자연스런 인간-로봇 상호작용을 위한 음성 신호의 AM-FM 성분 분해 및 순간 주파수와 순간 진폭의 추정에 관한 연구

AM-FM Decomposition and Estimation of Instantaneous Frequency and Instantaneous Amplitude of Speech Signals for Natural Human-robot Interaction

이 희 영*
Heyoung Lee

ABSTRACT

A Vowel of speech signals are multicomponent signals composed of AM-FM components whose instantaneous frequency and instantaneous amplitude are time-varying. The changes of emotion states cause the variation of the instantaneous frequencies and the instantaneous amplitudes of AM-FM components. Therefore, it is important to estimate exactly the instantaneous frequencies and the instantaneous amplitudes of AM-FM components for the extraction of key information representing emotion states and changes in speech signals.

In this paper, firstly a method decomposing speech signals into AM-FM components is addressed. Secondly, the fundamental frequency of vowel sound is estimated by the simple method based on the spectrogram. The estimate of the fundamental frequency is used for decomposing speech signals into AM-FM components. Thirdly, an estimation method is suggested for separation of the instantaneous frequencies and the instantaneous amplitudes of the decomposed AM-FM components, based on Hilbert transform and the demodulation property of the extended Fourier transform. The estimates of the instantaneous frequencies and the instantaneous amplitudes can be used for modification of the spectral distribution and smooth connection of two words in the speech synthesis systems based on a corpus.

Keywords: AM-FM decomposition, Instantaneous frequency, Instantaneous amplitude, Estimation, Multi-component signals, Modification of spectral distribution, Human-robot interaction, Emotion representation, Analytic signal

* 서울산업대학교 공과대학 제어계측공학과 교수

1. 서 론

음성은 일상생활에서 정보 교환 및 감정을 수반한 상호 교감 등에 사용되는 중요한 수단이므로 음성을 기반으로 한 부드러운 인간-로봇 상호작용을 갖는 로봇을 개발하기 위해서는 음성 기술과 로봇기술의 결합이 필요하다[6][8][9]. 장애인과 노약자를 위한 로봇이 상황에 적응하면서 때로는 부드럽게 때로는 즐겁게 감정을 실어 말을 할 경우 사용자의 정서 안정에 많은 도움을 줄 것이다. 또한 로봇이 노약자가 발성한 음성에서 감정 상태를 파악하고 상호작용하면서 필요한 서비스를 제공한다면 한층 더 인간적인 동반자로서의 역할을 수행할 것이다. 그러나 현재 개발되어 있는 기계음을 내는 음성 합성기 및 변조기를 바탕으로 한 부자연스런 인간-로봇 상호작용은 사용자가 곧바로 지루함을 느낄 것이고 오히려 스트레스를 주어 사용자의 정서 안정에 방해가 될 것이다. 즉 현재의 음성 처리 기술에 바탕을 둔 로봇은 음성을 통해 자연스런 감정을 충분히 표현하지 못할 뿐만 아니라 화자의 감정 상태를 파악하지 못하므로 인간에 친근한 인간 중심 로봇으로 사용하기에는 많은 문제점이 있다. 상기와 같은 이유로 음성을 기반으로 한 로봇의 감정 표현 및 의도 인식에 대한 연구의 필요성이 앞으로 증대될 것이다.

음성신호는 다중성분 신호(multicomponent signal)이며 스펙트럼 특성은 시간적으로 변한다[1][2][4][5][23]. 음성 신호의 모음 부분은 AM-FM(amplitude modulation - frequency modulation) 성분들의 합으로 모델링 할 수 있다[3][7][19][20]. 음성 신호의 중요 파라미터들은 시간 영역, 주파수 영역, 시간-주파수 영역 등에서 관찰할 수 있다. 진폭의 변화 즉 에너지의 시간적 변화는 시간 영역을 통해서 효과적으로 관찰할 수 있으며 스펙트럼의 분포 특성은 Fourier 주파수 영역에서 효과적으로 관찰할 수 있다. 그러나 음성 신호의 스펙트럼의 시간적 변화는 시간 영역 또는 주파수 영역만으로는 관찰할 수 없으며 시간-주파수 영역에서 즉 시간-주파수 표현을 통해서 관찰할 수 있다.

음성 신호에서 감정에 대한 핵심 정보가 어디에 있느냐 하는 문제, 즉 감정을 잘 반영하는 핵심 파라미터가 무엇이냐 하는 문제는 폭 넓고 심도 있게 연구 되었다[2][6][13][14][15][16]. <표 1>은 감정 인식 또는 표현에 사용되는 주요 파라미타를 정리한 것이다. 매우 많은 파라미터가 사용되고 있는데 파라미터들이 많은 경우 서로 독립적이지 않음을 알 수 있다. <표 2>는 6 가지의 전형적인 감정 형태와 주요 3 가지 형태의 음성 파라미터간의 관계를 요약한 것이다. <표 1>과 <표 2>에서 우리는 매우 많은 파라미터들이 복잡하게 연관되어 있음을 알 수 있다. <표 1> 및 <표 2>를 바탕으로 감정을 인식할 경우 약 70% 전후의 인식율을 얻는 것이 보고 되었다[13]. 더 높은 인식율을 얻기 위해서는 감정 인식 및 표현에 사용되는 파라미터들을 단순화 시킬 필요가 있다. 즉 단순하면서도 음성신호의 특성을 잘 나타내주는 서로 독립적인 특성을 갖는 파라미터를 정의하고 이를 바탕으로 감정 음성 인식 및 표현을 시도할 필요가 있다.

다중 성분 신호인 음성 신호를 각각의 AM-FM 성분 별로 분해하여 관찰할 경우 음성 신호의 시간적 변화 양상 및 특성들을 더욱 세밀하게 파악할 수 있다. 음성 신호의 모음은 각각의 AM-FM 성분들이 중첩되어진 신호이므로 시간 영역 중심의 처리만으로는 중첩 효과(superposition effect)로 인해 발생하는 여러 가지 문제 때문에 음성 신호 내부의 깊숙한 곳에 있는 정보를 추출하기가 매우 어렵다[2][3]. 음성신호를 시간-주파수 영역에서 관찰할 경우 시변 순간 주파수(instantaneous frequency)와

시변 순간 진폭(instantaneous amplitude)을 갖는 AM-FM 신호 성분들을 관찰할 수 있는데 각 성분들의 순간 주파수, 순간 진폭, 순간 대역폭(instantaneous bandwidth)등은 발성 시스템의 동적 특성의 변화 즉 감정 상태 및 변화에 대한 여러 가지 정보를 함유하고 있는 중요한 파라미터로 판단된다[2][3]. 즉 각각의 AM-FM 성분들의 시간적 변화, 좀더 정확하게 AM-FM 성분들의 순간 진폭, 순간 주파수 및 순간 대역폭들의 시간적 변화에 화자의 감정 상태의 변화가 반영되는 것으로 판단된다. 또한 시간-주파수 영역에서 포만트 주파수(formant frequency)의 시간적 변화를 관찰할 수 있는데 포만트 주파수의 시간적 변화는 각 성분들의 진폭의 시간적 변화 즉 발성 시스템의 동적 특성의 변화가 반영된 것으로 판단된다[2]. 음성 신호의 AM-FM 성분들의 순간 진폭, 순간 주파수 및 순간 대역폭과 같은 미시적 정보들은 음성 신호의 전체 에너지 변화와 같은 거시적 정보와 더불어 음성 인식 및 합성과 감정 인식 및 표현 필요한 중요한 수단을 제공할 것으로 판단된다. 이와 같은 관점에서 볼 때 음성 신호를 AM-FM 성분별로 분해하고 각각의 성분에 대한 순간 진폭과 순간 주파수의 추출은 매우 의미 있는 작업이 될 것이다. 본 논문은 목소리에 실린 감정 상태 및 변화에 대한 정보가 음성 신호의 에너지 변화와 각각의 AM-FM 성분들의 순간 주파수의 궤적 그리고 순간 진폭의 변화에 반영된다는 가정하에 각각의 AM-FM 성분들의 분해 및 분해된 각 성분들의 순간 주파수, 순간 진폭을 추정하는 방법을 다룬다.

표 1. 음성기반 감정인식 또는 표현에 사용되는 파라미터[6]

Measures relating to	parameters
Tune	tune duration, fit of tune to a quadratic function, number of inflections in fundamental frequency (F0) contour per tune
Spectrum	energy below 250 Hz
Intensity contour (excluding pauses)	mean intensity, median intensity, inter-quartile range of intensity distribution
Intensity at local extrema in the intensity contour	mean at maxima, inter-quartile range for intensities at maxima, mean at minima, inter-quartile range for intensities at minima
Magnitude of rises or falls in the intensity contour	inter-quartile range for magnitudes of rises, inter-quartile range for magnitudes of falls
Pitch of points in the F0 contour	number of contributing observations, mean, inter-quartile range
Pitch at local extrema in the F0 contour	inter-quartile range for pitch at maxima, inter-quartile range for pitch at minima, inter-quartile range for pitch at all local extrema
Magnitude of rises in the F0 contour	median, inter-quartile range
Durations of rises and falls in the intensity contour	median duration for rises, median duration for falls
Durations of level sections in the intensity contour ('plateaux')	inter-quartile range for plateaux at intensity peaks, upper limit (90%) of range for plateaux at intensity peaks, median for plateaux at intensity minima, inter-quartile range for plateaux at intensity minima
Durations of features in the F0 contour	median of silence durations, inter-quartile range for durations of silences, median duration of falls, median duration of plateaux at F0 maxima, inter-quartile range for duration of plateaux at F0 maxima

신호의 시간-주파수 표현에는 스펙트로그램, Cohen's Class, 스칼로그램(scalogram), Cone-Kernel Distribution과 같은 것들이 있는데 음성 신호에서는 스펙트로그램이 주로 사용된다[1][23]. 음성 신호의 스펙트로그램을 통한 해석은 크게 두 가지 문제점을 가지고 있다. 첫 번째 문제는 스펙트로그램의 해상도 문제로서 시간-주파수에 대한 불확정성 원리로 인하여 발생한다[1][23]. 두 번째는 역변환 문제이다[1][23]. 본 연구에서는 두 번째 문제인 역변환 문제를 다룬다. 음성 신호의 Short-time Fourier 변환 표현은 각 성분들의 관찰을 허용하지만 시간-주파수 표현으로부터 각각의 AM-FM 성분을 분해하고 순간 주파수 및 순간 진폭을 추출할 수는 없다. 즉 Short-time Fourier 변환 및 역변환의 경우, 시간-주파수 표현(time-frequency representation)에서 순간 주파수의 궤적을 따라 에너지가 집중된 신호를 분리하기 위하여 이진 마스크(binary mask)를 사용하여 시간-주파수 표현을 변형할 경우 변형된 시간-주파수 표현에 대응하는 시간 영역의 실제 신호(real signal)가 항상 존재하는 것을 보장할 수 없다[6]. 즉, 변형된 시간-주파수 표현(masked time-frequency representation)을 역 Short-time Fourier 변환하여 얻은 시간 영역 신호에는 허수부분(imaginary part)이 포함되어 있으므로 진짜 신호의 추정치로 볼 수 없다. 그러므로 Short-time Fourier 변환에 기반을 둔 시간-주파수 표현은 개략적인 정보의 추출이나 변화 추이의 관찰 등에는 사용할 수 있으나 음성 신호의 AM-FM 성분 분리 및 순간 주파수, 순간 진폭과 같은 미시적 정보 추출에는 사용할 수 없다. Wavelet 변환에 기반을 둔 스칼로그램은 역변환 문제는 없으나 주파수 축이 로그 눈금이므로 고주파 부분에서 해당도가 매우 떨어지는 문제점을 가지고 있다. 즉 Wavelet 역변환을 사용하여 순간 주파수 변화 범위가 큰 인접해 있는 두 AM-FM 성분을 분리할 경우 각 성분들 간의 간섭현상이 심하므로 분리 오차가 크다[1].

표 2. 6 가지 감정 형태와 음성 파라미터간의 관계[8]

Emotions		Anger	Surprise	Joy
기본주파수 (피치)		주파수의 시간적 변동, 엑센트가 있는 음절의 에너지 증가, 억양의 불규칙성, 피치의 평균값 증가	엑센트가 있는 음절의 급격한 에너지 증가 후 마지막 음절에서 에너지의 감쇠	불규칙한 간격으로 멜로디 상승
음향파라미터	평균 기본주파수	평균 증가	-	평균 증가
	기본주파수 범위	매우 넓어짐	범위는 넓어지고 중간 값은 정상이거나 높아짐	매우 넓어짐
	강도(intensity)	커짐	-	증가
	발성속도	고속 또는 저속	-	발성속도는 증가, 박자는 느려짐
	스펙트럼	비마찰음의 평균 스펙트럼이 높음	-	고주파수 에너지 증가
음질		긴장, 숨소리, 저음역 소리, 요란함	숨소리	긴장, 숨소리, 요란함

표 2. (계속)

Emotions		Fear	Disgust	Sadness
기본주파수 (피치)		패턴 붕괴, 방향이 심하게 변화	피치가 넓어짐, 마지막 억양이 강하됨	억양이 강하됨
음향파라미터	평균 기본주파수	평균 증가	매우 낮아짐	평균 이하
	기본주파수 범위	증가	약간 넓어짐	약간 좁아짐
	강도(intensity)	정상	낮아짐	감소
	발성속도	고속 또는 저속	매우 빨라짐	약간 느려짐
	스펙트럼	고주파수 영역에서의 에너지 증가	-	억양이 강하됨
음질		긴장, 불규칙한 유성음화	으르렁거림	이완, 공명음

본 연구는 음성의 AM-FM 성분 분해 및 각 성분의 순간 진폭과 순간 주파수 추정을 위한 방법을 다룬다. 참고문헌 [10]에서는 AM-FM 성분 분해에 사용한 가변대역폭 필터의 시변 차단 주파수를 스펙트로그램의 시각적 관찰을 바탕으로 수동으로 결정하였으며 추정된 순간 주파수에 고주파 잡음이 포함되어 있다. 본 논문에서는 이와 같은 문제점들을 해결하였다. 즉 AM-FM 모델을 사용하여 제 1 기본 주파수를 자동으로 추정하는 방법을 제안하였고 이를 바탕으로 각 성분들을 분리하였다. 또한 각 성분들의 순간 주파수 및 순간 진폭을 정확하게 추정하는 방법을 제안하였다.

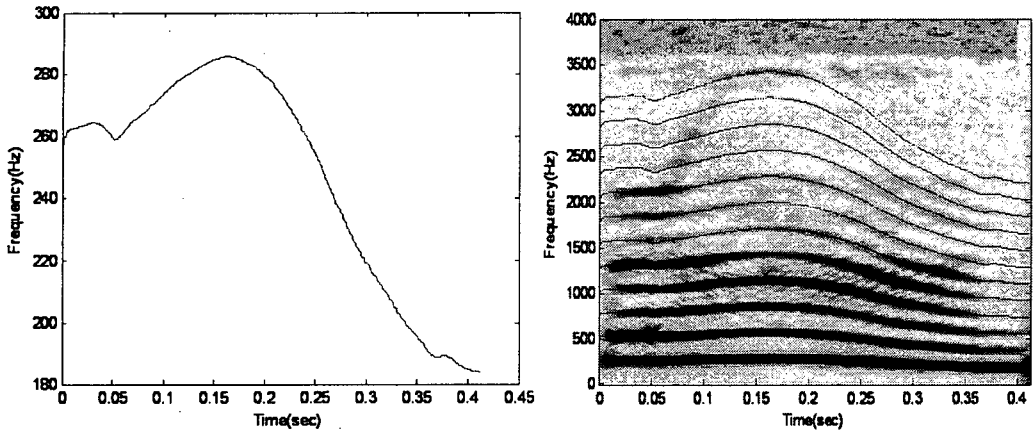
2. AM-FM 성분 분리

2.1 음성 신호의 AM-FM 모델

AM-FM 신호 $x_0(t) = a_0(t)\cos(\phi_0(t))$에서 순간 주파수 $\omega_0(t)$는 다음과 같이 정의된다[11].

$$\omega_0(t) = D\phi_0(t) \quad \text{rad/sec}$$

여기서 $D = d/dt$이다. $a_0(t)$는 순간 진폭이다. $\phi_0(t)$는 순간 위상이다. 시변 순간 주파수를 갖는 AM-FM 신호를 시간-주파수 평면에서 관찰할 경우 순간 주파수의 궤적을 따라 신호의 에너지가 집중되는 것을 알 수 있다. AM-FM 신호의 순간 대역폭은 순간 진폭과 순간 주파수의 시간적 변화량에 의해 결정된다[1]. 신호의 시간-주파수 표현으로부터 순간 주파수, 순간 대역폭을 구하기 위한 다양한 연구가 수행되었다. 일반적으로 신호의 시간-주파수 표현은 창 함수의 모양과 길이에 따라 달라지므로 시간-주파수 표현으로부터 직접 정확한 순간 주파수, 순간 진폭 및 순간 대역폭을 구하는 것은 매우 어렵다[1]. 그러나 시간-주파수 표현이 순간 주파수, 순간 진폭 및 순간 대역폭의 시간적 변화에 대한 유용한 정보를 제공함에는 틀림없다.



(a) (b)

그림 1. (a) 기본주파수
(b) 스펙트로그램(8 kHz 표본화, 12 bit 양자화된 20 대 여성 음성의 모음 부분)

음성 신호의 파형은 발성 시스템의 떨림 및 경계조건과 소리의 공진 등에 의하여 결정되는데 음성신호의 모음부분은 다중 성분 신호이며 다음과 같이 AM-FM 신호의 합으로 모델링할 수 있다.

$$x(t) = \sum_{n=0}^{N-1} a_n(t)\cos(\phi_n(t)) + N(t) \tag{1}$$

여기서 $a_n(t)$는 n 번째 AM-FM 성분의 순간 진폭이고 $\phi_n(t)$는 n 번째 AM-FM 성분의 순간 위상이다. 각 성분들의 순간 주파수는 다음과 같다.

$$\omega_n(t) = D\phi_n(t) \tag{2}$$

식 (1)은 음성 신호를 어느 정도 주기성을 갖는 준 주기적 신호(quasi-periodic signal)인 AM-FM 성분들과 잡음을 포함한 완전 비 주기적인 신호 $N(t)$ 나누었다. 일반적으로 시간-주파수 평면에서 모음 음성 신호의 에너지는 주로 AM-FM 성분의 순간 주파수 $\omega_n(t)$의 궤적에 집중되며 $N(t)$의 에너지는 전 영역에 분포하며 AM-FM 성분에 비하여 상대적으로 작다. 식 (1)과 (2)에서 $\omega_0(t)$는 제 1 AM-FM 성분의 순간 주파수 즉 기본 주파수(fundamental frequency)이다.

인간의 음성 발성 시스템의 특성상 서로 다른 순간 주파수들 간의 간격은 일정하지 않으며 고조파 순간 주파수들은 제 1 순간 주파수, 즉 기본 주파수 $\omega_0(t)$에 비례한다. 즉

$$\omega_n(t) \simeq b_n\omega_0(t) \tag{3}$$

여기서 b_n , $n=0,1,2,\cdots,N-1$는 상수이다. 식 (2) 및 (3)을 사용하면 식 (1)은 다음과 같이 표현된다.

$$x(t)=\sum_{n=0}^{N-1}a_n(t)\cos(\phi_n(t))+N(t)$$
$$=\sum_{n=0}^{N-1}a_n(t)\cos(b_n\int_0^t\omega_0(\tau)d\tau+\phi_n(0))+N(t) \quad (4)$$

여기서 $\phi_n(0)$는 초기 위상이다. <그림 1>은 음성신호의 AM-FM 모델이 타당함을 보여준다. <그림 1>의 (a)는 3 장에서 논의한 방법을 사용하여 추정한 기본 주파수 $\omega_0(t)$를 보여준다. <그림 1>의 (b)는 $b_n=n+1$으로 한 경우를 보여준다. 모두 12 개의 AM-FM 성분에 대한 순간 주파수를 나타냈는데 스펙트럼의 각 성분별 최대치와 정확히 일치함을 알 수 있다.

2.2 음성 신호의 AM-FM 성분 분리

일반적으로 음성 신호의 AM-FM 성분의 순간 주파수들은 일정하지 않고 시간에 따라 변하므로 LTI(linear time invariant) 필터와 같은 기존의 방법으로는 각 성분들을 분리할 수 없다. 음성 신호의 시간-주파수 표현은 순간 주파수에 대한 유용한 정보는 제공하나 순간 주파수를 추출하는 방법은 제공하지 못한다. 그러나 본 연구에서 다루는 확장 Fourier 변환을 사용할 경우 각 AM-FM 성분들을 분리할 수 있다. 확장 Fourier 변환(EFT: extended Fourier transform) 및 역 변환은 다음과 같이 정의된다[17][18].

$$F[x(t),g(t)]=\int_{-\infty}^{\infty}x(t)\frac{1}{g(t)}\,e^{-j\omega_g\int_0^t\frac{1}{g(t)}d\tau}dt$$
$$=X(\omega_g) \quad (5)$$

$$F^{-1}[X(\omega_g),g(t)]=\int_{-\infty}^{\infty}X(\omega_g)\,e^{j\omega_g\int_0^t\frac{1}{g(t)}d\tau}d\omega_g$$
$$=x(t) \quad (6)$$

여기서 ω_g는 주어진 함수 $g(t)$에 대한 확장 Fourier 주파수 변수이다. 확장 Fourier 변환을 바탕으로 가변 대역폭 필터(variable bandwidth filter)를 구현할 수 있다. 확장 Fourier 주파수 영역에서 표현된 다음과 같은 이상적인 저역 통과 가변 대역폭 필터(lowpass variable bandwidth filter)를 고려하자.

$$H(\omega_g)=\begin{cases}1 & \text{if}\ |\omega_g|<\alpha\\ 0 & else\end{cases}$$

여기서 α는 저역 통과 가변 대역폭 필터의 차단 주파수 이다. 상기 필터의 시간-주파수 평면에서의 차단 주파수의 궤적은 $\alpha/g(t)$이다. 이때 $x_l(t)=F^{-1}[X(\omega_g)H(\omega_g),g(t)]$는 이상적인 가변 대역폭 필터를 사용하여 저역 통과된 신호이다. 고역 통과 가변대역폭 필터, 대역 통과 가변대역폭 여과기, 대역 저지 가변대역폭 여과기 등이 있다.

다음은 확장 Fourier 변환에 대한 변조(modulation) 성질을 고찰하자. 기저 신호 $a_0(t)$를 $\cos(b_0\int_0^t 1/g(t)d\tau+\phi_0(0))$를 사용하여 변조한 신호 $x_0(t)$를 고려하자.

$$x_0(t)=a_0(t)\cos(b_0\int_0^t\frac{1}{g(\tau)}\,d\tau+\phi_0(0)) \tag{7}$$

여기서 b_0는 상수이고 $\phi_0(0)$는 초기치 이다. 식 (7)의 EFT를 구하면 다음과 같다.

$$\begin{aligned}X_0(\omega_g)&=\int_{-\infty}^{\infty}x_0(t)\frac{1}{g(t)}\,e^{-j\omega_g\int_0^t\frac{1}{g(t)}\,d\tau}dt\\&=\int_{-\infty}^{\infty}\frac{1}{2}a_0(t)(e^{j(b_0\int_0^t\frac{1}{g(\tau)}\,d\tau+\phi_0(0))}+e^{-j(b_0\int_0^t\frac{1}{g(\tau)}\,d\tau+\phi_0(0))})\frac{1}{g(t)}\,e^{-j\omega_g\int_0^t\frac{1}{g(t)}\,d\tau}dt\\&=\frac{1}{2}e^{j\phi_0(0)}\int_{-\infty}^{\infty}a_0(t)\frac{1}{g(t)}\,e^{-j(\omega_g-b_0)\int_0^t\frac{1}{g(t)}\,d\tau}dt+\\&\quad\frac{1}{2}e^{-j\phi_0(0)}\int_{-\infty}^{\infty}a_0(t)\frac{1}{g(t)}\,e^{-j(\omega_g+b_0)\int_0^t\frac{1}{g(t)}\,d\tau}dt\\&=\frac{1}{2}(A_0(\omega_g-b_0)e^{j\phi_0(0)}+A_0(\omega_g+b_0)e^{-j\phi_0(0)})\end{aligned} \tag{8}$$

여기서 $A_0(\omega_g)=F[a_0(t),g(t)]$이다. 기저 신호 즉 순간 진폭 $a_0(t)$를 변조 함수 $\exp(jb_0\int_0^t 1/g(\tau)d\tau)$로 진폭 변조할 경우 확장 Fourier 주파수 영역에서 변조된 함수는 기저함수의 EFT가 b_0 만큼 주파수 천이된 형태로 표현된다. 식 (4)의 EFT 표현을 얻기 위해 $g(t)$를 다음과 같이 놓자.

$$\frac{1}{g(t)}=\omega_0(t) \tag{9}$$

식 (8)을 사용하면 식 (4)의 EFT는 다음과 같이 표현된다. 모음의 경우 음성 발성 기관의 공진에 의한 신호가 대부분이므로 $N(t)$를 무시할 수 있다. 또한 비교적 좋은 환경에서 측정한 음성 신호의 경우는 외부 잡음을 무시할 수 있다.

$$\begin{aligned}X(\omega_g)&=F[x(t),g(t)]\\&=\frac{1}{2}\sum_{n=0}^{N-1}(A_n(\omega_g-b_n)e^{j\phi_n(0)}+A_n(\omega_g+b_n)e^{-j\phi_n(0)})\end{aligned} \tag{10}$$

여기서 $A_n(\omega_g)$는 각 AM-FM 성분의 기저 신호 즉 순간 진폭의 EFT로 다음과 같다.

$$A_n(\omega_g) = F[a_n(t), g(t)] \tag{11}$$

식 (10)에서 만일 $A_n(\omega_g - b_n)$, $n = 0, 1, 2, \cdots, N-1$의 대역이 서로 겹치지 않는다면 식 (5) 및 (6)을 바탕으로 한 가변 대역폭 필터를 사용하여 음성 신호의 각 AM-FM 성분들을 분리할 수 있다. <그림 3>은 <그림 2>의 (a) 신호의 EFT이다. <그림 3>으로부터 $A_n(\omega_g - b_n)$의 대역이 서로 겹치지 않음을 알 수 있다. 예를 들어 제 1 AM-FM 성분은 차단 주파수가 α=1.5인 저역 통과 가변대역폭 필터를 사용하여 분리할 수 있으며 제 2 AM-FM 성분은 상단 및 하단 차단주파수가 각각 1.5와 2.5인 대역 통과 가변대역폭 필터를 사용하여 분리할 수 있다.

<그림 4>는 가변대역폭 필터를 사용하여 분리한 AM-FM 성분들을 보여준다. 제 16 AM-FM 성분은 Aliasing 효과로 인해 발생한 오차를 많이 포함하고 있다. 즉 1000에서 1500번째 표본 사이에 Aliasing 오차 신호가 분포하고 있다. <그림 2>의 (b)는 다음과 같이 정의된 성분 분해 오차를 보여 준다.

$$e(t) = s(t) - \sum_{n=0}^{N-1} a_n(t)\cos(\phi_n(t))$$

여기서 $s(t)$는 AM-FM 성분 분해를 할 원래 신호이다. <그림 2>의 (b)는 제 15 AM-FM 성분까지 고려한 오차 신호를 보여준다.

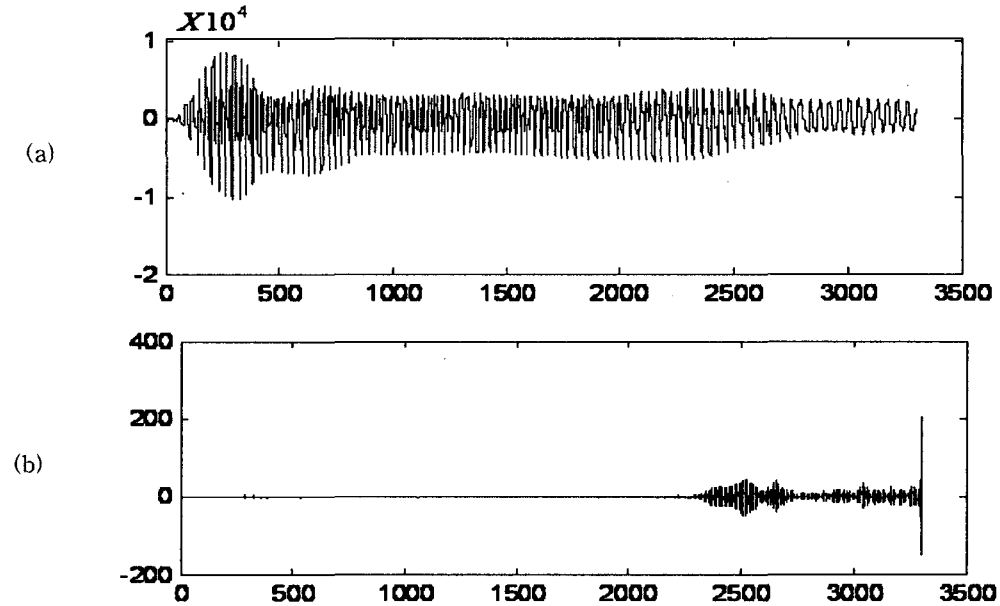


그림 2. (a) 그림 1의 스펙트로그램에 대응하는 음성신호 (b) AM-FM 성분 분해 오차 신호

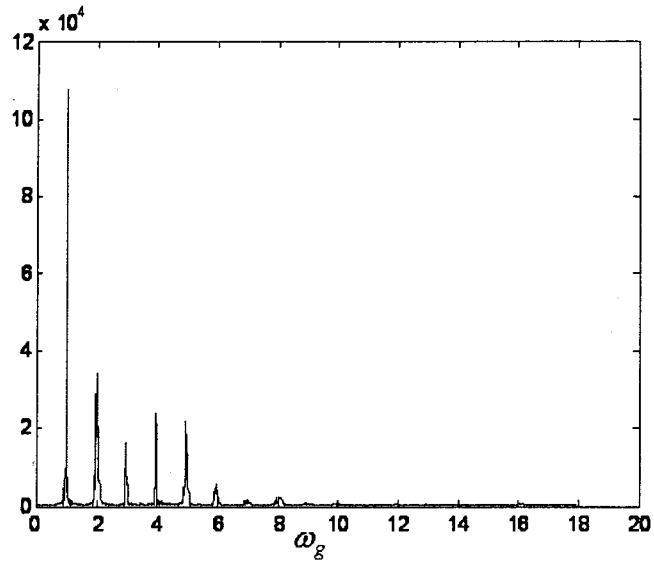
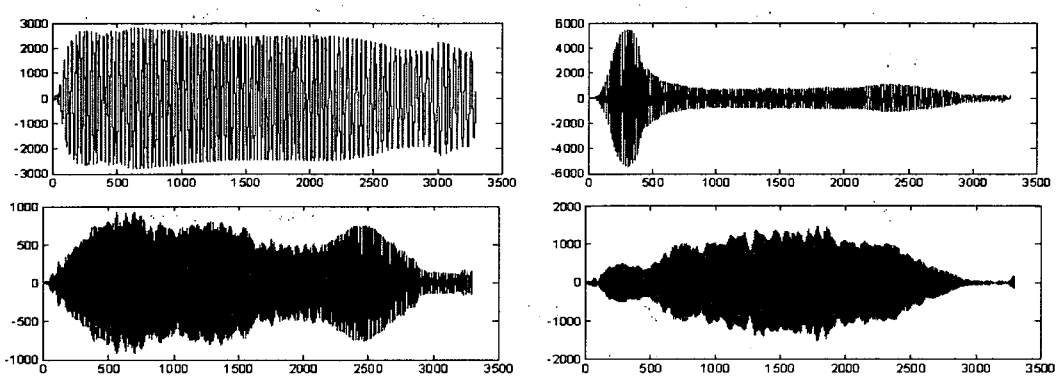


그림 3. 그림 2의(a) 음성 신호의 EFT.

3. AM-FM 성분 분리에 사용되는 순간 주파수 추정

가변대역폭 필터의 대역폭의 시간적 변화를 결정짓는 함수 $1/g(t)$는 첫 번째 AM-FM 성분 $a_0(t)\cos(\phi_0(t))$의 순간 주파수 $\omega_0(t)=D\phi_0(t)$로 선택하였다. 즉 $1/g(t)=D\phi_0(t)$로 하였다. $1/g(t)$는 스펙트로그램을 바탕으로 한 방법을 사용하여 구할 수 있다. 피치의 주기는 여성의 경우 대략적으로 3-4 msec 정도이고 남성의 경우는 이것의 약 두 배 정도가 된다. 피치 주기의 시간적 변화는 순간 주파수의 시간적 변화의 원인이 된다. 그러므로 시간-주파수 영역에서 여성음성의 기본주파수는 약 200 Hz에서 350 Hz 범위에 분포하고 남성 음성의 기본주파수는 대략적으로 100 Hz에서 175 Hz 범위에 분포한다.



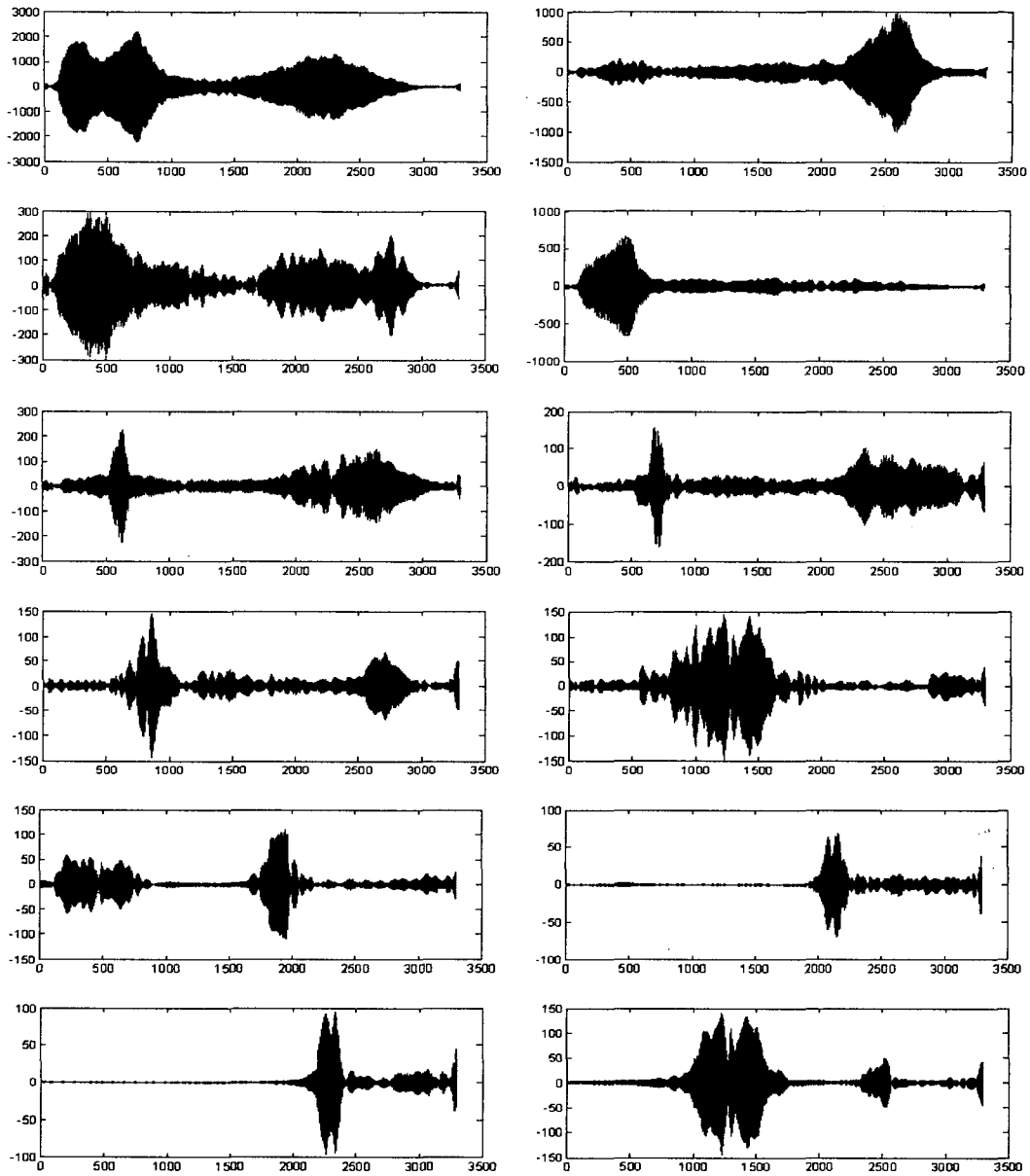


그림 4. 분해된 AM-FM 성분

상단 좌측에 제 1 성분, 상단 우측에 제 2 성분 순서로 그렸으며 하단 우측은 제 16 성분이다.

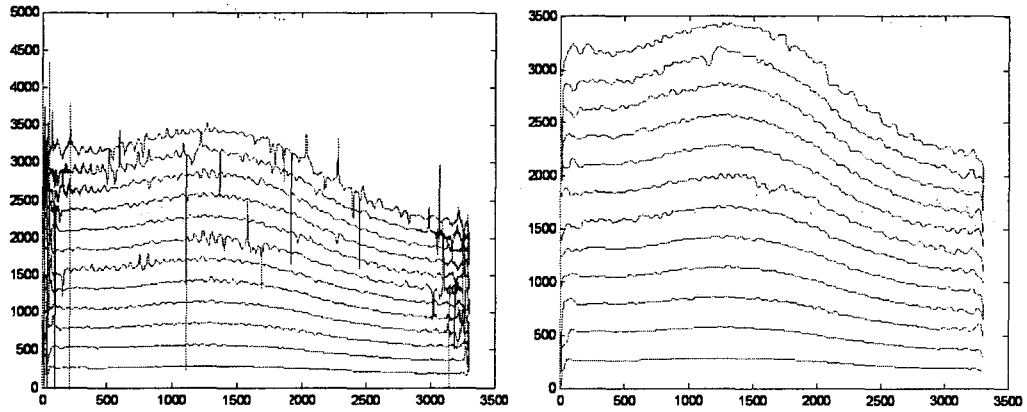
음성신호에서 고조파 순간 주파수의 변화는 매우 심하지만 기본주파수의 변화는 그렇게 심하지 않으므로 매 시간 순간마다 200 Hz에서 350 Hz(여성 음성의 경우) 주파수 범위에 있는 스펙트럼의 최대치 및 최대치 대응하는 주파수를 쉽게 구할 수 있다. 현실적으로 시간-주파수 표현으로 스펙트로그램을 많이 사용하므로 스펙트로그램에서 200 Hz에서 350 Hz(여성 음성의 경우) 범위에 있는

스펙트로그램의 최대치에 대응하는 주파수의 궤적을 기본주파수로 선택하였다. $1/g(t)$는 모든 t에 대해서 다음 관계식을 만족한다.

$$spgr(t, 2\pi/g(t)) = \max_{f_- < f < f_+} spgr(t, 2\pi f), \quad \text{모든 } t \text{ 에 대하여} \tag{12}$$

여기서 $spgr(t, 2\pi f)$는 음성신호의 스펙트로그램이고 f_-와 f_+는 남성음성의 경우는 100과 175로 선택하고 여성음성의 경우는 200과 350으로 선택한다.

스펙트럼을 바탕으로 기본주파수 즉 제 1 성분의 순간주파수를 추정할 경우 스펙트로그램의 최대치에 대응하는 주파수의 궤적이 $a_0(t)\cos(\phi_0(t))$의 Analytic 신호인 $a_0(t)\exp(j\phi_0(t))$의 순간주파수와 정확히 일치하지 않는 문제점을 가지고 있다. 즉 스펙트로그램의 최대치에 대응하는 주파수의 궤적이 $D\phi_0(t)$와 정확히 일치하지 않는다. 그러나 성분분리에 사용되는 가변대역폭필터의 시변 차단주차수의 궤적은 시간-주파수 평면에서 대략적으로 기본주파수와 제 1 고조파의 순간 주파수의 중간 정도에 위치하면 되므로 $D\phi_0(t)$를 정확히 추정할 필요는 없다. 정확한 기본주파수의 추정은 IV절에서 다룬다.



(a) (b)

그림 5. Hilbert 변환을 사용하여 추정한 순간 주파수와 잡음이 제거된 순간 주파수

4. 순간 진폭 및 순간 주파수 추정

AM-FM 신호 $x_0(t) = a_0(t)\cos\phi_0(t)$를 고려하자. 일반적으로 우리가 측정하는 신호는 순간 진폭과 순간 위상의 정현 함수가 곱하여진 $x_0(t)$이므로 이것으로부터 순간 진폭과 순간 위상 그리고 순간 주파수를 분리해야 한다. 측정된 $x_0(t)$를 만드는 $a_0(t)$와 $\phi_0(t)$의 조합은 무수히 많으므로 의

미 있는 조합을 찾아내는 것은 단순한 문제가 아니다[11][21][22]. $x_0(t)$는 어떤 물리적 현상에 의해 발생한 신호의 측정치이므로 $x_0(t)$의 순간 주파수와 순간 진폭은 대응하는 물리현상의 내부적 과정에 대한 정보를 포함하고 있을 것이다[11]. 즉 $a_0(t)$와 $\phi_0(t)$가 물리적 과정의 모델과 잘 일치한다면 의미있는 조합으로 받아들여질 수 있다[11][21]. AM-FM 신호 $x_0(t) = a_0(t)\cos(\phi_0(t))$의 순간 주파수와 순간 진폭을 추정하기 위하여 흔히 Hilbert 변환을 사용한다. Hilbert 변환은 다음과 같이 정의된다[11].

$$H[x_0(t)] = c.p.v. \int_{-\infty}^{\infty} (x_0(\tau)/\pi(t-\tau))d\tau \tag{13}$$

여기서 $c.p.v.$는 Cauchy 주치(principle value)이다. $a_0(t)$와 $\cos(\phi_0(t))$의 Fourier 스펙트럼들이 서로 겹치지 않고, $\cos(\phi_0(t))$의 Fourier 스펙트럼이 매우 좁은 영역에 분포할 경우 Hilbert 변환을 사용하여 $x_0(t)$로부터 $a_0(t)$와 $\phi_0(t)$를 분리할 수 있다. 즉 다음과 같은 식을 얻을 수 있다.

$$\begin{aligned} z_0(t) &= x_0(t) + jH[x_0(t)] \\ &= x_0(t) + ja_0(t)H[\cos(\phi_0(t))] \\ &= x_0(t) + ja_0(t)\sin(\phi_0(t)) \\ &= a_0(t)\exp(j\phi_0(t)) \end{aligned} \tag{14}$$

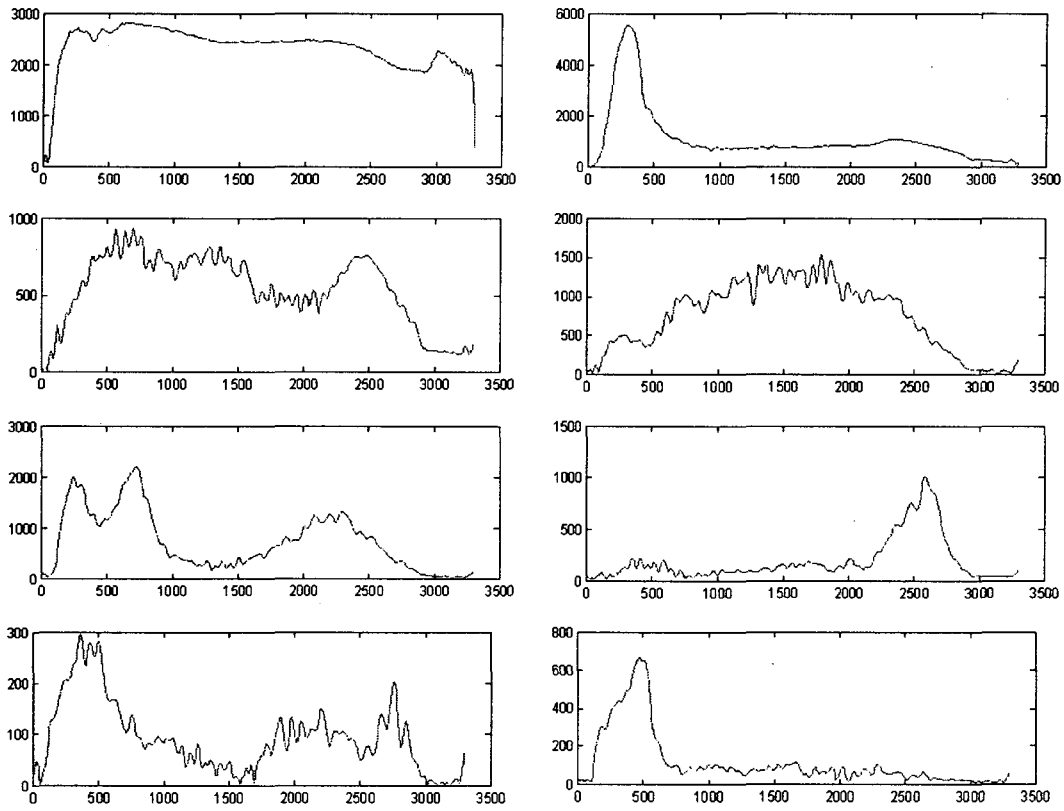
여기서 $z_0(t)$는 $x_0(t)$의 Analytic 신호이다[12][22]. 음성신호에서 각 AM-FM 성분의 순간 진폭의 시간적 변화는 매우 크지 않다. 또한 FM 변조 부분은 인간 발성 시스템의 특성상 변조 주파수의 변동이 매우 심하지 않다. 그러므로 순간 진폭 $a_0(t)$와 위상부분 $\cos(\phi_0(t))$의 Fourier 스펙트럼들이 서로 겹치지 않고, $\cos(\phi_0(t))$의 Fourier 스펙트럼이 매우 좁은 영역에 분포할 것이다. 그러므로 잡음이 없는 음성 신호의 AM-FM 성분에 대하여 식 (14)는 비교적 잘 동작할 것이다.

그러나 일반적으로 음성 신호는 잡음을 가지고 있다. 잡음이 섞인 실제 음성 신호에 대하여 식 (14)을 적용하여 얻은 순간 주파수는 잡음의 영향으로 인하여 고주파 성분이 섞인다. 즉 순간 주파수는 순간 위상의 시간 미분이므로 추정된 순간 위상에 잡음이 있다면 고주파 성분이 증폭되고 결과적으로 순간 주파수의 추정 오차가 클 것이며 더 나아가 순간 진폭의 추정 오차 또한 클 것이다. <그림 5>의 (a)는 Hilbert 변환을 사용하여 추정한 순간 주파수를 보여준다. 각각의 AM-FM 성분들에는 고주파 잡음이 섞여 있음을 관찰할 수 있다. 순간 주파수는 음성 발생 시스템의 물리적 특성이 반영된 것이므로 정상인의 발성 시스템에서 <그림 5>의 (a)와 같이 순간 주파수가 급격히 변하는 것은 불가능하다. 그러므로 순간 주파수의 급격한 변화는 잡음에 의해 발생한 것으로 추정된다. 잡음 효과는 식 (14)에 의해 얻어진 순간 진폭도 왜곡시킬 것이다. 이런 문제는 확장 Fourier 변환의 복조(demodulation) 성질을 사용하면 해결할 수 있다. 확장 Fourier 변환의 복조 성질은 다음과 같다.

신호 $x_0(t) = a_0(t)\cos\phi_0(t)$를 고려하자. $x_0(t)$에 $\cos\phi_0(t)$을 곱하여 정리하면 다음 수식을 얻을 수 있다.

$$\begin{aligned} s(t) &= x_0(t)\cos\phi_0(t) \\ &= a_0(t)\cos^2\phi_0(t) \\ &= a_0(t)\frac{1}{2}(1+\cos(2\phi_0(t))) \\ &= \frac{1}{2}a_0(t) + \frac{1}{2}a_0(t)\cos(2\phi_0(t)) \end{aligned} \tag{15}$$

$F[a_0(t), g(t)]$와 $F[a_0(t)\cos(2\phi_0(t)), g(t)]$의 스펙트럼이 서로 겹치지 않는다면 가변 대역폭 저역통과 필터를 사용하여 $a_0(t)$를 분리할 수 있다. 식 (15)는 $\phi_0(t)$를 알고 있을 경우 $a_0(t)$를 분리할 수 있다는 것을 의미한다. 그러나 일반적으로 $\phi_0(t)$를 알 수 없다. 식 (15)는 $\phi_0(t)$가 정확할수록 $a_0(t)$도 정확해진다는 것을 의미한다. $a_0(t)$와 $\phi_0(t)$의 추정치를 구하는 과정은 다음과 같다. 첫째 식 (14)를 사용하여 구한 순간 위상을 미분하여 순간 주파수를 구한다. 이 과정에서 얻어진 순간 주파수는 고주파 잡음을 포함하고 있으므로 필터를 사용하여 고주파 잡음을 제거한다.



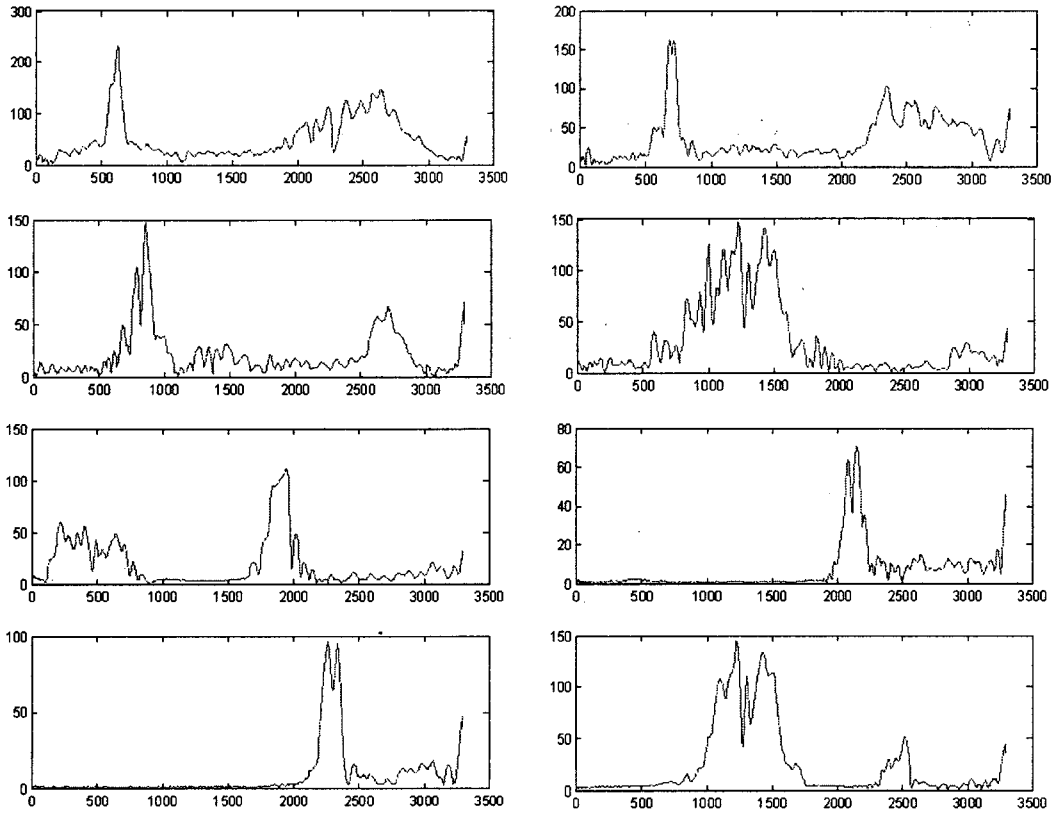


그림 6. AM-FM 성분들의 순간 진폭.
상단 좌측은 제 1 성분의 순간 진폭, 상단 우측은 제 2 성분의 순간 진폭이며
하단 우측은 제 16 성분의 순간 진폭이다.

다시 역으로 고주파 잡음이 제거된 순간 주파수를 적분하여 고주파 잡음이 제거된 순간 위상 즉 추정된 순간 위상을 구한다. 이때 추정된 순간 위상을 $\overline{\phi_0}(t)$ 라 하자. 즉 $\overline{\phi_0}(t)$를 $\phi_0(t)$의 추정치(estimate)라고 하자. 그러면

$$\phi_0(t) = \overline{\phi_0}(t) + n(t)$$

여기서 $n(t)$는 순간 위상의 추정 오차로 매우 작다고 가정 하자.

$a_0(t)$를 구하기 위하여 $x_0(t)$를 $\cos\overline{\phi_0}(t)$로 복조를 하자. 그러면 $x_0(t)\cos\overline{\phi_0}(t)$는 다음과 같다.

$$\begin{aligned} x_0(t)\cos\overline{\phi_0}(t) &= a_0(t)\cos\phi_0(t)\cos\overline{\phi_0}(t) \\ &= a_0(t)\cos(\overline{\phi_0}(t)+n(t))\cos\overline{\phi_0}(t) \\ &= \frac{1}{2}a_0(t)(\cos n(t) + \cos(2\overline{\phi_0}(t)+n(t))) \end{aligned}$$

한편 $n(t) \simeq 0$이고 $\cos n(t) \simeq 1$ 이므로

$$x_0(t)\cos\overline{\phi_0}(t) \simeq \frac{1}{2}a_0(t)(1+\cos(2\overline{\phi_0}(t))) \tag{16}$$

이때 $1/g(t)$를 다음과 같이 놓자.

$$\frac{1}{g(t)} = \frac{d}{dt}\overline{\phi_0}(t)$$
$$= \overline{\omega_0}(t)$$

만일 $F[a_0(t), g(t)]$와 $F[a_0(t)\cos(2\overline{\phi_0}(t)), g(t)]$의 스펙트럼이 서로 심하게 겹치지 않는다면 저역통과 가변대역폭 필터를 사용하여 $a_0(t)$의 추정치를 구할 수 있다. <그림 5>는 50 개의 표본을 사용한 메디안 필터를 사용하여 고주파 잡음을 제거한 순간 주파수들을 보여준다. 고주파 잡음이 효과적으로 제거되었음을 알 수 있다. <그림 6>은 <그림 5>의 순간 주파수와 EFT의 복조 성질을 사용하여 추정한 순간 진폭을 보여준다.

5. 결론 및 향후 연구

음성 신호의 모음 부분은 다중성분 신호이며 AM-FM 성분들의 합으로 모델링할 수 있다. 음성 신호를 시간-주파수 영역에서 관찰할 경우 시변 순간 주파수를 갖는 AM-FM 신호 성분들을 관찰할 수 있는데 각 성분들의 순간 주파수, 순간 진폭, 순간 대역폭 등은 음성 신호에 대한 여러 가지 정보를 함유하고 있는 단순하면서도 중요한 파라미터이다.

순간 주파수 및 순간 진폭이 시간적으로 변화는 경우 각 AM-FM 성분들의 Fourier 스펙트럼이 서로 겹쳐지므로 LTI 필터와 같은 기존의 방법들을 사용할 경우 AM-FM 성분들을 분리하기가 매우 어렵다. 본 연구에서는 음성신호의 모음 부분의 AM-FM 성분들을 가변 대역폭 필터를 사용하여 각 성분별로 분해하였다. 사용된 가변 대역폭 필터의 시변 대역폭을 결정하기 위하여 기본주파수를 추정하였다. 또한 Hilbert 변환과 확장 Fourier 변환의 복조 성질을 사용하여 각 AM-FM 성분들의 순간 진폭 및 순간 주파수를 추정하였다. Hilbert 변환만을 사용하여 순간 진폭 및 순간 주파수를 추정할 경우 고주파 잡음이 유입되는데 본 논문에서 제안한 방법을 사용할 경우 추정한 순간 진폭 및 순간 주파수에 고주파 잡음이 유입되지 않는다.

추정한 순간 진폭 및 순간 주파수를 변경할 경우 음성 신호의 스펙트럼 특성이 변할 것이다. 이런 성질은 코퍼스 기반 음성 합성 시스템에서 두 단어를 연결할 때 스펙트럼 특성이 서로 다름으로 인해서 발생하는 부자연스런 음을 제거하는데 유용하게 사용할 수 있을 것이다. 즉 연결할 두 단어의 에너지뿐 만 아니라 시간-주파수 표현 스펙트럼도 부드럽게 연결할 경우 합성음의 부자연스러

움이 많이 제거 될 것이다. 본 연구는 스펙트럼 변환을 바탕으로 한 감정 음성 표현을 위한 사전 연구의 일환으로 수행되었다.

참 고 문 헌

[1] Qian, Shie. & Chen, Dapang. 1996. *Joint Time-frequency Analysis: Methods and Applications*, Prentice-Hall.

[2] Quatieri, T. F. 2002. *Discrete-time Speech Signal Processing*, Prentice-Hall.

[3] Rao, A. & Kumaresan, R. 2000. "On Decomposing Speech into Modulated Components." *IEEE Tr. on Speech and Audio Processing*, Vol. 8, No. 3, 240-254.

[4] Rabiner, L. R. & Juang, B. H. 1993. *Fundamental of Speech Recognition*, Prentice-Hall.

[5] Allen, J. *et al.* 1987. *From Text to Speech*, Cambridge University Press.

[6] McGilloway, S. *et al.*. 2000. "Approaching Automatic Recognition of Emotion from Voice: a Rough Benchmark." *ISCA Workshop on Speech and Emotion.*

[7] Quatieri, T. F., Hanna, T. E. & O'Leary, G. C. 1997. "AM-FM Separation using Auditory Motivated Filters." *IEEE Tr. on Speech Audio Processing*, Vol. 5, No. 4, 465-480.

[8] De Silva, L. C. 2004. "Audiovisual Emotion Recognition." pp. 649-654, *IEEE International Conference on Systems, Man and Cybernetics.*

[9] Polzin, T. S. & Waibel, A. 2000. "Emotion-sensitive Human-computer Interfaces." *ISCA Workshop on Speech and Emotion.*

[10] 이희영, 송민. "가변대역폭 필터를 이용한 음성신호의 AM-FM 성분 분리에 관한 연구." *음성과학*, 제 8권 4호, pp. 45-58.

[11] Boashash, B. 1992. "Estimating and Interpreting the Instantaneous Frequency of a Signal-part I, II: Fundamentals." *Proc. IEEE*, Vol. 80, No. 4, 520-568.

[12] Bedrosian, E. 1962. "The Analytic Signal Representation of Modulated Waveforms." *Proc. IRE*, Vol. 50, No. 10, 2071-2076.

[13] Schroder, M. 2000. "Experimental Study of Affect Bursts." *ISCA Workshop on Speech and Emotion.*

[14] Moriyama, T. & Ozawa, S. 1999. "Emotion Recognition and Synthesis System on Speech." *IEEE International Conference on Multimedia Computing and Systems*, pp. 840-844.

[15] 김원구. "음성 신호를 사용한 감정인식의 특징 파라미터 비교." *전자공학회논문지* 제 40권 SP 제 5호, pp. 69-75.

[16] Iida, A. *et al.* 2000. "A Speech Synthesis System with Emotion for Assisting Communication." *ISCA Workshop on Speech and Emotion.*

[17] Lee, Hyoung. & Bien, Z. 1998. "Reconstruction of Signals with Known Instantaneous Frequency using Linear Time-varying Filter." *Electronics Letters*, Vol. 34, No. 24, 2312-2313.

[18] Lee Heyoung. & Bien, Z. 2004. "Bandpass Variable-bandwidth Filter for Reconstruction of Signals with Known Boundary in Time-frequency Domain.": *IEEE Signal Processing Letters*, Vol. 11, No. 2, 160-163.

[19] Yegnanarayana, B., d'Alessandro, C. & Darsinos, V. 1998. "An Iterative Algorithm for Decomposition of Speech Signals into Periodic and Aperiodic Components." *IEEE Tr. on*

Speech and Audio Processing, Vol. 6, No. 1, 1-11.

[20] d'Alessandro, C., Darsinos, V. & Yegnanarayana, B. 1998. "Effectiveness of a Periodic and Aperiodic Decomposition Method for Analysis of Voice Sources." *IEEE Tr. on Speech and Audio Processing*, Vol. 6, No. 1, 12-23.

[21] Picinbono, B. 1997. "On Instantaneous Amplitude and Phase of Signals." *IEEE Tr. on Signal Processing*, Vol. 45, No. 3, 552-560.

[22] Vakman, D. 1996. "On the Analytic Signal, the Teager-Kaiser Energy Algorithm, and Other Methods for Defining Amplitude and Frequency." *IEEE Tr. on Signal Processing*, Vol. 44, No. 4, 791-797.

[23] Cohen, L. 1989. "Time-frequency distributionsa review." *Proc. IEEE*, vol. 77, pp. 941-981.

접수일자: 2005. 11. 10
게재결정: 2005. 12. 10

▲ 이희영
서울시 노원구 공릉 2동 172번지 (우: 139-743)
서울 산업대학교 공과대학 제어계측공학과
Tel: 02-970-6545 Fax: 02-949-2654
E-mail: leehy@snut.ac.kr