한국음향학회지 제29권 제2호 pp. 141~147 (2010)

위상 모델 기반의 소프트 마스크를 이용한 단일 채널 음성분리

Single-Channel Speech Separation Using Phase Model-Based Soft Mask

이 윤 경*, 권 오 욱*
(Yun-Kyung Lee*, Oh-Wook Kwon*)

*충북대학교 제어로봇공학과
(접수일자: 2009년 12월 21일; 수정일자: 2010년 1월 22일; 채택일자: 2010년 1월 28일)

본 논문은 혼합 음성 신호로부터 크기와 위상 정보를 모두 고려하여 목표 음성 신호를 추출하고 향상하는 음성 분리 알고리듬을 제안한다. 기존 연구에서는 혼합된 음성 신호의 로그 전력 스펙트럼 값이 시간-주파수 영역에서 서로 독립이라고 가정한 통계적 모델을 적용하기 때문에 음성 분리 결과 파형에 불연속을 야기한다. 본 논문에서는 이러한 불연속을 감소시키기 위하여 시간-주파수 영역에서의 스무딩 필터를 적용한다. 음성 분리 성능을 더욱 향상시키기 위하여 음성 신호의 크기와 함께 위상 정보를 고려하는 통계적 모델을 제안한다. 실험 결과, 제안된 알고리듬이 기존의 크기 정보만을 사용한 알고리듬에 비하여 1.5 dB의 화자대간섭비 (SIR)를 개선하는 것으로 나타난다.
핵심용어: 위상 모델, 소프트 마스크, 음성 분리, 음성 향상
투고분야: 음성처리 분야 (2.3)

In this paper, we propose a new speech separation algorithm to extract and enhance the target speech signals from mixed speech signals by utilizing both magnitude and phase information. Since the previous statistical modeling algorithms assume that the log power spectrum values of the mixed speech signals are independent in the temporal and frequency domain, discontinuities occur in the resultant separated speech signals. To reduce the discontinuities, we apply a smoothing filter in the time-frequency domain. To further improve speech separation performance, we propose a statistical model based on both magnitude and phase information of speech signals. Experimental results show that the proposed algorithm improve signal-to-interference ratio (SIR) by 1.5 dB compared with the previous magnitude-only algorithms.
Keywords: Phase Modeling, Soft Mask, Speech Separation, Speech Enhancement
ASK subject classification: Speech Signal Processing (2.3)

I. 서 론

최근 디지털 신호처리 기술이 발전함에 따라 관련된 시장이 커지고 있으며 음성통신 및 음성인식 시스템을 활용한 음성 다이얼링, 잠금장치 등의 다양한 음성 서비스들이 보편화 되고 있다. 음성 정보처리에 있어 음성 신호를 향상시키고 보다 효과적으로 시스템에 적용하기 위해 잡음 요인의 제거 또는 잡음의 영향을 경감시키는 기술이 필수적으로 요구된다. 잡음 제거 기술은 음질을 향상시키고 음성 인식률을 개선시킬 뿐 아니라 음성의 명료도를 증가시켜 청각적 피로도를 감소시키는 효과가 있다.

단일 채널 음성 신호 분리 기술은 1개의 마이크를 통해 입력된 신호를 이용하여 음성 분리를 수행하는 것으로 전산 청각 장면 분석 (computational auditory scene analysis: CASA), 소프트 마스크 (soft mask) 등이 있다. CASA는 입력된 혼합 음성 신호로부터 사람의 청각 특성을 이용하여 동일 음원으로부터 발생한 음향 요소들을 찾아내는 방법으로 음성 신호를 분리하는 기술이다 [2]-[4]. 이 방법은 동적인 잡음 환경에서 성능이 우수하

책임저자: 권 오 욱 (owkwon@cbnu.ac.kr)
361-763 충북 청주시 흥덕구 성봉로 410 충북대학교 전자공학전공
(전화: 043-261-3374; 팩스: 043-268-2386)

다고 알려지고 있지만, 음원 분리 마스크로 이진 마스크(binary mask)를 사용하며, 음성학적 지식과 휴리스틱이 요구되는 단점이 있다.

소프트 마스크는 통계적 모델링 기반의 음성 분리 기술이다 [1][7]. 소프트 마스크는 입력된 혼합신호가 원하는 신호 (target signal)일 확률을 계산한 후 계산된 확률 값을 다시 혼합신호에 곱함으로써 원하는 음성신호를 추정하는 것이다. 이 방법은 음성학적 지식이 필요하지 않고 음원 분리 마스크를 확률에 의한 소프트 마스크를 사용하지만, 통계적 모델링에 의한 분리이기 때문에 인접한 음성 신호임에도 불구하고 다른 신호로 분리되는 비연속적인 경우가 종종 있다.

기존의 연구들은 보통 위상 성분을 무시하고 크기 성분(magnitude)만을 사용하여 음성신호를 분리한다. 위상 성분은 음성 신호에 대한 정보를 가지며, 위상 정보를 이용한 모델 역시 사람의 음성 인지 (human speech perception)와 음성 인식에 유용하다 [5][6]. 따라서 본 논문에서는 기존 방법들의 단점을 보완하고 음성 분리의 성능을 높이기 위해 통계적 모델링을 기반으로 한 음성 분리 기술에 시간-주파수 영역에서 인접한 신호들과의 유사도를 참조하도록 스무딩 필터를 적용하여 음성의 분리 과정에서 발생하는 비연속적인 경우를 보완하며, 음성학적 지식을 요구하지 않도록 한다. 또, 위상 성분을 이용하여 추정된 음성 신호를 조합하여 음성 신호의 크기와 위상 정보를 모두 고려함으로써 음성 분리의 정확도를 높인다.

본 논문의 구성은 다음과 같다. 제 2장에서는 음성 신호의 크기 모델 (magnitude model)을 위한 통계적 모델링 기반의 음성 분리 기술에 대해 설명한다. 제 3장에서는 위상 정보를 이용한 위상 모델 (phase model)을 설명하고, 위상과 크기 정보의 조합 방법을 설명한다. 제 4장에서는 음성 신호의 위상과 크기 정보를 이용하여 혼합 음성 신호를 분리한 결과를 비교 및 분석한다. 마지막으로 제 5장에서는 본 논문에 대한 결론을 맺는다.

II. 크기 모델 (magnitude model)

본 논문에서 제안된 음성 분리 알고리듬은 크기 모델과 위상 모델을 이용하여 계산된 크기와 위상 정보를 조합하여 원하는 음성 신호를 추정하는 기술로 그 구성도는 그림 1과 같다. 크기 모델은 통계적 모델링을 기반으로 하는 음성 분리 시스템의 방법 중 하나인 소프트 마스크를 사용하여 계산한다. 소프트 마스크는 음원 분리 마스크로 이진 마스크가 아닌 확률에 의한 마스크를 적용하여 음원을 분리한다. 혼합 음성 신호 $z(t)$의 로그 스펙트럼 벡터 z가 원하는 음성 신호일 확률을 계산하여 혼합 음성 신호의 로그 스펙트럼에 가중치로 적용함으로써 원하는 신호의 로그 스펙트럼을 추출한다.

이러한 방법으로 추출된 결과는 두 화자의 신호 x, y가 독립적인 신호라는 가정으로부터 얻어진 것이기 때문에, 혼합 음성 신호의 로그 스펙트럼 z를 x 또는 y의 둘 중 하나로 분리한다. 따라서 바로 옆 시간-주파수 대역에서의 로그 스펙트럼 분리 결과가 서로 다른 음성 신호로 계산이 될 수 있어 비연속적인 경우가 종종 발생한다. 이를 보완하기 위해 스무딩 필터를 사용하여 음성 신호에 스무딩을 적용한 후 확률을 계산함으로써, 인접한 시간-주파수 대역간의 연속성을 높이고 음성의 특성을 고려한다 [1].

2.1. 혼합 신호 모델링

단일 마이크를 통해 얻어진 화자 S_x, S_y의 입력 음성신호를 각각 $x(t)$, $y(t)$라고 할 때, 혼합된 음성신호 $z(t)$는 두 입력 음성신호의 합으로 얻어지며 다음과 같이 정의된다.

$$z(t) = x(t) + y(t) \tag{1}$$

여기서, $x(t)$와 $y(t)$는 서로 독립적인 신호라고 가정한다. $x(t)$, $y(t)$의 로그 파워 스펙트럼을 $x(\omega)$, $y(\omega)$라고 하면, 혼합 음성신호의 로그 스펙트럼 $z(\omega)$는 $x(\omega)+y(\omega)$로 계산되며 다음과 같이 정의 된다 [1].

$$z(\omega) = \max(x(\omega), y(\omega)) + e,$$
$$e = \log(1 + e^{\min(x(\omega),y(\omega)) - \max(x(\omega),y(\omega))}) \tag{2}$$

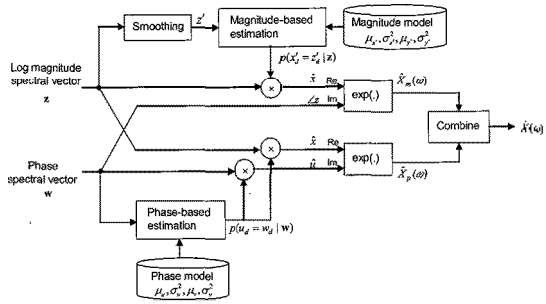

그림 1. 제안된 음성 분리 시스템의 구성도
Fig. 1. Block diagram of the proposed speech separation system

일반적으로 혼합 음성신호의 로그 스펙트럼은 두 입력 음성 신호의 로그 스펙트럼 중 더 큰 값을 가지는 로그 스펙트럼과 매우 유사한 값을 나타낸다. 따라서 로그-최대 근사법 (log-max approximation) [1]을 사용하여 식 2를 다음과 같이 근사화할 수 있다.

$$z(\omega) = \max(x(\omega), y(\omega)) \tag{3}$$

로그-최대 근사법에 따른 오차 e는 $x(\omega)$와 $y(\omega)$의 값이 같을 때 최대가 되며, 최대값은 $\log(2) = 0.69$이다. 로그 스펙트럼 벡터의 분포는 혼합 가우시안 밀도 (mixture Gaussian density)를 사용하여 계산하며, 각 혼합 가우시안에서 로그 스펙트럼 벡터들의 주파수 대역간은 서로 독립이라고 가정한다.

2.2. 스무딩이 적용된 크기 모델 기반의 소프트 마스크

식 3에서 정의한 로그-최대 근사 법에 따라 혼합 음성 신호의 로그 스펙트럼 벡터가 원하는 음성신호 x일 확률은 x의 로그 스펙트럼의 값이 y의 값 보다 클 확률과 같다. 즉, z의 d번째 차수의 로그 스펙트럼 값 z_d가 x_d일 확률은 x_d의 값이 y_d보다 클 확률로 계산되며 다음과 같이 정의된다.

$$p(x_d = z_d|z) = p(x_d > y_d|z) \tag{4}$$

시간-주파수 영역에서의 연속성을 높이기 위해 시간-주파수 영역에서 인접한 로그 스펙트럼 벡터 값을 참조하여 확률을 계산한다. 스무딩을 적용한 로그 스펙트럼 벡터를 x'_d, y'_d라 하면 스무딩이 적용된 소프트 마스크 필터는 다음과 같이 정의된다.

$$\begin{aligned} & p(x'_d = z_d|z) \\ &= \sum_{m_x,m_y} p(m_x,m_y|z) \frac{(x'_d = z_d, y'_d < z_d|m_x,m_y)}{p(z_d|m_x,m_y)} \\ &= \sum_{m_x,m_y} p(m_x,m_y|z) \frac{P_{x'}(z_d|m_x) C_{y'}(z_d|m_y)}{p(z_d|m_x,m_y)} \end{aligned} \tag{5}$$

여기서 m_x, m_y는 x, y의 혼합 가우시안 모델에서 가우시안의 개수 M_x, M_y에 해당하는 가중 합의 인덱스이다. 평균, 분산 등과 같은 파라미터 들은 학습 음성 데이터로부터 계산되며, 평균 μ_X와 분산 σ_X가 주어졌을 때의 가우시안 분포를 $N(x;\mu_X,\sigma_X^2)$라 하면, 확률 밀도 함수 $P_{x'}(z_d|m_x)$와 누적 확률 밀도 함수 $C_{y'}(z_d|m_y)$는 다음과 같다.

$$P_{x'}(z_d|m_x) = N(z_d;\mu_{x',m_x,d},\sigma^2_{x',m_x,d}) \tag{6}$$

$$C_{y'}(z_d|m_y) = \int_{-\infty}^{z_d} N(z_d;\mu_{y',m_y,d},\sigma^2_{y',m_y,d})dz_d \tag{7}$$

식 5을 사용하여 계산된 확률 값을 α_x라고 하면, 혼합 음성 신호로부터 화자 S_x로 추정된 로그 스펙트럼 벡터의 d번째 로그 스펙트럼 $\hat{x}_d$는 다음과 같이 정의된다.

$$\alpha_{x,d} = p(x'_d = z_d|z) \tag{8}$$

$$\hat{x}_d = \alpha_{x,d}\, z_d \tag{9}$$

분리된 음성 신호의 이산 푸리에 변환 (DFT)은 로그 스펙트럼 벡터 $\hat{x}_d$와 혼합 음성 신호의 위상 스펙트럼 벡터를 이용하여 계산하며, 혼합 음성신호 $z(t)$의 위상 성분이 $\angle z(w)$일 때 다음과 같이 계산된다.

$$\hat{X}_m(\omega) = \exp(\hat{x}_d + j\angle z(w)) \tag{10}$$

III. 위상 모델 (phase model)

기존의 음성 분리를 위한 연구들은 일반적으로 위상 성분을 고려하지 않고 크기 정보만을 이용한다. 하지만 위상 성분은 음성 신호의 정보를 가지며, 음성 인지와 음성 인식 시스템에서도 위상 정보가 유용하게 이용될 수 있다. 따라서 잡음을 효과적으로 제거하고 음성 분리 시스템의 성능을 향상시키기 위해서는 음성 신호의 위상 정보를 함께 고려해 주어야 한다.

본 논문에서는 혼합 음성신호 $z(t)$의 위상 스펙트럼 벡터 w가 원하는 음성 신호의 위상 정보일 확률을 계산하여 혼합 음성 신호에 곱해 줌으로써 원하는 음성 신호를 추정한다. 위상 정보를 이용하여 계산된 음성 신호와 크기 모델을 이용해 추정한 음성신호를 조합함으로써 음성 분리의 정확도를 향상시킨다.

3.1. 위상 신호 모델링

일반적으로 혼합 음성 신호의 크기 성분인 로그 스펙트럼이 두 입력 음성 신호의 크기 성분 중 더 큰 값과 유사한

것과 마찬가지로, 혼합 음성신호의 위상 성분 또한 크기 성분이 더 큰 신호의 위상과 매우 유사한 값을 나타낸다. 이는 그림 2의 음성신호 $x(t)$, $y(t)$와 혼합 음성신호 $z(t)$의 한 프레임의 일부에 대한 로그 스펙트럼과 위상 성분의 출력 예에서도 볼 수 있다. $x(t)$의 로그 스펙트럼이 $y(t)$의 로그 스펙트럼보다 큰 구간 'x'에서 위상 스펙트럼 역시 $x(t)$의 위상 스펙트럼과 유사하게 나타난다. 또, 구간 'y'에서 위상 스펙트럼은 $y(t)$의 위상 스펙트럼에 가깝게 출력된다.

두 화자의 음성 신호 $x(t)$, $y(t)$의 위상 스펙트럼 벡터를 각각 u, v라고 하고 혼합 음성 신호의 위상 스펙트럼 벡터를 w라고 할 때, 혼합 음성 신호의 위상은 둘 중 하나의 신호의 위상과 유사하다고 가정하며 다음과 같이 정의된다.

$$w(\omega)=\begin{cases}u(\omega)\ , & |w(\omega)-u(\omega)|<|w(\omega)-v(\omega)|\\ v(\omega)\ , & otherwise\end{cases} \tag{11}$$

3.2. 위상 모델 기반의 소프트 마스크

혼합 음성신호의 위상 스펙트럼 벡터 (phase spectrum)가 원하는 음성신호 x의 위상일 확률은 x의 위상 스펙트럼과의 유사도 값이 y의 위상 스펙트럼과의 유사도 값 보다 높을 확률과 같다. 즉, 위상 스펙트럼 값들 사이의 차이 정보를 이용하여 확률을 계산하며 다음과 같이 정의된다.

$$p(u_d=w_d|w)=p(|u_d-w_d|<|v_d-w_d|\,|w) \tag{12}$$

위 식을 정리하면 다음과 같다.

$$\begin{aligned}&p(u_d=w_d|w)\\&=\sum_{m_u,m_v}p(m_u,m_v|w)p(|u_d-w_d|<|v_d-w_d|\,|m_u,m_v)\\&=\begin{cases}\sum_{m_u,m_v}p(m_u,m_v|w)p(u_d=w_d,v_d<w_d|m_u,m_v)\ , & w_d>v_d\\ \sum_{m_u,m_v}p(m_u,m_v|w)p(u_d=w_d,w_d<v_d|m_u,m_v)\ , & v_d>w_d\end{cases}\end{aligned} \tag{13}$$

식 13으로부터, w_d가 원하는 신호의 위상 스펙트럼 u_d일 확률은 다음과 같다.

$$\begin{aligned}p(u_d=w_d|w)=&\sum_{m_u,m_v}p(m_u,m_v|w)P_u(w_d|m_u)\\&\times\max(C_v(w_d|m_v),1-C_v(w_d|m_v))\end{aligned} \tag{14}$$

위상 모델을 이용하여 구한 혼합 음성 신호의 위상 스펙트럼이 화자 S_x의 위상 스펙트럼일 확률을 β_x라고 하면, 혼합 음성 신호의 위상 정보로부터 추정된 로그 스펙트럼 벡터와 위상 스펙트럼 벡터의 d번째 값은 다음과 같이 정의된다.

$$\beta_{x,d}=p(u_d=w_d|w) \tag{15}$$

$$\hat{x}_d=\beta_{x,d}\,z_d \tag{16}$$

$$\hat{u}_d=\beta_{x,d}\,w_d \tag{17}$$

로그 스펙트럼 벡터와 위상 스펙트럼 벡터를 이용하여 추정된 음성 신호의 이산 푸리에 변환 $\hat{X}_p(\omega)$를 다음과 같이 계산한다.

$$\hat{X}_p(\omega)=\exp(\hat{x}_d+j\hat{u}_d) \tag{18}$$

3.3. 크기와 위상 정보를 고려한 조합 방법

혼합 음성신호로부터 원하는 음성 신호를 분리하는 과정에서 음성 신호의 크기 정보와 위상 정보를 모두 고려하기 위해 크기 모델과 위상 모델을 사용하여 계산된 음성 정보를 가중치 합하여 조합함으로써 원하는 음성 신호를 추정한다.

크기 모델로부터 추출된 음성신호의 이산 푸리에 변환을 $\hat{X}_m(\omega)$, 위상 모델로부터 얻어진 음성신호의 이산 푸리에 변환을 $\hat{X}_p(\omega)$라고 하면, 음성신호의 크기 정보와 위상 정보를 고려하여 분리된 음성신호는 다음과 같이 정의된다.

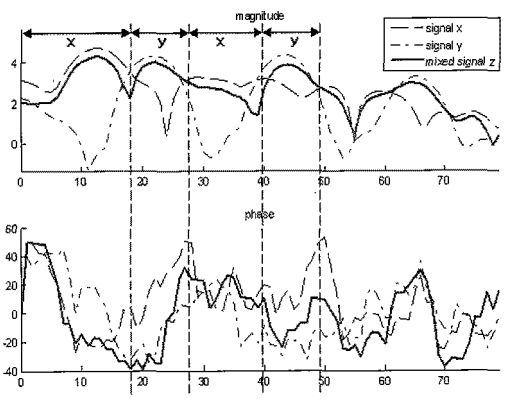

그림 2. 한 프레임 음성신호 스펙트럼의 일부
Fig. 2. Part of spectrum for a frame of speech signals

$$\hat{s}_x(\omega) = \frac{\alpha_x}{\alpha_x + \beta_x}\hat{X}_m(\omega) + \frac{\beta_x}{\alpha_x + \beta_x}\hat{X}_p(\omega) \quad (19)$$

$\hat{s}_x(\omega)$를 역변환한 후 오버랩-애드 (overlap-add) 방법을 사용하여 음성 신호를 복원한다.

IV. 실험 결과

음성 분리의 결과를 확인하고 성능의 정도를 측정하기 위하여 공개된 음성 데이터베이스에 대해 음성 분리 실험을 수행하였다.

4.1. 음성 데이터베이스

음성 분리 실험을 위해 사용된 음성 데이터는 Interspeech 2006 음성분리대회 (speech separation challenge) [8]에서 제공하는 데이터베이스에서 선택하였다. 혼합 음성신호는 두 화자의 음성 신호를 합산하여 사용하였으며, 학습을 위한 음성 데이터는 여성 화자 3명, 남성 화자 3명의 각 10문장을 사용하였다.

학습과 입력 음성데이터로 사용한 음성 신호는 샘플링 주파수를 25 kHz에서 16 kHz로 축소하여 사용하였으며, 평균 0, 분산 1을 갖도록 정규화 하였다. 정규화된 음성 데이터는 인접한 프레임들과 16 ms가 겹치도록 하여 32 ms크기의 프레임으로 나누어 해밍 윈도우를 적용하였다. 이산 푸리에 변환 (discrete Fourier transform: DFT)의 크기는 512-포인트로 하였으며, 푸리에 스펙트럼 결과로부터 257 차원의 스펙트럼 벡터를 분리하여 로그 스펙트럼 벡터 (크기 스펙트럼)와 위상 스펙트럼 벡터로 사용하였다. 혼합 가우시안의 밀도는 학습 데이터로부터 계산된다.

4.2. 음성분리 실험 결과

음성 분리 실험을 위한 혼합 음성 신호는 학습에 사용되지 않은 두 화자의 문장을 정규화 한 후 합산하여 사용하였다. 실험은 남성 화자+남성 화자, 남성 화자+여성 화자, 여성 화자+여성 화자의 세 가지 경우에 대하여 수행하였다. 음성 데이터는 혼합된 음성 신호가 -10, -5, 0, 5, 그리고 10 dB의 화자대간섭비 (speaker-to-interference ratio: SIR)를 갖도록 혼합하였다. 음성의 분리된 정도를 확인하기 위하여 두 화자의 음성신호와 혼합 음성신호, 분리된 후 음성신호의 파형과 스펙트로그램을 출력하였으며, 음성 분리의 결과를 수치적으로 보기 위하여 SIR를 계산하여 비교하였다.

4.2.1. 파형 및 스펙트로그램

그림 3과 그림 4에 화자 S_x, S_y의 음성신호 $x(t)$, $y(t)$와 혼합된 음성신호 $z(t)$, 그리고 음성 분리를 수행한 결과 파형과 스펙트로그램의 출력 예를 나타내었다. 분리가 잘 된 경우에도 스펙트로그램 결과에서 잡음 성분이 완전히 없어지는 것은 아니지만 눈에 띄게 옅어지며, 분리된 음성을 직접 들어봤을 때에도 잡음은 거의 들리지 않았다.

4.2.2. SIR

음성 신호 향상의 정도를 확인하기 위해 화자대간섭비 (SIR)를 계산하여 비교한다. 화자대간섭비는 화자 S_x와 분리된 음성신호 $\hat{s}_x$의 스펙트럼의 크기를 이용하여 계산하며 다음과 같다.

$$SIR = 10\log_{10}\left[\frac{|X|^2}{(|X| - |\hat{X}|)^2}\right] \quad (20)$$

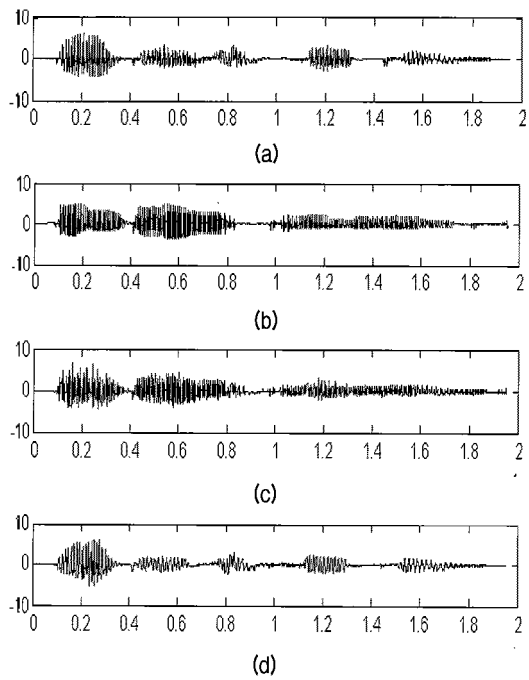

그림 3. 음성 신호의 파형 출력
(a) 화자 S_x, (b) 화자 S_y, (c) 혼합 음성신호
(d) 제안한 방법을 사용하여 분리된 음성신호

Fig. 3. Waveforms for
(a) speaker S_x, (b) speaker S_y, (c) mixed signal
(d) reconstructed signal with proposed method

여기서 $|X|$와 $|\hat{X}|$는 각각 화자 S_x와 분리된 음성신호 $\hat{s}_x$의 스펙트럼의 크기이다. 화자대간섭비는 화자의 음성신호와 간섭에 해당하는 잡음 신호의 비를 나타낸 것으로서, 화자대간섭비가 클수록 양호한 신호라고 할 수 있다.

표 1은 스무딩을 적용하지 않은 크기 모델과 스무딩을 적용한 크기 모델을 이용하여 음성분리를 수행한 후의 화자대간섭비를 나타낸 것이고, 표 2는 위상 모델을 이용하여 음성 분리를 수행한 후와 크기와 위상 정보를 모두 고려하여 음성을 분리한 후의 화자대간섭비를 나타낸 것이다. 화자대간섭비는 남성화자+남성화자, 남성화자+여성화자, 여성화자+여성화자의 각 경우에 대한 SIR의 평균을 계산한 것이다.

위상 정보만을 사용하여 음성을 분리한 경우 크기 정보만을 사용한 경우에 비하여 전체적으로 약 0.32 dB 낮게 측정되었으나 음성분리가 효과적으로 수행되었으며, 크기 모델과 결합되어 혼합 음성신호의 분리 성능을 높인다. 위상과 크기 정보를 모두 고려한 제안 알고리듬은 크기 정보만을 사용한 방법에 비해 약 1.5 dB의 SIR 증가를 보였다. 특히, 남성화자+여성화자와 같이 다른 성의 화자가 혼합된 신호의 경우 전체적으로 약 8.35 dB로 남성화자+남성화자의 혼합 신호 6.13 dB, 여성화자+여성화자의 혼합 신호 6.06 dB에 비해 분리가 잘 되었다. 또, 스무딩을 이용하지 않은 통계적 모델링 기반 방법(baseline)에 비해 모든 경우에서 그 분리 성능이 우수한 결과를 나타낸다 [7]. 소프트 마스크를 이용한 기존의 방법들이 이진 마스크에 비해 평균적으로 약 0.7~1 dB의 SIR 개선을 보이는 점에 비추어 보았을 때에도 효과적으로 음성이 분리되었다는 것을 볼 수 있다. 스무딩을 이용한 경우에 대한 성능 개선문제는 [1]에서 자세히 다루었다. 스무딩 필터는 스펙트럼 벡터가 아닌 확률 마스크에 적용할 수도 있으며, 이전의 스무딩 필터에 대한 음성 분리 실험을 수행하였을 때 마스크에 적용한 경우에 비해 스펙트럼 벡터에 스무딩을 적용한 경우 SIR의 증가가 높

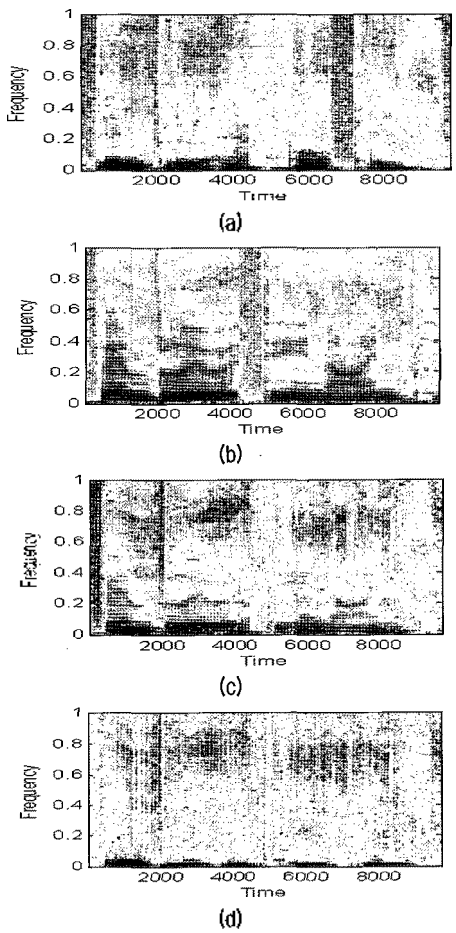

그림 4. 음성 신호의 스펙트로그램 출력
(a) 화자 S_x, (b) 화자 S_y, (c) 혼합 음성신호
(d) 제안한 방법을 사용하여 분리된 음성신호
Fig. 4. Spectrograms for
(a) speaker S_x, (b) speaker S_y, (c) mixed signal
(d) reconstructed signal with proposed method

표 1. 스무딩을 적용하지 않은 크기 모델과 스무딩이 적용된 크기 모델의 평균 SIR 계산 결과
Table 1. Average SIR (dB) for magnitude-only model without and with smoothing

SIR (dB)	Baseline		크기 모델	
	S_x	S_y	S_x	S_y
-10	1.49	7.24	1.80	8.66
-5	2.91	6.14	3.70	7.77
0	5.51	5.84	6.21	6.58
5	6.58	2.81	7.89	3.80
10	7.73	1.54	9.40	1.76
평균	4.84	4.71	5.80	5.71

표 2. 위상 모델과 크기와 위상 정보를 고려한 소프트마스크의 평균 SIR 계산 결과
Table 2. Average SIR (dB) of phase-only model not combined and combined with the magnitude-only model with smoothing

SIR (dB)	위상 모델		크기와 위상 정보를 고려한 소프트 마스크	
	S_x	S_y	S_x	S_y
-10	1.51	8.39	3.69	10.28
-5	3.43	7.55	5.20	8.56
0	5.87	6.23	7.81	7.14
5	7.55	3.39	9.25	4.02
10	8.86	1.55	10.61	2.00
평균	5.44	5.42	7.31	6.38

았기 때문에, 본 논문에서는 필터 폭이 균일한 9x9의 균일 마스크 필터 ($\Delta = 4$)를 스펙트럼 벡터에 스무딩 필터로 적용하였다.

VI. 결론

음성 처리기술의 사용이 증가하면서 잡음 요인을 제거하고 음성 신호를 향상시키기 위해 음성 분리기를 필요로 하고 있다. 본 논문에서는 음성의 크기와 위상 성분을 모두 고려한 통계적 모델링 기반의 단일 채널 음성 분리 시스템을 제안하였다. 제안한 방법에서는 통계적 모델에 의한 불연속의 경우를 보완하기 위해 스무딩 필터를 적용하여 음성의 연속적인 특징을 반영하였다. 또, 음성 분리의 단계에서 음성 신호의 크기 정보 뿐 아니라 위상 정보를 함께 고려하여 음성 분리의 성능을 높인다.

공개 음성 데이터베이스를 사용하여 음성 분리 실험을 수행한 결과, 제안된 방법을 사용하여 분리된 음성 신호의 파형과 스펙트로그램이 원하는 화자의 음성 신호와 가깝게 출력되었으며, 화자대간섭비를 계산한 결과에서도 기존의 크기 정보만을 이용한 경우에 비하여 전체적으로 1.5 dB의 SIR 증가를 보였다. 본 논문의 연구 결과로부터, 위상 정보의 유용성을 확인할 수 있었으며 음성 신호 처리의 분야에서 크기 정보와 함께 활용될 것으로 기대한다.

감사의 글

이 논문은 2009년도 충북대학교 학술연구지원사업의 연구비 지원에 의하여 연구되었음.

참 고 문 헌

1. 이윤경, 권오욱, "시간-주파수 스무딩이 적용된 소프트 마스크 필터를 이용한 단일 채널 음성 분리," *말소리*, 제67호, 195-216쪽, 2008.
2. Y.-K. Lee and O.-W. Kwon, "Application of shape analysis techniques for improved CASA-based speech separation," *IEEE Trans. Consumer Electronics*, vol. 55, no. 1, pp. 146-149, 2009.
3. G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, no. 4, pp. 297-326, 1994.
4. H. Runqiang, Z. Pei, G. Qin, Q. Zhiping, W. Hao, and W. Xihong, "CASA based speech separation for robust speech recognition," in *Proc. Interspeech*, pp. 2068-2071, 2006.
5. K. K. Paliwal, "Usefulness of phase in speech processing," in *Proc. IPSJ Spoken Language Processing Workshop, Gifu, Japan*, pp. 1-6, 2003.
6. F. Faubel, J. McDonough, and D. Klakow, "A phase-averaged model for the relationship between noisy speech, clean speech and noise in the log-Mel domain," in *Proc. Interspeech*, pp. 553-556, 2008.
7. A. M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1766-1776, 2007.
8. M. Cooke and T.-W. Lee, *Speech Separation and Recognition Competition*, http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm, 2006.

저자 약력

•이 윤 경 (Yun-Kyung Lee)

2007년 2월: 충북대학교 전자공학과 (공학사)
2009년 2월: 충북대학교 제어계측공학과 (공학석사)
2009년 2월~현재: 충북대학교 제어로봇공학과 (박사과정)
※ 주관심 분야: 음원분리, 음성인식, 반향제거, 음성 및 오디오 처리

•권 오 욱 (Oh-Wook Kwon)
한국음향학회지 제28권 제1호 참조