Sinusoidal Model을 이용한 Cochannel상에서의 음성분리에 관한 연구

박현규°, 신중인, 박상희
연세대 전기공학과

A Study on Speech Separation in Cochannel using Sinusoidal Model

Hyun-Gyu Park°, Joong-In Shin, Sang-Hee Park
Dept. of Electrical Eng., Yonsei Univ.

Abstract - Cochannel 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. Cochannel speech occurs in many common situations such as when two AM signals containing speech are transmitted on the same frequency or when two people are speaking simultaneously (e. g., when talking on the telephone).

In this paper, the method that separated the speech in such a situation is proposed.

Especially, only the voiced sound of few sound states is separated. And the similarity of the signals by the cross correlation between the signals for exactness of original signal and separated signal is proved.

1. 서 론

Cochannel상에서의 음성분리는 두 화자로부터 온 음성이 한 신호로 합해졌을 때, 그리고 그 복합신호로부터 한명 혹은 두명의 음성신호를 복원하고자 할 때 적용된다. Cochannel상에서 음성이 혼합될 수 있는 것은 많은 상황에서 발생할 수 있는데, 그 예로 음성신호를 포함하는 두 개의 AM 신호들이 같은 주파수대로 전송될 때이거나, 두명의 화자가 동시에 말을 하고 있을 때, 그리고 전화로 이야기 하고 있을 때 등등이다. 본 논문에서는 이러한 상황이 발생했다고 가정을 하고 임의로 2개의 음성을 뽑아서 각각을 분리해 내는 방법을 제시하였다. 특히 이 논문에서는 유성음만을 분리하고자 하였다.

음성신호를 표현하는 방식은 여러 가지가 있으나 여기서는 음성의 분리를 좀 더 쉽게 하기 위해서 Sinusoidal 모델을 사용하였다. Sinusoidal 모델이란 음성신호를 사인파의 합으로 나타내는 모델이다. 즉 임의의 크기와 주파수 그리고 위상을 가지는 몇 개의 사인파로 음성신호를 나타낼 수 있다는 것이다. 이 모델을 이용한 기존의 음성 분리 방법은 frequency sampling 방법이 있는데, 이 방법은 두 음성신호의 에너지 차가 커지면 커질수록 상대적으로 효과가 떨어지는 단점이 있다. 분리시에 서로 겹치는 주파수를 그대로 두기 때문이다. 이 논문에서 제안한 방식은 위의 단점을 보완하고 효과적으로 혼합된 두 음성신호를 분리하는 방식이다.

일단 1차적으로 분리된 음성신호들은 각각의 기본 주파수들의 최소공배수 만큼의 주파수 대에서는 정보를 상실하게 되기 때문에 분리된 신호를 살펴보면 처음의 인위적으로 혼합한 신호와 비교해 볼 때 상당한 차이가 있음을 알 수 있다. 이러한 단점을 보완하기 위해 1차적으로 분리된 각각의 음성신호를 선형예측법을 이용해서 LPC 스펙트럼를 구하는 방법을 제안한다. 즉, 스펙트럼을 구해서 이것을 토대로 분리시 상실한 주파수 쪽을 보강함으로써 좀 더 정확한 결과를 얻을 수 있었다.

2. 음성분리를 위한 음성파라미터의 추출

2.1 피치의 추출

2.1.1 ML(Maximum Likelihood) 피치예측

심한 잡음이 섞여 있어도 상대적으로 가장 효과적인 Maximum likelihood Estimation 방법을 이용한 피치추출(pitch detection)을 사용하였으며 식(1)과 같다.

$$g(P) \cong \frac{2P}{K_0} \sum_{k=1}^{N=1} \Phi_K(lP) \qquad (1)$$

2.2 LPC 스펙트럼

LPC 스펙트럼은 선형예측법(linear prediction method)에 의하여 선형 예측 계수를 구하고, 이 값들을 식(2)와 같은 음성 신호에 대한 자기회귀(auto-regressive; AR) 모델에 대입함으로써 쉽게 얻어진다.

$$P(f) = \frac{\varepsilon_{min}}{\left|1 + \sum_{i=1}^{p} a_i \exp(-j2\pi f_i T)\right|^2} \qquad (2)$$

$P(f)$; AR스펙트럼 T ; 음성신호 샘플링 주기

3. Sinusoidal 모델을 이용한 음성분리

3.1 혼합된 음성신호의 피치 추출

두 음성이 혼합(composite)되어 있을 때에는 먼저 크기가 큰 쪽의 피치를 먼저 구한 후에 기존의 주파수샘플링(frequency sampling)방법으로 큰 쪽의 신호를 구한 후에 나머지 신호를 구해서 나머지 피치를 구한다. 그림 3.1에 블록선도를 나타내었다.

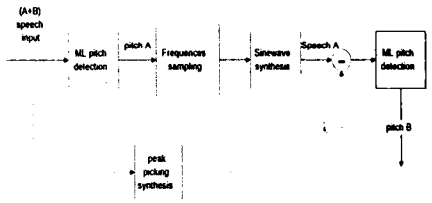


그림 3.1 피치 추출을 위한 블럭선도

3.2 Sinusoidal 음성 모델

3.2.1 Sinusoidal 모델의 두명의 화자의 음성 표현

한명의 화자의 음성표현의 Sinusoidal 음성모델은 두명의 화자의 경우에도 쉽게 일반화 할 수 있다. 두 명의 동시에 말하는 화자들에 의한 음성파형은 시변의 크기, 주파수 그리고 위상을 가진 각각의 사인파들의 합으로 식(3)과 같이 표현되어진다.

$$x(n) = x_a(n) + x_b(n) \qquad (3)$$

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

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

시퀀스 $x_a(n)$와 $x_b(n)$는 각각 화자 A와 화자 B의 음성을 나타낸다.

3.3 Sinusoidal 모델을 이용한 음성분리

3.3.1 밀접하게 붙어있는 주파수들의 문제 해결

$$S(w) = \sum_{k=1}^{M_a} a_k \exp(j\phi_{a,k}) W(w - w_{a,k}) + \sum_{k=1}^{M_b} b_k \exp(j\phi_{b,k}) W(w - w_{b,k}) \qquad (4)$$

그림 3.2는 파형 A와B를 가진 음성신호가 각각 주파수 w_1과 w_2를 가진 하나의 사인파로 구성된 것이다. 그 대응하는 STFT는 분석윈도우 $W(w)$의 두 이동된 형식 $X_a(w)$와 $X_b(w)$ 이고 합쳐진 STFT는 식(4) $S(w)$이다. 그림 3.2은 어떻게 $X_a(w)$와 $X_b(w)$의 메인로브(mainlobes)가 두 밀접한 주파수 w_1과 w_2에서 중첩(overlap)되었는가를 잘 보여주고 있다. 두 파형의 분리를 위해서 식(5)와 같은 매트릭스(matrix) 방정식을 풀어야 한다. 실제로는 이것을 확장시킨다.

$$\begin{bmatrix} 1 & W(\Delta\omega) \\ W(\Delta\omega) & 1 \end{bmatrix} \begin{bmatrix} X_a(\omega_1) \\ X_b(\omega_2) \end{bmatrix} = \begin{bmatrix} S(\omega_1) \\ S(\omega_2) \end{bmatrix} \qquad (5)$$

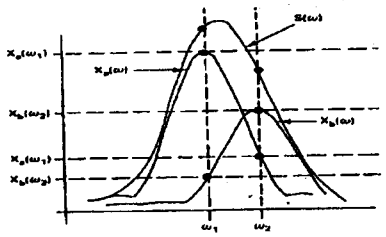


그림 3.2 두 사인파에 대한 Least-Square 해결

3.3.2 LPC 스펙트럼으로 상실된 주파수 보강

혼합된 음성신호를 1차적으로 분리한 후에 두 신호 사이의 기본주파수의 배수들의 최소공배수 주파수들은 상실되게 되어서 1차 분해 후의 파형을 조사해 보면 원신호의 파형과 많은 차이가 있게 된다. 그러므로 1차 분해 후의 신호를 가지고 LPC예측법으로 스펙트럼을 그려보면 신호의 전반적인 스펙트럼을 한눈에 볼 수 있게 된다. (그림 3.3)

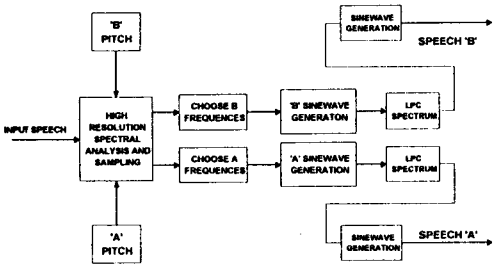


그림 3.3 제안한 방식의 전체 블록 선도

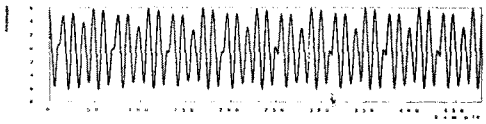
4. 실험 및 결과고찰

본 연구에서는 정상인 6명의 화자가 한국어 모음 5가지를 발음한 데이터를 이용하였다. 합성된 음성신호를 얻기 위해서 각각 따로 음성신호를 얻은 후에 선형적으로 합한 방식을 취하였다.

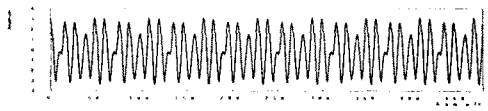
다음 결과는 피치주기가 samples 60인 화자 JSW의 '이' 발음과 피치주기가 samples 80인 화자 HKY의 '에' 발음을 혼합하여 크기가 작은 신호를 분리한 것이다. 각 음성의 에너지 차이는 약 23.4dB이다.

표 4.1 상관계수 분석

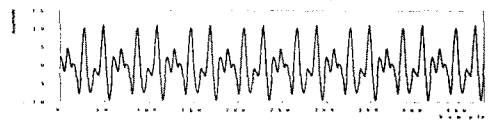
		기존 방식		제안 방식	
		원 신호	재합성 신호	원 신호	재합성 신호
상관계수	작은 신호	0.651276	0.678121	0.928320	0.944186
	큰 신호	0.878255	0.954754	0.871317	0.951601



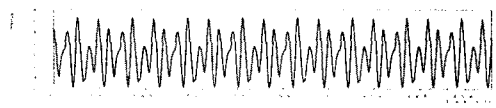
a) 화자 JSW의 '이' 발음 원 음성 샘플 파형



b) 화자 JSW의 '이' 발음 Sinusoidal로 재합성된 음성

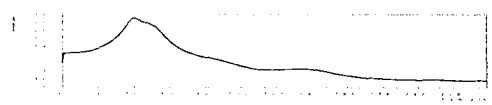


c) 기존 방식으로 분리한 화자 JSW의 '이' 의 음성

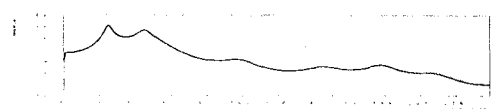


d) 제안한 방식으로 분리한 화자 JSW의 '이' 의 음성

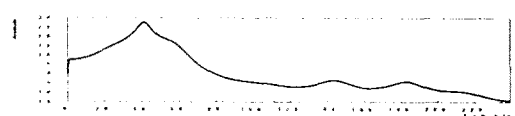
그림 4.1 분리된 음성 파형의 비교



a) 화자 JSW의 '이'의 원 LPC 스펙트럼



b) 기존 방식으로 분리한 화자 JSW의 '이' 의 LPC 스펙트럼



c) 제안한 방식으로 분리한 JSW의 '이' 의 LPC 스펙트럼

그림 4.2 LPC 스펙트럼 비교

5. 결 론

Cochannel상에서 혼합된 두 음성신호의 분리방법은 기존의 주파수 샘플링 방식에 비해서 음성신호의 상관계수를 비교해 보았을 때 0.2 ~ 0.3 우수함을 확인할 수 있었다. 1차 분리된 음성신호를 가지고 선형예측법을 사용하여 LPC 스펙트럼를 구하였다. 거기서 1차 형성음 주파수쪽에 근접한 상실된 주파수쪽을 보강하여 기존 방법보다 효과적인 결과를 얻을 수 있었다. 음성신호의 에너지차가 클 때는 기존의 방법으로도 에너지가 큰쪽의 신호는 쉽게 분리가 되지만 상대적으로 에너지가 작은 신호의 분리 결과는 상당히 좋지 않았으나 제안한 방식은 에너지차의 크기에 상관없이 전반적으로 효과적임을 알 수 있었다. 이 논문에서는 순순한 유성음만의 합인 것만을 다루었지만, 음성 중 유성음과 무성음이 동시에 존재하는 연속음이나 피치주기가 급격하게 변하는 부분에서의 연구가 지속적으로 이루어져야 하겠다.

*본 연구는 1995년 보건복지부에서 시행한 G7의료공학기술개발사업 2차년도 연구의 부분결과 (HMP-95-G-2-31)임을 밝힙니다.

[참 고 문 헌]

[1] 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.

[2] T. F. Quatieri and R. J. McAulay, "Speech transformation based on a sinusoidal representation. "Tech. Rep, TR-717, Lincoln Lab., M.I.T., May 16, 1986, and in IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-34, pp. 1449-1464, Dec, 1986.

[3] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection, "J. Acoust. Soc. Amer., no. 60, pp. 911-918, 1976.

[4] 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

[5] 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