독립 성분 분석과 스펙트럼 향상에 의한 잡음 환경에서의 음성인식

최승호(서울산업대학교)

<차 례>

<Abstract>

Speech Recognition in Noise Environment by Independent Component Analysis and Spectral Enhancement

Seung-Ho Choi

In this paper, we propose a speech recognition method based on independent component analysis (ICA) and spectral enhancement techniques. While ICA tris to separate speech signal from noisy speech using multiple channels, some noise remains by its algorithmic limitations. Spectral enhancement techniques can compensate for lack of ICA's signal separation ability. From the speech recognition experiments with instantaneous and convolved mixing environments, we show that the proposed approach gives much improved recognition accuracies than conventional methods.

* Keywords: Independent component analysis (ICA), Signal separation, Spectral enhancement, Speech recognition

1. 서 론

음성인식기는 일반 사무실과 같이 소음이 억제된 환경에서는 비교적 좋은 성능을 보이지만 실제 인식 환경에 적용된다면 성능이 급격히 저하될 수 있다. 음성인식을 수행하는 실제 환경은 주변 소음, 발성 거리, 마이크 특성, 채널 왜곡 및 화자의 변이 등 인식 성능을 저하시키는 요소들을 수반한다. 음성인식기의 인식률은 학습에 사용된 데이터베이스를 수집할 때와 유사한 환경에서 인식기가 사용될 때 좋은 성능을 보인다. 하지만 실제 환경은 인식기가 사용되는 장소와 시간에 따라 변하게 된다. 대부분의 경우 인식기를 훈련시킬 때는 잡음이 최대한 억제된 환경에서 얻은 음성 데이터베이스를 사용하고 실제 인식할 경우에는 음성과 잡음이 혼합된 신호로부터 잡음을 감쇄하거나 특징파라미터를 향상시키거나 인식 모델을 적응하는 방법 등을 사용한다. 본 연구는 이중에서 잡음을 감쇄하는 기술에 대한 것이다.

대부분의 인식기는 한 개의 마이크 (또는 채널)를 이용하여 음성 신호를 입력받는다. 한 개의 마이크로 입력된 신호로부터 잡음을 최대한 감쇄하고 깨끗한 음성 신호를 얻고자 하는 연구는 꾸준히 진행되어 왔으나 아직까지 인식 성능을 만족시키지 못하고 있다. 최근 들어 독립 성분 분석(ICA: Independent Component Analysis)을 이용하여 혼합된 신호를 분리하는 신호 분리 기술(BSS: Blind Signal Separation)이 연구되고 있다[1, 2, 3]. 이 기술은 신호들 간의 파워 비율에 따라 성능 차이가 크지 않아서 신호대 잡음비(SNR: Signal-to-Noise Ratio)가 낮은 경우에 특히 유용하다. 스펙트럼 차감(spectral subtraction)[4]과 같은 기존 방법들은 SNR이 낮은 경우 잡음 스펙트럼을 잘못 추정하여 원음성까지 손상시키는 경우가 자주 발생한다.

그러나 실제 환경에서 ICA 기술을 이용하여 분리한 신호가 우리가 원하는 만큼의 깨끗한 음질을 항상 보장하는 것은 아니다. 실제 환경에 적용할 때의 가장 큰 문제는 음원의 숫자와 잡음의 비정제성(nonstationarity)이다. 센서로서 사용할 수 있는 마이크의 개수는 한정되어 있고 실제 환경에서 고려해야 할 음원의 숫자는 이보다 클 수 있으며, 그 개수 또한 가변적이다. 잡음도 실험 환경에서 가정하듯 정제적(stationary)이지 않다. 그래서 실제 환경에서 분리해 낸 신호는 일정량의 인공적인 잡음을 포함하고 있는데 이를 후처리 해주는 기술이 필요하다.

본 연구에서는 이와 같이 기존의 스펙트럼 차감법 등의 단점과 ICA 알고리즘의 한계를 보상하기 위해 ICA 기술과 스펙트럼 향상(spectral enhancement)에 의한 잡음 감쇄(noise suppression) 기술을 결합하는 기법을 제안한다. 본 논문은 ICA 기술에 의한 신호 분리 기술을 2장에서 소개하고 3장에서 스펙트럼 향상법, 4장에서 ICA 기술과 스펙트럼 향상 결합 방식을 설명하고 5장에서는 음성인식 실험 결과와 고찰, 그리고 마지막으로 6장에서 결론을 맺는다.

2. 독립 성분 분석(ICA)에 의한 신호 분리

2.1. ICA와 신호 분리

ICA 기술은 두 개 이상의 다수의 센서로부터 얻어진 혼합 신호(mixuture)를 독립적인(independent) 성분들로 분리해 내는 통계적인 방법이다. ICA에 의한 신호 분리 과정을 설명하기 위해 우선 신호가 혼합되는 모델은 다음과 같다. 시간 t에서 각각의 차수 간에 상호 독립적인(mutually independent) 평균값이 0인 N차의 벡터 $s(t)=[s_1(t), s_2(t), \cdots, s_N(t)]^T$를 가정하자. 즉, 벡터 $s(t)$는 N 개의 독립적인 스칼라 값을 갖는 원신호들(source signals)이다. N 개의 원신호들은 N 개의 센서열로 입력되며, 이를 $x(t)=[x_1(t), x_2(t), \cdots, x_N(t)]^T$라고 하자. 이와 같이 입력된 신호는 원신호들이 필터링되어 혼합된 것으로서 다음 식과 같이 표현된다.

$$x_i(t)=\sum_{j=1}^{N}\sum_{k=0}^{M-1} a_{ijk}s_j(t-k) \tag{1}$$

식 (1)에서 j 번째 원신호와 i 번째 센서사이는 M 차의 필터 a_{ij}로서 모델링된다. BSS는 환경에 대한 사전지식 없이 $x(t)$로부터 원신호 $s(t)$를 추출하는 문제이다. 위 식은 컨벌루션 혼합(convolved mixture)을 나타내며, $M=1$인 경우가 식 (2)와 같이 즉시 혼합(instantaneous mixture)인 경우이다.

$$x_i(t)=\sum_{j=0}^{N-1} a_j s_j(t) \tag{2}$$

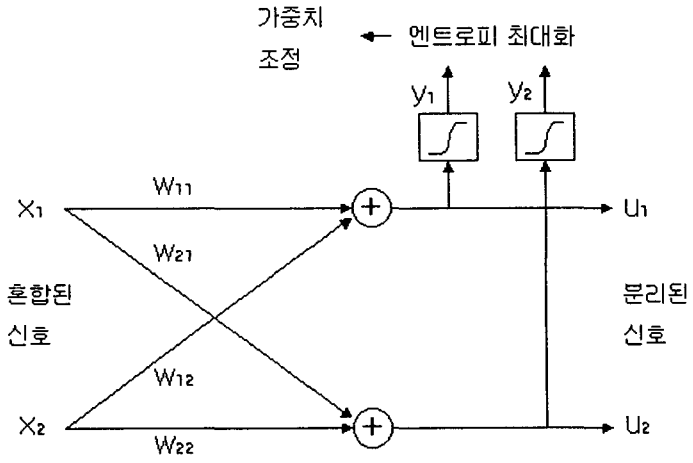
혼합된 신호들로부터 원신호를 추정하기 위해서는 식 (1)에서의 M 차 필터들 $\{a_{ij}\}$에 의해 컨벌루션 혼합된 신호들로부터 디컨벌루션(deconvolution) 과정을 밟아야 한다.

ICA 기술을 이용한 신호 분리 방식은 1995년 Bell과 Sejnowski에 의해 상호정보(mutual information)를 사용하는 정보 이론적인 접근법으로 제안되었고[2], 1996년 Torkkola는 시간 영역 되먹임(time-domain feedback) 구조의 알고리즘을 개발하였다[3]. 본 논문은 ICA에 기반을 둔 여러 가지 신호 분리 알고리즘들의 성능에 대한 자체 사전 연구에 의해 가장 좋은 성능을 보였던 시간 영역 되먹임 알고리즘을 기본으로 한다.

2.2. 시간 영역 되먹임 알고리즘

본 연구에서는 2개의 마이크를 사용하여 실험을 하였고, 이것은 식 (1)에서 $N=2$인 경우이다. 우선, 즉시 혼합의 경우에 대해서 신호 분리 알고리즘을 설명한다. <그림 1>은 혼합된 신호 $x_1(t)$와 $x_2(t)$들로부터 분리된 신호들을 얻기 위한 신호 분리 네트워크이다. 여기서 $y_1(t)=g(u_1(t))$와 $y_2(t)=g(u_2(t))$ 간의 상호 정보를 최소화하도록 분리행렬 $W=\begin{bmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \end{bmatrix}$를 학습 과정을 통해 구한다. 여기서 $g(u)=1/(1+e^{-u})$으로 비선형 함수이다. 학습 규칙은 다음 식 (3)과 같은 stochastic gradient ascent rule을 사용하였다. 여기서 γ는 학습률이다.

$$\Delta W=\gamma(\,1-2\,\boldsymbol{y}(t))\,\boldsymbol{x}(t)^{T}+[\,W^{T}]^{-1} \tag{3}$$



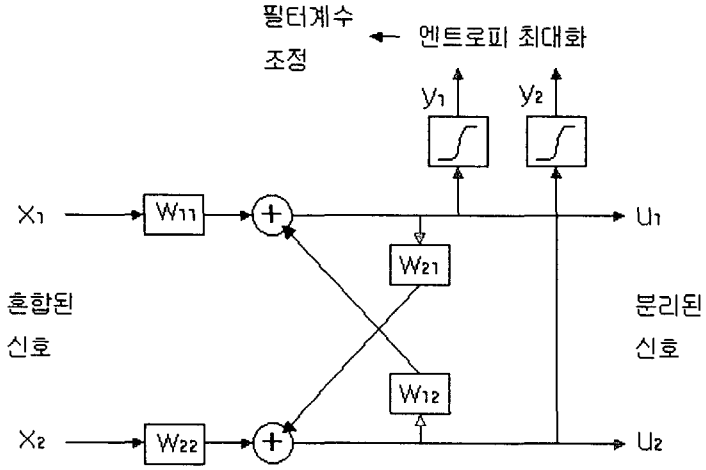
<그림 1> 즉시 혼합 신호 분리 네트워크

컨벌루션 혼합의 경우에는 방과 같은 음향 환경은 각각의 신호원과 마이크 간에 서로 다른 임펄스 응답(impulse response)를 가진다. <그림 2>는 컨벌루션 혼합된 신호들을 분리하기 위해 적응 필터(adaptive filter)들로 이루어진 되먹임 구조의 분리 네트워크이다. 여기에서 W_{11}, W_{21}, W_{12}, W_{22}는 각각 유한 임펄스 응답(FIR: Finite Impulse Response) 필터이다. 이 그림에서 각각의 필터 계수들은 시간 영역에서 다음 식 (4)와 같이 표현할 수 있다.

$$u_1(t) = \sum_{k=0}^{L_{11}} w_{1k1} x_1(t-k) + \sum_{k=1}^{L_{12}} w_{1k2} u_2(t-k),$$
$$u_2(t) = \sum_{k=0}^{L_{22}} w_{2k2} x_2(t-k) + \sum_{k=1}^{L_{21}} w_{2k1} u_1(t-k) \quad (4)$$

식 (4)에서의 각각의 필터 계수들을 학습하기 위한 학습규칙은 다음 식 (5)와 같다.

$$\Delta w_{i0i} = \gamma((1-2y_i)x_i + 1/w_{i0i},$$
$$\Delta w_{iki} = \gamma((1-2y_i)x_i(t-k), \quad (5)$$
$$\Delta w_{ikj} = \gamma((1-2y_i)u_j(t-k)$$



<그림 2> 컨벌루션 혼합 신호 분리 네트워크

3. 스펙트럼 향상법

잡음이 부가된 음성에서 잡음을 감쇄(suppression)하기 위한 방법은 크게 스펙트럼 차감(spectral subtraction) 방식, 모델에 기반한 방식, 음성의 주기성을 이용한 방식 등이 있다. 이중에서 구현이 용이하고 우수한 성능을 발휘하는 스펙트럼 차감 방식이 가장 널리 사용되고 있다[4, 5, 6]. 본 연구도 이러한 스펙트럼 차감방식을 응용한 스펙트럼 향상(spectral enhancement) 기법을 기본으로 한다. 또한 본 연구에서는 각 프레임을 음성 신호의 부재 또는 존재 중의 하나로 구분하는 경판정(hard decision)이 아닌 통계적 확률 모델에 근거하여 판정하는 전체적 연판정 (global soft decision) 기법을 사용한다. 여기서 전체적(global)이라는 것은 각 스펙트럼 성분에

서 독립적으로 판정을 하는 것이 아니라 주어진 프레임에서 전체적으로 판정을 한다는 것이다[6]. t번째 프레임에서 $Y_k(t)$를 k번째 스펙트럼 성분이라고 할 때 입력된 음성의 스펙트럼은 $Y(t)=[Y_1(t), Y_2(t), \cdots, Y_M(t)]$이라고 하자. 음성이 존재하지 않는다는 가정(hypothesis)과 존재한다는 가정을 각각 H_0와 H_1이라고 하면 다음 식 (6)과 같이 표현할 수 있다.

$$\begin{aligned} H_0&: Y(t)=N(t) \\ H_1&: Y(t)=X(t)+N(t) \end{aligned} \tag{6}$$

여기서 $N(t)=[N_1(t), N_2(t), \cdots, N_M(t)]$와 $X(t)=[X_1(t), X_2(t), \cdots, X_M(t)]$는 잡음과 음성의 스펙트럼이다. t번째 프레임에서 추정된 음성만의 스펙트럼을 $\widehat{X}(t)=[\widehat{X}_1(t), \widehat{X}_2(t), \cdots, \widehat{X}_M(t)]$이라고 할 때, Ephraim과 Malah[4]에 의해 제안된 스펙트럼 향상법은 다음 식 (7)과 같다.

$$\widehat{X}_k(t)=G(\alpha_k(t), \beta_k(t))Y_k(t) \tag{7}$$

여기에서 $\alpha_k(t)$와 $\beta_k(t)$는 a priori SNR과 a posteriori SNR이며, 다음 식 (8)에 의해 조정된다.

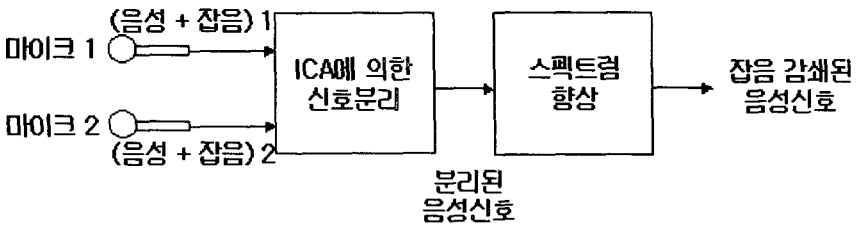
$$\begin{aligned} \tilde{\alpha}_k(t)&=p(H_1|Y(t))\alpha_k(t) \\ \tilde{\beta}_k(t)&=p(H_1|Y(t))\beta_k(t) \end{aligned} \tag{8}$$

여기에서 $p(H_1|Y(t))$는[6]에서의 global soft decision 방식에 의해 구해진다.

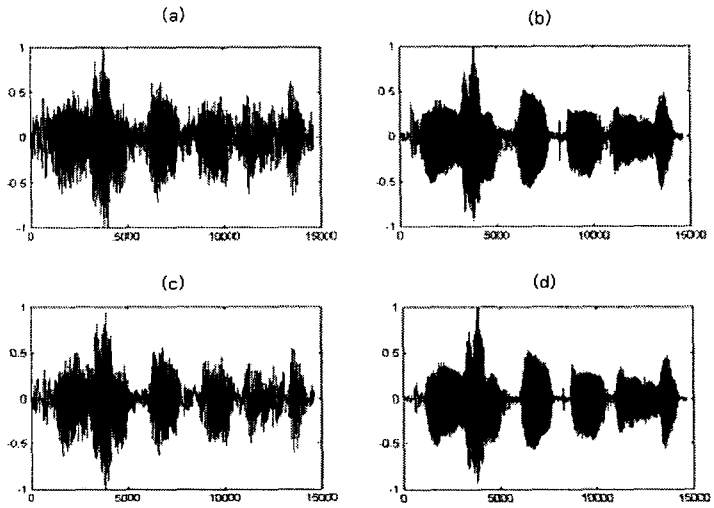
4. 독립 성분 분석과 스펙트럼 향상

스펙트럼 차감 등에 의한 잡음 감쇄 기법은 SNR이 낮은 경우 잡음 스펙트럼에 대한 추정 에러로 인해 원음성까지 손상되는 경우가 자주 발생한다. 그리고 ICA 알고리즘에 의한 신호 분리는 실제 환경에서는 원하는 만큼의 깨끗한 음질을 보장하지 못한다. 실제 환경에서는 마이크의 개수는 한정되어 있고 음원의 개수가 시간에 따라 가변적일 수 있다. 또한 잡음 신호는 실제로 정제적이지 않을 수 있으며, 어떤 시간 구간 동안에는 원하는 음성 신호보다 더 큰 음량을 갖기도 한다.

그래서 실제 환경에서는 ICA에 의해 분리된 음성 신호가 일정량의 잡음을 포함하고 있으며, 이를 스펙트럼 향상법 등에 의해 후처리 해주는 방식이 필요하게 된다.



<그림 3> 독립 성분 분석과 스펙트럼 향상에 의한 잡음 감쇄 기법



<그림 4> 잡음 감쇄 결과 비교
(a) 잡음 음성, (b) ICA 결과, (c) SE 결과, (d) ICA+SE 결과

<그림 3>은 본 연구에서 제안하는 잡음 처리 기법에 대한 블럭도이다. 음성 신호와 함께 마이크에 입력되는 잡음 신호를 감쇄하기 위해 우선 ICA 알고리즘을 적용하여 신호 분리를 수행한다. 그리고 신호 분리 기법에 의해 잡음이 감쇄된 음성 신호를 스펙트럼 향상 기법을 적용하여 비음성 구간에 존재하는 잡음을 최대한 감축한다. <그림 4>는 이러한 신호들의 잡음 감쇄 방법에 따른 비교를 보여 준다. (a)는 음성 신호와 잡음이 혼합된 것이다. 5장의 인식 실험 결과에서 보듯이 이렇게 잡음이 섞인 경우 인식률은 크게 낮아지게 된다. (b)는 ICA 기법을 적용한

경우이다. 그림에서 보이는 바와 같이 음성 구간에서는 SNR을 크게 향상되었다. 그러나 비음성 구간에서는 아직도 잡음의 잔재가 남아 있다. (c)는 (a)의 신호를 스펙트럼 향상법을 이용하여 잡음을 감쇄시킨 파형이다. 많은 양의 잡음이 감쇄되긴 하였지만, 아직도 잡음이 상당량 존재한다. (d)는 ICA로 분리한 후 이것을 스펙트럼 향상법에 의해 추가적으로 잡음을 감쇄한 것이다. 그림의 파형으로부터 상당한 양의 인식률 개선을 예측할 수 있다.

5. 음성인식 실험 및 고찰

본 연구에서 신호 분리 실험은 가로와 세로, 그리고 높이가 각각 3m, 5m, 3m 정도의 연구실 공간에서 수행되었다. 2개의 마이크를 사용하였으며, 마이크 간의 거리는 60cm이고 마이크와 신호 발생원과의 직선거리는 1m이다. 사용된 알고리즘은 시간 영역 되먹임(time-domain feedback) 알고리즘이며 필터의 길이는 128이다. ICA 학습은 최대 10회까지 하였고 학습률은 0.0001이다. 우선, 본 실험에서는 ICA 기법에 의한 신호 분리 성능을 검토하기 위한 음성인식 실험에 대하여 다룬다. 두 개의 마이크를 통해 혼합된 신호를 입력으로 하였을 때 잡음처리를 하지 않은 경우, ICA 만을 한 경우, SE 만을 한 경우, 그리고 ICA와 SE를 함께 사용한 경우 등 네 가지 방식의 비교 실험을 수행하였다. 혼합 방식은 각 시간에서의 샘플 단위로 더하는 즉시 혼합 방식과 방의 임펄스 응답(room impulse response)을 고려한 콘벌루션 혼합 방식을 사용하였다.

음성인식 시스템으로서 본 실험에서는 한국어 연속 숫자음을 대상으로 하였다. 화자는 훈련에 93명(남성 60명, 여성 33명)과 테스트에 47명(남성 30명, 여성 17명)을 사용한다. 각 화자는 40개의 숫자열을 발성하였으며, 각 숫자열은 3개에서 7개까지 랜덤하게 선택된다. 표본화 및 양자화 비율은 8kHz와 16bit이다. 음성인식을 위한 특징파파미터는 12차 멜켑스트럼(MFCC: Mel-Frequency Cepstral Coefficients), 12차 델타 MFCC, 에너지, 델타 에너지로 구성되는 총 26차 벡터이다. 인식 단위는 숫자별 단어 단위이며 연속 분포 은닉마코프모델(HMM: hidden Markov model)을 사용하고 HMM 상태별로 4개의 mixture로 출력 분포를 모델링 하였고, 학습은 Segmental K-means 알고리즘을 사용하고 인식은 One Stage Dynamic Programming 기법을 사용하였다[7]. 잡음 환경 시뮬레이션에 사용된 잡음 데이터는 100km/h의 차량 주행 소음, 백색 가우시안 잡음(white Gaussian noise), babble 잡음의 3가지로 하였다. 잡음이 섞이지 않은 환경에서 인식 결과는 92.67%이다.

시간 영역 되먹임 알고리즘을 사용하여 ICA 실험을 하였다. 즉시 혼합의 경우, 2x2 역행렬을 추정하는 것으로서 1회 학습으로 충분한 결과를 얻었으며 2회 이상의 학습에 의한 실험 결과는 1회 학습의 경우와 거의 변함이 없었다. 컨벌루

션 혼합의 경우, 학습 회수를 각각 1회, 3회, 5회에 대하여 실험 결과를 구하였다. 6회에서 10회까지의 실험 결과는 5회에 대한 결과와 거의 변함이 없었다. 우선 즉시 혼합의 실험은 음성과 잡음을 SNR이 0dB로 혼합된 경우에 대해서 인식 실험을 수행하였다. 매 발성음과 잡음의 파워를 구해서 이를 토대로 0dB로 혼합하기 위한 혼합 행렬을 구하였다. 표에서 None은 혼합된 신호를 그대로 사용한 경우, SE는 혼합된 신호에 대한 스펙트럼 향상법을 적용한 경우, ICA는 혼합된 신호에 대한 ICA 알고리즘 적용한 경우, ICA+SE는 혼합된 신호에 대한 ICA 알고리즘 적용 후에 스펙트럼 향상법을 적용한 경우를 나타낸다. <표 1>에서 보는 바와 같이 SE를 적용한 경우보다 ICA를 적용하거나 ICA와 SE를 동시에 적용한 경우가 매우 높은 인식률을 보임을 알 수 있다.

<표 1> 즉시 혼합시 인식 결과 (단위: %)

잡음	None	SE	ICA	ICA+SE
100km/h	53.49	70.95	87.73	85.46
white	12.29	24.66	66.05	76.01
babble	21.20	34.63	84.38	85.40
평균	28.99	43.41	79.39	82.29

다음으로 컨벌루션 혼합에 대한 실험이다. <표 2>, <표 3>, <표 4>는 ICA 알고리즘의 학습 회수가 각각 1회, 3회, 5회인 경우이다.

<표 2> 컨벌루션 혼합시 인식 실험 결과 (ICA 1회 학습) (단위: %)

잡음	None	SE	ICA	ICA+SE
100km/h	69.09	75.77	61.00	69.72
white	20.86	44.27	23.35	46.43
babble	8.81	12.84	10.61	14.48
평균	32.92	44.29	31.65	43.54

<표 3> 컨벌루션 혼합시 인식 실험 결과 (ICA 3회 학습) (단위: %)

잡음	None	SE	ICA	ICA+SE
100km/h	69.09	75.77	75.27	80.32
white	20.86	44.27	57.81	72.17
babble	8.81	12.84	16.88	21.19
평균	32.92	44.29	49.99	57.89

<표 4> 컨벌루션 혼합시 인식실험 결과 (ICA 5회 학습) (단위: %)

잡음	None	SE	ICA	ICA+SE
100km/h	69.09	75.77	78.25	81.64
white	20.86	44.27	72.90	79.20
babble	8.81	12.84	18.37	22.04
평균	32.92	44.29	56.51	60.96

인식 실험의 결과에서 보듯이 ICA 기법을 인식 전처리 단계에 포함시켰을 경우 인식률의 향상을 얻을 수 있었다. 특히, 스펙트럼 향상법 같은 기존의 방식과 비교했을 때도 성능 향상이 있었고 이들 두 가지 방식을 결합하였을 경우, 즉 ICA 기법에 의해 신호를 분리한 후에 스펙트럼 향상법을 적용한 경우에 가장 좋은 성능을 얻었다. 그러나 이러한 성능 우위는 ICA 학습 회수가 3회 이상일 경우에 뚜렷해지고 1회 학습만을 하였을 경우에는 기존 기법에 비해 성능 향상을 얻을 수 없었다.

6. 결 론

본 논문에서는 잡음 환경에서의 음성인식을 위해 한 개의 채널을 입력으로 하는 기존 방식의 한계를 극복하기 위하여 ICA에 의한 신호 분리 기술과 스펙트럼 향상에 의한 잡음 감쇄 기술을 결합하여 상호 단점을 보완하는 방식을 제안하였다. 음성인식 실험 결과, 즉시 혼합과 컨벌루션 혼합의 실험에서 ICA 기술의 우수성을 볼 수 있었으며, ICA와 스펙트럼 향상의 결합 기법이 기존 방식보다 상당히 향상된 인식률을 얻었다.

참 고 문 헌

[1] A. Hyvarinen, J. Karhunen, O. Erkki, *Independent Component Analysis,* John Wiley & Sons, 2001.

[2] A. J. Bell, T. J. Sejnowski, "An information maximisation approach to blind separation and blind deconvolution", Neural Computation, Vol. 7, No. 6, pp.1129-1159, 1995.

[3] K. Torkkola, "Blind separation of convolved sources based on information maximisation", *Proc. IEEE Workshop on Neural Networks for Signal Processing,* pp.423-432, 1996.

[4] Y. Ephraim, D. Malah, "Speech Enhancement using a minimum mean-squre error short-time spectral amplitude estimator", *IEEE Trans. Acoust., Speech, Signal Processing,* Vol. 32, pp.1109-1121, 1984.

[5] B. L. Sim, Y. C. Tong et al., "A parametric formulation of the generalized spectral subtraction method", *IEEE Trans. on Speech and Audio Processing,* Vol. 6, pp.328-337, 1998.

[6] N. S. Kim, J. H. Chang, "Spectral enhancement based on global soft decision", *IEEE Signal Processing Letters,* Vol. 7, No. 5, pp.108-110, 2000.

[7] L. R. Rabiner, B. H. Juang, *Fundamentals of Speech Recognition,* New Jersey: Prentice-Hall, 1993.

접수일자: 11월 15일
게재결정: 12월 12일

▶ 최승호(Seung-Ho Choi)
주소: 서울특별시 노원구 공릉동
소속: 서울산업대학교 전자정보공학과
전화: 02) 970-6461
FAX: 02) 979-7903
E-mail: shchoi@snut.ac.kr