2001년도 대한전기학회 하계학술대회 논문집 2001.7.18-20

정현파 모델과 사이코어쿠스틱스 모델을 이용한 음성 분리에 관한 연구

황선일, 한두진, 권철현, 신대규, 박상희
연세대학교 전기 전자 공학과

A Study on Speech Separation using Sinusoidal Model and Psycoacoustics Model

Sun-Il Hwang, Doo-Jin Han, Chul-Hyun Kwon, Dae-Kyu Shin, Sang-Hui Park
Dept.of Electrical and Electronic Eng. Yonsei University

Abstract - In this thesis, speaker separation is employed when speech from two talkers has been summed into one signal and it is desirable to recover one or both of the speech signals from the composite signal. This paper proposed the method that separated the summed speeches and proved the similarity between the signals by the cross correlation between the signals for exact between original signal and separated signal. This paper uses frequency sampling method based on sinusoidal model to separate the composite signal with vocalic speech and vocalic speech and noise masking method based on psycoacoustics model to separate the composite signal with vocalic speech and nonvocalic speech.

1. 서 론

음성 분리는 두 화자로부터 온 음성이 한 신호로 합해졌을 때, 그리고 그 복합신호로부터 특정한 화자 또는 두 화자 모두의 음성 신호를 복원하고자 할 때 적용된다. 음성 신호를 표현하는 방법은 여러 가지가 있으나, 본 논문에서는 효과적인 음성의 분리를 위해서 정현파(sinusoidal)모델을 사용하였다[2]. 정현파 모델이란 그 음성 신호를 사인파의 합으로 나타내는 모델이다. 즉 임의의 크기와 주파수 그리고 위상을 갖는 몇 개의 사인파로 음성 신호를 나타낼 수 있다는 것이다. 이 모델을 이용한 기존의 음성 분리 방법으로는 주파수 샘플링(frequency sampling) 방법[1]이 있다. 이 방법은 유성음과 유성음이 결합된 신호를 분리할 때에는 유성음이 사인파의 합으로 정확히 나타내어 질 수 있기 때문에 좋은 효과를 볼 수 있지만 유성음과 무성음이 결합된 신호를 분리할 때에는 무성음이 사인파의 합으로 정확히 모델링 되어 나타내어 질 수 없기 때문에 유성음과 유성음이 결합된 신호의 분리시보다 성능이 떨어지는 단점이 있다. 이것은 무성음의 특성에 기인한 것이며 무성음의 특성상 무성음은 사인파의 합보다는 단순한 백색 가우시안 잡음으로 나타내는 것이 더 원 신호와 근접함을 보인다. 본 논문에서는 무성음의 이러한 특성을 이용하여 기존 방법보다 좀 더 효과적으로 혼합된 두 음성 신호를 분리하는 방법을 제안하였다.

음성 신호 중 유성음은 사람에 따라서 고유한 피치를 가지고 있는데, 이것을 주파수 측면에서 기본 주파수(fundamental frequency)의 배수배 만큼 크기가 피크 치로 나타난다. 이 피크치만을 가지고 원래의 신호를 복원할 수 있는데, 이것을 피치가 다른 혼합된 음성 신호에 적용시켜보면 주파수 측에서 각 신호의 기본주파수를 배수로 해서 피크 치가 나타나는 정보를 이용해서 음성을 분리해 낸다. 그러기 위해서는 각 두 신호의 고유한 피치를 구해야 하는데, 본 논문에서는 maximum likelihood estimation 방법[2]으로 피치를 추출하였다. 음성 분리의 초기 연구는 스펙트럼의 주파수 정보를 미리 알고서, 즉 두 신호의 피치를 미리 알고서 분리해 내었지만, 본 논문에서는 피치 주기를 모르는 상태에서 음성 분리를 하고자 한다. 정현파 모델[3]을 이용한 주파수 샘플링 법은 무성음도 사인파의 합으로 어느 정도 모델링이 된다고 가정하고 임의로 설정한 피치를 가지고 정현파 모델을 적용시켜서 무성음을 분리하였다. 하지만 본 논문에서는 무성음을 백색 가우시안 잡음으로 가정한 뒤 사이코어쿠스틱스 모델[4][5]에 근거한 잡음 매스킹 법을 적용하여 유성음과 무성음이 혼합된 신호에서 유성음과 무성음을 분리해 내었다.

2. 본 론

2.1 정현파 음성 모델

음성 신호를 표현하는 방법 중 하나는 음성 신호를 성도의 공명 특색을 모델화 하는 시변형 선형 필터를 통해서 나오는 성음 여기(excitation)파형의 결과로 본다는 것이다. 음성 응용 측면에서 본다면, 성음 여기란 2가지 가능한 상태 중의 하나 일 수 있는데 그것은 유성음과 무성음에 대응된다.

성음 여기의 모델을 일반화하기 위해 본 논문에서는 다중 임펄스(multi impulse)를 사용해서 표현하기 보다 정현파 구성 요소들, 즉 크기, 주파수, 위상으로만 성음 여기 모델을 구성하였다[3]. 성도 필터의 시변형 임펄스 응답이 $h(\tau\,;\,t)$면 음성 신호 $s(t)$는 식 (2.1) 과 같이 표현되어진다.

$$s(t) = \int_0^t h(t-\tau\,;\,t)e(\tau)\,d\tau \qquad (2.1)$$

유성음과 무성음 여기 모델의 대안으로 일반적인 다중 임펄스 모델은 임의의 크기와 주파수와 위상을 가진 사인파의 합으로 표현되는 여기 신호이며, 그 모델은 식 (2.2)와 같다.

$$e(t) = \mathrm{Re}\sum_{l=0}^{L(t)} a_l(t)\exp\left\{j\left[\int_0^t \omega_l(\sigma)\,d\sigma + \Phi_l\right]\right\} \qquad (2.2)$$

여기서 l 번째 정현파 요소에 대해서 $a_l(t)$과 $w_l(t)$은 각각 크기와 주파수를 나타내고 ϕ_l은 사인파들의 위상이 모두 같지 않음을 나타내 주는 고정된 위상 offset을 보여주고 있다. 이 모델은 특히 음성 신호를 매우 간단하게 나타내 주고 있는데, 이것은 식 (2.3)과 같이 시변형 성도 전달 함수로 표현할 수 있다.

$$H(\omega\,;\,t) = M(\omega\,;\,t)\exp[j\Phi(\omega\,;\,t)] \qquad (2.3)$$

성음과 성도 크기, 위상의 영향을 혼합하여 더욱 간결하게 표현해 보면 식 (2.4)와 같다.

$$s(t) = \sum_{l=0}^{L(t)} A_l(t)\exp[j\psi_l(t)] \qquad (2.4)$$

이 때, $A_l(t) = a_l(t)\,M[\omega_l(t)\,;\,t]$ (2.5)

$$\psi_l(t) = \int_0^t \omega_l(\sigma)\,d\sigma + \Phi[\omega_l(t)\,;\,t] + \phi_l \qquad (2.6)$$

가 된다.

2.2 두명의 화자의 음성 표현

한 명의 화자의 정현파 음성 모델은 두 명의 화자의 경우에도 쉽게 일반화 할 수 있다. 두 명이 동시에 말하는 화자들에 의한 음성 파형은 시변의 크기, 주파수 그리고 위상을 가진 각각의 사인파들의 합으로 표현되어진다.
안정한 상태에서 성대와 성도 특성이 분석 윈도우 간격 동안 일정하다고 가정하면, 사인파 모델은 식 (2.7)과 같이 표현 될 수 있다.

$$x(n) = x_a(n) + x_b(n) \qquad (2.7)$$

이 때 $x_a(n) = \sum_{k=1}^{M_a} a_k \cos[\omega_{a,k} n + \phi_{a,k}]$

$x_b(n) = \sum_{k=1}^{M_b} b_k \cos[\omega_{b,k} n + \phi_{b,k}]$

이고, 시퀀스 $x_a(n)$와 $x_b(n)$는 각각 화자 A와 화자 B의 신호를 나타낸다. 화자 A와 관련된 크기와 주파수는 a_k와 $w_{a,k}$로 표시하고 위상 offset은 $\phi_{a,k}$이며, 화자 B에 대해서도 유사하게 적용된다. 식 (2.7)을 이용하여 두 명의 화자의 파형을 복원하기 위해 한 명의 화자의 경우에 디자인 된 분석·합성 시스템을 사용할 수 있다.

2.3 밀접하게 붙어있는 주파수들의 문제 해결

합해진 파형의 STFT를 조사해 보면 주파수 샘플링 방법이 가진 문제를 알 수 있다. $s_p(n)$이 두 시퀀스의 합의 시간 이동 p번째 윈도우 된 음성 부분을 나타낸다고 하면 식 (2.8)와 같이 나타낼 수 있다.

$$s_p(n) = w(n)[x_a(n + pL) + x_b(n + pL)] \qquad (2.8)$$

여기서 L은 세그먼트간의 시간 이동이고 분석 윈도우 $w(n)$은 유한기간 N동안 영이 아니다. 식 (2.8)를 가지고 $w \geq 0$에 대해 합해진 파형 $S_p(w)$가 분석 윈도우 $W(w)$의 scaled 이동된 것의 합으로 본다면 음성 신호는 식 (2.9)과 같이 표현된다.

$$S_p(w) = \sum_{k=1}^{M_a} a_k \exp(j\phi_{a,k}) W(w - w_{a,k}) + \sum_{k=1}^{M_b} b_k \exp(j\phi_{b,k}) W(w - w_{b,k}) \qquad (2.9)$$

그 대응하는 STFT는 분석 윈도우 $W(w)$의 두 이동된 형식 $X_a(w)$와 $X_b(w)$ 이고 합쳐진 STFT는 $S(w)$이다. 두 파형의 분리를 위해서 식 (2.10)의 matrix 방정식을 풀어야 한다.

$$\begin{bmatrix} 1 & W(\Delta\omega) \\ W(\Delta\omega) & 1 \end{bmatrix} \begin{bmatrix} X_{a(\omega_1)} \\ X_{b(\omega_2)} \end{bmatrix} = \begin{bmatrix} S(\omega_1) \\ S(\omega_2) \end{bmatrix} \qquad (2.10)$$

2.4 음성 향상을 위한 사이코어쿠스틱스 모델

다음의 분석은 음성과 잡음 신호들이 연속적인 시간이 아니고, 유한한 지속시간을 가졌다고 가정한다. 추가된 잡음의 경우, 잡음이 첨가된 음성 신호는 원래의 깨끗한 음성 신호와 잡음 성분의 합으로 구성된다. 즉,

$$y(n) = x(n) + d(n), \qquad (0 \leq n \leq N-1) \qquad (2.11)$$

(x(n)은 잡음이 없는 음성 신호, d(n)은 잡음 성분)
수식(2.11)은 주파수 영역에서 동일한 표현을 가진다. 대부분의 실제적인 상황에서는 짧은 시간(short-time) 스펙트라가 필요하기 때문에 각각 $Y_w(k,i)$와 $X_w(k,i)$로 주어진 윈도우(window)된 잡음이 첨가된 음성 신호와 깨끗한 음성 신호의 퓨리에 변환은 계산되어져야 한다. 즉,

$$Y_w(k,i) = \sum_{n=0}^{K-1} y(n + off_i) w(n) \; I_K^{kn} \qquad (2.12)$$

$$X_w(k,i) = \sum_{n=0}^{K-1} x(n + off_i) w(n) \; I_K^{kn} \qquad (2.13)$$

($I_K^{kn} = e^{-j(2\pi kn/K)}$, w(k)는 윈도우, K:퓨리에 변환 크기, i:시간 영역 윈도우 인덱스, off_i:오프셋)
그리고, 음성 신호는 오버랩된 시간의 윈도우를 사용해서 변환한다고 가정한다. 또, 각각에 대응하는 파워 스펙트라는 $Y_p(k,i)$와 $X_w(k,i)$로 주어진다. 즉,

$$Y_p(k,i) = |Y_w(k,i)|^2 \quad 0 \leq k \leq K-1 \qquad (2.14)$$

$$X_p(k,i) = |X_w(k,i)|^2 \quad 0 \leq k \leq K-1 \qquad (2.15)$$

사이코어쿠스틱 신호의 향상 기술의 기본 원리는 가청 잡음에 영향을 미치는 스펙트럴 성분을 압축하는 것이다. 이러한 성분들은 T(k,i)로 정의된 깨끗한 신호의 가청 매스킹 한계(AMT)의 추정으로부터 얻어질 수 있다. AMT는 매스커 신호의 존재하에서 그 값 아래에서 매스크 되어지는 모든 주파수 성분의 스펙트럴 크기 한계를 결정한다. 결국, 이 한계값 아래에 존재하는 잡음이 첨가된 음성 신호는 음성 신호의 영향으로 들을 수 없을 것이다.

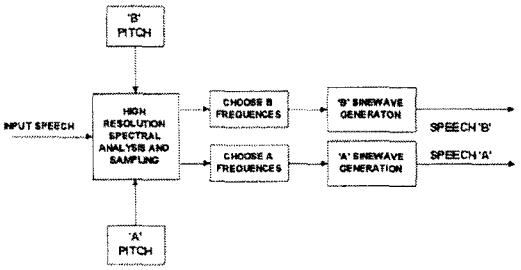


그림 2.1 주파수 샘플링 방식의 전체 블록 선도

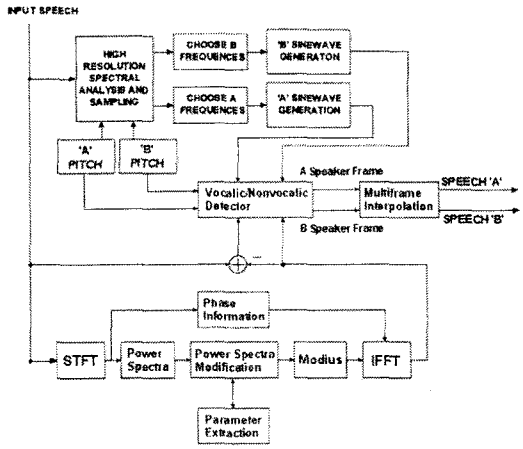


그림 2.2 제안한 방식의 전체 블록 선도

3. 실험 및 결과 고찰

3.1 실험

본 논문에 사용할 데이터의 수집을 위해서 음성데이터를 마이크로 통해 받는다. 마이크를 통과한 음성 신호는 저주파 필터를 통과하여 Data Translation사의 16채널, 12bit 해상도를 갖는 DT-2821을 사용하여 A/D 변환하였다. 이때의 샘플링 주파수는 10,000Hz로 하였다.

3.2 결과 고찰

화자 A의 "내가 사는 곳은" 문장과 화자 B의 "안타깝게도" 문장을 혼합하였다. 화자 A의 발음의 전체 음성 중 무성음 비율은 약 17% 이고 화자 B의 발음의 전체 음성 중 무성음 비율은 약 37% 이다.
표 4.1은 분리된 결과인데, 보는 바와 같이 기존의 방법에 비해서 제안한 방법이 더 효과적임을 볼 수 있다. B의 음성 신호의 무성음 비율이 A의 음성 신호의 그것보다 높기 때문에 더욱 효과적임을 볼 수 있다.

표 3.1 상관 계수 분석

		화자 A	화자 B
상관 계수	기존 방식	0.825312	0.651301
	제안 방식	0.842473	0.739462
전체 음성 중 무성음 비율 (%)		17.28	37.25

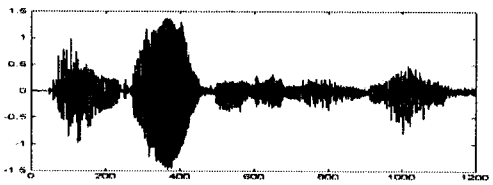
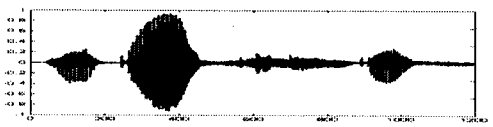
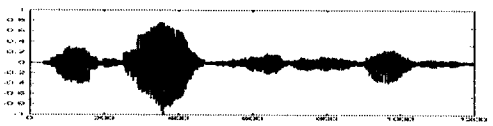


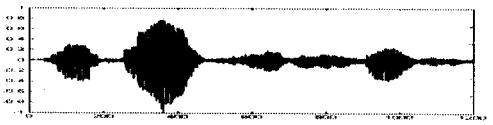
그림 3.1 화자 A와 화자 B의 혼합된 음성 12000개 샘플 파형



a) 화자 A의 발음 원 샘플 파형

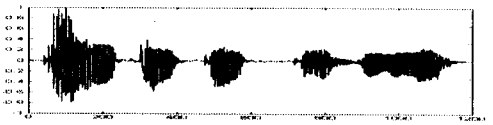


b) 기존 방식으로 분리한 화자 A의 음성 파형

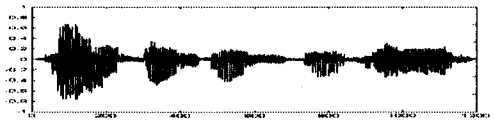


b) 제안한 방식으로 분리한 화자 A의 음성 파형

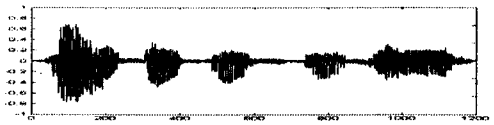
그림 3.2 화자 A의 분리된 음성 파형의 비교



a) 화자 B의 발음 원 샘플 파형



b) 기존 방식으로 분리한 화자 B의 음성 파형



b) 제안한 방식으로 분리한 화자 B의 음성 파형

그림 3.3 화자 B의 분리된 음성 파형의 비교

4. 결 론

본 논문에서는 임의의 두 화자의 음성이 결합된 신호에서 각각의 화자의 음성 신호를 분리해 내는 방법을 제시하였다. 혼합된 음성 신호에서 먼저 각각의 음성 신호의 피치를 구한 후에 그 정보를 이용하여 유성음과 유성음의 결합인지 유성음과 무성음의 결합인지를 판단하고 유성음과 유성음의 결합인 경우에는 피치로부터 기본 주파수를 구하여 기본 주파수의 배수만큼의 주파수만을 선택해서 각 신호를 복원해 내며 유성음과 무성음의 결합인 경우에는 사이코어쿠스틱스 모델에 근거한 잡음 매스킹을 이용하여 각 신호를 복원한다. 그리고 원 신호와 분리된 신호와의 정확성을 위해서 두 신호 사이의 상관 계수를 구하여 정량적으로 그 신호의 유사성을 확인해 보았다.
본 논문에서는 유성음과 무성음이 결합된 신호의 분리시에 기존의 주파수 샘플링 방식이 아닌 잡음 매스킹 기술을 사용하였고, 그 결과 기존 방식에 비해 더 좋은 효과를 얻을 수 있었다. 음성 신호 내에 무성음의 존재 비율에 따라 분리된 신호의 상관 계수도 달라짐을 확인할 수 있었다. 무성음의 존재 비율이 커짐에 따라 상관 계수도 떨어졌지만 반비례 관계는 아니었다. 기존 방법이나 제안한 방법 모두 무성음의 존재 비율에 영향을 받았지만 제안한 방법이 기존 방법에 비해 영향을 덜 받음을 확인할 수 있었다.

〔참 고 문 헌〕

[1] J. Naylor and S. F. Boll, "Techniques for supression of an interfering talker in co-channel speech, " in Proc. Int. Conf. Acoust., Speech, Signal Processing, vol. 1, Dallas. TX, Apr. 1987, pp. 205-208
[2] J. D. Wise, J. A. Capro, and T. W. Parks, "Maximum Likelihood Pitch Estimation," IEEE Trans. Acoustics, Speech, and Sig. Proc., vol. ASSP-24, no. 5, pp. 418-423, Oct. 1976
[3] R. J. McAulay and T. F. Quatieri, "Speech analysis - synthesis based on a sinusoidal representation, "Lincoln Lab., M.I.T., Lexington, MA, Tech. Rep. 693, May 1985; also IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-34, pp. 744-754, Aug. 1986.
[4] Dionysis E, Tsoukalas, John N. Mourjopoulos, and George Kokkinakis, "Speech enhancement based on audible noise suppression," IEEE Trans. Speech, Audio Processing, vol. 5, No. 6, Nov. 1997.
[5] E. Zwicker and H. Fastl, Psychoacoustics, Facts and Models. New York: Springer-Verlag, 1990.