論文2000-37CI-6-3

음성의 특징 단계에 독립 요소 해석 기법의 효율적 적용을 통한 잡음 음성 인식

(Independent Component Analysis on a Subband Domain for Robust Speech Recognition)

朴 亨 敏 *, 丁 護 榮 **, 李 泰 遠 ***, 李 壽 永 *

(Hyung-Min Park, Ho-Young Jung, Te-Won Lee and Soo-Young Lee)

요 약

본 논문에서는 잡음이 섞인 음성 신호로부터 특징을 추출하는 과정에서 잡음의 영향이 배제된 음성의 특징을 추출하는 방법을 제안한다. 이 방법은 여러 개의 마이크로폰으로 녹음된 잡음 음성 신호에 독립 요소 해석 (Independent Component Analysis) 기법을 사용한 암묵 신호 분리를 적용하여 잡음 성분을 제거하게 된다. 또한, 새로운 스펙트럼 분석법을 제안하여 음성 인식을 위한 특징에 가까운 단계에서 독립 요소 해석 기법을 효율적으로 적용할 수 있도록 한다. 이 스펙트럼 분석법은 기존의 대역 에너지 계산 방법을 수정하여 하나의 대역을 몇 개의 영역으로 구분하고 그 영역내의 Fast Fourier Transform (FFT) 포인트 값들의 평균을 먼저 구한 후 대역 에너지를 계산하게 된다. 음성과 잡음에 대한 대역 에너지의 표본 분산을 사용한 해석과 인식 실험을 통해 이 스펙트럼 분석법이 잡음에 둔감한 방법임을 보였다. 또, 실세계에서 녹음된 잡음 음성 신호에 대해 새로운 스펙트럼 분석법에 독립 요소 해석 기법을 적용한 방법은 인식 성능을 크게 향상시켰으며, 특히 낮은 신호 대 잡음비에 대하여 효과적이었다. 이 방법은 음성 인식을 위한 특징 단계에 독립 요소 해석 기법을 효율적으로 적용 가능할 수 있도록 하는 방안을 제시한다.

Abstract

In this paper, we propose a method for removing noise components in the feature extraction process for robust speech recognition. This method is based on blind separation using independent component analysis (ICA). Given two noisy speech recordings the algorithm linearly separates speech from the unwanted noise signal. To apply ICA as closely as possible to the feature level for recognition, a new spectral analysis is presented. It modifies the computation of band energies by previously averaging out fast Fourier transform (FFT) points in several divided ranges within one mel-scaled band. The simple analysis using sample variances of band energies of speech and noise, and recognition experiments showed its noise robustness. For noisy speech signals recorded in real environments, the proposed method which applies ICA to the new spectral analysis improved the recognition performances to a considerable extent, and was particularly effective for low signal-to-noise ratios (SNRs). This method gives some insights into applying ICA to feature levels and appears useful for robust speech recognition.

* 正會員, 韓國科學技術院 電子電算學科 및 腦科學硏究센터

(Department of Electrical Engineering & Computer Science and Brain Science Research Center, Korea Advanced Institute of Science and Technology)

** 正會員, 韓國電子通信硏究院 音聲言語팀

(Spoken Language Processing Team, Electronics and Teleommunications Rescarch Institute)

*** 正會員, Institute for Neural Computation, University of California, San Diego

接受日字:2000年5月18日, 수정완료일:2000年10月5日

Ⅰ. 서 론

음성 인식 분야에서 해결해야 할 다양한 문제 중에 잡음에 둔감한 인식 기술의 개발은 가장 중요한 문제라고 할 수 있는데, 그 이유는 잡음 환경 하에서 인식 성능의 저하가 심각한 수준으로 인식 기술의 상용화를 위해서는 필수적으로 해결해야 하는 문제이기 때문이다. 지금까지 잡음에 둔감한 인식 성능을 얻기 위한 많은 방법들이 제안되었으나, 대부분의 방법들은 하나의 마이크로폰만을 이용하여 잡음에 둔감한 특징을 추출하거나[1, 2], 특정한 잡음 환경을 가정하고 인식기의 모델 파라미터나 음성의 특징 벡터를 변환하는 방법을 사용하였다[3, 4]. 그러나, 이러한 방법들은 여러 개의 마이크로폰을 이용할 수 있는 경우에 이를 충분히 활용할 수 없다는 단점을 가질 수밖에 없다. 최근에 여러 개의 마이크로폰을 이용하여 잡음에 둔감한 성능을 얻고자 하는 방법들이 제안되고 있는데, 이들은 beamforming이나 암묵 신호 분리 (blind signal separation) 기법들을 이용하여 음성 신호를 개선한 후에 인식에 사용하는 수준에 머물러 있다[5]. 그러나, 음성 인식을 위한 관점에서 보면 음성의 특징 추출에 필요한 정보만 잡음을 제거하고 사용할 수 있다면 그것으로 충분하다. 이렇게 함으로써 잡음 제거 이후의 음성 신호 표본 하나 하나를 다시 복원할 필요 없이 인식에 필요한 정보만을 얻어내어 직접 인식에 사용할 수 있고, 잡음 제거를 위한 네트워크와 추정해야 하는 파라미터의 개수도 줄일 수 있다.

본 논문에서는 음성 신호로부터 인식에 사용하는 음성 정보를 얻어내는 과정에서 섞여있는 잡음 성분을 효과적으로 제거하여 잡음 환경에서 인식 성능을 향상시키는 방법을 제안한다. 제안한 알고리즘은 잡음 성분을 제거하기 위해 여러 개의 마이크로폰에서 받은 신호를 이용하여 원래 음원 신호를 복원해 내는 독립 요소 해석 기법을 사용하며, 새롭게 제안한 "소대역" 스펙트럼 분석법을 도입하여 이 독립 요소 해석 기법이 특징 추출 단계에서 수행되도록 한다. 이 스펙트럼 분석법은 기존의 대역 에너지 계산 방법을 수정하여 하나의 대역을 몇 개의 주파수 영역으로 구분하고 그 영역내의 FFT 포인트 값들의 평균을 먼저 구한 후 대역 에너지를 계산한다. 음성과 잡음에 대한 대역 에너지의 표본 분산 (sample variance)을 이용한 해석을 통해 이 스펙트럼 분석법이 기존의 대역 에너지 계산 방법에 비하여 잡음에 둔감한 방법임을 보이고, 고립 단어 음성 인식 실험을 통해서 이를 확인한다. 또, 이 스펙트럼 분석법은 소대역의 총합 값과 입력 음성 신호 간에 선형적인 관계를 제공함으로써 잡음 성분을 제거하기 위해 각 소대역에 하나의 분리 네트워크만을 사용할 수 있도록 하고, 특징 추출 단계에 독립 요소 해석 기법의 도입이 가능하도록 한다. 따라서, 이러한 방법은 음성 인식의 관점에서 음성 표본 각각을 개선하기 위해 각 FFT 포인트에 하나의 분리 네트워크를 필요로 하는 기존의 독립 요소 해석 기법보다 계산량을 크게 줄일 수 있다. 고립 단어 음성 인식 실험에서 잡음과 음성 신호의 단순 가중치 혼합 신호에 대한 인식률은 잡음 신호가 없는 음성 신호의 인식률과 거의 비슷한 성능을 나타내었으며, 실세계에서 녹음된 잡음이 섞인 음성 신호에 대해서 제안한 방법은 인식 성능을 크게 개선하였다.

Ⅱ. 독립 요소 해석

독립 요소 해석은 알지 못하는 채널을 통해서 섞인 음원들의 혼합 신호를 입력으로 받아들여 상호 독립적인 음원 신호를 복원해 내는 문제이다. 미지의 확률적으로 독립인 N개의 음원 신호 $s(t)=[s_1(t), s_2(t), \cdots, s_N(t)]^T$가 단순 가중치 혼합을 통해 N개의 센서로 측정된다고 가정한다면, 센서로 측정한 신호 $x(t)=[x_1(t), x_2(t), \cdots, x_N(t)]^T$는 다음과 같이 나타낼 수 있다.

$$x(t) = A \cdot s(t) \tag{1}$$

여기서, A는 미지의 가역 행렬로 "혼합 행렬"이라고 부른다. 결국 문제는 혼합 행렬 A의 역행렬로부터 순서와 크기가 변환된 "분리 행렬" W를 추정하여 원래의 음원 신호를 복원하는 것이 된다.

분리 행렬 W를 추정하기 위해 Bell과 Sejnowski, 그리고 Amari 등은 정보 이론적 접근 방법을 제안하였다[6, 7]. 이 접근 방법들은 추정되는 음원 신호간의 통계적 독립성이 최대가 되도록 분리행렬을 학습하는 것인데, 이를 위해 Bell과 Sejnowski는 정보량 최대화를 이용한 다음과 같은 알고리즘을 제안하였다.

$$\Delta W \propto [W^T]^{-1} - 2\tanh(u)\, x^T \quad (2)$$

여기서 분리 신호 $u(t) = W \cdot x(t)$이다. 이와는 다른 방법으로 Amari 등은 Kullback-Leibler divergence라는 비용 함수와 natural gradient라는 개념을 사용하여 다음과 같은 학습 법칙을 제안하였다.

$$\Delta W \propto [I - f(u)\, u^T] W \quad (3)$$

여기서 $f(x) = \frac{3}{4}x^{11} + \frac{25}{4}x^9 - \frac{14}{3}x^7 - \frac{47}{4}x^5 + \frac{29}{4}x^3$는 비선형 활성 함수를 나타낸다. 이 학습 법칙은 다른 형태의 활성 함수에 대해서도 잘 적용되는데, 이 비선형 활성 함수의 출력이 무한한 값으로 커질 수 있으므로 그 대신에 $\tanh(\cdot)$를 사용하도록 한다[8].

실세계에서는 단순 가중치 합으로 이루어진 혼합 신호를 좀처럼 찾기 힘들며, 전파에 따른 시간 지연과 주위 환경에 의해 convolution의 형태로 섞인 혼합 신호를 센서로 받아들인다. 이 convolved 혼합 신호는 다음과 같은 식으로 표현할 수 있다.

$$x_i(t) = \sum_{j=1}^{N} \sum_{k=0}^{K-1} a_{ij}(k)\, s_j(t-k) \quad (4)$$

여기서 $x_i(t)$는 센서의 측정 신호를, $s_j(t)$는 N개의 음원 신호를 나타내고, $a_{ij}(k)$는 길이 K의 혼합 여파기 계수를 의미한다. convolved 혼합 신호로부터 원래의 음원 신호를 추정하기 위해 주파수 영역에서 분리하는 알고리즘이 제안되었다[8]. 위의 혼합 신호를 푸리에 변환을 통해 주파수 영역으로 변환하면 각 주파수에서 원소간의 곱의 형태로 표현되므로 다음과 같은 식으로 나타낼 수 있다.

$$X_f(t) = A_f \cdot S_f(t), \quad \forall f \quad (5)$$

여기에서 $X_f(t)$와 $S_f(t)$는 각각 센서와 음원 신호의 푸리에 변환을 통해 주파수 f에서 얻은 값으로 이루어진 벡터를 나타내며, A_f는 혼합 여파기의 푸리에 변환을 통해 주파수 f에서 얻은 값을 원소로 갖는 행렬을 나타낸다. 이 식은 식 (1)과 같은 형태를 취하고 있으므로 주파수 영역에서 convolved 혼합 신호는 결국 각 주파수에서의 단순 가중치 혼합 신호의 집합으로 나타낼 수 있음을 알 수 있다. 따라서, 모든 FFT 포인트에서 단순 가중치 합으로 된 혼합 신호의 분리 알고리즘을 적용함으로써 convolved 혼합 신호의 분리가 가능하다. 유일한 차이는 입력 신호가 실수가 아닌 복소수의 형태를 취하고 있다는 점이다. 따라서, 활성 함수가 복소수 영역에서 정의되도록 다음과 같이 새롭게 제안되었다[8].

$$\varphi(z) = \tanh(Re\{z\}) + j \cdot \tanh(Im\{z\}) \quad (6)$$

이 때, Amari의 natural gradient를 이용한 학습 알고리즘은 다음과 같이 바뀌게 된다.

$$\Delta W \propto [I - \varphi(u)\, u^H] W \quad (7)$$

수렴 속도가 빠른 natural gradient의 장점을 이용하기 위해 본 논문에서는 식 (7)을 학습 법칙으로 사용하기로 한다. 주파수 영역에서 신호 분리를 통해 시간 영역에서의 convolution을 주파수 영역에서는 각 주파수에서의 푸리에 계수간의 곱셈으로 해결함으로써 많은 양의 계산을 줄일 수 있다.

Ⅲ. 독립 요소 해석 기법을 이용한 암묵 신호 분리

독립 요소 해석 기법은 "칵테일 파티 문제"를 해결할 수 있는 하나의 방안이 될 수 있다. 여기에서는 임의의 혼합 필터를 사용한 혼합 신호와 실세계에서 녹음된 혼합 신호의 분리에 관한 실험을 수행하였다. 임의의 혼합 필터를 사용한 혼합 신호의 분리에서는 음성과 기타 두 음원 신호가 다음과 같은 혼합 필터를 통해서 혼합되고 이를 다시 분리하였다.

$$A_{11}(z) = 0.9 + 0.5z^{-1} - 0.3z^{-2} \quad (8)$$

$$A_{12}(z) = -0.7z^{-10} - 0.3z^{-11} - 0.2z^{-12} \quad (9)$$

$$A_{21}(z) = 0.2z^{-10} + 0.1z^{-11} + 0.05z^{-12} \quad (10)$$

$$A_{22}(z) = 1.5 - 0.6z^{-1} \quad (11)$$

그림 1에 음성과 기타 신호로부터 혼합된 두 혼합 신호와 분리된 두 음원 신호를 나타내었다.

실세계에서 녹음된 혼합 신호에 대한 분리 실험을 수행하기 위해서 혼합 신호를 16kHz의 표본화 주파수로 일반적인 사무실 환경에서 녹음하였다. 두 개의 마이크로폰과 스피커를 그림 2에서처럼 한 변이 60cm인 정방 위치에 놓고, 두 스피커에서 각각 음성 신호과 잡

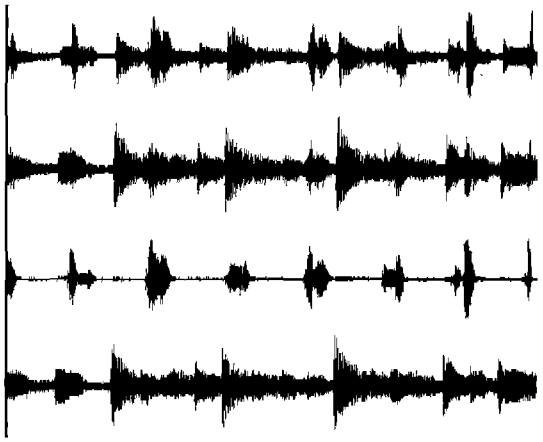

그림 1. 임의의 혼합 필터를 사용한 혼합 신호와 분리된 음원 신호
(위로부터 두 혼합 신호, 분리된 음성과 기타 신호를 나타낸다.)

Fig. 1. Two mixtures using arbitrary mixing filters and two recovered source signals.
(From the top, two mixtures, speech signal, and guitar signal are displayed.)

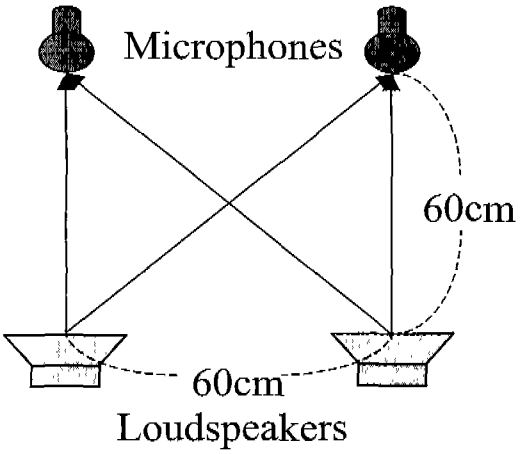

그림 2. 잡음 음성 신호를 녹음하기 위한 마이크로폰과 스피커의 배치

Fig. 2. The placement of loudspeakers and micro phones for recording noisy speech signals.

음 신호가 나도록 하여 두 마이크로폰으로 두 잡음 음성 신호를 녹음하였다. 그림 3에 잡음 신호가 F-16 전투기 소리일 때, 실세계에서 녹음된 두 잡음 음성 신호와 분리된 두 음원 신호를 나타내었다.

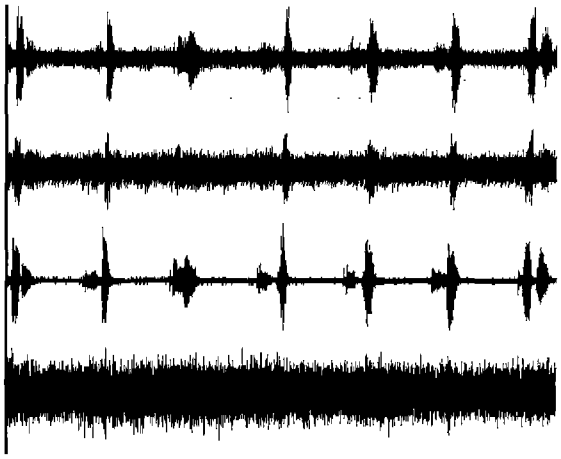

그림 3. 실세계에서 녹음된 혼합 신호와 분리된 음원 신호
(위로부터 두 혼합 신호, 분리된 음성과 잡음 신호를 나타낸다.)

Fig. 3. Two real-recorded noisy speech signals and two recovered source signals.
(From the top, two mixtures, speech signal, and noise signal are displayed.)

Ⅳ. 소대역 스펙트럼 분석

음성 신호의 특징을 추출하기 위해서 대부분의 음성 인식 시스템들은 표본화된 음성 신호에 스펙트럼 분석을 수행한다. FFT를 이용하여 스펙트럼 분석을 수행하는 경우에 k번째 대역 에너지 $y(k)$는 다음과 같이 나타낼 수 있다.

$$y(k)=\sum_{n=f_k}^{l_k}\{Re(\tilde{X}(n))^2+Im(\tilde{X}(n))^2\},\quad k=1,\cdots,K \tag{12}$$

여기에서 $\tilde{X}(n)$는 n번째 FFT 포인트의 값을 나타내고, f_k와 l_k는 각각 k번째 대역의 시작과 끝 포인트를 나타내며, K는 대역의 개수를 나타낸다.

각 대역을 몇 개의 소대역으로 나누고, 각 소대역 내의 FFT 포인트 값들은 에너지를 계산하기 전에 총합을 먼저 구하는 경우를 살펴본다. 이 경우에 k번째 대역 에너지 $y(k)$는 다음과 같이 기존의 대역 에너지 계산 방법과 차이를 보이게 된다.

$$y(k)=\sum_{i=1}^{I}([\sum_{n=f_{ki}}^{l_{ki}}Re(\tilde{X}(n))]^2+[\sum_{n=f_{ki}}^{l_{ki}}Im(\tilde{X}(n))]^2),$$
$$k=1,\cdots,K \tag{13}$$

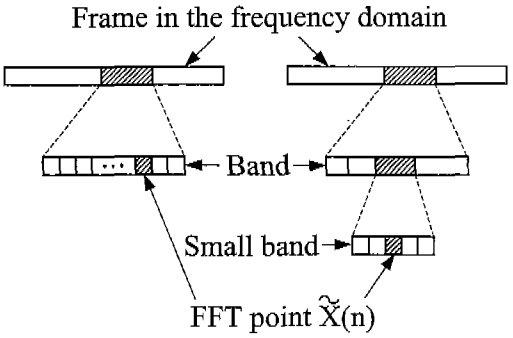

그림 4. 주파수 영역에서 하나의 프레임과 이를 구성하는 FFT 포인트의 관계
Fig. 4. The hierarchy from FFT points to a frame in the frequency domain.

여기에서 I는 k번째 대역에 속한 소대역의 개수를 나타내고, f_{s_i}와 l_{s_i}는 각각 i번째 소대역의 FFT 시작과 끝 포인트를 나타낸다. 그림 4는 주파수 영역에서 하나의 프레임과 이를 구성하는 FFT 포인트의 관계를 나타낸다. 이 방법은 소대역 내의 FFT 포인트 값의 종합을 통해서 각 FFT 포인트에서의 변화 성분들을 많이 경감시킴으로써 특히 잡음이 존재하는 환경에서 인식 성능을 향상시킬 수 있다. 물론, 인식 성능의 저하가 오지 않을 만큼 충분한 개수의 소대역을 사용하여야 한다. 또한, 이와 같은 방법으로 소대역의 종합 값은 입력 음성 신호와 선형적인 관계를 유지하고 있으므로 각 소대역에 하나의 신호 분리 네트워크만을 필요로 하게 되어 분리 네트워크의 개수를 크게 줄일 수 있다.

1. 소대역 스펙트럼 분석법의 잡음에 대한 둔감도 해석

소대역 스펙트럼 분석법의 두 가지 극단적인 경우를 살펴본다. 그 하나는 각 소대역이 하나의 FFT 포인트만으로 이루어져서 대역 속에 FFT 포인트 개수만큼의 소대역을 가지게 되어 기존의 대역 에너지 계산 방법과 동일한 결과를 얻을 수 있는 "경우 1"이며, 다른 하나는 각 대역이 하나의 소대역만으로 이루어진 "경우 2"이다. 주파수 영역에서 잡음이 섞인 음성 입력 신호 $\tilde{X}(n)$은 다음과 같이 음성과 잡음 성분으로 나누어 나타낼 수 있다.

$$\tilde{X}(n)=(S_{r_n}+N_{r_n})+j\cdot(S_{i_n}+N_{i_n}) \tag{14}$$

여기서 S_{r_n}과 N_{r_n}은 각각 FFT 포인트 n에서 음성과 잡음의 실수 값을 나타내고, S_{i_n}과 N_{i_n}은 허수 값을 나타낸다. 경우 1과 경우 2에 대해 k번째 대역 에너지를 각각 $E_1(k)$와 $E_2(k)$라고 하면 다음과 같이 나타낼 수 있다.

$$E_1(k)=\sum_{n=f_k}^{l_k}[(S_{r_n}+N_{r_n})^2+(S_{i_n}+N_{i_n})^2] \tag{15}$$

$$E_2(k)=[\sum_{n=f_k}^{l_k}(S_{r_n}+N_{r_n})]^2+[\sum_{n=f_k}^{l_k}(S_{i_n}+N_{i_n})]^2 \tag{16}$$

k번째 대역 내의 S_{r_n}으로부터 구한 표본 평균(sample mean)과 표본 분산을 각각 m_{S_r}, v_{S_r}이라 하고, 같은 방법으로 m_{S_i}, v_{S_i}, m_{N_r}, v_{N_r}, m_{N_i}, v_{N_i}를 정의하여 각 경우에 대한 대역 에너지를 나타내면 다음과 같다.

$$\begin{aligned}E_1(k)&=\sum_{n=f_k}^{l_k}(S_{r_n}^2+N_{r_n}^2+2S_{r_n}N_{r_n}+S_{i_n}^2+N_{i_n}^2+2S_{i_n}N_{i_n})\\&\approx N(m_{S_r}^2+m_{N_r}^2+2m_{S_r}m_{N_r}+m_{S_i}^2+m_{N_i}^2+2m_{S_i}m_{N_i})\\&\quad+(N-1)(v_{S_r}+v_{N_r}+v_{S_i}+v_{N_i})\end{aligned} \tag{17}$$

$$\begin{aligned}E_2(k)&=[N(m_{S_r}+m_{N_r})]^2+[N(m_{S_i}+m_{N_i})]^2\\&=N^2(m_{S_r}^2+m_{N_r}^2+2m_{S_r}m_{N_r}+m_{S_i}^2+m_{N_i}^2+2m_{S_i}m_{N_i})\end{aligned} \tag{18}$$

식 (17)에서 근사식은 음성과 잡음이 서로 uncorrelate되어 있다고 가정하여 표본 상호 분산(sample covariance)이 0에 가까운 값일 때, $\sum_{n=f_k}^{l_k}S_{r_n}N_{r_n}\approx Nm_{S_r}m_{N_r}$이고, $\sum_{n=f_k}^{l_k}S_{i_n}N_{i_n}\approx Nm_{S_i}m_{N_i}$라는 성질을 이용하였다. 각 대역 에너지를 정규화 하면 두 식에서 N과 N^2의 영향을 무시할 수 있으므로 두 식의 차이는 식 (17)의 근사식에서 표본 분산으로 이루어진 두 번째 항에 의해 나타난다. 이 중 v_{S_r}과 v_{S_i}는 음성에 기인한 성분으로 인식에 필요한 정보로 작용할 수 있는 반면, v_{N_r}과 v_{N_i}는 잡음에 기인한 성분으로 인식 성능을 저하시키는 요소로 작용할 수 있다.

2. 실제 음성 데이터를 통한 표본 분산 비교

백색 잡음이 섞인 실제 음성 신호에 대해서 음성과 잡음 성분의 표본 분산을 살펴보았다. 그림 5는 음성 인식에 필요한 정보로 작용할 수 있는 성분의 손실 정도를 알아보기 위해, 각 소대역이 하나의 FFT 포인트만으로 이루어진 경우를 기준으로 $v_{S_r}+v_{S_i}$를 정규화

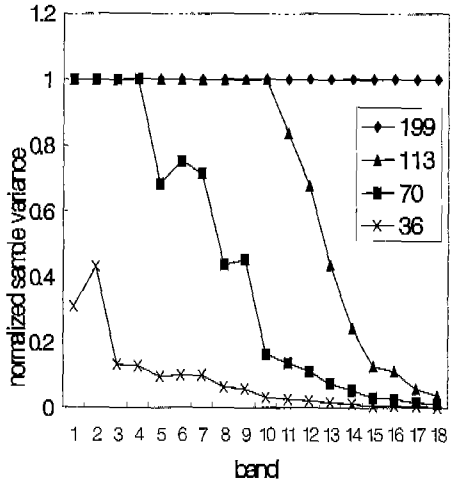

그림 5. 음성의 대역별 정규화된 표본 분산
Fig. 5. Normalized sample variances of speech.

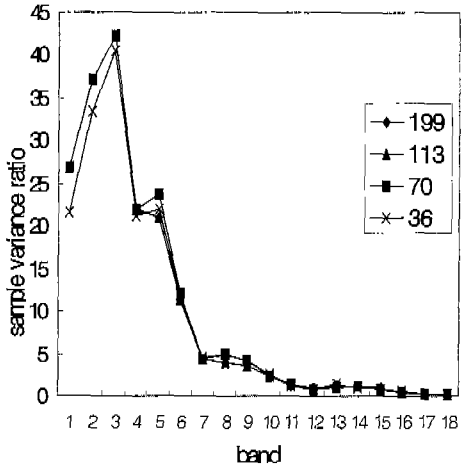

그림 6. 음성과 잡음의 대역별 표본 분산 비
Fig. 6. Ratio of sample variances of speech and noise.

하여 각 대역별로 나타내었다. 대역은 하나의 프레임을 mel 주파수에 따라 구분하였으며[9], mel 주파수 분석 방법은 하나의 프레임에 해당하는 256개의 FFT 포인트 중에서 6번째부터 204번째까지의 FFT 포인트 값을 이용한다. 범례에 있는 숫자는 전체 소대역의 개수를 나타내는 것으로 199는 각 소대역이 하나의 FFT 포인트만으로 이루어진 경우를 나타낸다. 또, 113, 70, 36은 각각 하나의 대역이 8개, 4개, 2개 이하의 소대역으로 이루어진 경우에 전체 소대역의 개수를 나타낸다. 소대역의 개수가 줄어듦에 따라 인식에 필요한 정보가 더 많이 손실됨을 알 수 있으나, 켑스트럼을 추출할 때 큰 영향을 미치는 저주파 영역에서의 정보량 손실 정도는 고주파 영역에서의 손실 정도에 비하여 상대적으로 적다. 이는 mel 주파수에 따라 FFT 포인트를 각 대역에 배정하고 이를 몇 개의 소대역으로 나눌 때, 저주파 영역에서 하나의 소대역에 포함되는 FFT 포인트의 개수가 고주파 영역에 비해 상대적으로 적기 때문이다[9].

그림 6은 신호 대 잡음비가 5dB일 때, 위와 같은 네 가지 전체 소대역 개수에 대해 잡음에 대한 둔감도를 나타낼 수 있는 음성과 잡음의 표본 분산 비 $\frac{v_{S_i}+v_{S_j}}{v_{N_i}+v_{N_j}}$ 를 각 대역별로 나타내었다. 전체 소대역의 개수가 70개인 경우에 다른 경우에 비하여 모든 대역에서 표본 분산 비가 작지 않은 값을 갖는다는 것을 알 수 있다. 따라서, 전체 소대역의 개수가 70개인 경우가 다른 경우에 비하여 상대적으로 잡음에 의한 인식 성능의 저하가 가장 적을 것이라고 추론해 볼 수 있다.

3. 인식 실험

고립 단어 음성 인식 실험을 통해서 제안한 스펙트럼 분석법의 잡음에 대한 둔감도를 확인하였다. 조용한 사무실 환경에서 녹음된 75개의 한국어 고립 단어를 이용하여 연속 분포 Hidden Markov Model (HMM)[10]을 학습시키기 위해 38명의 화자가 말한 데이터를 사용하였고, 테스트에 10명의 화자가 말한 데이터를 사용하였다. 또, 소대역 스펙트럼 분석법을 이용하여 추출한 Mel-Frequency Cepstral Coefficients (MFCC)를 음성의 특징으로 사용하였다. 이 실험에서는 실험의 정확성을 기하기 위해 학습에 참여한 화자의 집합과 테스트에 참여한 화자의 집합을 4가지로 바꾸어 가면서 실험을 하여 그 평균을 인식률로 사용하였다. 표 1은 음성 신호에 백색 Gaussian 잡음이 섞여 있을 때, 몇 가지 신호 대 잡음비에 대해 전체 소대역의 개수에 따른 인식률을 나타낸 것이다. 전체 소대역의 개수가 199개, 18개일 때는 각각 1의 경우 1과 경우 2에 해당하며, 70개일 때는 하나의 대역이 4개 이하의 소대역으로 이루어진 경우에 해당한다.

전체 소대역의 개수가 199개인 경우에 다른 두 경우에 비하여 잡음이 섞인 음성 신호에 대한 인식 성능의 저하가 더 크게 나타남을 알 수 있다. 또, 전체 소대역의 개수가 18개인 경우에는 잡음의 영향이 클 때 상대적으로 효과적이지만, 잡음의 영향이 작을 때 인식 성능의 저하가 상당히 큰 데 이는 그림 5에 나타내었듯이 소대역의 개수가 적어질수록 음성 정보의 손실 정

표 1. 백색 잡음이 섞여 있는 음성 신호에 대한 전체 소대역의 개수에 따른 인식률(%)

Table 1. The recognition rates of speech with white Gaussian noise according to the numbers of total small bands(%).

소대역 개수	∞	20dB	15dB	10dB	5dB
18	87.4	76.1	65.7	47.2	24.2
70	**94.8**	**88.2**	**77.5**	**53.8**	**24.6**
199	94.3	85.9	70.1	43.6	17.8

도가 커지기 때문이다. 잡음이 없을 때 전체 소대역의 개수가 70개인 경우에 199개인 경우와 비슷한 인식 성능을 보인다는 점에서 전체 소대역의 개수가 70개일 때 손실되는 고주파 영역에서의 음성 정보가 인식 성능에 큰 영향을 미치지 못한다는 것을 유추해 볼 수 있다. 실험한 모든 신호 대 잡음비에 대해 전체 소대역의 개수가 70개인 경우에 다른 두 경우에 비하여 가장 높은 인식 성능을 나타내었으며, 이는 그림 6으로부터 추론한 것과 같은 맥락에서 이해할 수 있다. 따라서, 적절한 개수의 소대역을 사용함으로써 극단적인 두 경우의 장점을 모두 취하여 잡음의 영향에 관계없이 가장 높은 인식 성능을 얻을 수 있음을 알 수 있다. 표 2는 F-16 전투기 잡음에 대한 인식 성능을 나타내고 있으며, 백색 Gaussian 잡음에서와 동일한 경향을 나타냄을 알 수 있다.

표 2. F-16 전투기 잡음이 섞여 있는 음성 신호에 대한 전체 소대역의 개수에 따른 인식률 (%)

Table 2. The recognition rates of speech with F-16 fighter noise according to the numbers of total small bands (%).

소대역 개수	∞	20dB	15dB	10dB	5dB
18	87.4	79.2	67.4	50.1	27.4
70	**94.8**	**90.9**	**84.5**	**68.1**	**38.3**
199	94.3	90.1	81.2	60.2	30.7

V. 수정된 대역 에너지에 독립 요소 해석의 적용

독립 요소 해석 기법에 관한 연구는 전통적으로 음원 신호의 분리에 초점이 맞춰져 진행되어 왔다. 즉, 시간 영역에서든 주파수 영역에서든 모든 입력 신호의 각 표본 값에 대해 신호 분리 네트워크를 적용하여 음원 신호의 복원을 꾀하였다. 그러나, 음성 인식을 위한 관점에서 보면 음성의 특징 추출에 필요한 정보만이 필요하고 제안한 스펙트럼 분석법은 소대역의 총합 값과 입력 음성 신호가 선형적인 관계를 유지하고 있으므로, 각 소대역에 하나의 신호 분리 네트워크만을 사용하여 혼합된 잡음의 영향을 제거할 수 있다. 이렇게 함으로써 잡음 제거 이후의 음성 신호를 다시 복원할 필요 없이 인식에 필요한 정보만을 직접 얻어내어 인식에 사용할 수 있고, 잡음 성분을 제거하기 위한 신호 분리 네트워크와 추정해야 하는 파라미터의 개수도 크게 줄일 수 있다.

독립 요소 해석 기법을 도입한 소대역 스펙트럼 분석법의 인식 실험을 위해 3절에서와 동일한 음성 인식 시스템을 사용하였다. 잡음을 발생하는 음원의 개수를 하나로 설정하여 음성 신호를 발생하는 음원과 함께 2개의 음원이 사용되므로 각 신호 분리 네트워크는 입력 2개, 출력 2개로 구성하였다.

1. 단순 가중치 혼합 신호에 대한 인식 실험

표 3은 잡음 신호가 F-16 전투기 소리일 때, 음성과 잡음의 단순 가중치 혼합 신호에 대하여 독립 요소 해석 기법을 적용하였을 때와 적용하지 않았을 때의 인식 결과를 나타내고 있다. 하나의 입력 신호는 -5dB의 신호 대 잡음비를 사용하였고, 나머지 하나의 입력 신호는 첫 번째 열에 표시된 신호 대 잡음비를 사용하였다. 분리 행렬의 학습을 위해서 인식할 단어와 그 외 임의의 9개 단어를 사용하였다. 다른 개수의 전체 소대역을 갖는 각각의 경우에 대해 독립 요소 해석 기법을

표 3. 단순 가중치 혼합 신호에 대한 인식률

Table 3. The recognition rates of instantaneous mixtures (%).

SNR (dB)	199		70		18	
	ICA 비적용	ICA 적용	ICA 비적용	ICA 적용	ICA 비적용	ICA 적용
∞	93.9	94.1	95.6	**95.8**	87.8	86.1
15	82.6	93.9	87.6	**95.7**	67.7	86.9
10	62.3	94.3	71.1	**95.6**	51.1	87.2
5	30.8	89.9	38.1	**93.5**	26.8	86.0

적용하지 않았을 때에는 잡음의 영향이 커짐에 따라 인식 성능이 급격히 저하되지만, 독립 요소 해석 기법을 적용하게 되면 신호 대 잡음비에 관계없이 잡음이 없는 음성 신호의 인식률과 거의 비슷한 성능을 보인다. 이는 독립 요소 해석 기법을 통해서 음성에 섞인 잡음 성분이 거의 완벽하게 제거되었음을 의미한다.

2. 실제 환경에서 녹음된 신호의 인식 실험

실제 환경에서 녹음된 잡음이 섞인 음성 신호에 대한 인식 실험을 수행하기 위해 마이크로폰과 스피커를 그림 2와 동일하게 배치하고 잡음 음성 신호를 녹음하였다. 하나의 스피커에서는 음성 신호가, 다른 스피커에서는 잡음 소리가 나도록 하였다. 표 4와 표 5는 각각 잡음원으로 F-16 전투기 소리와 speech babbling 소리를 사용하였을 때, 실제 환경에서 녹음된 잡음이 섞인 음성 신호의 인식 결과를 나타낸 것이다. 각 표의 첫 번째 열에 두 마이크로폰에서 측정된 신호 대 잡음비 중 큰 값을 나타내었다. 모든 신호 대 잡음비에 대해 독립 요소 해석 기법을 적용하였을 때의 인식률이 적용하지 않았을 때에 비하여 크게 향상되었음을 알 수 있다. 그러나, 전체적으로 독립 요소 해석 기법을 적용하였을 때의 인식률이 표 3의 비슷한 신호 대 잡음비에서의 인식률보다 낮게 나타났다. 이는 스피커에서 마이크로폰까지의 먼 거리와 스피커와 마이크로폰의 이상적이지 못한 특성에 기인한 것으로 추정된다. 실험한 모든 신호 대 잡음비에 대하여 전체 소대역의 개수가 70개인 경우가 그 외의 경우에 비하여 인식 성능이 가장 높게 나타났다.

표 4. 실세계에서 녹음된 잡음 음성 신호에 대한 인식률 (%)
(F-16 전투기 소리를 잡음원으로 사용)

Table 4. The recognition rates of noisy speech recorded in real environments with F-16 fighter noise (%).

SNRs (dB)	199		70		18	
	ICA 비적용	ICA 적용	ICA 비적용	ICA 적용	ICA 비적용	ICA 적용
14.96	79.9	82.1	87.8	**90.8**	87.1	89.5
9.95	63.3	74.2	68.9	**87.9**	64.7	86.3
5.26	36.5	61.2	37.0	**79.9**	32.7	74.0

표 5. 실세계에서 녹음된 잡음 음성 신호에 대한 인식률 (%)
(Speech babbling 소리를 잡음원으로 사용)

Table 5. The recognition rates of noisy speech recorded in real environments with speech babbling noise (%).

SNRs (dB)	199		70		18	
	ICA 비적용	ICA 적용	ICA 비적용	ICA 적용	ICA 비적용	ICA 적용
14.92	72.8	73.5	87.2	**89.8**	85.2	86.3
10.01	55.5	74.4	69.7	**86.8**	68.6	86.5
5.03	31.6	65.3	44.7	**82.6**	42.1	79.3

VI. 결 론

본 논문에서는 특징 추출 과정에서 음성에 섞인 잡음 성분을 제거하는 방법을 제안하였다. 제안한 방법은 잡음 성분을 제거하기 위해서 독립 요소 해석에 기반한 암묵 신호 분리 방법을 사용하며, 소대역 스펙트럼 분석법을 통하여 특징을 추출하는 과정에서 이를 수행할 수 있도록 하였다. 독립 요소 해석 기법을 도입한 소대역 스펙트럼 분석법은 각 소대역에 하나의 신호 분리 네트워크만을 필요로 하게 되고 음성 신호의 복원 없이 인식에 필요한 정보만을 직접 얻어낼 수 있으므로 기존의 신호 분리를 위해 사용하는 독립 요소 해석 기법보다 많은 장점을 가지고 있다. 소대역 스펙트럼 분석법은 그 자체만으로도 기존의 스펙트럼 분석법에 비하여 잡음에 더욱 둔감한 특성을 나타내었으며, 여기에 독립 요소 해석 기법을 도입하여 고립 단어 인식 실험에서 인식 성능을 크게 향상시켰고, 혼합된 잡음 성분을 근본적이고 효율적으로 제거할 수 있었다. 특히 잡음 제거를 위해서 독립 요소 해석 기법이라는 신호 분리 기법을 사용함으로써 잡음의 특성에 관계없이 인식 성능을 크게 향상시킬 수 있다. 또, 이 방법은 특징 단계에서 독립 요소 해석 기법을 효율적으로 적용 가능하도록 하는 방안을 제시하고 있다.

Acknowledgements

본 연구는 과기부의 뇌과학 연구사업으로부터 지원받았습니다. 그리고, Te-Won Lee는 미국 National

Science Foundation Grant CCR-9902961로부터 지원 받았습니다.

참 고 문 헌

[1] D.-S. Kim, S.-Y. Lee, and R.-M. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," IEEE Trans. Speech and Audio Processing, Vol. 7, No. 1, pp. 55-69, Jan. 1999.

[2] H. Hermansky, N. Morgan, and H. Hirsch, "Recognition of speech in additive and convolutional noise based on RASTA spectral processing," in Proc. ICASSP, Vol. 2, pp. 83-86, Apr. 1993.

[3] S. F. Boll, "Suppression of acoustic noise in speech using the spectral subtraction," IEEE Trans. Acoust., Speech, Signal Processing, Vol. 27, No. 2, pp. 113-120, Apr. 1979.

[4] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in Proc. ICASSP, pp. 845-848, Apr. 1990.

[5] T.-W. Lee, A. J. Bell, and R. Orglmeister, "Blind source separation of real world signals," in Proc. ICNN, pp. 2129-2135, 1997.

[6] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," Neural Computation 7, pp. 1129-1159, 1995.

[7] S. Amari, A. Cichocki, and H. Yang, "A new learning algorithm for blind signal separation," Neural Information Processing Systems 8, pp. 757-763, 1996.

[8] P. Smaragdis, "Information Theoretic Approaches to Sources Separartion," Masters Thesis, MIT Media Arts and Sciences Dept., 1997.

[9] J. R. Deller, J. G. Proakis, and J. H. Hanson, Discrete-Time Processing of Speech Signals, Macmillan Publishing Company, 1993.

[10] X. Huang, Y. Ariki, and M. Jack, Hidden Markov Models for Speech Recognition, Edinburgh University Press, 1990.

저 자 소 개

朴 亨 敏(正會員)

1975년 4월 6일생. 1997년 2월 한국과학기술원 전기 및 전자공학과 졸업(공학사). 1999년 2월 한국과학기술원 전기 및 전자공학과 졸업(공학석사). 1999년~현재 한국과학기술원 전자전산학과 박사과정 재학중. 주관심 분야는 잡음하 음성 인식, 독립 요소 해석 기법 알고리즘 개발 및 신호 분리에의 응용 등.

李 泰 遠(正會員)

1995년 3월 독일 the University of Tecjmp;pgu Berlin, Electrical Engineering 졸업(공학사). 1997년 10월 독일 the University of Technology Berlin, Electrical Engineering 졸업(공학박사). 1995년~1997년 Max-Planck Institute (Fellow). 현재 The Salk Institute, the Computational Neurobiology Laboratory (Research Associate) 및 the University of California, San Diego, the Institute for Neural Computation (Research Assistant Professor). 주관심 분야는 자율 학습 (Unsupervised Learning) 알고리즘, 신경회로망 및 Bayesian 확률 이론의 신호 처리에의 응용 등.

丁 護 榮(正會員)

1993SUS 2월 경북대학교 공과대학 전자공학과 졸업(공학사). 1995년 2월 한국과학기술원 전기 및 전자공학과 졸업(공학석사). 1999년 8월 한국과학기술원 전기 및 전자공학과 졸업(공학박사). 1999년~현재 한국전자통신연구원 선임연구원. 주관심분야는 음성 인식과 독립 요소 해석 기법 등

李 壽 永(正會員)

1975년 서울대학교 전자공학과 졸업(공학사). 1977년 한국과학원 전기전자공학과 졸업(공학석사). 1984년 Polytechnic Institute of New York 졸업(공학박사). 1977년~1980년 대한엔지니어링(주) (대리). 1982년~1985년 미국 Gencral Physics Corp. (Staff/Senior Scientist). 1986년~현재 한국과학기술원 전기 및 전자공학과 교수. 주관심 분야는 음성 인식, 신경회로망 및 인공 시스템 등