논문 2010-47SP-5-12

선형 보간법을 이용한 시간과 주파수 조합영역에서의 피치 추정 방법

(Pitch Estimation Method in an Integrated Time and Frequency Domain by Applying Linear Interpolation)

김 기 출*, 박 성 주**, 이 석 필**, 김 무 영***

(Kichul Kim, Sung-Joo Park, Seok-Pil Lee, and Moo Young Kim)

요 약

본 논문은 피치를 추출하는 방법으로 자기상관을 이용하였다. 시간과 주파수 영역의 자기상관은 서로 다른 특성을 가지고 있으며, 각각 피치주기와 기본주파수에 대응된다. 본 논문에서는 시간과 주파수 영역에서의 자기상관을 결합하는 방법을 이용하였다. 이 방법은 자기상관에서 발생하는 피치 doubling과 having 에러를 크게 개선시킬 수 있었다. 하지만, 시간과 주파수 영역에서 유성음의 주기적 특성인 피치주기와 기본주파수는 서로 역수 관계이며, 특히 기본주파수의 에러는 FFT의 분해능에 의하여 발생된다. 이러한 영향을 줄이기 위하여 시간 영역과 주파수 영역에서의 자기상관 결합에 보간법을 적용함으로써 피치 검출율을 향상 시킬 수 있었다. 자기상관을 결합할 때 시간영역에서 찾은 피치후보들에 대해서만 주파수영역의 자기상관을 구함으로써 계산량은 감축될 수 있었다. 또한, 선형보간을 이용하여 기존방법 보다 FFT 계수를 8배 줄일 수 있었다. 그 결과, FFT 연산량과 주파수영역의 자기상관 계산량을 크게 감축하여 기존 방법 대비 알고리즘 처리시간을 약 9.5배 줄일 수 있었다.

Abstract

An autocorrelation method is used in pitch estimation. Autocorrelation values in time and frequency domains, which have different characteristics, correspond to the pitch period and fundamental frequency, respectively. We utilize an integrated autocorrelation method in time and frequency domains. It can remove the errors of pitch doubling and having. In the time and frequency domains, pitch period and fundamental frequency have reciprocal relation to each other. Especially, fundamental frequency estimation ends up as an error because of the resolution of FFT. To reduce these artifacts, interpolation methods are applied in the integrated autocorrelation domain, which decreases pitch errors. Moreover, only for the pitch candidates found in a time domain, the corresponding frequency-domain autocorrelation values are calculated with reduced computational complexity. Using linear interpolation, we can decrease the required number of FFT coefficients by 8 times. Thus, compared to the conventional methods, computational complexity can be reduced by 9.5 times.

Keywords : Pitch, Fundamental Frequency, Autocorrelation, Speech Signal Processing

* 학생회원, *** 정회원, 세종대학교 정보통신공학과
(Dept. of Information and Communication Engineering, Sejong University)
** 정회원, 전자부품연구원 디지털미디어연구센터
(Korea Electronics Technology Institute)
※ 이 논문은 2010년도 정부(지식경제부)의 재원으로 정보통신 산업원천기술사업의 지원을 받아 수행된 연구임 (No. 2010-S-001-01). 또한, 논문 작성 시 다양한 조언을 해준 임종욱 연구원께 감사합니다.
접수일자: 2010년7월5일, 수정완료일: 2010년8월13일

I. 서 론

유성음은 성대의 진동에 의해 발생하고, 이러한 진동은 음성신호의 주기적 특성으로 나타난다. 이와 같은 주기적 특성을 피치라고 하며, 이는 short-time spectrum 분석에 있어서 기본주파수에 대응된다.

피치의 특성은 음의 높낮이에 해당된다. 이러한 피치

는 효율적인 음성압축을 위하여 음성 부호화기에 사용되며[1], 피치정보를 이용한 화자인식, 음성인식 방법에도 적용되어 왔다[2~3]. 특히, 근래에는 사용자의 허밍에 의한 음악 검색 시스템에 피치를 이용하는 등, 피치검출 방식은 음성 신호처리 시스템에서 매우 파급효과가 큰 연구 분야이다[4~6].

본 논문은 발성자의 노래나 허밍 등으로부터 음 높낮이 정보인 피치추출을 목적으로 하고 있다. 일반적으로 G.729[7]와 같은 음성 부호화기에서는 기본주파수가 약 55~400Hz에서 나타나지만, 노래나 허밍의 경우 일반적으로 80~800Hz 사이에서 나타나게 된다[8]. 따라서 피치의 탐색구간을 다르게 적용할 필요가 있다.

이러한 피치추출 방식은 음성신호의 분석 및 응용에 있어서 기본적인 연구 분야이며 오래 전부터 여러 방법을 이용하여 연구되어 왔다. 가장 기본적으로 시간영역에서의 주기적 특성을 분석하는 자기상관법[7, 9~10], Average Magnitude Differential Function (AMDF)를 이용한 YIN[11] 등이 있으며 주파수영역에서는 청각모델을 이용한 방법[12], 켑스트럼 기반의 방법 등이 있다[13].

시간영역에서 피치추출 방법의 단점은 피치 doubling 에러이며, 주파수영역에서 피치추출 방법의 단점은 having 에러이다. 본 논문에서는 피치추출 방법으로 자기상관을 이용하였으며, 시간과 주파수영역을 고려하여 doubling과 having 에러를 줄이는 방법을 제안한다. 시간과 주파수영역의 분해능 불일치 문제는 보간법을 적용하여 해결하였으며, 보간법을 이용하여 계산량 또한 감축할 수 있었다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 제안한 피치추출 알고리즘으로 시간과 주파수영역의 자기상관을 보간하여 결합하는 방법을 설명한다. Ⅲ장에서는 기존 알고리즘과 제안된 알고리즘에 대한 비교 실험 결과가 나타나 있다. 마지막으로 Ⅳ장에는 최종 결론이 포함된다.

Ⅱ. 피치추정을 위한 제안하는 선형보간 알고리즘

본 논문에서는 시간영역에서의 자기상관과 주파수영역에서의 자기상관을 결합함으로써, 각각의 영역에서 발생하는 자기상관의 단점을 극복하고자 한다. 발성자의 음성이 저음인 경우, 시간영역의 자기상관은 피치 doubling 에러를 발생시킨다. 반면 고음으로 발성된 경우, 주파수영역의 자기상관은 피치 having 에러를 발생시킨다. 각각의 영역에서 자기상관의 단점은 서로 다르며 두 영역의 자기상관을 결합함으로써 피치 doubling과 having 에러를 줄일 수 있다[9].

두 영역의 피치와 기본주파수는 서로 역수 관계이므로 각각의 자기상관은 샘플링과 FFT 계수에 의하여 표현할 수 있는 분해능이 한정되어 있다. 특히, 주파수영역의 분해능은 프레임 길이, 창함수의 길이와 종류에 따라서 달라진다. 긴 길이의 프레임과 창함수를 사용하는 경우 주파수영역에서 피치검출이 용이하다. 하지만, 분석 프레임 구간내에서 피치가 변화하는 경우 피치검출은 어렵게 되며, 또한 많은 계산량이 필요한 단점이 있다. 본 논문에서는 일반적으로 많이 사용하는 해밍 윈도우를 사용하였으며, 프레임의 사이즈는 32.5ms를 이용하였다.

본 논문에서 제안하는 방법은 시간과 주파수 영역에서의 자기상관을 결합할 때, 보간법을 적용하는 것이다. 각각의 영역에서 발생하는 피치 doubling 에러와 having 에러는 시간과 주파수영역의 자기상관을 고려함으로써 해결하였다. 또한 보간법을 적용하여 주파수영역에서 필요한 FFT의 차수를 줄임으로써 기존의 방법보다 성능과 계산량 측면에서 효율적인 알고리즘을 개발하였다.

1. 시간과 주파수영역에서의 자기상관법

시간영역에서의 자기상관법 (Time domain Autocorrelation Function, TACF)을 이용한 피치추출 방법은

$$R_t(\tau) = \frac{\sum_{n=0}^{N_t-\tau-1} x[n]x[n+\tau]}{\sqrt{\sum_{n=0}^{N_t-\tau-1} x^2[n] \sum_{n=0}^{N_t-\tau-1} x^2[n+\tau]}} \quad (1)$$

과 같다. 여기서 $x[n]$, τ, N_t는 각각 n번째 음성신호, 피치에 대응되는 딜레이, 프레임길이를 나타낸다. 시간영역에서 추출한 피치는 $R_t(\tau)$의 최대 피크 딜레이를 의미한다.

주파수영역에서의 자기상관법 (Frequency domain Autocorrelation Function, FACF)을 이용한 피치추출 방법은

$$R_f(\omega_\tau) = \frac{\sum_{n=0}^{N_f/2-\omega_\tau-1} X[w_n]X[\omega_n+\omega_\tau]}{\sqrt{\sum_{n=0}^{N_f/2-\omega_\tau-1} X^2[\omega_n] \sum_{n=0}^{N_f/2-\omega_\tau-1} X^2[\omega_n+\omega_\tau]}} \quad (2)$$

과 같다. 여기서 $X[\omega_n]$, ω_τ, N_f은 각각 ω_n번째 음성 신호의 log magnitude, 기본주파수에 대응되는 딜레이, FFT 계수를 나타낸다. 주파수영역의 기본주파수는 시간영역의 피치에 대응되며, 윈도우에 의한 하모닉들의 간격을 의미한다. 따라서 기본주파수는 $R_f(w_\tau)$의 최대 피크를 의미한다. $X[\omega_n]$는 log magnitude로써 음성의 에너지가 많은 부분을 제외하면 하모닉이 잘 표현되지 않는 불필요한 정보이므로 평균값으로 클리핑하여 적용되었다.

식 (1)과 식 (2)에서 $R_t(\tau)$와 $R_f(\omega_\tau)$는 신호의 크기를 고려하여 최대 1로 정규화 하였다. 이와 같은 정규화 과정은 시간과 주파수영역의 자기상관을 결합하기 위함이다.

2. 보간법을 적용한 자기상관의 결합

TACF와 FACF를 결합하는 방법은

$$R_{tf}(\tau) = \beta R_t(\tau) + (1-\beta)R_f(\hat{\omega_\tau}) \quad (3)$$

과 같이 계산되며, 그림 1과 같다. 여기서 β는 시간과 주파수영역의 자기상관 값의 가중치를 나타내며, 피치의 doubling과 having 에러를 조절할 수 있는 파라미터이다. 본 연구에서는 β를 0.5로 설정하였다. $\hat{\omega_\tau}$는 시간영역 τ와 가장 인접한 ω_τ 값을 나타낸다 ($\hat{\omega_\tau}$= $\lceil N_f/\tau \rceil$, $\lceil \bullet \rceil$: round).

그림 1은 TACF와 FACF를 결합한 그림을 나타내며 각각의 그림에서 *는 해당 자기상관의 최대 피크를 의미한다. 그림 1에서는 TACF에서 발생하는 doubling 에러를 FACF와 결합함으로써 해결할 수 있음을 보여준다. TACF는 실제 피치의 정수배 τ를 피치로 추출하게 된다. 그림 1 (b)에서 TACF는 실제 피치의 2배에 해당하는 doubling 에러가 발생함을 보여준다. 하지만 그림 1 (c)의 FACF와 결합함으로써 그림 1 (d)에서처럼 피치 doubling 에러를 수정할 수 있다.

기존에는 시간과 주파수영역의 자기상관을 결합하는 방법으로 시간영역 τ와 가장 인접한 ω_τ의 자기상관 값을 식 (3)과 같이 이용하는 방법[9]을 사용하였다 (Time-Frequency domain Autocorrelation, TFAC). 하

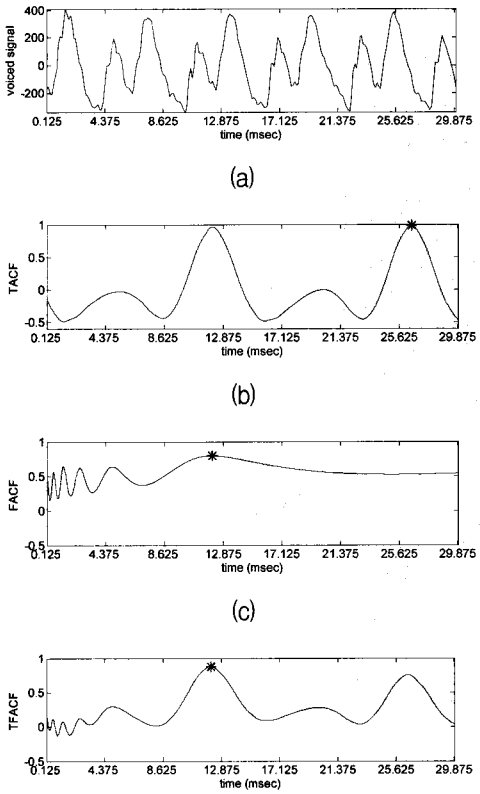

그림 1. 시간과 주파수영역의 자기상관 결합방법
(a) 시간영역에서의 유성음 신호 (b) 시간영역에서의 자기상관 값 (c) 주파수영역의 자기상관 값을 시간영역에서 표현 (d) 시간과 주파수영역의 자기상관 값을 결합

Fig. 1. Integrated autocorrelation in time and frequency domains.
(a) voice activity in time domain,
(b) autocorrelation in time domain,
(c) autocorrelation in frequency domain, which is represented in time domain, and (d) integrated autocorrelation in time and frequency domains

지만 본 논문에서는 자기상관을 결합하는 과정에 보간법을 적용하였다.

보간법은 여러 방법이 있으나, 본 논문에서는 spline[14]과 계산량이 적은 선형보간을 적용하였다. 보간법을 적용하는 목적은 크게 두 가지로 구분된다. 첫 번째 이유는 시간과 주파수영역의 분해능 불일치를 줄이는 것이다. 식 (1)과 식 (2)에서 τ와 ω_τ는 서로 역수 관계이다. 시간영역의 경우 샘플링에 의하여 신호의 분해능이 결정되며, 주파수영역의 경우 FFT 계수 N_f에 의하여 결정된다. 예를 들어 8kHz, 1028-FFT는 시간과 주파수영역에서 다른 해상도를 갖는다. 시간영역의 샘플들은 주파수영역에서 N_f/n인 반비례 간격으로 표

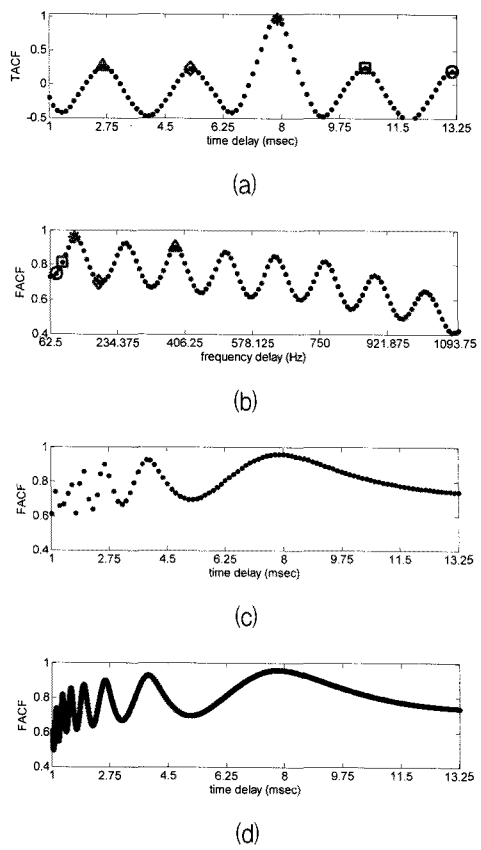

그림 2. 보간법이 적용된 자기상관
(a) 시간영역에서의 자기상관 (b) 주파수영역의 자기상관 (c) 주파수영역의 자기상관을 시간영역에서 표현 (d) 보간을 적용한 주파수영역의 자기상관을 시간영역에서 표현

Fig. 2. Interpolated autocorrelation.
(a) autocorrelation in time domain,
(b) autocorrelation in frequency domain,
(c) autocorrelation in frequency domain, which is represented in time domain, and (d) interpolated autocorrelation in frequency domain, which is represented in time domain

현 되지만, 주파수영역의 샘플들은 정비례 간격으로 표현된다. 따라서 각각의 영역은 250Hz를 기준으로 250Hz이하는 시간영역에서 자기상관이 잘 표현되며, 250Hz이상에서는 주파수영역의 자기상관이 잘 표현되는 특징이 있다. 그림 2 (a)는 TACF를 나타낸 것이며, 그림 2 (b)는 FACF를 나타낸 것이다. 그림 2 (c)는 FACF를 시간영역에서 나타낸 것이며, 그림 2 (d)는 보간을 적용한 주파수영역의 자기상관을 시간영역에서 나타낸 것이다. 그림 2 (a)와 그림 2 (b)에서 ○, □, *, ◇, △들은 TACF의 피크값 인덱스를 나타낸다. 그림 2 (a)는 8kHz 샘플링에 의하여 시간영역에서 정비례 간격으로 자기상관이 표현된다. 자기상관 값의 피크들은 신호의 주기적 특성에 의하여 정비례 간격으로 표현된다. 하지만 그림 2 (b)에서는 TACF의 피크 인덱스들은 주파수영역에서 역수관계로 매핑되는 것을 확인할 수 있다. 그림 2 (c)는 FACF를 시간영역으로 표현한 것으로, 시간인덱스 앞부분에서는 명료하게 표현되지 못하는 것을 확인 할 수 있다. 하지만 그림 2 (c)에 보간을 추가적으로 적용하면 그림 2 (d)에서처럼 시간인덱스의 앞부분에 해당하는 고주파영역이 잘 표현되는 것을 볼 수 있다.

보간을 사용하는 두 번째 이유는 보간법 적용으로 계산량을 감축하기 위함이다. 보간법을 이용하면 FFT 계수를 줄여도 성능저하를 막을 수 있었다. 주파수영역의 분해능은 FFT 계수 N_f에 의하여 결정된다. 주파수영역에서 세밀하게 표현하기 위하여 큰 N_f의 FFT를 수행한다. 이처럼 큰 N_f를 이용하면 보간법을 적용하지 않아도 되지만, FFT와 자기상관에서 많은 계산량이 요구되는 단점이 있다. FFT에서 계산량은 길이가 N_f인 신호에 대하여 대략적으로 $N_f \log_2(N_f)$가 된다. 하지만 보간법을 적용하면 $N_f/8$의 길이를 사용해도 성능의 저하가 거의 없었다. 예를 들어 N_f=2048인 경우, 두 영역을 결합한 자기상관 (TFAC)의 FFT 계산량은 22528개의 곱셈이 필요하지만, 보간법을 적용하면 FFT 계산량은 2048개의 곱셈으로 약 11배 정도 감축될 수 있다. 또한, N_f가 8배 줄어들기 때문에 식 (2)의 FACF 계산량은 더욱 감축된다.

본 논문에서 제안하는 알고리즘의 기본이 되는 시간영역의 자기상관 (TACF)과 주파수영역의 자기상관 (FACF)을 결합한 자기상관은 보간하는 방법에 따라, 다음과 같이 3가지로 구분된다.

첫 번째는 시간과 주파수영역 모두 고려하여 보간법을 적용하는 방법이다. 시간과 주파수는 서로 분해능이 달라서 8kHz, 1024 point FFT의 경우, 250Hz이하에서는 FACF를 보간을 적용하고 250Hz이상에서는 TACF를 보간하여 시간과 주파수영역의 자기상관을 결합하였다. 즉, TACF와 FACF에 모두 보간을 적용하는 것으로 비교적 정확한 보간법인 spline을 이용하였다 (Interpolated Autocorrelation Function 1, IACF1).

두 번째 방법은 FACF의 계산량을 감축하기 위한 방법이다. 우선 시간영역의 자기상관으로부터 후보 피치들을 찾는다. 후보 피치는 자기상관에서 발생한 피크 인덱스에 해당한다. 주파수영역의 자기상관은 시간

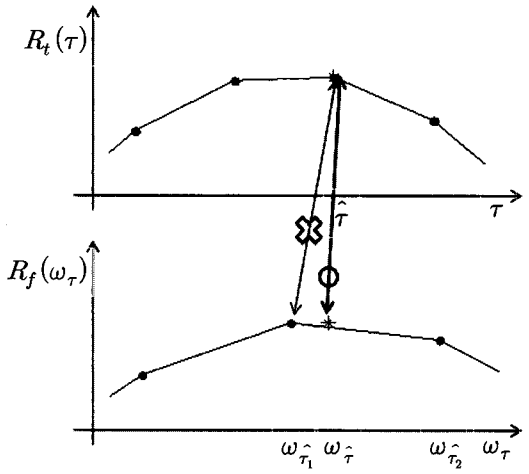

그림 3. 시간과 주파수영역의 자기상관을 결합하기 위한 선형보간법

Fig. 3. Integrated autocorrelation in time and frequency domains using linear interpolation.

영역에서 찾은 후보 피치의 인접 인덱스에서만 수행된다. 그러면 주파수영역에서 자기상관을 모두 수행 할 필요가 없어서 계산량을 크게 감축할 수 있는 장점을 갖는다. 여기서 사용된 보간법은 spline 방법이다 (Interpolated Autocorrelation Function 2, IACF2).

세 번째 방법은 IACF2의 복잡한 보간법 대신 계산량이 적은 선형보간을 적용하는 방법이다 (Interpolated Autocorrelation Function 3, IACF3). 선형보간을 적용하는 방법은 다음과 같이 계산되며, 그림 3과 같다.

$$R_f(\omega_{\hat{\tau}}) = R_f(\omega_{\hat{\tau}_1}) + \frac{R_f(\omega_{\hat{\tau}_2}) - R_f(\omega_{\hat{\tau}_1})}{\omega_{\hat{\tau}_2} - \omega_{\hat{\tau}_1}} \left(\omega_{\hat{\tau}} - \omega_{\hat{\tau}_1}\right) \qquad (4)$$

여기서 $\hat{\tau}$와 $R_t(\hat{\tau})$는 각각 시간영역의 자기상관 피크 인덱스와 자기상관 값을 나타낸다. $\omega_{\hat{\tau}}$와 $R_f(\omega_{\hat{\tau}})$는 각각 $\hat{\tau}$의 주파수영역에 대응되는 인덱스와 자기상관 값을 나타낸다. $\omega_{\hat{\tau}_1}$, $\omega_{\hat{\tau}_2}$는 $\omega_{\hat{\tau}}$의 인접한 인덱스들을 의미하며 자기상관 값은 각각 $R_f(\omega_{\hat{\tau}_1})$, $R_f(\omega_{\hat{\tau}_2})$이다.

본 논문에서는 보간법에 따라서 IACF1, IACF2, IACF3을 제안하였으며, 그 중에서 성능과 계산량 관점에서 가장 우수한 방법은 IACF3이다. IACF1은 TACF와 FACF에 보간법을 각각 적용하였지만, 시간영역의 후보피치에서만 FACF를 보간하여 결합하는 방법인 IACF2와 IACF3에 비해 성능차이가 거의 없었다. 또한 IACF2와 IACF3에서는 각각 보간법의 성능이 비교적 좋은 spline방법과 계산량이 적고 간단한 선형보간 방법을 적용하였으나, 피치추출에 있어서 성능차이가 발생하지 않았기 때문에 가장 우수한 방법은 계산량이 적은 IACF3이다.

Ⅲ. 실 험

실험을 위하여 Jang's collection의 singing/humming 데이터베이스를 사용하였다[15]. Jang's collection은 허밍 질의에 의한 오디오 검색 시스템 (query by singing humming system) 데이터베이스로써, 8초, 8kHz로 녹음된 음성파일이 2797개이며, 피치정보를 32msec마다 세미톤 형태로 제공한다. 실험에 사용된 모든 방법들은 분석프레임을 10msec씩 이동시키며 기본주파수를 추출하여 데이터베이스에서 제공된 피치정보와 비교하였다.

또한, 본 논문에서 제안하는 방법이 배경 잡음환경에 강인한지 확인하기 위하여 Jang's collection에 NOISEX92의 volvo, babble, white 잡음을 각각 0, 10, 20, 30dB로 섞어서 사용하였다.

추출한 피치의 정확성을 측정하기 위하여 데이터베이스에서 제공된 정확한 기본주파수 (reference F0)를 기준으로 gross F0 error rate (GER)를 측정하였다. 본 논문의 실험에서는 GER-10%를 측정하여 알고리즘의 성능을 평가하였다. 또한 피치 doubling과 having의 영향을 확인하기 위하여 too low, too high 에러를 측정하였다. 예를 들어 GER-10%의 경우, 추출한 기본주파수의 값이 정확한 기본주파수를 기준으로 기본주파수의 상위 10% (+ error rate), 하위 10% (- error rate)일 경우 인식성공 (acc), 상위 10% 이상일 경우 too high, 10% 이하일 경우 too low로 정의하며 이것을 그림 4에 도식화하였다. 여기서 too low는 피치의 doubling 현상을, too

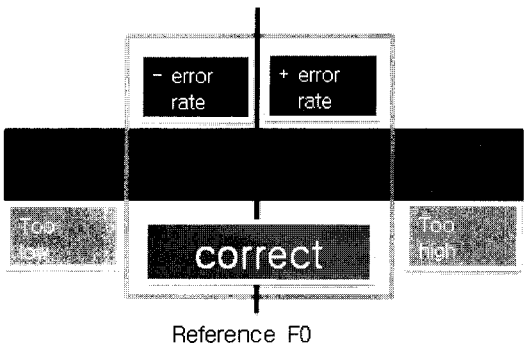

그림 4. 피치추출의 정확성을 평가하는 방법

Fig. 4. GER of accuracy of extracted pitch.

표 1. 여러 배경잡음 환경에서의 G.729, YIN, TACF, FACF, TFAC의 성능 평가 (GER-10%)
Table 1. Evaluation of G.729, YIN, TACF, FACF, and TFAC under background noise environments (GER-10%).

Environmental Conditions		G.729			YIN			TACF			FACF			TFAC		
		low	high	acc	low	high	acc	low	high	acc	low	high	acc	low	high	acc
clean		4.0	1.2	94.7	1.5	1.2	97.3	23.1	0.7	76.2	1.8	8.6	89.7	1.5	2.1	96.3
babble	30dB	4.4	1.3	94.4	1.6	1.2	97.2	23.5	0.7	75.8	1.8	9.6	88.5	1.6	2.2	96.2
	20dB	6.6	1.5	91.9	2.4	1.3	96.3	26.4	0.7	72.9	2.4	12.8	84.8	2.7	2.2	95.1
	10dB	17.3	2.8	80.0	16.3	2.3	81.4	33.7	1.3	64.9	6.4	21.6	71.9	10.6	2.8	86.6
	0dB	44.0	10.3	45.7	35.3	11.7	53.0	44.1	7.3	48.7	21.8	35.9	42.3	36.6	8.6	54.9
volvo	30dB	4.1	1.2	94.7	1.5	1.2	97.3	22.2	0.7	77.1	1.8	8.0	90.1	1.5	2.2	96.3
	20dB	4.2	1.2	94.6	1.5	1.3	97.2	18.2	0.8	81.0	2.1	6.8	91.1	1.4	2.2	96.4
	10dB	5.3	1.1	93.6	2.7	1.6	95.7	11.6	1.5	86.9	2.9	5.8	91.4	1.5	2.5	96.0
	0dB	15.5	1.2	83.3	14.4	5.5	80.1	11.8	8.6	79.6	4.7	6.7	88.5	4.2	5.5	90.3
white	30dB	4.1	1.3	94.6	1.5	1.2	97.3	23.2	0.7	76.1	1.8	9.5	88.7	1.6	2.2	96.3
	20dB	4.6	1.4	94.0	1.9	1.3	96.8	24.5	0.8	74.7	1.9	12.3	85.8	1.7	2.4	95.9
	10dB	10.2	2.0	87.7	12.0	2.0	85.9	31.8	1.1	67.1	2.1	20.6	77.4	2.6	3.6	93.8
	0dB	33.3	6.1	60.7	25.8	20.9	53.3	45.9	2.0	52.1	3.0	47.1	49.8	8.2	16.1	75.7
total average		12.1	2.5	85.4	9.1	4.1	86.8	26.2	2.1	71.8	4.2	15.8	80.0	5.8	4.2	90.0

high는 피치의 having 현상에 대한 정보를 알려준다. 제안하는 알고리즘의 비교 평가를 위하여 기존에 잘 알려진 피치 추출 방법인 G.729[7]와 YIN[11]을 이용하였다.

실험은 크게 4가지로 구성된다. 첫 번째는 시간과 주파수영역에서 정규화된 자기상관 (TACF, FACF)과각각의 자기상관을 결합한 방법 (TFAC[9])을 비교한다. 두 번째는 G.729, YIN, TFAC의 성능을 비교한다. 세 번째는 TFAC와 보간을 적용 (IACF1, IACF2, IACF3)한 실험을 비교한다. 마지막으로 비교적 작은 FFT 계수를 이용한 IACF3의 성능을 측정한다.

첫 번째 실험에서 TFAC는 TACF와 FACF를 결합함으로써 시간과 주파수영역의 자기상관 에러를 줄이는 것으로, 기존 방법인 G.729, YIN과 비교한 결과는 표 1과 같다.

TACF는 식 (1)을 이용하였고 FACF는 식 (2)를 이용하였으며 TFAC는 식 (1)과 식 (2)에서 구한 자기상관을 식 (3)을 이용하여 결합하였다. 시간영역에서 피치를 찾는 방법으로 G.729, YIN, TACF에서는 too low error (doubling 에러에 대응)가 많이 발생하였다. 이와 반대로 주파수영역에서 피치를 찾는 방법으로 FACF에서는 too high error (having 에러에 대응)가 많이 발생한 것을 확인 할 수 있었다. 이러한 TACF와 FACF, 두 가지 방법을 결합한 TFAC는 too low error와 too high error 모두 줄어든 것을 확인 할 수 있다. 또한, TFAC 방법이 모든 배경잡음에 대하여 G.729보다 우수하며, 10dB 이하의 배경잡음에 대하여 YIN보다 우수한 성능을 나타낸다.

두 번째 실험은 기존에 널리 알려진 G.729, YIN 방법과 TFAC의 정밀한 피치 추출을 비교하는 것으로 결과는 표 2와 같다. 좀 더 정밀한 피치 추출을 확인하기 위하여 GER-5%를 측정하였다. 모든 배경잡음 환경에 대하여 TFAC가 G.729, YIN의 방법에 비해 우수한 성능을 나타낸다. 특히, 10dB이하의 배경잡음 환경에서는 시간영역만 고려된 G.729, YIN 보다 주파수영역까지 고려된 TFAC가 우수하였다.

표 2. 여러 배경잡음 환경에서의 G.729, YIN, TFAC의 성능 평가 (GER-5%)
Table 2. Evaluation of G.729, YIN, and TFAC under background noise environments (GER-5%).

Environmental Conditions		G.729	YIN	TFAC
clean		92.3	85.7	95.4
babble	30dB	91.8	86.1	95.2
	20dB	88.6	87.5	94.0
	10dB	74.3	76.8	84.5
	0dB	37.2	43.5	47.5
volvo	30dB	92.3	85.8	95.4
	20dB	92.2	86.7	95.4
	10dB	91.3	89.4	94.9
	0dB	81.1	76.5	88.1
white	30dB	92.0	86.1	95.3
	20dB	90.7	88.2	94.9
	10dB	82.0	82.8	92.4
	0dB	50.6	48.5	71.1
total average		81.3	78.7	88.0

표 3. 여러 배경잡음 환경에서의 TFAC, IACF1, IACF2, IACF3의 성능 평가 (GER-10%)

Table 3. Evaluation of TFAC, IACF1, IACF2 and IACF3 under background noise environments (GER-10%).

Environmental Conditions		TFAC	IACF1	IACF2	IACF3
clean		96.3	97.7	97.8	97.8
babble	30dB	96.2	97.6	97.7	97.7
	20dB	95.1	96.5	96.6	96.6
	10dB	86.6	88.1	88.0	88.0
	0dB	54.9	55.8	55.7	55.7
volvo	30dB	96.3	97.7	97.9	97.9
	20dB	96.4	97.7	97.9	97.9
	10dB	96.0	97.0	97.5	97.4
	0dB	90.3	87.3	91.7	91.6
white	30dB	96.3	97.7	97.8	97.8
	20dB	95.9	97.4	97.5	97.5
	10dB	93.8	95.8	96.0	96.0
	0dB	75.7	85.2	85.2	85.7
total average		90.0	91.7	92.1	92.1

표 4. 여러 배경잡음 환경에서의 GER-10%와 처리시간 (sec)

Table 4. Evaluation of TFAC-2048, IACF3-512 and IACF-256 under background noise environments (GER-10%) and processing time (sec).

Environmental Conditions		TFAC 2048	IACF3 512	IACF3 256
clean		97.8	97.8	97.7
babble	30dB	97.7	97.7	97.5
	20dB	96.5	96.5	96.2
	10dB	87.9	87.8	86.9
	0dB	55.6	55.5	54.8
volvo	30dB	97.8	97.8	97.7
	20dB	97.9	97.9	97.8
	10dB	97.4	97.4	97.4
	0dB	91.6	91.6	91.2
white	30dB	97.8	97.8	97.6
	20dB	97.5	97.5	97.3
	10dB	96.0	95.9	95.8
	0dB	85.7	85.7	85.2
total average		92.1	92.1	91.8
time (sec)		1404	188	147

GER-10%에서 YIN의 경우는 G.729보다 좋은 성능을 나타내었고 일부 TFAC보다 좋은 성능을 나타내었지만, 정밀한 피치추출의 성능은 뒤떨어지는 것을 확인할 수 있다.

세 번째 실험은 모두 1024 point FFT를 적용 하였으며, TFAC, IACF1, IACF2, IACF3의 성능을 표 3에서 나타낸다. 실험 결과에서 TFAC에 보간을 적용한 IACF1, IACF2, IACF3 방법들은 보간을 적용하지 않은 TFAC에 비해 성능이 향상되는 것을 확인 할 수 있었다. 시간과 주파수영역 모두를 보간한 방법 (IACF1)은 IACF2와 IACF3에 비하여 계산량도 많지만, 일부 잡음에 대하여 IACF2와 IACF3에 비해 성능이 뒤떨어진다. 이는 잡음의 특성에 따라 FACF의 성능이 저하되기 때문이다. IACF2와 IACF3의 경우에는 시간영역에서 추출한 후보 피치에 대하여만 주파수영역에서 각각 spline과 선형보간을 적용한 결과로써, 다른 방법들에 비하여 성능이 가장 우수하며, 후보 피치에서만 자기상관을 계산하기 때문에 계산량도 가장 우수하다. 특히, IACF3의 경우에는 선형보간을 사용하여 spline을 사용한 IACF2보다 계산량이 우수하며 성능저하가 미미하였다.

표 4의 마지막 실험은 선형보간을 이용하여 성능저하 없이 FFT 계수를 줄여서 계산량을 더욱 감축할 수 있음을 나타낸다. 계산량의 측정은 TACF에서 찾은 후보 피치에 따라서 FACF의 계산량 차이가 발생하기 때문에 전체 발성음의 처리 수행 시간으로 측정하였다. 사용한 컴퓨터의 사양은 Intel Quad Core 2.66 GHz (6 GB RAM)이다. TFAC-2048은 TFAC방법에서 FACF의 FFT 계수를 2048로 적용하였으며, IACF3-512와 IACF3-256은 각각 FACF의 FFT 계수를 512, 256으로 적용하였을 경우를 나타낸다. 여기서 대부분의 경우 IACF3-256은 TFAC-512와 TFAC-2048에 비하여 성능 차이가 거의 없음을 확인 할 수 있었다.

계산량의 경우에는 IACF3-256이 가장 우수하다. IACF3-256의 FFT 계수는 TFAC-2048에 비하여 8배 가량 줄어들었다. 이렇게 줄어든 FFT 계수로 인하여 FFT와 자기상관에서의 계산량은 대폭 감축하였다. 또한, TACF의 후보피치에 대해서만 FACF를 수행하게 되어서 계산량은 더욱 감축하였다. IACF-2048은 큰 FFT 계수로 인하여 약 1404초의 처리시간이 필요했다. IACF3-512와 IACF3-256은 줄어든 FFT 계수로 인하여 FFT 계산과 FACF의 계산량이 크게 감축하여 각각 약 188초, 147초 정도의 처리시간이 필요하였다. 이처럼 선형보간을 적용시키면 적은 계산량으로도 우수한 성능을 얻을 수 있다.

Ⅳ. 결 론

본 논문에서 제안하는 알고리즘은 시간과 주파수영역의 자기상관을 보간하여 결합하는 방법이다. 시간과 주파수영역의 자기상관을 결합하면 피치 doubling 에러와 having 에러를 개선시킬 수 있다. 하지만 시간과 주파수영역에서 피치주기와 기본주파수는 샘플링과 FFT로 인하여 분해능이 서로 다르다. 본 논문에서는 시간과 주파수영역의 자기상관 결합에 선형 보간법을 적용하여 분해능 불일치를 해결함으로써 피치 검출율을 향상 시킬 수 있었다. 또한, 선형보간법을 적용하여 FFT 차수를 기존 방법대비 8배 줄였으며, 이로 인하여 FFT와 자기상관의 계산량을 효과적으로 감축할 수 있었다. 그리고, 시간영역에서 찾은 후보피치에만 주파수영역의 자기상관을 구함으로써, 계산량을 더욱 감축하여 알고리즘 처리시간이 기존 방법대비 약 9.5배 줄었다.

현재 개발되어 있는 여러 피치추출 알고리즘은 배경잡음에 의한 낮은 SNR (Signal to Noise Ratio)에서 성능이 우수하지 못하며, 특히 배경잡음이 babble noise와 같은 non-stationary noise인 경우 성능이 우수하지 못하다. 따라서, 향후에는 낮은 SNR과 non-stationary 배경잡음에 대한 연구가 진행 되어야 할 것이다.

참 고 문 헌

[1] 손상목, 홍성훈, 배명진, "IMBE VOCODER의 피치검색시간 단축에 관한 연구," *대한전자공학회 학술대회 논문집*, vol. 10, no. 1, pp. 271-274, 1997.

[2] Y. J. Kim and J. H. Chung, "Pitch synchronous cepstrum for robust speaker recognition over telephone channels," *IET Electronics letters*, vol. 40, no. 3, pp. 207-209, 2004.

[3] H. Singer and S. Sagayama, "Pitch dependent phone modelling for HMM based speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, pp. 273-276, 1992.

[4] S. -P. Heo, M. Suzuki, A. Ito, and S, Makino, "An effective music information retrieval method using three-dimensional continuous DP," *IEEE Trans. MULTIMEDIA*, vol. 8, no. 3, pp. 633-639, 2006.

[5] J.-S. R. Jang and H. -R. Lee, "A general framework of progressive filtering and its application to Query by Singing/Humming," *IEEE Trans. Audio, Speech, Language process.*, vol. 16, no. 2, pp. 350-358, 2008.

[6] 박호종, 윤제열, "오디오 신호의 다중 피치 검출 기술," *대한전자공학회 전자공학회지*, vol. 37, no. 1, pp. 63-72, 2010.

[7] ITU-T Recommendation G.729, Coding of Speech at 8 kbit/s using Conjugate-Structure Algebraic-
Code-Excited Linear-Prediction (CS-ACELP).

[8] M. Antonelli and A. Rizzi, "A Correntropy-based voice to MIDI transcription algorithm," in *Proc. IEEE int. Multimedia Signal Processing Workshop*, pp. 978-983, 2008.

[9] Y. D. Cho, M. Y. Kim, and S. R. Kim, "A spectrally mixed excitation (SMX) vocoder with robust parameter determination," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, pp. 601-604, 1998.

[10] 한민수, 강동규, "유성음의 프레임별 피치검출," *대한전자공학회 학술대회 논문집*, vol. 9, no. 1, pp. 491-494, 1996.

[11] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimation for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1917-1930, 2002.

[12] A. Klapuri, "Multipitch analysis of polyphonic music and speech signals using an auditory model," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 2, pp. 255-266, 2008.

[13] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 44, no. 6, pp. 1585-1968, 1968.

[14] C. DeBoor, "A Practical Guide to Splines", New York: Springer-Verlag, 1978.

[15] J.-S. R. Jang, "QBSH: A Corpus for designing QBSH (query by singing/humming) systems", Available at the "QBSH Corpus for Query by Singing/Humming" Link of the "Corpus page" at the organizer's homepage. [Online]. Available: http://www.cs.nthu.edu.tw/~jang

저 자 소 개

김 기 출(학생회원)

2009년 세종대학교
정보통신공학과 학사 졸업
2009년~현재 세종대학교
정보통신공학과 석사과정
<주관심분야 : 화자인식, 음질향상, 음악정보검색>

박 성 주(정회원)

1997년 경북대학교
전자공학과 석사 졸업
1997년~1999년 대우전자
영상연구소 연구원
2000년~2002년 디지털앤디지털
선임연구원
2002년~2004년 LSI Logic Korea 선임연구원
2004년~현재 KETI 디지털미디어연구센터
선임 연구원
<주관심 분야 : 디지털방송통신융합시스템, A/V 신호처리>

이 석 필(정회원)

1990년 연세대학교
전기공학과 학사 졸업
1992년 연세대학교
전기공학과 석사 졸업
1997년 연세대학교 전기전자
공학과 박사 졸업
1997년~2002년 대우전자 영상연구소 선임연구원
2002년~현재 KETI 디지털미디어연구센터
센터장
<주관심 분야 : 디지털방송통신융합시스템, A/V 신호처리>

김 무 영(정회원)-교신저자

1993년 연세대학교
전자공학과 학사 졸업
1995년 연세대학교
전자공학과 석사 졸업
1995년-2000년 삼성종합기술원
전문연구원
2001년~2004년 Royal Institute of Technology (KTH, 스웨덴) Dept. Signals, Sensors, Systems, 박사
2004년~2005년 Royal Institute of Technology (KTH, 스웨덴) Dept. Signals, Sensors, Systems, PostDoc
2005년~2006년 Ericsson Research (스웨덴)
Senior Research Engineer
2006년~현재 세종대학교 정보통신공학과 조교수
<주관심분야 : 음성/오디오/비디오 신호처리, 패턴인식, 정보이론>