시간-주파수 스무딩이 적용된 소프트 마스크 필터를 이용한 단일 채널 음성 분리*

이윤경(충북대), 권오욱(충북대)

<차 례>

<Abstract>

Single-Channel Speech Separation Using the Time-Frequency Smoothed Soft Mask Filter

Yun-Kyung Lee, Oh-Wook Kwon

This paper addresses the problem of single-channel speech separation to extract the speech signal uttered by the speaker of interest from a mixture of speech signals. We propose to apply time-frequency smoothing to the existing statistical single-channel speech separation algorithms: The soft mask and the minimum-mean-square-error (MMSE) algorithms. In the proposed method, we use the two smoothing filter. One is the uniform mask filter whose filter length is uniform at the time-frequency domain, and the other is the mel-scale filter whose filter length is mel-scaled at the time domain. In our speech separation experiments, the uniform mask filter improves speaker-to-interference ratio (SIR) by 2.1 dB and 1 dB for the soft mask algorithm and the MMSE algorithm, respectively, whereas the mel-scale filter achieves 1.1 dB and 0.8 dB for the same algorithms.

* Keywords: Soft mask, Speech separation.

* 이 논문은 2008년도 충북대학교 학술연구지원사업의 연구비지원에 의하여 연구되었음.

1. 서 론

대부분의 음성 인식 시스템들은 주변 잡음이 없거나 무시할 수 있을 정도인 경우에는 높은 성능을 보이고 있지만, 잡음이 포함되는 경우에는 성능이 급격히 감소하게 된다. 그러나 음성 인식 시스템이 실제로 사용되는 환경은 여러 가지 잡음을 포함하고 있는 경우가 대부분이다. 따라서 동적인 잡음이 존재하는 환경에서의 음성 인식 성능의 향상을 위하여 독립적인 여러 개의 음원이 동시에 제시될 때 잡음 요인의 제거 또는 잡음의 영향을 경감시키는 기술이 필요하며, 이에 대한 연구가 넓은 범위에서 이루어져 왔다[1]-[3]. 잡음 신호를 제거하기 위한 접근 방법의 하나인 음성 신호 분리 기술은 2개의 마이크를 통해 입력된 신호를 이용하는 2채널 음성 분리 기술과 1개의 마이크를 통해서 얻는 입력신호를 이용하는 단일 채널 음성 분리 기술이 있다. 기존의 대다수 음성처리 시스템에서는 입력신호가 1개의 마이크를 통해서 얻어지므로, 본 논문에서는 단일 채널 입력에 대한 잡음 제거 기술을 연구한다.

단일 채널 음성신호 분리 기술에는 전산 청각 장면 분석(computational auditory scene analysis: CASA), 최소평균자승오류(minimum-mean-squared error: MMSE), 소프트 마스크(soft mask) 등이 있다. CASA는 사람의 청각 특성을 이용하여 음성신호를 분리하는 기술이다. 음성신호로부터 피치, 자기상관(autocorrelation), 온셋/오프셋(onset/offset) 특성을 계산하여 동일 음원으로부터 발생한 음향요소들을 찾아내는 방법이다[1]. 이 방법은 동적인 잡음 또는 복잡한 잡음 환경에서 성능이 우수하다고 알려지고 있지만, 음성학적 지식과 휴리스틱이 요구되는 단점이 있다. 최소평균자승오류와 소프트 마스크는 통계적 모델링 기반의 음성분리 기술이다. 최소평균자승오류[5]는 추출된 음성신호와 원하는 신호간의 최소자승오류(mean square error)를 최소화하도록 음성분리 시스템을 모델링하는 것이고, 소프트 마스크[4]는 입력된 혼합신호가 원하는 신호일 확률을 계산하여 그 확률 값을 혼합신호에 곱해 줌으로써 원하는 음성 신호를 추정한다. 통계적 모델링 기반의 음성분리 기술은 별도의 지식이 필요하지 않은 장점이 있으나, 단순한 통계적 결과에 의한 분리이기 때문에 인접한 음성신호임에도 불구하고 다른 신호로 계산되는 비연속적인 경우가 종종 있다. 또한 학습 음성 데이터가 많을수록 음성의 특성이 뭉뚱그려져 음성분리의 성능이 떨어지는 원인이 된다.

본 논문에서는 기존의 방법들의 단점을 보완하기 위해 통계적 모델링을 기반으로 한 음성 분리 기술에 스무딩 필터를 적용하여 음성의 분리 과정에서 발생하는 비연속적인 경우를 보완하였으며, 음성학적 지식을 요구하지 않도록 한다. 혼합신호로부터 원하는 음성신호를 추정할 때 신호 각각의 통계적 확률만을 계산하지 않고, 시간-주파수 영역에서 인접한 신호들과의 유사도를 계산하여 참조함으로써 음성신호의 연속성을 고려한다. 또, 이웃 신호와의 연속성을 고려함으로써 원

하는 신호의 확률과 잡음 요소 신호의 구분의 정확도를 높인다. 스무딩 필터는 시간-주파수 영역에서의 필터 폭이 균일한 3x3 마스크 필터와 시간 영역에서의 필터 폭을 멜-스케일(mel-scale)로 조정한 멜-스케일 필터의 두 필터를 사용한다. 스무딩 필터를 적용하여 혼합 음성 신호를 분리한 결과, 파형을 출력하였을 때 파형이 튀거나 비연속적인 부분 없이 원하는 음성 신호와 유사하게 출력되었다. 분리된 음성 신호의 화자대간섭비(signal-to-interference ratio: SIR)를 계산하여 음성 신호와 잡음 요소의 음성 신호의 비를 나타내었을 때에도 스무딩을 적용하지 않고 음성 분리를 수행한 후 계산한 SIR보다 높게 나타났다.

본 논문은 2장에서 소스 모델링에 대하여 간략히 소개하고, 3장에서 소프트 마스크 필터와 MMSE를 설명한다. 그리고 4장에서는 스무딩 필터를 적용한 소프트 마스크 필터와 MMSE에 대하여 설명하였다. 5장에서 각 필터의 음성분리 실험결과를 제시하고 6장에서 결론을 맺는다.

2. 소스 모델링

2.1. 혼합 모델

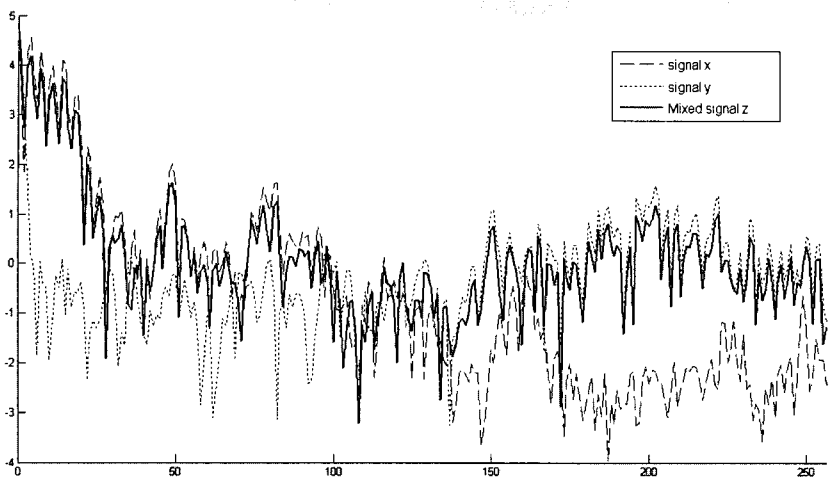
1개의 마이크를 통해 얻어진 화자 S_x, S_y의 입력 음성신호를 각각 $x(t)$, $y(t)$라고 할 때, 혼합된 음성신호 $z(t)$는 두 입력 음성신호의 합으로 얻어지며 다음과 같이 정의된다.

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

여기서 $x(t)$와 $y(t)$는 서로 독립적인 신호라고 가정한다. 일반적으로 음성 분리의 수행은 2개의 음원이 서로 독립적으로 발생되는 것을 대상으로 하므로 $x(t)$, $y(t)$의 로그 파워 스펙트럼을 $x(\omega)$, $y(\omega)$라고 하면, 혼합된 음성신호의 로그 스펙트럼 $z(\omega)$는 $x(\omega)+y(\omega)$로 계산된다. 혼합된 음성신호의 로그 스펙트럼은 다음과 같이 재정의된다.

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

일반적으로, 혼합된 음성신호의 로그 스펙트럼은 두 음성신호의 로그 스펙트럼 중 더 큰 값을 가지는 로그 스펙트럼과 매우 유사한 값을 나타낸다. 이는 <그림 1>의 음성신호 $x(t)$, $y(t)$와 혼합된 음성신호 $z(t)$의 한 프레임(32 ms)에 대한 로



<그림 1> 음성의 한 프레임에 대한 로그 스펙트럼

그 스펙트럼의 출력 예에서도 볼 수 있다. 두 음성신호의 로그 스펙트럼 중 더 큰 값을 가지는 로그 스펙트럼과 실제로 계산된 혼합 음성신호의 로그 스펙트럼의 오차 e는 $x(\omega)$와 $y(\omega)$의 값이 같을 때 최대가 되며, 최대값은 $\log(2)=0.69$이다. 로그-최대 근사법(log-max approximation)[4]을 사용하여 식 (2)를 다음과 같이 근사화할 수 있다.

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

본 논문에서는 음성분리를 수행하기 위하여 샘플링 주파수가 16 kHz인 음성데이터를 학습과 입력 음성데이터로 사용하였다. 음성 데이터들은 인접한 프레임들과 16 ms를 겹치도록 하여 32 ms 크기의 프레임으로 나누었다. 각 프레임에 32 ms의 해밍 윈도우를 적용하여 512-포인트의 이산 푸리에 변환(discrete Fourier transform: DFT)을 계산한다. 계산된 푸리에 스펙트럼 결과로부터 257 차원의 스펙트럼 벡터를 분리한 후 로그를 취하여 로그 스펙트럼 벡터를 계산한다.

2.2. 가우시안 소스 모델

로그 스펙트럼 벡터의 분포를 혼합 가우시안 밀도(mixture Gaussian density)를 사용하여 계산한다[4]. 각 혼합 가우시안에서, 로그 스펙트럼 벡터들의 주파수 대역간은 서로 독립이라고 가정한다. x, y가 화자 S_x, S_y의 로그 스펙트럼 벡터일 때, 화자 S_x의 로그 스펙트럼 x에 대한 분포는 다음과 같이 정의된다.

$$p(x) = \sum_{i=1}^{M_x} P_x(i) \prod_{d=1}^{D} N(x_d; \mu_{x,i,d}, \sigma^2_{x,i,d}) = \sum_{i=1}^{M_x} P_x(i) \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\sigma^2_{x,i,d}}} exp\left(- \frac{(x_d - \mu_{x,i,d})^2}{2\sigma^2_{x,i,d}} \right) \quad (4)$$

여기서 M_x는 혼합 가우시안에서의 가우시안의 개수이고, $P_x(i)$는 i번째 가우시안의 선험적 확률(*a priori* probability)이다. D는 로그 스펙트럼 벡터 x의 차원이고 x_d는 x의 d번째 차원이며, $\mu_{x,i,d}$과 $\sigma^2_{x,i,d}$는 각각 i번째 가우시안의 평균과 분산에서 d번째 차수에 해당하는 평균, 분산값을 말한다. $N(x_d; \mu_{x,i,d}, \sigma^2_{x,i,d})$는 평균 $\mu_{x,i,d}$와 분산 $\sigma^2_{x,i,d}$가 주어졌을 때, x_d에서의 가우시안 밀도이다. S_x, S_y의 평균, 분산 등과 같은 파라미터 들은 학습 음성 데이터로부터 계산되며, 화자 S_y의 로그 스펙트럼 y에 대한 분포는 다음과 같이 정의된다.

$$p(y) = \sum_{j=1}^{M_y} P_y(j) \prod_{d=1}^{D} N(y_d; \mu_{y,j,d}, \sigma^2_{y,j,d}) = \sum_{j=1}^{M_y} P_y(j) \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\sigma^2_{y,j,d}}} exp\left(- \frac{(y_d - \mu_{y,j,d})^2}{2\sigma^2_{y,j,d}} \right) \quad (5)$$

3. 통계적 모델링 기반 필터

소프트 마스크와 최소평균자승오류(MMSE)는 통계적 모델링을 기반으로 하는 단일 채널 음성 분리 시스템의 방법 중 하나이다. 소프트 마스크는 혼합 신호의 로그 스펙트럼에 가중치를 적용하여 원하는 신호의 로그 스펙트럼을 추출하는 것이고, 최소평균자승오류는 원하는 신호의 로그 스펙트럼 벡터를 추정할 때, 추정 결과의 오류가 최소가 되도록 추정하는 것이다. 본 논문에서는 소프트 마스크[4] 또는 최소평균자승오류[5]에 스무딩을 적용하여 음성분리 시스템을 모델링하였다.

3.1. 소프트 마스크

혼합 음성신호 $z(t)$의 로그 스펙트럼 벡터 z가 원하는 음성신호일 확률을 구한다. 혼합 음성신호의 로그 스펙트럼 벡터는 입력 음성신호의 로그 스펙트럼 벡터 x, y중 큰 값을 가지는 로그 스펙트럼과 유사한 값을 가지므로, 식 (3)에서 정

의한 로그-최대 근사법에 따라 혼합 음성신호의 로그스펙트럼 벡터가 원하는 음성신호 x일 확률은 x의 로그 스펙트럼 값이 y의 로그 스펙트럼 값 보다 클 확률과 같다. 즉, z의 d번째 차수의 로그 스펙트럼 값 z_d가 x_d 일 확률은 x_d의 값이 y_d보다 클 확률로 계산되며 다음과 같이 정의된다.

$$p(x_d = z_d|z) = p(x_d > y_d|z) \tag{6}$$

위 식을 정리하면 다음과 같다.

$$p(x_d > y_d|z) = \sum_{M_x, M_y} p(M_x, M_y|z)p(x_d > y_d|z_d, M_x, M_y) \tag{7}$$

로그 스펙트럼 z_d는 x_d와 y_d의 값 중에서 더 큰 값으로 계산되므로, 로그 스펙트럼 x_d의 값이 y_d의 값 보다 클 확률을 x_d의 값이 z_d와 같고 y_d가 z_d보다 작을 확률을 사용하여 계산한다. 베이즈의 정리를 이용하여 다음과 같이 정리된다.

$$p(x_d > y_d|z_d, M_x, M_y) = \frac{p(x_d = z_d, y_d < z_d|M_x, M_y)}{p(z_d|M_x, M_y)} \tag{8}$$

식 (6)~(8)을 조합하여, z_d가 원하는 신호 x_d일 확률을 구하는 계산식을 다음과 같이 정리하여 사용한다.

$$p(x_d = z_d|z) = \sum_{M_x, M_y} P(M_x, M_y|z) \frac{N(z_d; \mu^{x_d}_{M_x,d}, \sigma^{2x_d}_{M_x,d}) \times \int_{-\infty}^{z_d} N(z_d; \mu^{y_d}_{M_y,d}, \sigma^{2y_d}_{M_y,d})}{p(z_d|M_x, M_y)} \tag{9}$$

여기서, $p(x_d = z_d, y_d < z_d|M_x, M_y)$과 $p(z_d|M_x, M_y)$은 각각 다음과 같다.

$$\begin{aligned} p(x_d = z_d, y_d < z_d|M_x, M_y) &= \left(\frac{1}{\sqrt{2\pi\sigma^2_{x_d, M_x, d}}} exp\left(- \frac{(x_d - \mu_{x_d, M_x, d})^2}{2\sigma^2_{x_d, M_x, d}} \right) \right) \\ &\quad \times \left(\int_{-\infty}^{z_d} \frac{1}{\sqrt{2\pi\sigma^2_{y_d, M_y, d}}} exp\left(- \frac{(y_d - \mu_{y_d, M_y, d})^2}{2\sigma^2_{y_d, M_y, d}} \right) \right) \\ &= N(z_d; \mu_{x_d, M_x, d}, \sigma^2_{x_d, M_x, d}) \times \int_{-\infty}^{z_d} N(z_d; \mu_{y_d, M_y, d}, \sigma^2_{y_d, M_y, d}) \end{aligned} \tag{10}$$

$$\begin{aligned} p(z_d|M_x, M_y) &= N(z_d; \mu_{x_d, M_x, d}, \sigma^2_{x_d, M_x, d}) \times \int_{-\infty}^{z_d} N(z_d; \mu_{y_d, M_y, d}, \sigma^2_{y_d, M_y, d}) \\ &\quad + N(z_d; \mu_{y_d, M_y, d}, \sigma^2_{y_d, M_y, d}) \times \int_{-\infty}^{z_d} N(z_d; \mu_{x_d, M_x, d}, \sigma^2_{x_d, M_x, d}) \end{aligned} \tag{11}$$

식 (11)은 z_d의 확률밀도가 $x_d = z_d$이고 $y_d < z_d$일 확률과 $y_d = z_d$이고 $x_d < z_d$일 확률의 결합 확률의 합으로 계산된다는 것을 나타낸다. 이를 이용하여 z의 가우시안 조건부 확률을 다음과 같이 계산할 수 있다.

$$p(M_x, M_y|z) = \frac{p_x(M_x)p_y(M_y)p(z|M_x, M_y)}{p(z)} \tag{12}$$

여기서 $p(z|M_x, M_y)$과 $p(z)$는 각각 다음과 같다.

$$p(z|M_x, M_y) = \prod_{d=1}^{D} p(z_d|M_x, M_y) \tag{13}$$

$$\begin{aligned} p(z) &= \sum_{M_x, M_y} p(M_x, M_y)p(z|M_x, M_y) \\ &= \sum_{M_x, M_y} p(M_x)p(M_y)\prod_{d=1}^{D} p(z_d|M_x, M_y) \end{aligned} \tag{14}$$

소프트 마스크를 사용하여 구한 혼합 음성 신호가 화자 S_x일 확률, $p(x_d = z_d|z)$을 m_x라고 하면, 혼합 음성 신호가 화자 S_y일 확률은 $1 - m_x$가 된다. 혼합 음성 신호로부터 화자 S_x 신호로 추정된 로그 스펙트럼 벡터의 d번째 로그 스펙트럼 값, $\hat{x}_d$는 다음과 같이 정의된다.

$$\hat{x}_d = m_{x,d} \cdot z_d - H(z_d, m_{x,d}) \tag{15}$$

여기서 $m_{x,d}$는 m_x의 d번째 값이고, $H(z_d, m_{x,d})$는 정규화를 위한 것으로서 혼합 음성 신호의 파워 스펙트럼과 두 화자의 파워 스펙트럼의 합이 같도록 하기 위해 계산하며 다음 식과 같다.

$$H(z_d, m_{x,d}) = \log(\exp(-z_d m_{x,d}) + \exp(z_d(m_{x,d} - 1))) \tag{16}$$

3.2. 최소평균자승오류(MMSE)

최소평균자승오류는 음성 분리를 수행하여 추출된 음성 신호와 원하는 음성 신호의 평균제곱오차(MSE)를 최소화시키는 것으로, 추출된 음성 신호를 Y, 원하는 음성 신호(기준 신호)를 X라고 할 때, 다음 식으로 정의된다.

$$\hat{X} = \arg\min_Y E[\|Y - X\|^2] \tag{17}$$

평균제곱오차가 최소가 되는 경우는 혼합 음성 신호로부터 음성 분리를 수행하여 얻어진 추출 음성 신호가 원하는 음성신호 x_d일 때이므로, 혼합 음성 신호의 로그 스펙트럼 벡터 z가 주어졌을 때 원하는 음성신호의 로그 스펙트럼벡터 x_d일 확률의 조건부 기대값을 이용하여 원하는 음성신호에 가장 가까운 음성 신호를 추정한다. 조건부 기대값을 이용한 최소평균자승오류는 다음과 같이 정의된다.

$$\hat{x}_d = E[x_d|z] = \int_{x_d} x_d p(x_d|z) dx_d \tag{18}$$

여기서, $p(x_d|z)$는 다음과 같다.

$$p(x_d|z) = \sum_{M_x, M_y} p(x_d|z_d, M_x, M_y) p(M_x, M_y|z) \tag{19}$$

혼합 가우시안의 공분산 행렬들이 대칭행렬이므로 $p(x_d|z_d, M_x, M_y)$는 혼합 음성 신호의 로그 스펙트럼 z의 d번째 차수의 로그 스펙트럼 벡터 값, z_d에 의해서만 영향을 받으며 다음과 같이 정리된다[5].

$$p(x_d|z_d, M_x, M_y) = \begin{cases} \dfrac{N(x_d;\mu_{x,M_x,d},\sigma^2_{x,M_x,d}) \times N(z_d;\mu_{y,M_y,d},\sigma^2_{y,M_y,d})}{p(z_d|M_x,M_y)} \\ \quad + \dfrac{N(z_d;\mu_{x,M_x,d},\sigma^2_{x,M_x,d}) \times \int_{-\infty}^{z_d} N(y_d;\mu_{y,M_y,d},\sigma^2_{y,M_y,d}) dy_d \times z_d}{p(z_d|M_x,M_y)} & , \text{if } x_d \le z_d \\ 0 & , otherwise \end{cases} \tag{20}$$

$N(x_d;\mu_{x,M_x,d},\sigma^2_{x,M_x,d})$는 다음과 같다.

$$\begin{aligned} &\int_{-\infty}^{z_d} x_d N(x_d;\mu_{x,M_x,d},\sigma^2_{x,M_x,d}) dx_d \\ &\quad = \mu_{x,M_x,d} \times \int_{-\infty}^{z_d} N(x_d;\mu_{x,M_x,d},\sigma^2_{x,M_x,d}) dx_d - \sigma^2_{x,M_x,d} \times N(z_d;\mu_{x,M_x,d},\sigma^2_{x,M_x,d}) \end{aligned} \tag{21}$$

$\hat{x}_d$는 식 (18)~(21)을 조합하여 다음과 같이 나타낸다.

$$\begin{aligned} x_d = \sum_{M_x, M_y} \frac{p(M_x,M_y|z)}{p(z_d|M_x,M_y)} [&N(z_d;\mu_{y,M_y,d},\sigma^2_{y,M_y,d}) \times \int_{-\infty}^{z_d} x_d N(x_d;\mu_{x,M_x,d},\sigma^2_{x,M_x,d}) dx_d \\ &+ N(z_d;\mu_{x,M_x,d},\sigma^2_{x,M_x,d}) \times \int_{-\infty}^{z_d} N(y_d;\mu_{y,M_y,d},\sigma^2_{y,M_y,d}) dy_d] \end{aligned} \tag{22}$$

3.3. 분리된 음성 신호 재합성

소프트 마스크와 최소평균제곱오류의 방법을 사용하여 분리된 음성신호의 로그 스펙트럼 벡터를 재합성한다. 분리된 음성신호 스펙트럼의 위상 성분은 혼합 음성 신호의 위상 성분과 같은 값으로 계산한다. 혼합 음성신호 $Z(t)$의 푸리에 변환을 $Z(\omega)$라고 하면, 위상 성분은 $\angle Z(\omega)$로 계산된다. 분리된 음성 신호의 로그 스펙트럼 벡터를 $\hat{x}$라 하면, 분리된 음성신호의 이산푸리에변환은 다음과 같다.

$$\hat{X}(\omega) = \exp(\hat{x} + i \angle Z(\omega)) \tag{23}$$

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

4. 스무딩이 적용된 소프트 마스크 필터

통계적 모델링을 기반으로 하여 혼합 음성 신호가 원하는 신호일 확률을 계산한 결과는 혼합 음성 신호의 로그 스펙트럼 z_d를 x_d 또는 y_d 둘 중 하나로 분리한다. 따라서 바로 옆 시간-주파수 대역의 로그 스펙트럼의 분리 결과가 서로 다른 음성 신호로 계산이 될 수 있어 비연속적인 경우가 종종 있다. 또, 학습을 위한 음성 데이터의 수가 많아짐에 따라 음성 특성들이 뭉뚱그려져, 음성 분리가 되지 않는 경우도 있다. 특히, 혼합 음성이 남성 화자+남성 화자 또는 여성 화자+여성화자의 같은 성별로 혼합되었을 경우 음성 신호와 잡음 요소의 음성신호의 특성이 비슷하여 음성 분리에 어려움이 있다. 이를 보완하기 위하여 확률을 계산하기 전, 클러스터링을 하는 과정에서 음성 신호에 스무딩을 적용하여 클러스터링을 함으로써 인접한 시간-주파수 대역간의 연속성을 높여 음성 신호의 연속성을 고려하였다. z_d가 x_d일 확률과 y_d일 확률이 비슷한 경우에도 인접한 시간-주파수 대역의 분리 결과를 참조함으로써 원하는 음성신호와 잡음 요소의 음성신호 구분의 정확도를 높였다.

4.1. 스무딩 필터

시간-주파수 영역에서의 연속성을 높이기 위하여 필터 폭이 균일한 3x3의 균일 마스크 필터와 필터 폭이 멜-스케일로 조정된 멜 스케일 필터의 두 가지 필터를 스무딩 필터로 사용하였다.

4.1.1. 균일 마스크 필터

혼합 음성 신호의 로그 스펙트럼 벡터 z가 x일 확률을 계산할 때, 시간-주파수 영역에서 인접한 로그 스펙트럼 벡터 값을 참조하여 계산한다. 시간 영역에서 이전 채널과 다음 채널, 그리고 주파수 영역에서 이전 주파수 채널과 다음 주파수 채널의 로그 스펙트럼 벡터 값을 사용하여 스무딩을 적용한다. 스무딩은 참조된 로그 스펙트럼 벡터 값들의 평균으로 계산된다. 시간 영역에서 이전 프레임과 다음 프레임을 참조하여 스무딩을 적용한 로그 스펙트럼 벡터를 $\tilde{x}$라고 할 때, 로그 스펙트럼 벡터 $\tilde{x}$의 k번째 프레임에서 d 번째 차원의 값 $\tilde{x}_{k,d}$는 다음 식으로 정의된다.

$$\tilde{x}_{k,d} = \frac{1}{3}(x_{k-1,d} + x_{k,d} + x_{k+1,d}) \tag{24}$$

$\tilde{x}_{k,d}$를 사용하여 주파수 영역에서 이전 주파수 채널과 다음 주파수 채널의 로그 스펙트럼 벡터 값을 참조하여 스무딩을 적용한다. 즉, 시간-주파수 영역에서 스무딩을 적용한 로그 스펙트럼 벡터 $x'_{k,d}$는 다음과 같이 정의된다.

$$x'_{k,d} = \frac{1}{3}(\tilde{x}_{k,d-1} + \tilde{x}_{k,d} + \tilde{x}_{k,d+1}) \tag{25}$$

로그 스펙트럼 벡터 y에 대하여 시간 영역에서 스무딩을 적용한 로그 스펙트럼 벡터 $\tilde{y}$의 k번째 프레임에서 d번째 차원의 로그 스펙트럼 값 $\tilde{y}_{k,d}$와 시간-주파수 영역에서 스무딩을 적용한 로그 스펙트럼 벡터 $y'_{k,d}$는 다음과 같다.

$$\tilde{y}_{k,d} = \frac{1}{3}(y_{k-1,d} + y_{k,d} + y_{k+1,d}) \tag{26}$$

$$y'_{k,d} = \frac{1}{3}(\tilde{y}_{k,d-1} + \tilde{y}_{k,d} + \tilde{y}_{k,d+1}) \tag{27}$$

스무딩을 적용한 로그 스펙트럼 벡터 $x'_{k,d}$와 $y'_{k,d}$을 이용하여 클러스터링을 한다.

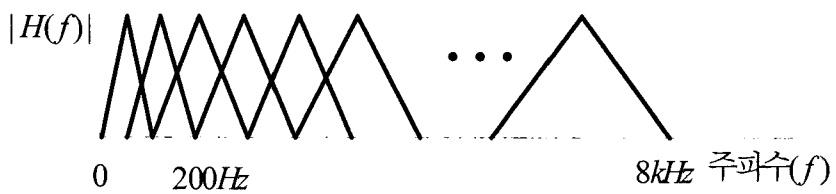
4.1.2. 멜 스케일 필터

스무딩 필터의 필터 폭을 멜 스케일로 조정하여 스무딩을 적용한 후 클러스터링 한다. 필터 폭을 균일하게 조정하지 않고 사람의 달팽이관을 본뜬 멜-스케일로 조정함으로써 사람의 청각 신경 특징을 적용하였다. 사람의 청각은 저주파에 민감

하고 고주파에서는 상대적으로 감지를 잘하지 못한다. 음성 분리를 수행할 때 정량적으로는 성능이 개선되지만 실제로 청취하였을 때는 그렇지 않은 경우가 종종 있다. 멜-스케일 필터는 필터의 폭을 멜-스케일 필터로 조정함으로써 청각 특징에 맞추어 음성 분리가 수행되므로 청취하였을 때 그 결과가 더 좋아질 수 있다. 멜-스케일 필터는 주파수 영역에서만 이웃한 로그 스펙트럼 벡터 값을 참조한다. 멜 스케일로 조정된 로그 스펙트럼 벡터의 채널을 $mel(d)$라고 할 때, 다음과 같이 정의된다.

$$mel(d) = 25 \times \ln(1 + d/22), \qquad d = 0,1,...256. \tag{28}$$

<그림 2>에 차원이 63인 멜-스케일 필터의 예를 나타내었다. 257 차원을 가진 로그 스펙트럼 벡터에 멜-스케일 필터를 사용하여 스무딩을 적용하였을 때, 각 멜-스케일 필터의 필터 밴드 하나당 1 개의 계수가 계산되어 전체적으로 63개의 계수가 계산된다. 스무딩은 멜-스케일 필터의 필터 폭 안에 속하는 로그 스펙트럼 벡터들의 평균으로 계산된다. 로그 스펙트럼 벡터 x, z에 멜-스케일 필터를 이용하여 스무딩을 적용한 후 계산된 63 차원의 특징 벡터를 각각 $\tilde{x} = \{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_{63}\}$, $\tilde{z} = \{\tilde{z}_1, \tilde{z}_2, ..., \tilde{z}_{63}\}$라 하면 혼합 음성 신호의 로그 스펙트럼 벡터 z_d가 x_d일 확률, $p(x_d = z_d|z)$는 $p(\tilde{x}_d = \tilde{z}_d|z)$로 계산한다.



<그림 2> 멜-스케일 필터의 예

균일 마스크 필터를 사용하여 스무딩을 적용한 로그 스펙트럼 벡터의 차원이 스무딩을 적용하기 전의 로그 스펙트럼 벡터의 차원과 같은 반면, 멜-스케일 필터를 사용하여 스무딩을 적용한 로그 스펙트럼 벡터의 차원은 스무딩을 적용하기 전의 로그 스펙트럼 벡터의 차원에 비하여 줄어든다. 따라서 멜-스케일 필터를 적용하여 분리한 음성을 복원하기 위해, 계산된 소프트 마스크 값을 멜-스케일 필터들의 가중치에 따라 로그 스펙트럼 영역에서 분포시킨다.

4.2. 스무딩이 적용된 소프트 마스크 필터

$x'_{k,d}$의 평균과 분산을 $\mu_{x',M_x,d}$, $\sigma^2_{x',M_x,d}$ 그리고 y'_d의 평균과 분산을 $\mu_{y',M_y,d}$, $\sigma^2_{y',M_y,d}$라고 할 때, 스무딩이 적용된 소프트 마스크 필터는 다음과 같이 정의된다.

$$p(x'_d = z_d|z) = \sum_{M_x,M_y} p(M_x,M_y|z) \frac{N(z_d;\mu_{x',M_x,d},\sigma^2_{x',M_x,d}) \times \int_{-\infty}^{z_d} N(y_d;\mu_{y',M_y,d},\sigma^2_{y',M_y,d})dy_d}{p(z_d|M_x,M_y)} \quad (29)$$

마찬가지로, 시간-주파수 영역에서 스무딩이 적용된 최소평균자승오차는 다음과 같이 정의된다.

$$\begin{aligned}\hat{x}'_d = \sum_{M_x,M_y} & \frac{p(M_x,M_y|z)}{p(z_d|M_x,M_y)} \\ & \times [N(z_d;\mu_{y',M_y,d},\sigma^2_{y',M_y,d}) \times \int_{-\infty}^{z_d} x'_d N(x'_d;\mu_{x',M_x,d},\sigma^2_{x',M_x,d})dx'_d \\ & + N(z_d;\mu_{x',M_x,d},\sigma^2_{x',M_x,d}) \times \int_{-\infty}^{z_d} N(y'_d;\mu_{y',M_y,d},\sigma^2_{y',M_y,d})dy'_d]\end{aligned} \quad (30)$$

스무딩 필터를 적용하여 음성 분리를 수행할 때의 계산량은 스무딩을 적용하지 않는 기존 방법의 계산량에서 각 스무딩 필터를 계산하는 것만큼이 추가된다. 즉, 균일 마스크 필터를 사용하여 스무딩 필터를 계산할 때는 매 프레임마다 257의 계수 각각에 3x3 크기의 마스크의 평균을 구하기 위한 6 번의 합과 6 번의 곱의 계산이 추가된다. 또, 스무딩 필터를 멜-스케일 필터를 사용하여 계산할 때는 각 필터에 속하는 로그 스펙트럼 벡터들의 평균을 계산하기 위한 계산량과 다시 257의 차원을 맞춰주기 위해 수행하는 재배열을 위한 계산량이 추가된다. 균일 마스크 필터와 멜-스케일 필터 모두 기존의 방법에서 요구되는 계산량에서 추가로 요구되는 계산량은 많지 않다.

5. 실험결과

음성 분리의 결과를 확인하기 위하여 소프트 마스크와 MMSE 각각에 스무딩 필터를 적용하여 음성 분리 실험을 수행하였다.

5.1. 음성 데이터베이스

음성 분리 실험을 위해 사용된 음성 파일은 Interspeech 2006 음성분리대회 (speech separation challenge)[6]에서 제공하는 데이터베이스에서 선택하였다. 혼합 음성신호는 두 화자의 음성 신호를 합산하여 사용하였다. 학습을 위한 음성 데이터는 여섯 명의 화자(여성 화자 3, 남성 화자 3)로부터의 각 10 문장을 사용하였다. 음성은 GRID 형식으로 <command:4> <color:4> <preposition:4> <letter:25> <number:10> <adverb:4>와 같이 6 가지의 단어로 표현되어 있으며, 각 단어의 구성은 다음과 같다.

command : bin, lay, place, set
color : blue, green, red, white
preposition : at, by, in, with
letter : W를 제외한 A부터 Z까지의 알파벳
number : 0, 1부터 9까지의 숫자
adverb : again, now, please, soon

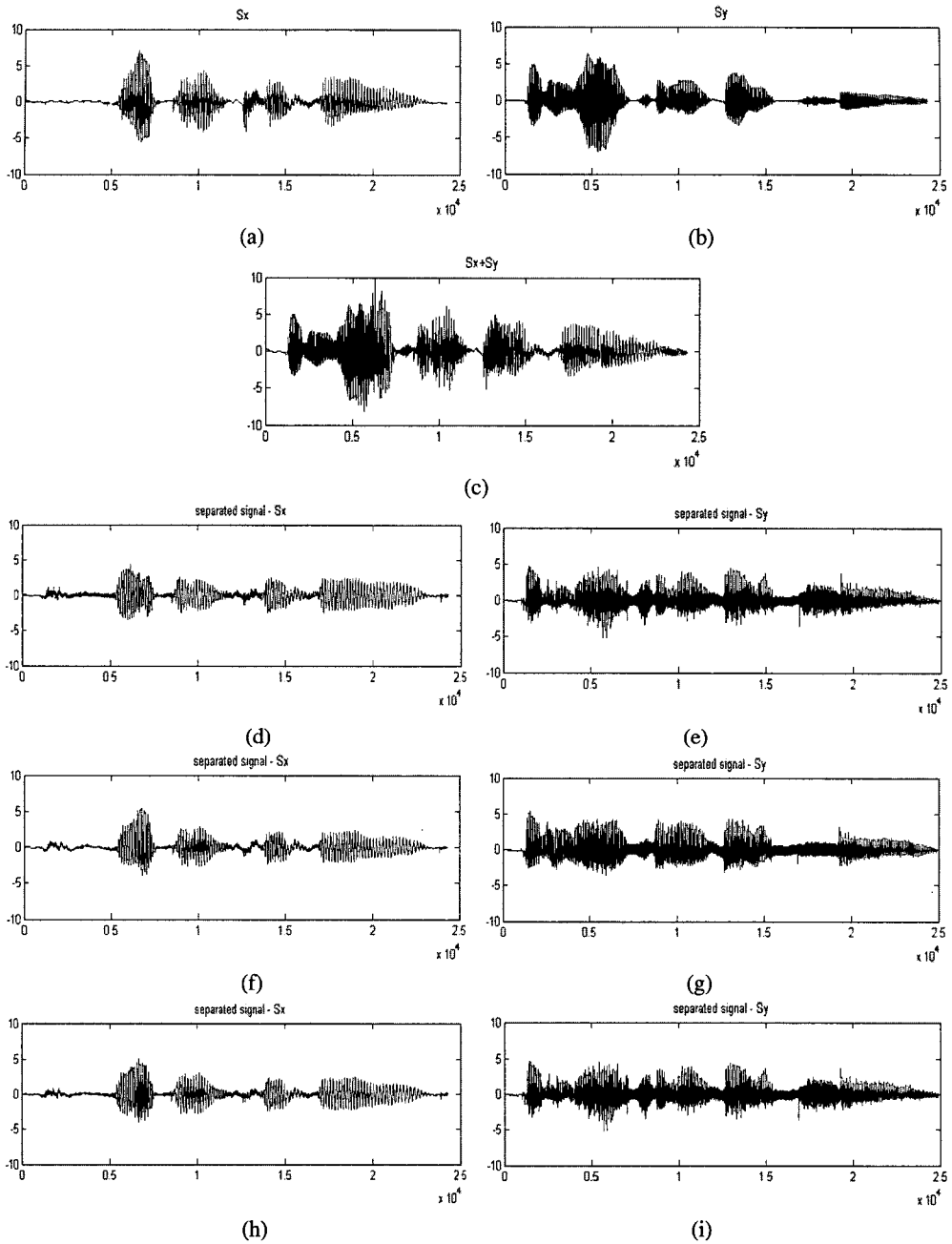
샘플링 주파수가 25 kHz인 음성분리대회에서 제공되는 음성 데이터를 16 kHz로 축소하여 사용하였으며 실험에 사용된 음성 데이터는 평균 0, 분산 1을 갖도록 정규화하였다. 정규화된 음성 데이터를 32 ms의 크기의 윈도우의 이산 푸리에 변환을 512 차수로 계산하고 앞부분 257개 계수를 분리하여 로그를 취해 로그 스펙트럼 벡터로 사용하였다. 혼합 가우시안의 밀도는 학습 데이터로부터 계산된다.

5.2. 음성분리 실험 결과

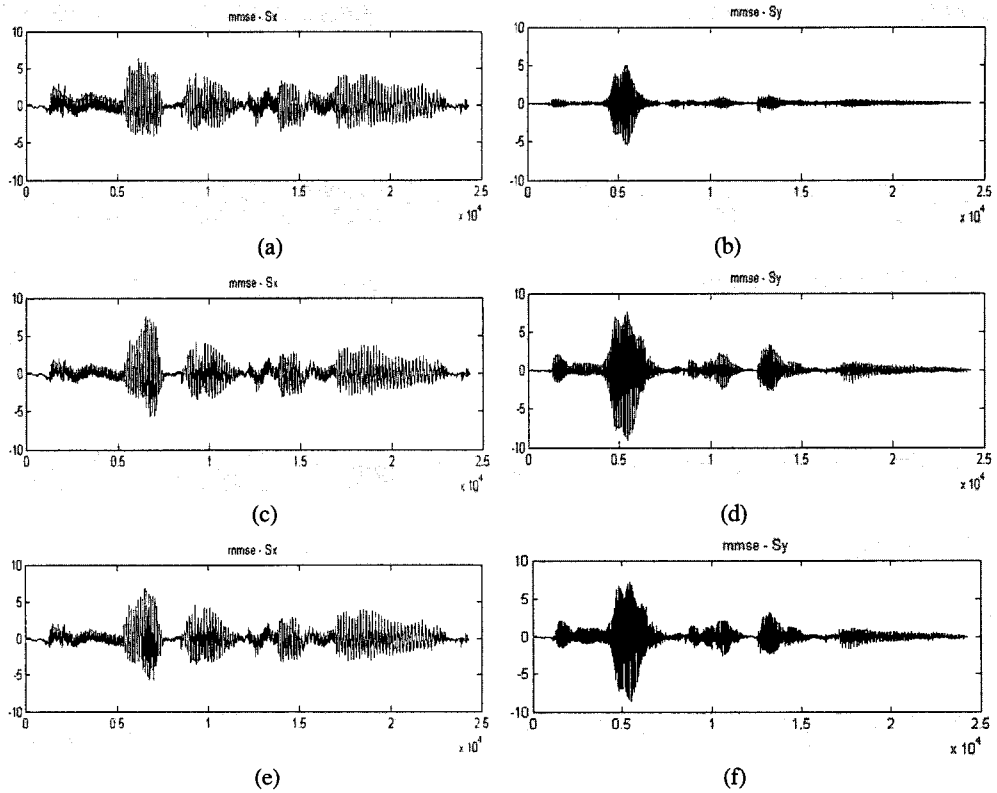
음성 분리 실험을 위한 혼합 음성 신호는 정규화된 두 화자의 음성 데이터를 합산하여 사용하였으며, 남성 화자+남성 화자, 남성 화자+여성 화자, 여성 화자+여성 화자의 세 가지 경우로 구분하여 실험을 하였다. 혼합된 음성 데이터가 -10, -5, 0, 5, 그리고 10 dB에서의 5개의 화자대간섭비(speaker-to-interference ratio: SIR)를 갖도록 음성 데이터를 혼합하였다. 음성의 분리된 정도를 확인하기 위하여 두 화자의 음성신호와 혼합된 음성신호, 분리 된 후의 두 화자의 음성신호를 파형과 스펙트로그램을 출력하였으며, 음성 분리의 결과를 수치적으로 보기 위하여 SIR를 계산하여 비교하였다.

5.2.1. 파형 및 스펙트로그램

<그림 3>와 <그림 4>, 그리고 <그림 5>에 화자 S_x, S_y와 혼합된 음성신호



<그림 3> 소프트 마스크를 사용하여 분리된 음성 신호; (a) 화자 S_x의 음성 신호, (b) 화자 S_y의 음성 신호, (c) 혼합된 음성 신호, (d) 스무딩을 적용하지 않고 분리한 S_x 음성 신호, (e) S_y 음성 신호, (f) 균일 마스크 필터를 적용하여 분리한 S_x 음성 신호, (g) S_y 음성 신호, (h) 멜 스케일 필터를 적용하여 분리한 S_x 음성 신호, (i) S_y 음성 신호

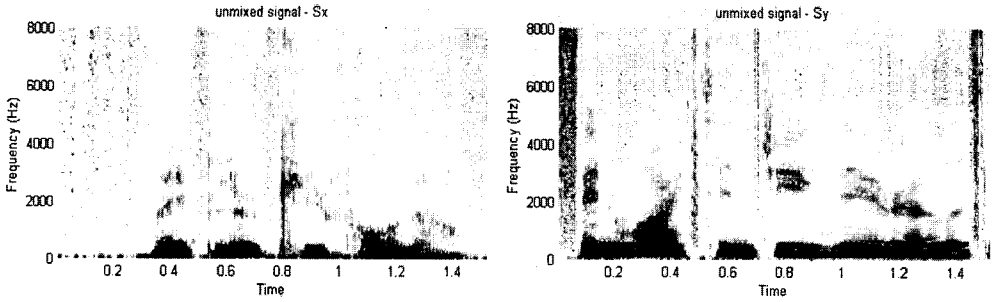


<그림 4> 최소평균자승오차를 사용하여 분리된 음성 신호; (a) 스무딩을 적용하지 않고 분리한 S_x 음성 신호, (b) S_y 음성 신호, (c) 균일 마스크 필터를 적용하여 분리한 S_x 음성 신호, (d) S_y 음성 신호, (e) 멜 스케일 필터를 적용하여 분리한 S_x 음성 신호, (f) S_y 음성 신호

$z(t)$, 그리고 음성 분리를 수행한 결과 파형과 스펙트로그램의 출력을 나타내었다. 소프트 마스크와 최소평균자승오차를 사용한 경우에 대하여 파형과 스펙트로그램 출력을 비교한 결과, 소프트 마스크 알고리듬으로 음성을 분리하였을 때 원하는 음성신호 S_x에 좀 더 가깝게 출력되었다. 스무딩 필터를 적용하여 음성 분리를 수행한 경우, 멜 스케일 필터를 사용하였을 때 보다 균일 마스크 필터를 사용하였을 때 원하는 음성신호에 가깝게 출력되었으며 균일 마스크 필터와 멜 스케일 필터 모두 스무딩 필터를 적용하지 않고 음성 분리를 수행하였을 때 보다 음성 분리에 효과적이었다.

5.2.2. SIR

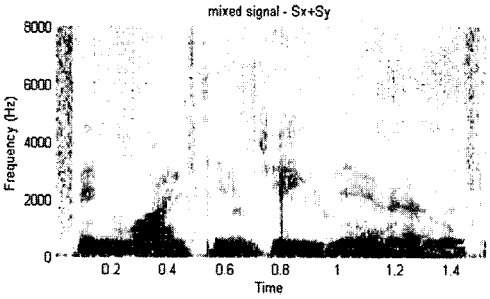
음성 분리 실험의 결과를 수치적으로 확인하기 위하여 분리된 음성 신호의


unmixed signal - Sx
Frequency (Hz)
Time

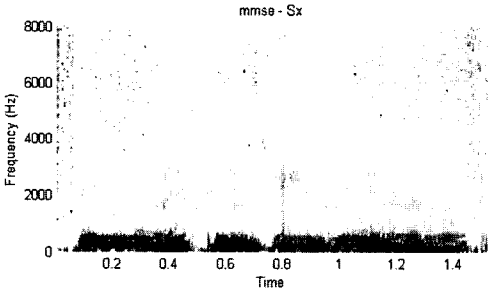
(a)

unmixed signal - Sy
Frequency (Hz)
Time

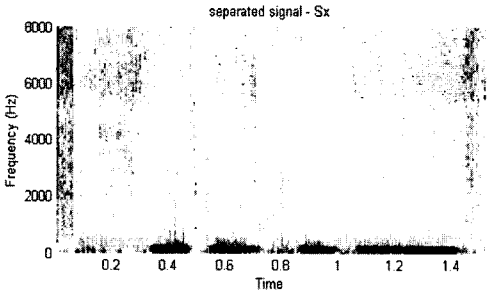
(b)


mixed signal - Sx+Sy
Frequency (Hz)
Time

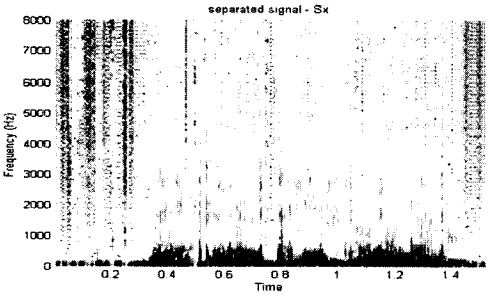
(c)


mmse - Sx
Frequency (Hz)
Time

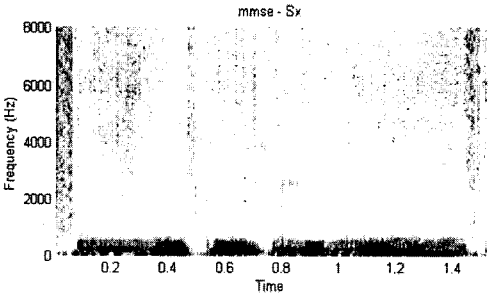
(d)


separated signal - Sx
Frequency (Hz)
Time

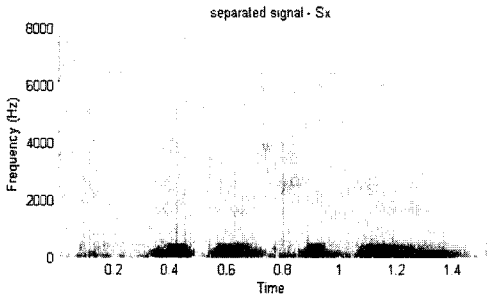
(e)


separated signal - Sx
Frequency (Hz)
Time

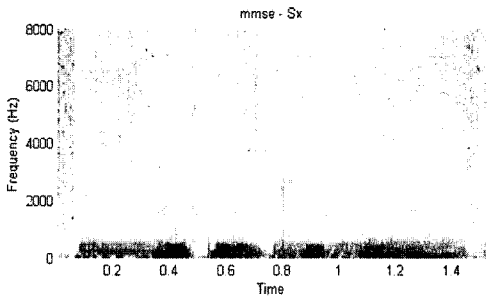
(f)


mmse - Sx
Frequency (Hz)
Time

(g)



(h)



(i)

<그림 5> (a) 화자 S_x의 음성 신호, (b) 화자 S_y의 음성 신호, (c) 혼합된 음성 신호, (d) 소프트 마스크를 사용하여 분리된 S_x 음성 신호, (e) 최소평균제곱오차를 사용하여 분리된 S_x 음성 신호, (f) 균일 마스크 필터가 적용된 소프트 마스크를 사용하여 분리된 S_x 음성 신호, (g) 균일 마스크 필터가 적용된 최소평균자승오차를 사용하여 분리된 S_x 음성 신호, (h) 멜 스케일 필터가 적용된 소프트 마스크를 사용하여 분리된 S_x 음성 신호, (i) 멜 스케일 필터가 적용된 최소평균제곱오차를 사용하여 분리된 S_x 음성 신호

SIR를 계산하였다.

$$SIR = 10\log_{10}\left[\frac{|X|^2}{\left|e^{\log|X|+i\angle Z} - e^{\log|\hat{X}|+i\angle Z}\right|^2}\right] = 10\log_{10}\left[\frac{|X|^2}{(|X|-|\hat{X}|)^2}\right] \tag{31}$$

여기서 $|X|$는 화자 S_x의 음성 신호(혼합되지 않은 신호) 스펙트럼의 크기이고, $|\hat{X}|$는 음성 분리를 수행하여 추출된 음성 신호 스펙트럼의 크기이다. $\angle Z$는 혼합 음성 신호의 위상(phase) 성분을 나타낸다.

테스트 음성 데이터는 남성 화자 3, 여성 화자 3의 여섯 명의 화자로부터의 학습에 사용되지 않은 각 10문장을 혼합하여 사용하였으며, -10, -5, 0, 5, 그리고 10 dB에서의 5개의 SIR를 가진다. 혼합된 음성에 따라서 남성 화자+남성 화자, 남성 화자+여성 화자, 여성 화자+여성 화자로 구분된다. <표 1>과 <표 2>에 스무딩이 적용되지 않은 소프트 마스크와 최소평균자승오차를 사용하여 음성 분리를 수행한 후의 SIR를, <표 3>과 <표 4>에 균일 마스크 필터를 적용한 소프트 마스크와 최소평균자승오차를 사용한 음성 분리 후의 SIR를, <표 5>와 <표 6>에 멜 스케일 필터를 적용한 소프트 마스크와 최소평균자승오차를 사용하여 음성 분리를 수행한 후의 SIR를 나타내었다.

SIR을 계산한 결과, 소프트 마스크를 사용하여 음성 분리를 수행한 경우, 스무

딩을 적용하지 않고 음성 분리를 수행한 후의 SIR에 비하여 균일 마스크 필터를 적용하였을 때는 화자1(S_x)에서 약 2.24 dB, 화자2(S_y)에서 약 1.9 dB의 증가가 있었으며 멜 스케일 필터를 적용하였을 때는 화자1(S_x)에서 약 0.9 dB, 화자2(S_y)에서 약 1.3 dB의 증가가 있었다. 최소평균자승오차를 사용하여 음성 분리를 수행한 경우 균일 마스크 필터를 적용하였을 때는 화자1(S_x)에서 약 0.5 dB, 화자2(S_y)에서 약 1.5 dB의 증가가 있었으며 멜 스케일 필터를 적용하였을 때는 화자1(S_x)에서 약 0.3 dB, 화자2(S_y)에서 약 1.3 dB의 증가를 보였다.

<표 7>에 가우시안의 개수가 2, 4, 8, 16, 32일 때, 0 dB의 SIR를 가지는 혼합된 음성의 분리 후의 SIR를 나타내었다. 테스트 음성 데이터는 남성 화자+남성 화자, 남성 화자+여성 화자, 여성 화자+여성 화자의 세 가지 경우를 사용하였으며, 결과 SIR은 실험 후 계산된 SIR들의 평균으로 나타내었다. 가우시안의 개수를 늘려 실험한 결과, 일관되게 성능이 개선되었으나 균일 마스크 필터를 적용한 경우 가우시안의 개수가 1인 경우보다 가우시안의 개수가 2일 때 화자1(S_x)에서 약 0.07 dB, 가우시안의 개수가 2인 경우보다 4일 때 약 0.12 dB, 4인 경우보다 8일 때 약 0.07 dB, 8인 경우보다 16일 때 약 0.02 dB, 그리고 가우시안의 개수가 16인 경우보다 32인 경우 약 -0.01 dB의 증가를 보였다. 또, 멜-스케일 필터를 적용한 경우 가우시안의 개수가 1인 경우보다 2인 경우 화자1(S_x)에서 약 0.16, 2인 경우보다 4일 때 약 0.07, 4인 경우보다 8일 때 약 0.06, 8인 경우보다 16인 경우 약 0.04, 그리고 16인 경우보다 32인 경우 약 0.01의 증가를 보여 기존 알고리듬에 처음 균일 마스크 필터와 멜 스케일 필터를 적용하였을 때 SIR 증가량인 2.24 dB, 0.5 dB에 비하여 성능 개선이 크지 않았다. 또, 가우시안의 개수가 늘어남에 따라 계산량이 비례적으로 증가한다. 본 실험에서는 편의상 가우시안의 개수는 계산량이 적으면서 성능의 개선이 큰 1인 경우로 고정하였다.

전체적인 결과를 살펴보면 스무딩을 적용하지 않았을 경우보다 스무딩을 적용하여 음성 분리를 수행하였을 경우에 음성 분리가 잘 되었으며, 멜 스케일 필터를 적용하였을 때 보다 균일 마스크 필터를 적용하였을 때 시간과 주파수 영역에서 음성의 특징을 모두 참조하면서도 음성을 분리하기 위한 음성의 시간과 주파수 대역의 채널을 많이 줄이지 않아 음성 분리가 잘 되었다. 또, 같은 성의 화자의 음성 신호와 혼합된 음성 신호일수록 음성 신호와 잡음 요소의 음성 신호의 특성이 비슷하여 음성의 분리가 어려웠고 다른 성의 화자가 잡음 요소의 음성신호로 혼합된 음성 신호일수록 음성의 분리가 잘 되었다.

<표 1> 소프트 마스크 SIR 계산 결과

	남성+남성		남성+여성		여성+여성		평균	
	남성1	남성2	남성	여성	여성1	여성2	화자1(S_x)	화자2(S_y)
-10	0.74	7.17	2.74	11.76	0.56	8.86	1.35	9.26
-5	2.18	6.12	5.85	10.23	1.83	7.12	3.29	7.82
0	5.98	5.38	8.62	8.05	5.49	5.24	6.70	6.22
5	6.88	1.33	10.86	5.03	7.67	1.24	8.47	2.53
10	8.39	0.84	12.34	2.09	9.28	0.64	10.00	1.19
평균	4.83	4.17	8.08	7.43	4.97	4.62	5.96	5.40

<표 2> 최소평균자승오차 SIR 계산 결과

	남성+남성		남성+여성		여성+여성		평균	
	남성1	남성2	남성	여성	여성1	여성2	화자1(S_x)	화자2(S_y)
-10	0.63	9.00	3.42	12.35	0.63	8.52	1.56	9.96
-5	2.12	7.65	5.98	10.73	2.12	7.12	3.41	8.50
0	5.87	5.96	9.46	8.73	6.13	5.24	7.15	6.64
5	7.86	2.87	11.45	5.09	7.86	2.47	9.06	3.48
10	10.46	1.02	12.61	2.58	9.76	0.81	10.94	1.47
평균	5.39	5.30	8.58	7.90	5.30	4.83	6.42	6.01

<표 3> 균일 마스크를 적용한 소프트 마스크 SIR 계산 결과

	남성+남성		남성+여성		여성+여성		평균	
	남성1	남성2	남성	여성	여성1	여성2	화자1(S_x)	화자2(S_y)
2	0.86	9.93	3.43	15.99	0.60	12.63	1.63	12.85
4	2.61	7.74	7.61	13.52	2.36	10.09	4.19	10.45
8	7.27	6.78	11.90	11.36	8.53	7.19	9.23	8.44
5	8.47	1.71	14.88	6.59	12.54	1.97	11.96	3.42
10	10.22	0.87	16.54	2.34	15.16	0.87	13.97	1.36
평균	5.89	5.41	10.87	9.96	7.84	6.55	8.20	7.30

<표 4> 균일 마스크를 적용한 최소평균자승오차 SIR 계산 결과

	남성+남성		남성+여성		여성+여성		평균	
	남성1	남성2	남성	여성	여성1	여성2	화자1(S_x)	화자2(S_y)
-10	0.82	10.52	3.59	15.35	0.66	11.04	1.69	12.30
-5	2.22	8.11	6.68	13.56	2.43	9.23	3.78	10.30
0	6.44	6.32	11.10	11.53	7.62	6.84	8.39	8.23
5	8.52	2.58	13.23	6.51	9.45	2.62	10.40	3.90
10	11.26	1.10	14.35	2.80	11.24	1.20	12.28	1.70
평균	5.85	5.73	9.79	9.95	6.28	6.19	7.31	7.29

<표 5> 멜 스케일 필터를 적용한 소프트 마스크 SIR 계산 결과

	남성+남성		남성+여성		여성+여성		평균	
	남성1	남성2	남성	여성	여성1	여성2	화자1(S_x)	화자2(S_y)
-10	0.77	10.63	2.85	15.68	0.66	10.54	1.43	12.28
-5	2.25	8.23	6.14	12.90	2.02	8.08	3.47	9.74
0	6.04	6.53	9.14	10.81	5.73	6.15	6.97	7.83
5	7.99	1.62	11.50	6.23	8.97	1.63	9.49	3.16
10	9.61	0.88	13.03	2.14	9.49	0.88	10.71	1.30
평균	5.33	5.58	8.53	9.55	5.37	5.46	6.41	6.86

<표 6> 멜 스케일 필터를 적용한 최소평균자승오차 SIR 계산 결과

	남성+남성		남성+여성		여성+여성		평균	
	남성1	남성2	남성	여성	여성1	여성2	화자1(S_x)	화자2(S_y)
-10	0.65	10.92	3.55	15.65	0.65	10.64	1.62	12.40
-5	2.22	8.49	6.30	13.73	2.23	8.95	3.58	10.39
0	6.25	6.69	10.10	11.47	6.53	6.59	7.63	8.25
5	8.19	2.44	11.96	6.01	8.07	2.57	9.41	3.67
10	10.84	1.12	13.07	2.61	9.84	1.11	11.25	1.61
평균	5.63	5.93	9.00	9.89	5.46	5.97	6.70	7.26

<표 7> 0 dB에서 가우시안 개수에 따른 화자1(S_x)의 SIR(dB) 계산 결과

	기존 알고리듬		균일 마스크 필터		멜-스케일 필터	
	소프트 마스크	MMSE	소프트 마스크	MMSE	소프트 마스크	MMSE
1	6.70	7.15	9.23	8.39	6.97	7.63
2	6.80	7.22	9.36	8.51	7.13	7.77
4	6.92	7.34	9.45	8.60	7.20	7.86
8	7.01	7.41	9.52	8.69	7.26	7.91
16	7.06	7.43	9.53	8.71	7.30	7.93
32	7.05	7.44	9.52	8.68	7.31	7.91

6. 결 론

본 논문에서는 통계적 기반 음성 분리기의 성능을 높이기 위하여 스무딩을 적용하였다. 제시한 방법에서는 3x3의 크기를 가진 균일 마스크 필터와 멜 스케일 필터를 사용하여 음성 분리 과정에서 음성의 특징을 반영하였다. 음성 신호 분리 성능을 평가하기 위하여 Interspeech의 음성 데이터를 사용하여 음성 분리 실험을 수행하였다. 혼합된 음성 신호의 분리 결과는 추출된 음성 신호의 파형과 스펙트로그램 출력으로 확인하였으며, 음성 분리 성능을 정량적으로 보기 위하여 SIR를 계산하였다. 균일 마스크 필터를 적용하여 음성 분리 실험을 수행한 결과 소프트 마스크는 스무딩을 적용하지 않았을 때에 비하여 전체적으로 약 2.1 dB의 증가를 보였으며, 최소평균자승오차는 약 1 dB의 증가를 보였다. 멜 스케일 필터를 적용한 경우 소프트 마스크는 약 1.1 dB의 증가를 보였으며, 최소평균자승오차는 약 0.8 dB의 증가를 보였다. 또, 스무딩을 적용하여 음성 분리를 수행한 후 추출된 음성 신호의 파형과 스펙트로그램 출력에서도 스무딩을 적용하지 않았을 경우보다 원 신호에 가깝게 출력되었다. 향후 음성 인식 전처리 평가용 데이터베이스를 사용한 음성인식 실험을 수행하여 음성 인식률의 향상 관점에서의 성능을 평가하는 것이 필요하다.

참 고 문 헌

[1] G. J. Brown, M. Cooke, "Computational auditory scene analysis", *Computer Speech and Language,* Vol. 8, No. 4, pp. 297-326, 1994.

[2] T. Kristjansson, J. Hershey, P. Olsen, S. Rennie, R. Gopinath, "Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system", *Proc. Interspeech,* pp. 1775-1778, 2006.

[3] T. Virtanen, "Speech recognition using factorial hidden Markov models for separation in the feature space", *Proc. Interspeech*, pp. 1850-1853, 2006.

[4] A. M. Reddy, 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.

[5] M. H. Radfar, R. M. Dansereau, "Single-channel speech separation using soft mask filtering", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 8, pp. 2299-2310, 2007.

[6] M. Cooke, T.-W. Lee. *Speech Separation and Recognition Competition*, Available at http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm.

접수일자: 2008년 8월 11일
게재결정: 2008년 9월 22일

▶ 이윤경(Yun-Kyung Lee)
주소: 361-763 충청북도 청주시 흥덕구 성봉로 410 (개신동)
소속: 충북대학교 제어계측공학과
전화: 043)261-3374
E-mail: yklee@cbnu.ac.kr

▶ 권오욱(Oh-Wook Kwon) : 교신저자
주소: 361-763 충청북도 청주시 흥덕구 성봉로 410 (개신동)
소속: 충북대학교 전기전자컴퓨터공학부
전화: 043)261-3374
E-mail: owkwon@cbnu.ac.kr