하여 문맥과 화자에 종속적인 켑스트럼 추출 방법과 추출된 켑스트럼에서 화자 정보의 손실을 최소화하는 채널 정규화 방법을 제안하였다. 제안된 추출 방법은 화자의 고유한 피치를 이용한 피치 동기 분석 방법에 기반을 두어 켑스트럼을 추출한다. 따라서 일명 피치 동기 켑스트럼 (PSC)은 유성음 구간에서 성도의 임펄스 응답을 보다 정확하게 표현할 수 있다. 또한 피치는 채널 환경에서 스펙트럼에 비해 강인하므로 피치 동기 켑스트럼은 채널에 의한 스펙트럼의 왜곡을 보상할 수 있다. 제안된 채널 정규화 방법인 포먼트 평활화 피치 동기 켑스트랄 평균 차감법 (FBPSCMS)은 포먼트 평활화 켑스트랄 평균 차감법을 PSC에 적용하여 프레임 내 처리의 정확도를 개선시킨다. 제안된 방법들의 화자 인식 성능을 비교하기 위해 남자 112명과 여자 56명에 대해 TIMIT과 전화선 환경의 NTIMIT을 이용한 화자 식별을 수행하였다. 실험 결과 피치 동기 LPCC는 기존 단구간 켑스트럼과 비교하여 에러 감소율을 최대 7.7%까지 향상시켰고, FBPSCMS는 극점 필터링 CMS에 비해 보다 안정되고 낮은 에러율을 나타내었다.
핵심용어: 특징 추출, 피치 동기 분석, 켑스트럼, 채널 정규화, 화자 인식
투고분야: 음성처리 분야 (2.5)

In this paper, a context- and speaker-dependent cepstrum extraction method and a channel normalization method for minimizing the loss of speaker characteristics in the cepstrum were proposed for a robust speaker recognition system over the channel. The proposed extraction method creates a cepstrum based on the pitch synchronous analysis using the inherent pitch of the speaker. Therefore, the cepstrum called the "pitch synchronous cepstrum" (PSC) represents the impulse response of the vocal tract more accurately in voiced speech. And the PSC can compensate for channel distortion because the pitch is more robust in a channel environment than the spectrum of speech. And the proposed channel normalization method, the "formant-broadened pitch synchronous CMS" (FBPSCMS), applies the Formant-Broadened CMS to the PSC and improves the accuracy of the intraframe processing. We compared the text-independent closed-set speaker identification on 56 females and 112 males using TIMIT and NTIMIT database, respectively. The results show that pitch synchronous LPCC improves the error reduction rate by up to 7.7% in comparison with conventional short-time cepstrum and the error rates of the FBPSCMS are more stable and lower than those of pole-filtered CMS.
Keywords: *Feature extraction, Pitch synchronous analysis, Cepstrum, Channel robustness, Speaker recognition*
ASK subject classification: *Speech signal processing (2.5)*

I. 서 론

음성을 통한 화자 (話者, speaker) 인식 기술은 일반적으로 음성 인식 기술로 대표되는 어의 (語義, lexical meaning) 인식 기술과 더불어 쉽고 편리한 인터페이스 기술을 위해 함께 발전해왔다. 그러나 기본적으로 음성 인식 기술이 음성의 화자내 (intraspeaker) 변이와 화자간 (interspeaker) 변이를 모두 최소화하려는 반면, 화자 인식은 화자내 변이를 최소화하지만 화자간 변이는 최대

책임저자: 김유진 (egkim@ieee.org)
402-751 인천광역시 남구 용현동 253
인하대학교 전자공학과
(전화: 032-860-7420; 팩스: 032-868-3654)

화시켜야 한다는 점에서 음성 인식 기술과 다르다.

지금까지의 화자 인식을 위한 음성 특징 및 분석 방법은 대부분 단구간 (short-time) 분석법에 기반을 두고 있다. 단구간 분석법은 10~30 msec의 짧은 구간의 음성이 비교적 안정된 특성을 보인다고 가정하고 고정된 분석 율과 분석 길이를 적용한다. 따라서 단구간 분석법은 문맥 또는 화자에 따라 변화하는 피치를 반영하지 못하고, 고정된 길이의 피치를 가정하는 일률적인 분석의 결과를 낳게 된다. 즉 문맥에 따라 전이구간 또는 복수의 유성음 구간을 포함하거나, 화자와 성별에 따라 음향학적 단위의 빠른 변화를 포함하는 구간에 대해서 모두 동일하게 평탄화된 스펙트럼 정보로서 표현하게 된다.

음성 내에서 문맥과 화자의 특성을 고려한 음향학적 단위를 표현할 수 있는 잘 알려진 기준은 기본 주파수 또는 피치로 불리는 F_0라고 할 수 있다. 피치는 화자의 특징을 가장 많이 포함하는 동시에 시간 축에서는 파형의 기본 단위이며 주파수축에서는 스펙트럼의 기본 주파수이기 때문이다. 결국 피치 단위의 분석, 즉 피치 동기 분석은 가장 기본적인 음향학적인 단위에 대한 세밀한 분석과 화자의 특성에 종속적인 분석을 가능케 한다. 피치 동기 분석은 지금까지 음성 합성 또는 코딩 분야에서 보다 정확한 음성 재생을 위해 주로 사용되어왔다[1,2]. 그림 1은 남자와 여자의 20 msec 길이의 유성음 구간에 대한 스펙트럼을 나타낸 것이다. 그림에서 연속된 피치 구간에 대한 스펙트럼이 변화되는 것으로 나타났으며, 단구간 스펙트럼은 피치 구간 스펙트라의 평균된 형태인 것으로 나타났다. 결과적으로 단구간 분석에 의한 스펙트럼은 짧은 음성 구간에서 변화하는 화자의 특징을 효과적으로 추출하지 못하며, 특히 화자의 정보를 많이 포함한 유성음 구간에서의 포먼트 정보를 정확하게 표현하지 못하는 단점을 확인할 수 있다.

단구간 분석법에 기반을 둔 초기의 화자 인식 시스템은 화자 고유의 특징에 대한 장시간 평균을 음성 특징으로 사용하였으며, 1969년 Luck은 현재의 음성 인식 및 화자 인식 시스템에서 가장 많이 사용되는 대표적인 특징인 켑스트럼을 최초로 화자 인식에 적용하였다[3]. 이후 화자의 특징을 가장 잘 나타낸다고 알려진 4개의 포먼트에 대한 정보는 정확한 검출의 기술적 어려움에도 불구하고 화자 인식을 위한 가장 뚜렷한 특징으로 알려졌다[4]. Atal의 1972년 연구는 기본 주파수 또는 피치를 단구간에 대해 검출하고 사용한 최초의 연구이다[5]. 한편 피치는 스펙트럼에 비해 채널 영향에 강인한 특성을 보이지만, 단독으로 사용되었을 경우 사칭자의 흉내에 약하고 화자 내 변이를 고려해야하는 단점을 가진 것으로 나타났다[5-7]. 1975년 Rosenberg의 연구에서는 포먼트 정보를 표현하기 위해 선형 예측 코딩 (LPC; Linear Predictive Coding)에 의한 전극점 성도 모델을 사용하였으며 피치와 이득 등을 조합하여 음성 특징을 구성하였다[8]. LPC는 단순하고 정확한 성도 모델링 특성을 바탕으로 음성 신호 처리 분야에 널리 적용되었고, 1976년 Atal은 LPC로부터 변환된 각종 음성 특징들 가운데 LPCC (Linear Predictive Cepstral Coefficient)가 가장 우수한 화자 인식 성능을 나타냄을 보여주었다[9]. 한편 1994년 Reynolds는 실제 전화선 환경에서의 음성 특징의 성능을 평가하기 위해 기존의 LPCC와 함께 음성 인식에서 주목할만한 성능을 보여준 MFCC (Mel-Frequency Cepstral Coeffi-

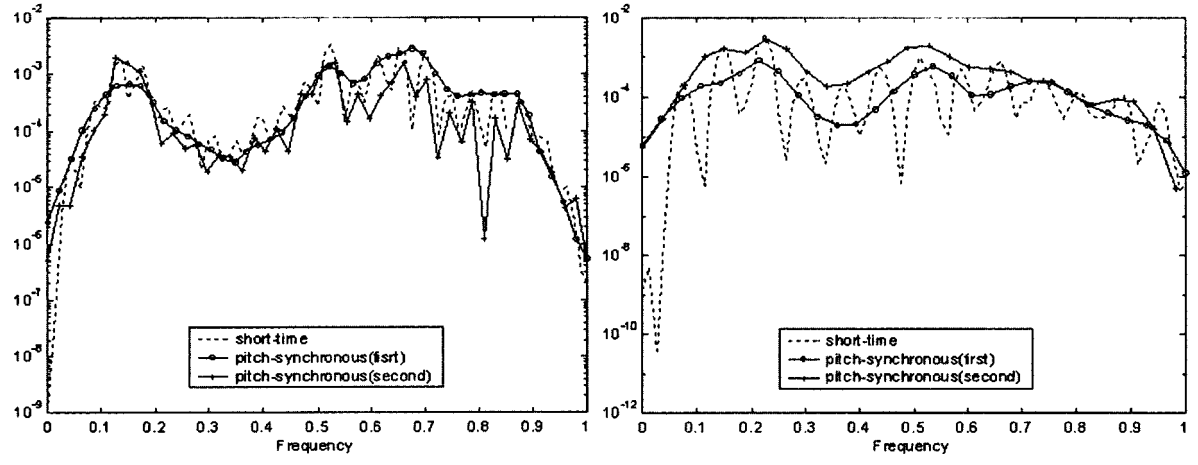

그림 1. 단구간과 피치구간의 스펙트럼 비교 (좌: 남성, 우: 여성) (점선: 20 ms 단구간 스펙트럼, 실선: 연속된 피치구간의 스펙트럼)
Fig. 1. Comparison between spectra of short-time and adjacent pitch periods (Top:male, Bottom: female) (dotted: short-time of 20 ms, solid: adjacent pitch periods).

cient)를 비교하였다. 이 연구 결과는 전화선 대역폭을 고려한 대역제한 MFCC가 가장 우수함을 보여주었고 그 후 화자 인식 시스템에서 지속적으로 사용되었다[10-12]. 현재 사용되는 MFCC는 고속 푸리에 변환 (FFT; Fast Fourier Transform) 기반의 주파수축 스펙트럼 분석 방법이므로 피치와 포먼트 정보를 함께 표현할 수 있는 장점을 가진다.

화자 인식 분야에서 기존 단구간 분석의 한계를 극복하고자 했던 연구는 이미 몇 차례로 시도되었다. 1995년 Jankowsky는 피치 동기 분석을 바탕으로 포먼트의 '미세구조 (fine-structure)'를 표현한 포먼트 AM, FM를 제안하였다. 그러나 제안된 특징만으로는 향상된 결과를 제시하지 못했으며 기존 MFCC와 결합하여 부분적인 향상을 보여주었다. 또한 2000년에 Ezzadi는 피치 동기 분석에 의해 포먼트의 포락선과 순시 주파수 (instantaneous frequency)를 추출하였으며, 역시 기존 MFCC와 결합하여 성능이 향상됨을 제시하였다. 결과적으로 이들 연구는 피치 동기 분석에 기반을 두어 제안된 특징들이 단구간 분석법에 의한 스펙트랄 포락선을 표현하는 기존 MFCC를 능가하지 못하고, 기존 켑스트럼의 정보가 매우 효과적임을 역설적으로 보여주었다[13,14].

본 논문에서는 화자 인식에서 새로운 음성 특징을 제안하기보다는, 기존 단구간 분석이 지닌 한계를 극복하고 음성 특징으로서 켑스트럼의 효과를 유지할 수 있는 피치 동기 켑스트럼 추출 방법을 제안하였다. 제안된 추출 방법은 문맥과 화자에 따라 변화하는 피치를 효과적으로 검출하기 위하여 제안된 2단계 피치 검출 알고리즘과 검출된 피치 구간에 종속적이고 효과적인 켑스트럼 추출하기 위한 방법으로 구성된다. 또한 본 논문에서는 채널 환경에 대한 강인함을 특징에 추가하기 위하여 제안된 피치 동기 켑스트럼과 채널 정규화 방법으로서 개선된 포먼트 평활화 켑스트랄 평균 차감법 (Formant-Broadened CMS; FBCMS)을 결합하였다.

본 논문의 구성은 다음과 같다. 2장에서는 제안된 피치 동기 분석의 기반 알고리즘인 Medan의 유사도 모델을 요약하고, 3장에서는 이를 이용한 2단계 피치 검출 방법을 설명한다. 4장에서는 피치 동기된 음성 구간의 분석 방법 및 특징 추출에 관하여 설명하고, 5장에서는 수정된 포먼트 평활화 켑스트랄 평균 차감법을 설명한다. 제안된 피치 동기 분석의 인식 실험 및 그 결과를 6장에서 분석하며 마지막으로 7장에서 결과 및 향후 연구 방향을 제시한다.

II. 피치 검출을 위한 유사도 모델

음성으로부터 보다 정확한 피치, T_0 또는 기본 주파수, F_0를 검출하기 위한 노력은 오랫동안 계속되어 왔다. 전통적인 피치 검출 방법으로서 선형 예측 분석을 통한 잔여 신호의 자기 상관 계수값을 이용한 SIFT (Simple Inverse Filter Tracking) 방법부터 웨이블릿을 이용한 주파수 분석법 등 다양한 방법이 제안되었다[15].

본 연구에서는 피치 검출 알고리즘의 정확도에 대한 평가 및 인식 성능의 오차를 줄이려는 목적에서 기존에 제안된 알고리즘 가운데 비교적 적은 연산량으로 정확한 피치 검출이 가능한 방법을 선택하였다. Medan에 의해 제안된 상호 상관도를 이용한 피치 검출 알고리즘은 잡음 환경에서도 다른 방법에 비해 상대적으로 우세한 성능을 나타내었으며, 복잡도의 측면에서도 SIFT방법과 웨이블릿을 이용한 주파수 분석법에 비해 높지 않아 인식 시스템의 전처리부에 결합되었을 때 연산량을 크게 증가시키지 않는 것으로 나타났다[16].

3장의 피치 동기 분석 방법의 전개를 돕기 위해 간단히 Medan의 피치 검출 알고리즘의 기본적인 수식을 요약하고자 한다.

2.1. 정규화된 상호 상관도 및 피치 검출

Medan은 유성음 구간의 준 주기적인 패턴, 즉 피치 구간을 정규화된 상호 상관도에 의한 유사도 모델을 통해 검출할 수 있음을 보여주었다. 음성 신호 $s[n]$의 샘플 $n=n_0$에서 피치 검출을 위한 인접한 길이 l의 음성 신호 $x_l[n, n_0]$과 $y_l[n, n_0]$는 다음과 같이 정의되며

$$x_l[n, n_0] = s[n] \cdot w_l[n-n_0], \quad n_0 \le n_0 + l \tag{1}$$

$$y_l[n, n_0] = s[n] \cdot w_l[n-n_0-l], \quad n_0 + l \le n_0 + 2l \tag{2}$$

이때 창함수 $w_l[n]=1,\ 0 \le n < l$이고 그 외의 n에 대해서는 $w_l[n]=0$이다. 유사도 모델은 음성 신호의 길이 l에 대한 정규화된 상호 상관도 ρ_l이며 다음과 같이 정의된다.

$$P = \arg\max_l \rho_l(x_l[n, n_0], y_l[n, n_0]) \tag{3}$$

$$\rho_l(x_l[n, n_0] y_l[n, n_0]) = \frac{C_l(x_l[n, n_0], y_l[n, n_0])}{\sqrt{R_l(x_l[n, n_0]) R_l(y_l[n, n_0])}} \tag{4}$$

$$C_l(x[n], y[n]) = \sum_{n=0}^{l-1} x[n] \cdot y[n] \quad (5)$$

$$R_l(x[n]) = C_l(x[n], x[n]) = \sum_{n=0}^{l-1} x^2[n] \quad (6)$$

유사도 모델은 음성 신호의 길이 l이 피치와 일치할 때 상호 상관도가 최대값이 되도록 한다. 다시 말해 유성음 구간 내에서 만약 피치가 l_P이고 연속된 값 l_P-1 그리고 l_P+1에 대해 사각 창함수를 적용하여 계산한 상호 상관도가 각각 ρ_{l_P}, $\rho_{l_{P-1}}$ 그리고 $\rho_{l_{P+1}}$ 이라면 다음과 같은 관계를 가진다.

$$\rho_{l_P} > \rho_{l_{P-1}} \quad (7a)$$

$$\rho_{l_P} > \rho_{l_{P+1}} \quad (7b)$$

$$\rho_{l_P} > \mathrm{TH} \quad (7c)$$

이때 TH는 검출된 피치의 신뢰도를 결정하는 값으로서 무성음 구간인 경우 문턱값보다 낮은 값을 나타내고 유성음 구간인 경우 문턱값보다는 큰 값을 가진다. 한편 식 (3) 및 (4) 계산은 최소 피치 길이 $P_{\min}$부터 최대 피치 길이 $P_{\max}$까지 길이 l을 변화시키며 수행된다.

2.2. 검출된 피치 길이의 보정

유성음 구간에서 식 (7)과 같은 관계를 만족시키는 값은 2개 이상 검출될 수 있다. 즉 피치 구간의 정수 배에 해당하는 조화 주파수에서도 식 (7)의 관계를 만족할 수 있다. 또는 피치 구간 내에서 유사한 패턴이 발견되면 피치 구간보다 작은 구간에서 큰 상호 상관도 값을 가지는 경우도 발생한다. 따라서 여러 개의 피치 후보들 중에서 정확한 피치를 결정하는 과정이 필요하게 된다.

우선 여러 개의 피치 후보들 가운데 상호 상관도값이 가장 큰 값을 가지는 구간을 피치로 선택한 다음 결정된 피치를 보정함으로써 정확한 피치를 얻을 수 있다. 선택한 피치가 정확하게 검출된 피치라면 정수 배에 해당하는 구간의 상호 상관도도 문턱값보다 큰 값을 갖게 된다. 즉 단구간 내에서 가능한 모든 정수 배의 구간을 대상으로 상호 상관도 값을 조사하면 문턱값보다 큰 상호 상관도 값을 나타낸다. 문턱값보다 작은 상호 상관도값을 가지는 정수 배의 구간이 발생하면 선택한 피치 구간이 실제 피치 구간의 정수 배인 경우이거나 실제 피치 구간보다 작은 구간을 피치 구간으로 선택한 것이다. 따라서 다른 피치 후보 값들 중에서 하나를 선택하여 다시 확인한다. 만약 피치 길이의 k 배에 해당하는 길이가 피치로 결정되었다면 현재 피치의 $1/k$ 배에 해당하는 길이의 상호 상관도값도 문턱값보다 큰 값을 가진다.

III. 피치 동기 분석을 위한 피치 검출

피치 동기 분석의 목적은 피치를 정확하게 검출하여 피치에 종속적인 가변 길이의 특징 추출을 위한 분석창을 구성하는 것이다. 그러나 일반적으로 피치는 화자 및 문맥에 따라 최대 25% 정도의 변이를 보이는 것으로 알려졌으며[16], 묵음 및 무성음 구간에서는 이론적으로 존재하지 않는다. 따라서 피치 동기 분석법은 다양한 문맥과 화자의 피치 변화에 동적으로 대응하고, 특징 추출의 전 단계에서의 연산량 부담을 최소화하는 알고리즘이어야 한다.

본 논문에서는 이러한 목적을 이루기 위해 최대피치를 고려한 주요 피치 검출과 세부 피치 검출의 2단계로 구성된 알고리즘을 제안하였다.

3.1 주요 피치 검출 (Principal Pitch Detection)

피치 검출을 위한 식 (3)은 $P_{\min}$과 $P_{\max}$의 범위에 따라서 피치의 신뢰도와 연산량에 큰 영향을 미치므로 정확하고 화자와 문맥에 적응적인 피치 검출이 반드시 필요하다. 그러나 식 (3)을 유성음이 시작되는 전이구간에 적용하면, 식 (7c)를 만족하는 피치를 검출하더라도 안정된 값으로 사용하기 어렵다. 따라서 유성음의 안정 구간에 이르기 전까지 항상 최대 피치를 고려한 검출을 수행해야 한다. 그리고 묵음 또는 무성음 후의 피치의 급격한 변화가 일어나는 구간인 경우에도, 기존의 피치에 적응적인 $P_{\min}$과 $P_{\max}$의 설정은 오히려 정확한 피치 검출을 제한할 수 있다. 또한 묵음 또는 무성음 구간이 길 경우, 피치 검출을 위한 분석창 이동은 검출된 피치 길이에 의존적이어야 하므로 불필요한 연산을 수행해야 한다. 이러한 문제점은 짧은 피치의 음성에서 더욱 두드러지게 된다.

이러한 문제점들을 해결하기 위해 최대 피치의 2배를 고려한 길이의 구간에서 $P_{\min} \le l \le P_{\max}$의 길이인 주요 피치를 검출한다. 따라서 주요 피치 P_0는 다음 식에 의해서 결정된다.

$$P_0 = \arg\max_l \rho_{P_{\max}}(x_{P_{\max}}[n, n_0], x_{P_{\max}}[n, n_0 + l]) \quad (8)$$

$$\alpha = \max_l \rho_{P_{\max}}(x_{P_{\max}}[n, n_0], x_{P_{\max}}[n, n_0 + l]) \quad (9)$$

식 (8) 및 (9) 계산은 $P_{\max}$ 길이의 두 신호 $x_{P_{\max}}$가 인

접한 경계 지점을 기준으로 최소 피치 길이 P_{min}부터 최대 피치 길이 P_{max}까지 길이 l을 통해 중첩 구간을 변화시키며 수행된다. 식 (3)는 상호 상관도의 계산 구간이 변수 l에 의해 변화되지만, 식 (8)은 계산 구간이 P_{max}로 고정되고 인접된 피치 구간의 중첩 길이가 l에 의해 변화된다. 본 연구에서는 P_{min}과 P_{max} 값을 각각 2.5 msec, 20 msec 로 정하였다. 그림 2는 식 (8) 및 (9)의 계산 과정을 나타낸 그림이다.

따라서 주요 피치 검출 알고리즘은 20 msec의 최대 피치의 두 배에 해당하는 40 msec의 구간에 대해서 수행된다. 40 msec의 길이는 전이 구간과 함께 안정 구간을 포함할 수 있는 충분한 길이이므로 비교적 안정된 피치를 검출할 수 있으며, 항상 최대 피치 20 msec를 고려하므로 피치의 급격한 변화에도 적응할 수 있다. 검출된 피치는 식 (3)의 P_{min}과 P_{max}를 설정하기 위해 사용되며, 주요 피치로부터 큰 변화가 없다고 가정되는 정확한 세부 피치를 검출한다. 또한 무성음 구간인 경우 무성음에 대한 특징 분석을 수행한 후 다시 새로운 구간에 대한 주요 피치 검출을 수행한다. 결국 주요 피치 검출은 안정된 피치를 검출할 수 있는 길이에 대해 묵음 또는 무성음 구간으로부터 안정된 유성음 구간을 찾을 때까지 수행된다.

유무성음 구간의 판단은 정규화된 상호 상관도값 α에 따라 결정된다. 본 연구에서는 문턱값 0.55와 비교하여 낮을 경우 검출된 피치는 무성음 구간에서 검출된 것으로 판단하였다. 문턱값보다 클 경우 유성음 구간으로 판단하고 검출된 피치를 이용하여 세부 피치를 검출한다.

3.2. 세부 피치 검출 (Particular Pitch Detection)

세부 피치는 주요 피치를 기준으로 검출된 정확한 피치를 말한다. 피치가 짧은 화자, 특히 여자 화자의 경우에는 3~4 msec의 짧은 피치를 가지므로, 검출된 주요 피치는 분석창내에 10회이상 반복되면서 피치가 변화를 보인다. 짧은 피치에서의 변이는 특징 분석을 위한 길이에 큰 영향을 주기 때문에, 정확한 세부 피치 검출이 반드시 요구된다.

세부 피치 검출은 식 (3)을 적용하며 검출된 피치와 상호 상관도는 다음과 같이 정의될 수 있다.

$$P_k = \arg\max_{l_k} \rho_{l_k}(x_{l_k}[n, n_k], y_{l_k}[n, n_k]) \qquad (10)$$

$$\beta_k = \max_{l_k} \rho_{l_k}(x_{l_k}[n, n_k], y_{l_k}[n, n_k]) \qquad (11)$$

이때 $P_{k-1} - \delta \le l_k \le P_{k-1} + \delta$ 이고 $n_k = n_{k-1} +$

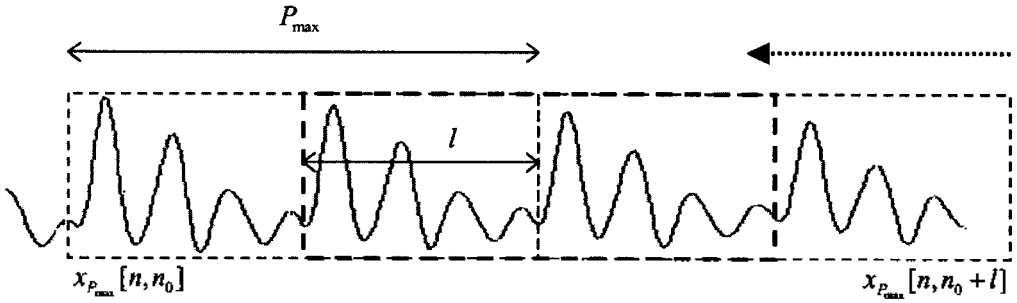

그림 2. 주요 피치의 검출
Fig. 2. Detection of principal pitch.

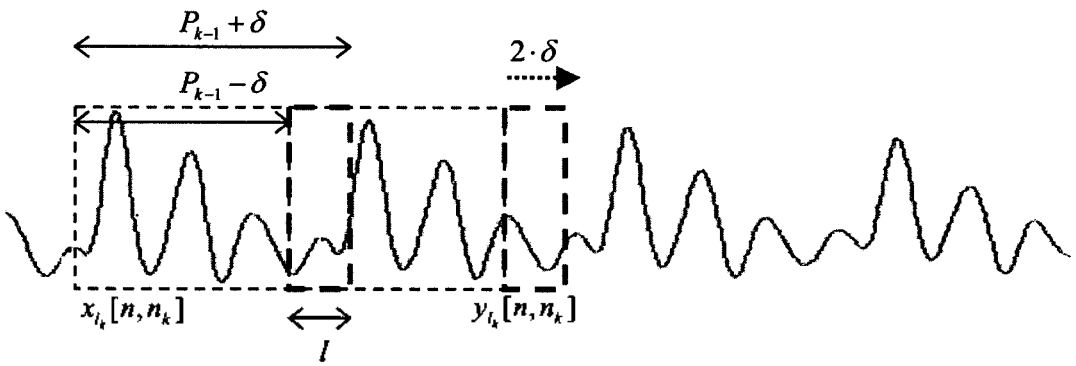

그림 3. 세부 피치의 검출
Fig. 3. Detection of particular pitch.

P_{k-1}, $k=1\cdots K$이다. 식 (10) 및 (11)의 계산은 $n=n_k$인 위치를 기준으로 주요 피치의 변이를 고려하여 l_k를 변화시키며 수행된다. 본 연구에서는 δ를 5 샘플로 정하였다. 그림 3은 식 (10) 및 (11)의 계산 과정을 나타낸 그림이다.

한편 β는 검출된 피치의 신뢰도를 결정하며 동시에 연속된 피치가 더 이상 나타나지 않는 무성음 구간으로의 전이를 검출하기 위해 사용된다. 본 연구에서는 문턱값 0.4와 비교하여 클 경우 지속적인 유성음 구간으로 판단하고 검출된 피치를 이용하여 특징 추출을 위한 분석 프레임을 구성하고, 다음 세부 피치 검출을 위해 n_k와 l_k를 결정한다. 문턱값보다 낮을 경우 검출 구간은 무성음 구간으로의 전이 구간으로 판단하고 다시 유성음 구간을 찾기 위해 주요 피치 검출 과정으로 돌아간다.

IV. 특징 추출을 위한 피치 동기 분석

검출된 피치를 이용한 특징 추출을 위한 분석창은 유성음과 무성음 구간에 따라 달리 구성된다. 우선 검출된 피치는 평탄화 함수를 거친 평균 피치와 비교하게 되고, 검출 오류를 보정한다. 유성음 구간인 경우 피치에 동기된 분석창을 구성하게 되며 검출된 피치에 근거하여 연속된 세부 피치 검출과 특징 추출 분석을 반복한다. 무성음 구간의 경우 고정 길이 창함수를 적용하여 특징을 추출하며, 고정 율에 따라 분석창은 이동된다.

4.1. 유성음 구간에서의 피치 동기 분석창 구성

세부 피치가 검출된 유성음 구간인 경우, 특징 분석을 위한 분석창은 피치 단위로 이동하며 피치 길이로 구성된다. 하지만 이러한 피치 동기된 분석을 위해서는 두 가지의 문제점을 해결해야 한다.

첫째, 일반적으로 음성 분석을 위한 사각 창함수는 주파수 응답 특성에서 나타나는 측엽 (側葉, sidelobe)의 이득 때문에 스펙트랄 누설 (spectral leakage)을 발생시킨다. 이를 완화시키기 위해 해밍창을 사용할 수 있다. 하지만 길이 M의 해밍창의 주파수 해상도는 $8\pi/M$로서, 주파수 해상도가 $4\pi/(M+1)$인 사각 창함수의 약 2배에 해당한다. 따라서 사각 창함수와 동일한 주파수 해상도를 유지하기 위해 피치 길이의 2배에 해당하는 창함수를 사용해야한다. 결국 한 주기의 피치구간에 대한 주파수 해상도를 유지하기 위해서는 두주기의 피치구간에 해밍창을 씌운 후 분석해야 한다. 이를 위해서는 동일한 피치를 반복 사용하거나, 두 주기에 해당하는 연속된 피치를 사용한다.

둘째, 실제 피치는 짧은 시간 내에서도 변화하므로 연속된 두 주기의 피치를 고려할 때 변화를 지속적으로 고려해야 한다. 따라서 두 주기 피치구간의 분석창을 구성하기 위해서는 두 번째 피치를 검출한 후, 첫 번째 피치와 연결시켜야 한다.

따라서 유성음 구간의 $n=n_k$에서 특징 추출을 위한 분석창 $v[n]$은 다음과 같이 정의된다.

$$v[n] = (x_{P_k}[n, n_{k-1}] + y_{P_{k-1}}[n, n_k]) \cdot w^H_{P_k+P_{k-1}}[n-n_{k-1}] \quad (12)$$

이때 창함수 $w^H_l[n]=0.54-0.46\cos(2\pi n/M)$, $0\le n\le l$이고 그 외의 n에 대해서는 $w^H_l[n]=0$ 이다. 한편 다음 특징 프레임을 위한 분석창은 첫 번째 피치 P_k만큼 이동하여 구성된다. 그림 4는 식 (12)을 나타낸 그림이다. 무성음 구간이 시작되면 특징 추출을 위한 분석창은 기존의 단구간 분석 방법을 따르며 다음과 같이 정의된다.

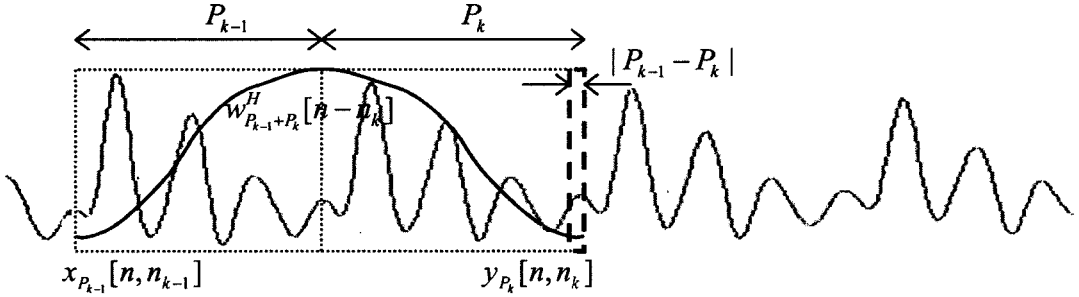

그림 4. 특징 추출을 위한 분석창 구성
Fig. 4. A frame construction for the feature extraction.

$$v[n] = s[n] \cdot w_L''[n - n_k] \quad (13)$$

이때 L은 단구간 분석을 위한 분석창 길이이며 다음 특징 프레임을 위한 분석창은 고정된 값 R만큼 이동하여 구성된다.

4.2. 특징 추출

기존 단구간 분석법과 피치 동기 분석법은 분석창 이동율과 길이의 가변성에서 기본적으로 다르다. 이에 따른 기존 특징 추출 방법에서 큰 변화들은 다음과 같다.

첫째, 분석창의 길이가 가변적이므로 창함수는 항상 새롭게 적용되어야 하며, 분석창의 이동률 또한 가변적이므로 동일한 길이의 음성 신호로부터 다른 길이의 특징 벡터가 구성된다. 따라서 동일한 문장을 동일한 길이로 발성해도 피치 변화에 따라 특징 벡터의 길이가 달라질 수 있다. 이는 구현시 매번 창함수를 계산해야하는 단점이 있지만, 화자 고유의 특징인 피치에 종속적인 특징을 추출하는 효과를 얻을 수 있다.

둘째, 두주기의 피치 구간을 고려할 때, 여자 화자의 경우 최소 5~6.25 msec (40~50 샘플)의 매우 짧은 분석 길이를 가지게 된다. 따라서 FFT기반의 스펙트럼은 물론 LPC기반의 스펙트럼 추정이 부정확해지는 문제점이 발생한다. 이를 해결하기 위해서 분석창의 길이가 지나치게 짧은 경우 피치 구간을 추가함으로써, 이른바 다중 피치 분석창을 구성하여 보완하였다.

셋째, 화자 인식을 위한 대표적인 특징인 켑스트럼은 LPC에 의한 전극점 모델로부터 유도되는 LPCC와 FFT기반의 로그 스펙트럼으로부터 유도되는 MFCC으로 나눌 수 있다. 가변 길이 분석창에 의해서 LPC 기반의 LPCC는 큰 변화가 없지만, MFCC는 가변 길이에 따라 FFT의 크기가 변화하고, 그에 따른 필터뱅크가 바뀌어야한다. 특히 FFT의 크기가 달라짐으로써, 각 특징 프레임의 주파수 분석 해상도가 일정하지 않은 문제가 발생된다. 또한 FFT의 크기를 고정하더라도 각 프레임마다 다른 길이의 음성 신호를 분석함으로써 비슷한 결과를 초래한다. 이러한 문제에 대한 가장 이상적인 해결책은 이산 푸리에 변환 (DFT; Discrete Fourier Transform)을 적용한 후 동일한 해상도로 표본화 (decimation)하거나 보간 (interpolation)하는 것이지만 연산량이 크게 증가하는 단점이 있다[2]. 또한 LPC기반의 분석에서 쌍일차 (bilinear) 변환을 적용한 Mel-LP 켑스트럼을 사용함으로써, FFT를 사용하지 않고 Mel 주파수 분석의 효과를 거둘 수 있다. 그러나 구현 방법의 차이에 의해서 동일한 효과를 얻을 수는 없는 것으로 알려져 있다[17]. 본 연구에서는 다중 피치 분석창을 구성하여 가변 길이에 대한 영향을 최소화하고 불가피한 경우 필터뱅크를 새롭게 적용하였다.

V. 포먼트 평활화 피치 동기 켑스트랄 평균 차감법

켑스트랄 평균 차감법 (Cepstral Mean Subtraction; CMS)은 채널 환경의 음성으로부터 채널 영향을 정규화하기 위해 널리 사용되는 방법이다. 그러나 CMS에 의한 켑스트랄 평균 성분은 유성음 구간의 포먼트 성분에 의해 편향되며, 화자 인식의 경우 화자 정보가 손실되는 것으로 나타났다[18]. Naik에 의해 제안된 PFCMS는 선형 예측 코딩에 의한 전극점 모델의 극점을 필터링한 후 평균함으로써 채널 성분에서 포먼트 성분을 감쇄시키는 방법이다. 그러나 이 방법은 극점의 차수에 해당하는 약 10차 이상의 다항식을 계산해야하는 단점을 가진다. 연산량을 줄이는 방법으로서 전극점 모델의 스펙트럼 피크를 일률적으로 평탄화시키는 방법이 제안되었으며, 켑스트럼에 직접 적용할 수 있는 간단한 구현 방법에도 불구하고 포먼트 성분이 감쇄된 채널 성분 스펙트럼을 효과적으로 평탄화하여 인식률을 향상시키는 것으로 나타났다[19].

최근의 연구에서는 켑스트럼으로부터 포먼트 정보를 직접 얻어내고 이에 해당하는 극점만을 필터링함으로써, Naik의 PFCMS와 비교할만한 결과를 얻을 수 있는 포먼트 평활화 켑스트럴 평균 차감법이 제안되었다. FBCMS는 켑스트럼으로부터 변환된 로그 스펙트럼에서 포먼트 위치를 쉽게 찾을 수 있고, 포먼트는 전극점 모델로 표현되는 성도 전달 함수의 우세 극점에 대응된다는 사실에 근거한다. 포먼트에 해당하는 우세 극점만으로 선택적으로 처리할 경우, 인수 분해의 연산량을 줄일 수 있다. 그러나 FBCMS는 포먼트 성분이 존재하는 유성음 구간을 검출하지 않고 적용함으로써 무성음 구간의 우세 극점들도 동일하게 처리되는 단점을 가진다. 또한 평탄화된 로그 스펙트럼으로부터의 극점 추정 방법은 부정확한 포먼트 검출의 가능성이 있다[20].

따라서 본 논문에서는 좀 더 효과적인 포먼트 성분의 검출 및 채널 켑스트럼의 추정을 위해서 피치 동기 켑스트럼을 이용하는 개선된 FBCMS를 제안하였다. 피치 동기 켑스트럼과 FBCMS가 결합된 포먼트 평활화 피치 동기 (Formant-Broadened Pitch Synchronous) CMS는 다음과 같은 효과를 얻을 수 있다. 첫째 피치 동기 분석의

유무성음 정보를 이용하여 유성음 구간에서만 포먼트를 평탄화시킴으로써 장구간 평균 켑스트럼에 나타나는 포먼트 성분에 의한 편향을 효과적으로 감쇄시킬 수 있다. 둘째 피치 동기 분석을 통해 유성음의 성도 모델 추정이 정확해지고, 따라서 포먼트 추정 및 평탄화가 정확해 진다. 마지막으로 추정된 근의 오류 등으로 인한 불완전한 스펙트럼의 평탄화를 간단한 PFCMS-γ 방법에 의해 보완할 수 있다.

일반적으로 선형 예측 시스템의 선형 예측 계수 a_n들을 변환한 켑스트럼은

$$\hat{c}[n] = a_n + \sum_{k=1}^{K} \left(\frac{k}{n}\right) \hat{c}[k] a_{n-k} \quad , \quad 1 \le n \qquad (14)$$

와 같이 나타낼 수 있다. 이때 유성음 구간의 켑스트럼으로부터 역 푸리에 변환에 의하여 구해진 로그 스펙트럼에서 포먼트로서 추정된 k번째 포먼트의 대역폭, 주파수가 각각 $\hat{B}_k$와 $\hat{\omega}_k$이라면 대응하는 근과 협대역 문턱값 B_{TH}에 의해 대역확장 된 후의 근들은 다음과 같이 정의된다.

$$\hat{z}_k = |e^{-\pi \hat{B}_k}| e^{j\hat{\omega}_k} \qquad (15)$$

$$\tilde{z}_k = |e^{-\pi B_{TH}}| e^{j\hat{\omega}_k} \qquad (16)$$

그리고 이 근들을 이용한 포먼트 평활화된 켑스트럼은

$$\tilde{c}[n] = \hat{c}[n] + \frac{1}{n}\left(\sum_{k=1}^{K} \tilde{z}_k^{\,n} - \sum_{k=1}^{K} \hat{z}_k^{\,n} \right) , \quad 1 \le n \qquad (17)$$

으로 정의된다. 한편 무성음 구간인 경우 켑스트럼은 직접 스펙트럼 평탄화를 적용한다. 결국 유무성음 정보를 고려한 FBPSCMS에 의해 처리되는 켑스트럼은 다음과 같이 표현할 수 있으며

$$\bar{c}[n] = \begin{cases} \gamma_v^{\,n} \cdot \tilde{c}_n & , \quad \text{voiced} \\ \gamma_{uv}^{\,n} \cdot \hat{c}_n & , \quad \text{unvoiced} \end{cases} \qquad (18)$$

이때 평탄화 상수 γ는 유무성음에 따라 달리 설정될 수 있다.

그림 5는 지금까지의 과정을 블록선도와 순서도로 나타낸 것이다. 기존 FBCMS 과정에서 유무성음 판단 과정과 스펙트럼 평탄화 과정이 추가되었다. 무성음인 경우 켑스트럼에 직접 적용하는 스펙트럼 평탄화 과정만을 수행하므로 기존 FBCC에 비해서 연산량이 크게 감소된다. 유성음 구간에서는 기존 FBCC의 과정을 거친 후, 스펙트럼 평탄화를 통해서 포먼트 검출 오류에 의한 부정확성을 줄이는 효과를 얻을 수 있다. 한편 그림 6은 유성음 구간에서 FBPSCMS에 의한 켑스트럼과 기존 채널 정규화 방

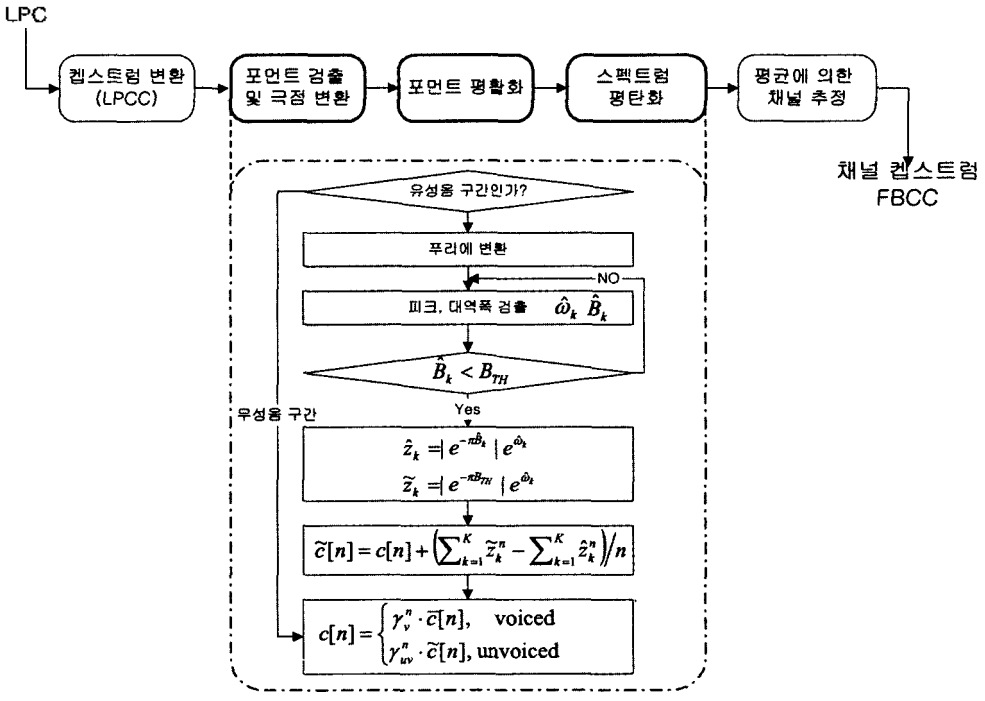

그림 5. 포먼트 평활화 피치 동기 켑스트랄 차감법의 블록 선도 및 순서도
Fig. 5. Block-diagram and flowchart of FBCMS.

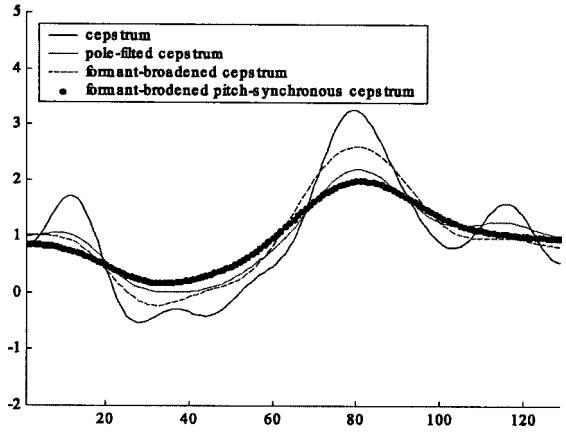

그림 6. 채널 정규화된 켑스트럼의 주파수 응답 비교
Fig. 6. Frequency response comparison between channel normalized cepstra.

법에 의한 스펙트럼을 비교한 것이다. 극점 필터링된 켑스트럼과 비교할 때, 세 번째 포먼트에서의 감쇄가 두드러지고 전체적으로 향상된 평탄화 결과를 보이는 것을 확인할 수 있다.

VI. 실험 및 결과

제안된 피치 동기 분석 및 특징 추출에 의한 화자 인식 성능을 확인하기 위해 TIMIT 데이터베이스와 이를 실제 전화선 채널에 통과시켜 구축된 NTIMIT 데이터베이스를 이용한 문장 독립 폐집단 화자 식별을 수행하였다. 모든 실험은 비교를 위해 기존 단구간 분석법과 제안된 피치 동기 분석법에 대해서 각각 수행되었으며 또한 각 분석법에 의한 음성 특징 추출 알고리즘의 영향을 고찰하기 위해 널리 사용되는 LPCC와 MFCC를 실험, 비교하였다. 다시 말해 실험은 기본적으로 단구간 분석과의 화자 식별 성능을 비교하는 동시에 채널 환경과 음성 특징에 따른 영향을 고찰하기 위해 구성되었다.

6.1. 특징 추출 및 화자 모델

훈련 및 인식용 데이터는 신뢰할 수 있는 실험 결과를 얻기 위해 기존의 연구 결과를 바탕으로 구성되었다. 따라서 사용된 음성 데이터는 TIMIT과 NTIMIT의 훈련용 데이터이며, 총 8 지역의 여자 56명과 남자 112명으로서 구성된 총 168명에 대해서 각각 10회의 발성들로 구성되어있다. 각 화자는 2회의 동일한 문장과 8회의 서로 다른 문장을 발성하여 녹음하였으며, 실험에서는 모든 음성을 8 Khz로 다운 샘플링하여 사용하였다. 발성 회수를 변화시키는 실험을 제외한 기본적인 실험에서 동일한 발성 2회를 포함 최대 7회의 발성을 ($168\times7=1176$회) 훈련에 사용하였고, 나머지를 ($168\times3=504$회) 인식에 사용하였다[11].

특징 추출을 위한 단구간 분석은 기본적으로 20 msec의 분석창을 10 msec마다 이동하며 수행되었다. 피치 동기 분석은 유성음 구간에 대해서는 최소 두 주기의 피치를 포함한 20 msec이하의 다중 피치를 선택하고 피치만큼 이동하며 분석되었다. 무성음 구간에서는 기존 단구간 분석을 동일하게 수행하였다. 한편 피치 동기 분석에서 일부 남자의 경우 두 주기 피치의 조건으로 인해 20 msec 이상의 음성 구간을 포함하는 경우도 발생되었다. 음성 특징 추출 알고리즘은 LPC기반의 LPCC와 FFT기반의 MFCC를 사용하였으며 각각 12차 계수를 추출하여 특징 벡터를 구성하였다. 또한 전화선 환경에서의 강인함을 비교하기 위해 일반적인 전화선 채널의 통과대역 300~3400 Hz만을 고려한 대역 제한 MFCC를 추출하여 실험에 적용하였다[11].

한편 피치에 의한 인식률 향상 및 분석법에 따른 영향을 고찰하기 위해 로그 함수를 거친 피치를 기본 특징 벡터에 추가하여 13차 특징 벡터를 구성하였다. 기존 단구간 분석에서는 주요 피치로부터 평탄화된 피치를 사용하였고, 피치 동기 분석에서는 세부 피치를 각각 사용하였다.

화자 식별을 위한 인식 시스템은 HTK 3.1을 기반으로 구성되었으며 각 화자 모델은 32개의 독립된 가우시안 분포로 구성된 GMM으로 표현하였다. 화자 모델은 음성의 레이블 정보를 이용하여 묵음을 제외한 음성 구간만을 선택하여 일반적인 EM (Expectation and Maximization) 알고리즘을 통해 훈련되었다.

6.2. 로그 피치의 영향

그림 7은 2종류의 데이터베이스와 3종류의 음성 특징에 대해서 12차 기본 특징 벡터와 로그 피치를 추가한 13차 특징 벡터를 추출하고 화자 식별을 수행한 결과이다. 화자 모델의 훈련과 인식을 위해 각각 7회와 3회의 발성을 사용하였다. 실험 결과는 기본 12차 음성 특징에 로그 함수를 거친 피치를 추가한 특징 벡터를 사용한 경우 최대 12.2%의 에러 감소율을 보여주었다. 이는 알려진 대로 피치가 화자 고유의 정보를 포함하는 중요한 특징임을 보여주는 결과라고 할 수 있다.

피치 추가 후의 에러 감소율을 나타낸 표 1의 결과에 의하면, NTIMIT보다는 TIMIT, MFCC보다는 LPCC 그리

고 기존 단구간 분석법보다는 피치 동기 분석법에서의 향상 정도가 상대적으로 큰 것으로 나타났다.

데이터베이스에 따른 차이는 TIMIT에서의 피치 검출의 성능이 NTIMIT에 비해서 정확하고 안정되기 때문인 것으로 사료된다.

또한 음성 특징에 따른 차이는 FFT기반의 MFCC가 화자의 피치 정보, 즉 기본 주파수, F_0를 포함하기 때문에 피치 추가 후에 LPCC에 비해 상대적으로 적은 향상을 보이는 것으로 판단된다. 이는 피치를 추가하기전의 에러율에서 MFCC가 LPCC에 비해 TIMIT에서 분석법에 따라 각각 최대 1.7%, 1.5% 낮은 것을 통해 확인할 수 있다. 그러나 NTIMIT에서는 MFCC의 에러율이 오히려 높게 나타난다. 이는 불안정한 채널 잡음을 포함한 대역의 영향으로 판단되며, 채널 잡음 대역을 제외한 MFCC(BLMF)의 성능이 높게 나타난 것을 통해 확인할 수 있다. 그러나 피치 추가 후의 성능은 역시 LPCC가 BLMF에 비해 높은 것으로 나타났는데, 이는 FFT 결과에서 채널의 통과대역만을 고려한 MFCC는 채널 성분과 동시에 화자의 성도의 특징도 재거한 반면, LPCC는 상대적으로 성도 정보를 손실 없이 표현하는 동시에 피치의 정보가 추가되었기 때문으로 사료된다.

마지막으로 피치 추가 후 기존 단구간 분석에 비해 피치 동기 분석의 성능 향상이 큰 이유는 추가된 피치와 음성 특징 분석창의 동기 여부에 의한 것으로 판단된다. 기존 단구간 분석은 최대 피치를 고려한 40 msec의 분석창에서 검출된 주요 피치를 사용하고 일정하게 20 msec의 음성 특징 분석창을 고려하지만, 피치 동기 분석은 인접한 세부 피치를 검출하고 해당하는 길이에 음성 특징 분석창을 동기 시키기 때문이다. 따라서 이러한 결과는 제안된 피치 동기 분석이 피치의 민감한 변화를 검출할 수 있고 그에 따라 보다 정확히 성도 모델을 표현함을 보여준다고 할 수 있다.

6.3. 훈련을 위한 발성 횟수 변화

표 2는 훈련을 위한 발성 횟수를 증가시키면서 각 데이터베이스와 음성 특징에 대해서 피치를 추가한 특징벡터를 추출하고 화자 식별 실험을 수행한 결과이다. 기존 단구간 분석과 제안된 피치 동기 분석의 인식률과 함께 성능 비교를 위해 단구간 분석에 대한 피치 동기 분석의 에러 감소율을 나타내었다. 실험 결과로부터 피치 동기 분석에 의한 성능은 LPCC을 사용한 NTIMIT에서의 실험 결과가 상대적으로 뛰어남을 알 수 있다.

구체적으로 말해서, 전체 발성 횟수에 대한 에러 감소율의 평균에서 LPCC를 사용한 실험은 TIMIT과 NTIMIT에서 각각 5.7%, 7.7%의 단구간 분석법에 대한 향상을 보여주었다. 상대적으로 TIMIT의 MFCC에 의한 결과는

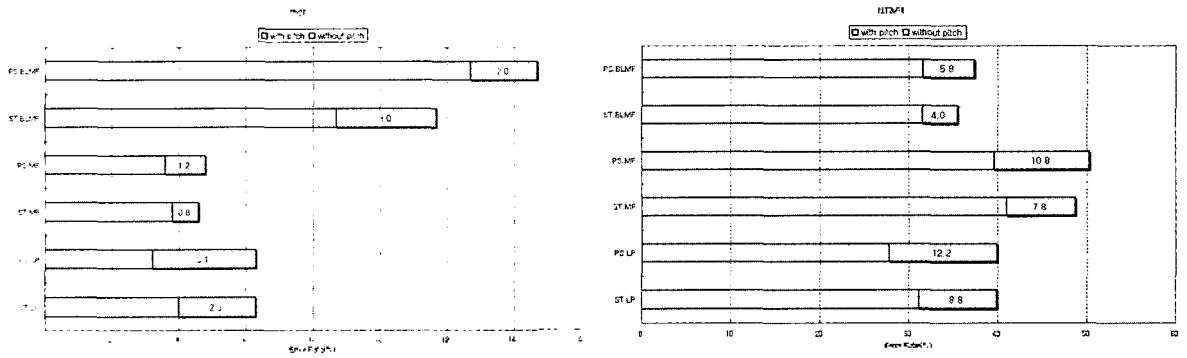

그림 7. 로그 피치에 의한 에러
Fig. 7. Error rates by log pitch.

표 1. 로그 피치에 의한 에러 감소율
Table 1. Error reduction rates by log pitch.

	TIMT						NTIMIT					
	ST.LP	PS.LP	ST.MF	PS.MF	ST.BLMF	PS.BLMF	ST.LP	PS.LP	ST.MF	PS.MF	ST.BLMF	PS.BLMF
without	93.7	93.7	95.4	95.2	88.3	85.3	60.0	60.0	51.2	49.6	64.4	62.6
with	96.0	96.8	96.2	96.4	91.3	87.3	68.8	72.2	59.0	60.4	68.4	68.4
ERR	**36.5%**	**49.2%**	**17.4%**	**25.0%**	**25.6%**	**13.6%**	**22.0%**	**30.5%**	**16.0%**	**21.4%**	**11.2%**	**15.5%**

오히려 기존 단구간 분석법에 비해 11.3% 저하되는 것으로 나타났다. 이는 4.2절에서 언급한 가변 길이에 대한 FFT 분석의 영향으로 사료된다.

이를 검증하기 위해 피치 동기 분석에서의 특징 추출 분석창의 길이를 단구간 분석과 동일하게 20 msec로 고정시켜 실험하였다. 이 실험을 통해서 고정 길이 분석창과 가변 길이 분석창의 FFT 분석에 의한 차이를 고찰할 수 있다. 표 3에 나타난 평균 에러 감소율은 TIMIT에서 MFCC의 경우 가변 길이 분석의 결과인 -11.3%에서 (표 2) 0.8%로서 큰 향상을 보인다. 그러나 LPCC의 경우 5.7%에서 -9.7%로 오히려 큰 폭으로 감소함을 볼 수 있다. 한편 NTIMIT에서는 분석창 길이의 변화가 큰 영향을 미치지 않는 것으로 나타났으며, 결론적으로 제안된 피치 동기 분석은 FFT기반의 MFCC보다는 LPC기반의 LPCC에 적합한 것으로 사료된다.

표 2의 결과에서 TIMIT과 NTIMIT의 에러 감소율 평균을 비교하면, LPCC, MFCC 그리고 BLMF의 결과가 각각 2%, 12.3% 그리고 15.6% 씩 높게 나타남으로써, 채널 환

표 2. 훈련을 위한 발성 횟수에 따른 에러
Table 2. Error rates with respect to the number of utterances for training.

TIMIT, w/ pitch

훈련 파일수	ST.LP	PS.LP	ERR	ST.MF	PS.MF	ERR	ST.BLMF	PS.BLMF	ERR
3	84.9	85.9	6.6%	80.0	78.8	-5.9%	57.1	61.9	11.1%
4	89.7	90.3	5.8%	89.3	85.3	-37.1%	73.6	75.4	6.8%
5	90.9	91.7	8.8%	92.3	91.9	-5.1%	81.2	80.8	-2.1%
6	95.2	94.6	-12.6%	95.6	95.0	-13.5%	85.9	85.7	-1.4%
7	96.0	96.8	20.0%	96.2	96.4	5.3%	91.3	87.3	-46.0%
Average	91.3	91.9	**5.7%**	90.7	89.3	**-11.3%**	77.8	78.2	**-6.3%**

NTIMIT, w/ pitch

훈련 파일수	ST.LP	PS.LP	ERR	ST.MF	PS.MF	ERR	ST.BLMF	PS.BLMF	ERR
3	48.0	52.4	8.5%	32.6	34.8	3.3%	36.6	40.8	6.6%
4	55.4	57.0	3.6%	40.8	39.0	-3.0%	42.8	49.8	12.2%
5	61.4	63.0	4.1%	46.4	48.4	3.7%	51.4	57.8	13.2%
6	64.8	68.8	11.4%	55.8	54.8	-2.3%	61.2	66.8	14.4%
7	68.8	72.2	10.9%	59.0	60.4	3.4%	68.4	68.4	0.0%
Average	59.7	62.7	**7.7%**	46.9	47.5	**1.0%**	52.1	56.7	**9.3%**

표 3. 훈련을 위한 발성 횟수에 따른 에러 (고정길이 피치 동기 분석창)
Table 3. Error rates with respect to the number of utterances for training (fixed length frame).

TIMIT, w/ pitch

훈련 파일수	ST.LP	PS.LP	ERR	ST.MF	PS.MF	ERR	ST.BLMF	PS.BLMF	ERR
3	84.9	82.7	-14.5%	80.0	79.8	-1.0%	57.1	61.7	10.7%
4	89.7	88.7	-9.6%	89.3	86.7	-24.1%	73.6	75.6	7.5%
5	90.9	91.5	6.6%	92.3	92.9	7.8%	81.2	80.6	-3.1%
6	95.2	94.4	-16.8%	95.6	95.6	0.0%	85.9	85.7	-1.4%
7	96.0	95.4	-14.0%	96.2	97.0	21.6%	91.3	88.1	-36.8%
Average	91.3	90.6	**-9.7%**	90.7	90.4	**0.8%**	77.8	78.3	**-4.6%**

NTIMIT, w/ pitch

훈련 파일수	ST.LP	PS.LP	ERR	ST.MF	PS.MF	ERR	ST.BLMF	PS.BLMF	ERR
3	48.0	54.0	11.5%	32.6	37.2	6.8%	36.6	42.4	9.1%
4	55.4	55.8	0.9%	40.8	39.4	-2.4%	42.8	52.4	16.8%
5	61.4	62.4	2.6%	46.4	47.6	2.2%	51.4	57.8	13.2%
6	64.8	70.2	15.3%	55.8	55.0	-1.8%	61.2	65.4	10.8%
7	68.8	71.0	7.1%	59.0	59.6	1.5%	68.4	69.6	3.8%
Average	59.7	62.7	**7.5%**	46.9	47.8	**1.3%**	52.1	57.5	**10.7%**

경에서 로그 피치 및 피치 동기 분석에 의한 향상이 두드러진 것을 알 수 있다. 이것은 채널의 영향으로 감쇄된 스펙트럼을 보상하는 효과에 의한 것으로 판단되며, 기존의 연구 결과와 일치하는 것이라 할 수 있다[3,13,14]. 구체적으로 말해 특징 추출을 위한 고정 길이 분석창을 사용하는 단구간 분석은 채널에 의해 성도 모델의 스펙트럼, 특히 기본 주파수가 일정하게 감쇄되지만, 제안된 피치 동기 분석 및 특징 구성은 채널의 영향에 의해 감쇄된 피치 정보를 보상하는 것으로 사료된다. 따라서 안정된 피치 검출이 가능하다면 피치에 동기된 분석법은 채널 영향에 민감한 성도의 스펙트럼 정보를 보완할 수 있는 효과를 얻을 수 있는 것으로 사료된다.

표 4와 5는 동일한 실험을 남자와 여자화자에 대해 독립적으로 수행한 결과이다. 각각의 결과를 통해 여자의 성능 향상이 남자의 성능 향상에 비해 큰 것으로 나타났으며, 이는 상대적으로 피치가 짧은 여자 화자의 피치 동

표 4. 훈련을 위한 발성 횟수에 따른 에러 (여자)
Table 4. Error rates with respect to the number of utterances for training (Female).

TIMIT, w/ pitch

훈련 파일수	ST.LP	PS.LP	ERR	ST.MF	PS.MF	ERR	ST.BLMF	PS.BLMF	ERR
3	84.5	89.9	34.6%	83.3	78.6	-28.6%	62.5	57.7	-12.7%
4	89.9	91.1	11.8%	91.7	84.5	-85.8%	75.0	66.7	-33.3%
5	91.7	92.3	7.1%	91.7	90.5	-14.3%	81.6	73.8	-42.0%
6	95.2	95.2	0.0%	96.4	95.2	-33.3%	85.7	81.6	-29.1%
7	95.2	97.0	37.4%	95.8	95.2	-14.1%	89.9	85.7	-41.2%
Average	91.3	93.1	**18.2%**	91.8	88.8	**-35.2%**	78.9	73.1	**-31.7%**

NTIMIT, w/ pitch

훈련 파일수	ST.LP	PS.LP	ERR	ST.MF	PS.MF	ERR	ST.BLMF	PS.BLMF	ERR
3	46.1	53.9	14.6%	38.2	38.8	1.0%	36.4	33.3	-4.8%
4	51.5	57.0	11.2%	41.8	45.5	6.2%	50.3	46.7	-7.3%
5	62.4	61.8	-1.6%	47.3	47.9	1.2%	51.5	52.1	1.2%
6	64.2	70.3	16.9%	58.8	55.8	-7.4%	61.8	63.6	4.8%
7	67.3	72.7	16.7%	57.6	60.6	7.1%	67.9	63.0	-15.1%
Average	58.3	63.2	**11.6%**	48.7	49.7	**1.6%**	53.6	51.8	**-4.2%**

표 5. 훈련을 위한 발성 횟수에 따른 에러 (남자)
Table 5. Error rates with respect to the number of utterances for training (Male).

TIMIT, w/ pitch

훈련 파일수	ST.LP	PS.LP	ERR	ST.MF	PS.MF	ERR	ST.BLMF	PS.BLMF	ERR
3	85.4	84.2	-8.2%	78.6	78.9	1.4%	61.6	64.3	7.0%
4	89.9	90.5	5.9%	88.7	86.3	-21.0%	81.0	80.4	-3.1%
5	91.1	92.0	10.0%	92.6	92.6	0.0%	85.7	84.2	-10.4%
6	95.2	94.4	-18.7%	95.2	94.9	-6.3%	87.8	88.1	2.5%
7	96.4	96.7	8.4%	96.4	97.0	16.5%	92.0	88.1	-48.0%
Average	91.6	91.6	**-0.5%**	90.3	89.9	**-1.9%**	81.6	81.0	**-10.4%**

NTIMIT, w/ pitch

훈련 파일수	ST.LP	PS.LP	ERR	ST.MF	PS.MF	ERR	ST.BLMF	PS.BLMF	ERR
3	49.3	52.2	5.9%	31.0	34.0	4.3%	40.9	45.1	7.1%
4	57.6	57.3	-0.7%	40.6	36.1	-7.5%	50.5	52.2	3.6%
5	61.2	63.9	6.9%	46.6	49.0	4.5%	59.4	60.9	3.7%
6	65.4	68.4	8.6%	54.6	54.3	-0.7%	66.3	68.4	6.2%
7	69.9	72.2	7.9%	60.0	60.6	1.5%	69.0	71.3	7.7%
Average	60.7	62.8	**5.7%**	46.6	46.8	**0.4%**	57.2	59.6	**5.6%**

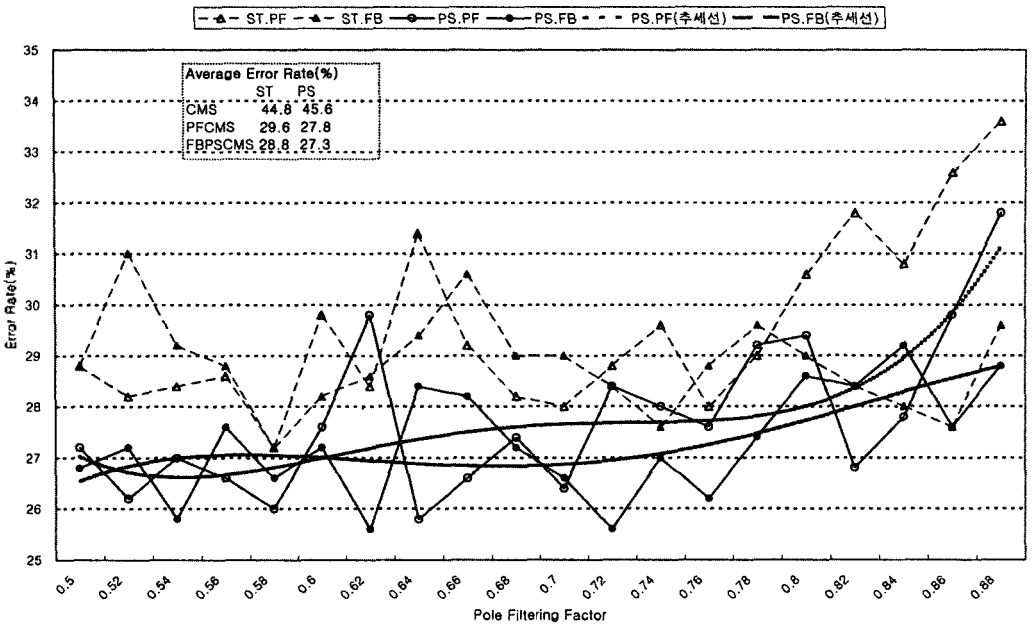

그림 8. 채널 정규화 기법의 에러율 비교
Fig. 8. Error rate comparison between channel normalization methods.

기 분석이 단구간 분석에 비해 유리하기 때문인 것으로 사료된다.

6.4. 채널 정규화 방법

그림 8은 채널 환경에서의 강인한 화자 식별을 위해 CMS, PFCMS 그리고 FBPSCMS를 각각 NTIMIT을 이용한 실험에 적용한 결과이다. 실험에서 PFCMS의 극점 필터링 상수 γ는 0.5부터 0.88까지 변화시키면서 수행했으며 FBPSCMS의 포먼트 평탄화 상수값 B_{TH}는 0.85를 실험적으로 선택하였다.

실험 결과에 의해 기존 CMS는 채널 정규화 이후 오히려 인식률이 낮아지는 것으로 나타났으며, 이는 화자의 정보가 함께 제거되었음을 보여주는 결과로 사료된다. 한편 단구간 분석에 비해 피치 동기 분석에 적용된 채널 정규화 방법이 거의 모든 γ값에 대해서 우세한 것으로 나타났다. 이는 피치 동기 분석이 유성음과 무성음 그리고 전이구간 등을 구별하여 처리하고, 결과적으로 성도 스펙트럼을 정확하게 표현하기 때문에 PFCMS 및 FBPSCMS의 포먼트를 감쇄시키는 처리가 정확해지는 것으로 사료된다. 한편 유성음 구간에 대해서 극점 필터링 처리를 추가한 FBPSCMS는 PFCMS에 비해 두드러진 향상을 보이지 않았다. 하지만 4차 다항식으로 표현한 추세선에서 나타나듯이 PFCMS에 비해 안정된 성능을 나타내었다.

VII. 결론 및 향후 연구 과제

본 논문에서는 문맥과 화자에 종속적인 피치에 동기된 음성 구간만을 고려한 피치 동기 분석 방법을 제안하고 화자 식별을 통해 성능을 비교하였다. 켑스트럼과 로그 피치를 추가한 켑스트럼을 비교한 실험에서는 분석 방법에 관계없이 피치에 의해 큰 향상을 나타내었다. 그리고 기존 단 구간 분석과의 비교 실험에서는 남자에 비해 여자의 결과가 그리고 TIMIT에 비해 NTIMIT의 결과가 우세한 것으로 나타났다. 결론적으로 피치 동기 분석은 기존 단구간 분석에 비해 화자에 종속적인 특징을 추출하는 것으로 확인되었으며, 특히 채널 환경에서 감쇄된 피치 정보를 보상해주는 것으로 나타났다.

또한 모든 실험에서 LPCC의 결과가 뛰어난 향상을 보인 반면, FFT 기반의 MFCC는 피치 정보를 포함하는 장점을 가졌지만 채널 잡음의 영향에 매우 민감한 것으로 나타났다. 채널의 통과 대역만을 고려한 MFCC의 경우 채널 잡음과 함께 피치 정보도 감쇄되므로 성능이 저하되는 것으로 나타났다. 전체적으로 로그 피치와 결합된 LPCC는 피치 동기 분석에 가장 적합한 음성 특징으로 판단된다.

한편 기존의 CMS의 경우 정규화 과정에서 화자의 정보가 동시에 제거되어 성능을 저하시키고, 이를 보완하기 위해 채널 성분의 추정 과정에서 포먼트 성분을 감쇄시키는 PFCMS, FBPSCMS 등의 정규화 방법이 큰 성능 개선을

나타내는 것으로 확인되었다. 또한 피치 동기 분석은 포먼트 성분을 감쇄시키는 PFCMS와 MFBCMS의 처리의 정확도를 향상시키는 것으로 확인되었다. 그러나 FBPSCMS는 PFCMS에 비해 두드러진 성능 차이를 보이지 못했으며, 이는 충분히 긴 음성의 장구간 평균에 의해 MFBCMS의 효과가 드러나지 않은 것으로 추정된다. 따라서 짧은 음성 또는 실시간 채널 정규화 방법에서의 효과를 고찰할 필요가 있는 것으로 판단된다. 앞으로 화자 확인 실험에 적용하여 성능을 검증하고, 잡음 환경에서의 성능 향상을 위한 피치 동기 분석 방법을 연구할 예정이다.

참 고 문 헌

1. L. C. Wood, and D. J. B. Pearce, "Excitation synchronous formant analysis," *IEE Proceedings*, **36** (2), 110-118, April 1989.
2. Y. Medan and E. Yair, "Pitch synchronous spectral analysis scheme for voiced speech," *IEEE Trans. Acoustics, Speech, and Signal Processing*, **37** (9), 1321-1328, Sep. 1989.
3. J. E. Luck, "Automatic speaker verification using cepstral measurements," *J. Acoustical Society of America*, **46** (4) (part 2), 1026-1031, April 1969.
4. R. W. Schaler and L.R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoustical Society of America*, **47** (2) (Part 2), 634-648, Feb. 1970.
5. B. S. Atal, "Automatic speaker recognition based on pitch contours," *J. Acoustical Society of America*, **52** (6) (Part 2), 1687-1697, July 1972.
6. R. C. Lummis and A. E. Rosenberg, "Test of an automatic speaker verification method with intensive trained mimics," *J. Acoustical Society of America*, **51** (1) (Part 1), 131 (A), Jan. 1972.
7. A. E. Rosenberg and M. R. Sambur, "New techniques for automatic speaker verification," *IEEE Trans. Acoustics, Speech, and Signal Processing*, **23** (2), 169-176, April 1975.
8. A. E. Rosenberg, "Automatic speaker verification: a review," *Processing. of the IEEE*, **64** (4), 475-487, April 1976.
9. B. S. Atal, "Automatic recognition of speakers from their voices," *Proc. of the IEEE*, **64** (4), 460-475, April 1976.
10. D. A. Reynolds, "Experimental evaluation of Features for robust speaker identification," *IEEE Trans. Speech and Audio Processing*, 2 (4), 639-643, Oct. 1994.
11. D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE. Trans. Speech and Audio Processing*, **3** (1), 72-83, Jan. 1995.
12. D. A. Reynolds, M. A. Zissman, T. F. Quatieri, G. C. O'Leary and B. A. Carlson, "The effects of telephone transmission degradations on speaker recognition performance," *Processing. ICASSP*, 329-332, May 1995.
13. C. R. Jankowsky Jr., T. F. Quatieri and D. A. Reynolds, "Measuring fine structure in speech: application to speaker identification," *Processing. ICASSP*, 325-328, May 1995.
14. H. Ezzaidi and J. Roual, "Comparison of MFCC and pitch synchronous AM, FM parameters for speaker identification," *Processing. ICSLP*, 2, 318-321, Oct. 2000.
15. P. Veprek and M. S. Scordilis, "Analysis, enhancement and evaluation of live pitch determination techniques," *Speech Communication*, **37** (3-4), 249-270, July 2002.
16. Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination speech signals," *IEEE Trans. Signal Processing*, **39** (1), 40-48, Jan. 1991.
17. H. Matsumoto, and M. Moroto, "Evaluation of MEL-LPC cepstrum in a large vocabulary continuous speech recognition," *Processing. ICASSP*, **1**, 7-11, May 2001.
18. D. Naik, "Pole-filtered cepstral mean subtraction," *Processing. ICASSP*, 157-160, May 1995.
19. R. P. Ramachandran and K. R. Farrell, "Fast pole-filtering for speaker recognition," *Proc. ISCAS*, **5**, 49-52, May 2000.
20. 정혜경, 김유진, 정재호, "켑스트럼으로부터 변환된 로그 스펙트럼을 이용한 포먼트 평활화 켑스트럴 평균 차감법," 한국음향학회지, 21 (4), 361-373, 2002년 5월.

저자 약력

● 김 유 진 (Yu-Jin Kim)

1995년 2월: 인하대학교 전자공학과 (공학사)
1997년 2월: 인하대학교 전자공학과 (공학석사)
2004년 2월: 인하대학교 전자공학과 (공학박사)
1995년 6월~1996년 12월: 한국전자 통신연구소 음성언어처리연구실 위촉 연구원
1997년 2월~1998년 5월: LG반도체 System Device 연구소 DSP그룹 연구원

● 정 재 호 (Jae-Ho Chung)

1982년: University of Maryland (BSEE)
1984년: University of Maryland (MSEE)
1990년: Georgia Institute of Technology (Ph. D)
1984년~1985년: 미국 국방성 산하 해군 연구소, 신호처리실 연구원
1991년~1992년: AT&T Bell Labs. 음성신호처리 연구실 연구원 (MTS)
1992년~현재: 인하대학교 공과대학 전자공학과, (현)정교수