■ '99 정보과학 논문경진대회 수상작

독립성분 분석을 이용한 강인한 화자식별

(Robust Speaker Identification using Independent Component Analysis)

장 길 진 † 오 영 환 ††

(Gil-Jin Jang) (Yung-Hwan Oh)

요 약 본 논문에서는 독립성분분석을 이용한 음성의 특징 벡터 변환방법을 제안한다. 제안한 방법은 여러 환경에서 수집된 음성신호의 켑스트럼 벡터를 다수의 특징 함수들의 선형결합으로 가정하고, 독립성분분석을 이용하여 분리된 켑스트럼 벡터를 학습과 인식에 사용한다. 변환된 벡터 영역에서는 반복적으로 나타나는 화자의 특징 정보는 강조되고 임의로 나타나는 채널 왜곡은 억제되는 효과를 볼 수 있다. 제안된 방법의 유효성을 검증하기 위해 실제 전화음성으로 문장독립형 화자식별 실험을 수행하였으며, 결과를 통해 독립성분분석을 이용한 특징벡터의 변환이 채널 환경 변화에 대해 보다 강인함을 보였다.

Abstract This paper proposes feature parameter transformation method using independent component analysis (ICA) for speaker identification. The proposed method assumes that the cepstral vectors from various channel-conditioned speech are constructed by a linear combination of some characteristic functions with random channel noise added, and transforms them into new vectors using ICA. The resultant vector space can give emphasis to the repetitive speaker information and suppress the random channel distortions. Experimental results show that the transformation method is effective for the improvement of speaker identification system.

1. 서 론

음성정보처리 기술의 진보와 컴퓨터들의 계산 능력의 향상으로 이전에는 불가능하였던, 대규모 음성 자료의 처리와 다양한 음성의 변이를 수용할 수 있는 확률 모델들의 구축이 가능해졌다. 이에 따라, 화자인식 기술도 비약적인 발전을 거듭하여 제한된 조건하에서는 인간을 능가하는 결과도 보이고 있다[1,2]. 하지만, 화자인식의 주요 응용 분야인 전화음성에서는 대역폭의 제한과 전송선에 의한 채널 응답 특성으로 인해 예측 불가능한 비선형적인 왜곡이 반영되기 때문에, 정확한 화자모델 추정 및 왜곡보상이 불가능하고, 이로 인하여 심각한 성능저하가 발생하게 된다.

기존의 채널 왜곡을 보상하는 방법들은 한 통화 내에서는 그 채널 특징이 변하지 않는다고 가정하고, 긴 구간의 음성에서 그 왜곡을 고정된 하나의 값으로 추정한다. 대표적으로는 켑스트럼 평균 차감법(CMS; cepstral mean subtraction) [3], 최대우도법으로 채널 왜곡을 추정하는 신호 편차 제거법(SBR; signal bias removal) [4], 그리고 채널 왜곡을 선형변환의 형태로 추정하는 켑스트럼 선형 변환법(ATC; affine transform of cepstrum) [5] 등이 있다. 하지만, 이 방법들은 긴 구간의 음성에서 고정적인 왜곡을 추정하기 때문에, 갑작스러운 왜곡 함수의 변화에 취약하며, 음성의 길이 및 왜곡의 정도에 크게 영향을 받는 단점이 있다[3,5].

독립성분분석(ICA; independent component analysis)이란, 특징이 상이한 여러 가지 신호들이 선형적으로 혼합되어 있을 때, 이를 통계적인 방법을 이용하여 효과적으로 분리하는 방법을 통칭하며, 최근에 많은 연구자들의 노력에 의해 여러 분야에 적용되어 좋은 결과들을 보이고 있다. 독립성분분석이란 용어는 1986년 Herault와 Jutten에 의해 처음 제안되었으며 기존의 주

† 비 회 원 : 한국과학기술원 전산학과
jangbal@bulsai.kaist.ac.kr
†† 종신회원 : 한국과학기술원 전산학과 교수
yhoh@cs.kaist.ac.kr

성분분석(PCA; principal component analysis)과 유사하다는 의미로 이름 지어졌다. 독립성분분석이 이와 구분되는 것은 신호들간의 관련도(correlation)뿐만이 아니라 의존성(dependency)까지 최소가 되도록 분리한다는 점이다. 신호들의 통계적인 의존성은 고차통계(HOS; higher-order statistics), 정보 이론(information theory) 등을 이용하여 정의할 수 있으며, 선형적인 가중치를 추정하는 신경회로망의 학습 방법에 의해 그 독립성분들을 추정하고 분리할 수 있다. 이때, 의존성을 정의하는 방법에 따라 추정되는 독립성분들의 특성이 조금씩 다른 형태로 나타나게 된다[6,7]. 독립성분분석의 대표적인 응용분야들로 잡음 제거 및 분리, 음질 개선 등의 음성 처리 분야와 여러 가지 영상 처리 분야 및, 컴퓨터 단층 촬영 영상을 분석하는 의학 분야 등을 예로 들 수 있다. 또한, 복잡한 패턴에서 대표적인 특징값을 추출하거나, 구분이 모호한 성분들을 강조하는 방법 등의 일반적인 패턴 인식에 관련되어 응용되기도 한다.

본 논문에서는 독립성분분석을 이용하여 음성의 특징벡터를 새로운 특징 공간으로 사상하는 방법으로, 왜곡이 심한 전화음성을 입력으로 하는 화자 식별기의 성능을 향상시키는 방법을 제안하고 그 시스템의 구현에 대하여 기술한다. 먼저, 화자의 특징을 나타내는 켑스트럼 벡터 공간을 채널 왜곡에 의한 공간의 분별력 감소에 강인한 공간으로 변환한다. 새롭게 변환된 공간에서는 음성의 특징은 강조되며, 화자간의 구분이 뚜렷하지 않은 분포들은 서로 분리되는 특성을 보일 수 있다. 전화음성과 같이 왜곡이 심한 음성의 경우, 손실된 정보에 의한 화자공간의 분별력 감소에 강인한 특성을 가지는 새로운 특징 공간을 얻을 수 있다. 본 논문의 구성은 다음과 같다. 2절에서는 일반적인 화자식별 방법 및 특징 추출 방법을 설명하고 전화음성 화자식별의 문제점을 소개한다. 3절에서는 제안한 방법에 적용된 독립성분분석 방법에 대해서 그 개념을 설명하고 추정 방법을 기술한다. 그리고, 4절에서는 실험결과를 보이고 제안된 방법의 유효성을 검증하며, 5절에서는 결론 및 앞으로의 연구 방향을 제시한다.

2. 잡음환경에서의 화자식별

음성에서 화자의 특징을 나타내는 특징 파라미터를 추출하고, 벡터공간을 분류할 수 있는 화자 모델링을 통하여 화자식별을 수행할 수 있다. 하지만, 전화음성과 같이 입력음성의 왜곡이 심할 때, 학습 자료와 인식 음성간의 비선형적인 차이가 발생하여 인식률을 저하시키게 된다. 학습자료의 분포를 반영하는 화자 모델은 특징벡터 공간에서 학습환경의 채널특성에 의해 변이된 음성을 표현하게 되므로 다른 환경에서 수집된 음성이 입력될 경우, 그 특징의 변이를 흡수하지 못하므로 다른 환경 음성에 대해서는 모델의 정확도가 떨어지기 때문이다. 임의의 잡음 환경에서의 화자 식별을 위해서는 가능한 모든 환경의 음성을 학습자료로 사용하여 화자 모델에 환경에 의한 영향을 흡수해야 하지만, 이러한 학습자료의 수집은 실제적으로 불가능하다. 따라서, 잡음 환경에서 화자 인식 시스템의 성능을 안정적으로 얻기 위해서는, 특징 파라미터에서 환경에 의한 음성의 왜곡을 보상하는 방법이 필요하다. 본 장에서는 화자식별에서 일반적으로 사용되는 켑스트럼 파라미터에서 환경에 의한 영향을 제거하는 기존의 채널왜곡 보상방법들에 대해 알아본다.

2.1 전화 음성의 왜곡과정

한 화자가 발성한 음성은 마이크에 의해 수집되고 전화선을 통해 다른 화자에게로 전송된다. 이 과정에서 음성 신호를 필터링하는 효과가 발생하며 이는 주파수축에서 스펙트럼 기울기(spectral tilt) 등으로 나타난다. 이에 따라 다른 환경에서 채집된 음성으로 학습과 인식을 수행할 경우 특징 파라미터의 차이가 가산적이 아닌 비선형적으로 나타나게 되어 심각한 인식률의 저하를 유발한다.

$$S(\omega) = G(\omega)H(\omega)M(\omega)\prod_{i=1}^{r} T_i(\omega) \quad (1)$$

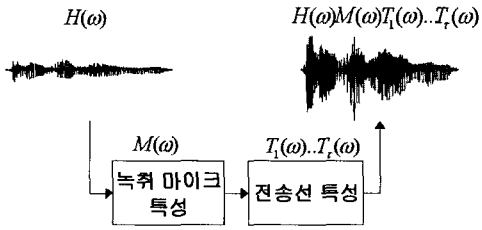


그림 1 전화음성의 왜곡과정

전화음성의 스펙트럼은 그림 1과 같은 과정을 거쳐 위의 식과 같이 음성의 스펙트럼에 전송선에 의한 전달함수가 곱해진 형태로 나타난다. $G(\omega)$, $H(\omega)$, $M(\omega)$, $T_i(\omega)$는 각각 성대의 기본진동, 성도의 특성, 채집 마이크의 특성, 전송선의 특징을 나타내는 전달함수들이다. 켑스트럼 영역에서는 이러한 필터함수들의 곱이 식 2와 같이 선형적인 합으로 나타나게 된다.

$$\begin{aligned}\mathbf{c}[n] &= FFT^{-1}(\log S(\omega)) \\ &\cong FFT^{-1}\left(\log H(\omega)M(\omega)\prod_{i=1}^{P}T_i(\omega)\right) \\ &= \mathbf{h}[n]+\mathbf{m}[n]+\sum_{i=1}^{P}\mathbf{t}_i[n]\end{aligned} \quad (2)$$

켑스트럼 분석 과정에서 성대의 기본진동(glottal pulse)을 나타내는 $G(\omega)$는 제외된다. 음성의 특징은 성도의 전달함수인 $\mathbf{h}[n]$에 의해서만 표현되므로, 학습음성과 인식음성간의 차이를 줄이기 위해서는 전송환경의 특징 함수, $\mathbf{t}_i[n]$을 억제해야 한다.

2.2 기존의 채널왜곡 보상방법들

전화망의 컨볼류션 왜곡을 나타내는 전달함수 $\mathbf{t}_i[n]$은 한번 연결된 통화에 대해서는 거의 변하지 않으며 매 통화마다 바뀐다고 알려져 있다[3]. 대부분의 채널 특성 정규화 방법들은 이 특성을 이용하여 전체 음성구간에서 상수 채널 왜곡을 추정하고 차감하는 방법들을 사용한다. 이러한 방법들에는 크게 채널의 특성만을 정규화하는 방법과, 학습자료와 입력음성의 채널 특성을 일치시키는 방법으로 크게 분류된다. 전자의 대표적인 방법으로는 켑스트럼 평균 차감법[3,5], 후자로는 최대우도 추정법에 의한 신호 편차 제거법[4] 등이 있다.

2.2.1 켑스트럼 평균 차감법 (CMS)

화자가 발성한 문장들에 포함된 음소들의 종류가 균등하게 분포가 되면, 그 문장들에 분석 결과인 켑스트럼 파라미터들도 균등하게 분포된다. 이러한 음성의 켑스트럼의 평균은 모두 같다고 가정할 수 있다. 전송선의 전달함수에 의해 켑스트럼이 왜곡된다면, 채널의 특징은 거의 변하지 않으므로 그 평균치만이 영향받게 된다. 이 방법에서는 왜곡되지 않은 음성의 켑스트럼 벡터의 평균을 0이라고 가정하고, 긴 구간 음성의 켑스트럼 벡터 평균을 채널 왜곡으로 추정한다. 그리고 이를 차감한 켑스트럼 벡터를 학습과 인식에 사용한다. 이 방법은 벡터 공간에서 파라미터의 정적인 분포는 무시하고, 전송함수에 의해 왜곡되지 않는 평균과의 차이만을 특징 파라미터에 반영한다.

$$\hat{\mathbf{c}}[n] = \mathbf{c}[n]-\frac{1}{T}\sum_{t=1}^{T}\mathbf{c}[t] \quad (3)$$

그림 2는 켑스트럼 평균 차감법의 처리과정을 나타낸다. 식 3과 같이 학습과 인식 모두 동일하게 평균을 차감하는 과정을 거친 켑스트럼을 사용한다. 켑스트럼 평균 차감법은 계산량이 많지 않고 간단하며, 일반적인 전화음성 처리에서 안정적인 결과를 보인다. 또한, 음성의 길이가 긴 문장독립형 화자인식에 경우에 보다 좋은 결과를 보일 수 있다.

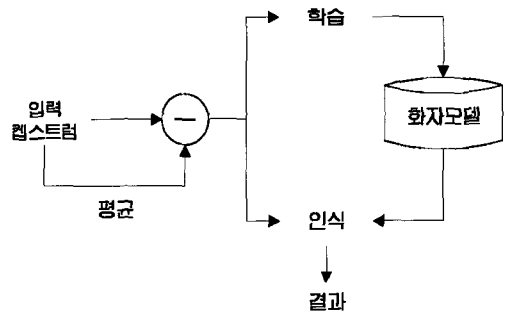


그림 2 켑스트럼 평균 차감법(CMS)

2.2.2 신호 편차 제거법 (SBR)

켑스트럼 평균 차감법과 같이 채널 왜곡에 의해 상수만큼의 편차가 학습자료와 인식음성간에 발생하였다고 가정하고, 최대우도 추정법(MLE; maximum likelihood estimation)에 의해 인식기의 학습에 사용된 음성자료와 특징 벡터의 분포가 가장 유사하도록 입력 벡터에서 상수 채널 편차를 추정하고 이를 차감한다. 실제로 이산 HMM을 이용한 전화음성 인식기에 적용되어 높은 성능 향상을 보인 바 있다.

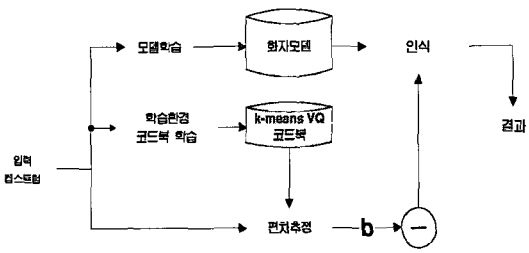


그림 3 신호 편차 제거법(SBR)

채널 왜곡이 켑스트럼 영역에서 고정적이라고 가정하면 전체 학습자료의 코드북에 대하여 채널왜곡을 최소화하는 상수 편차 벡터 $\mathbf{b}^*$의 값은 다음과 같다.

$$\mathbf{b}^* = \arg\min_{\mathbf{b}}\frac{1}{T}\sum_{t=1}^{T}\|(\mathbf{c}[t]-\mathbf{b})-\mathbf{z}(\mathbf{c}[t]-\mathbf{b})\|^2 \quad (4)$$

$\mathbf{z}(\cdot)$: 학습자료의 코드북에 의한 입력의 중심벡터

최대우도 추정법에 의해 반복적으로 채널왜곡 $\mathbf{b}^*$의 추정값 $\hat{\mathbf{b}}^*$을 찾는다.

$$\begin{aligned}\mathbf{b}_{i+1} &= \frac{1}{T}\sum_{t=1}^{T}(\hat{\mathbf{c}}_i[t]-\mathbf{z}(\hat{\mathbf{c}}_i[t])) \\ \hat{\mathbf{c}}_{i+1}[n] &= \hat{\mathbf{c}}_i[n]-\mathbf{b}_{i+1} \\ \hat{\mathbf{b}} &= \mathbf{b}_k, \quad \text{where } \|\mathbf{b}_k-\mathbf{b}_{k-1}\|<\varepsilon\end{aligned} \quad (5)$$

그림 3은 신호편차 제거법의 처리과정을 나타낸다. 켑스트럼 평균 차감법에서와는 달리 학습과 인식에 사용되는 파라미터는 서로 다른 처리 과정을 거친다. 학습은 화자 모델의 학습과, 이에 사용된 켑스트럼 코드북 학습의 두가지 과정으로 나뉘어 진다. 화자 모델은 정규화되지 않은 켑스트럼으로 학습되고, 같은 학습 자료로 k-means 알고리즘에 의해 학습환경의 코드북을 구축한다. 인식과정에서는 학습자료에 대한 입력 켑스트럼의 왜곡을 최소화하는 최적의 편차를 학습환경 코드북으로부터 구하고, 이 편차를 차감한 켑스트럼을 인식에 사용한다. 입력 음성의 출력확률은 공간상에서의 수평, 수직 변이를 허용하는 최적의 값으로 나타나며, 그 왜곡의 크기가 적을수록 정확한 추정이 가능하다.

2.3 기존의 잡음 보상 방법들의 문제점

기존의 보상 방법들은 전체 통화 내에서 채널 왜곡 함수가 변하지 않는다고 가정하고, 전체 통화에 대하여 동일한 채널 왜곡 보상 파라미터를 적용한다. 하지만, 화자의 특징은 개인의 발성 습관이나 사투리, 개인어와 같은 동적 특성 이외에 성도, 성대, 비강 특성과 같은 정적 특성들은 채널 특성과 마찬가지로 전체 발성에 대해서 일정하게 나타난다. 채널 특성을 제거하는 과정에서 화자 고유의 정적 특성의 손실이 발생하며, 시간에 따라 변이하는 채널 왜곡의 제거는 불가능하다.

특히 켑스트럼 평균 차감법의 경우, 입력음성의 통계적인 특성으로부터 채널 왜곡을 추정하므로, 입력음성의 길이가 짧거나 음성에서 음소의 분포가 고르지 않을 경우 추정값의 신뢰도가 떨어지며, 음성에서 채널의 영향이 크지 않을 경우에는 화자 고유의 정적인 정보가 손실되므로 오히려 성능저하를 유발한다. 신호 편차 제거법은 학습자료에 가장 적합한 채널함수를 추정하므로 입력음성의 길이에 의한 영향이 적어 보다 정확한 추정이 가능하므로, 정규화에 의한 정보의 손실을 켑스트럼 평균 차감법보다 줄일 수 있다. 하지만, 역시 변하는 왜곡을 추정하는 것은 불가능하며, 입력음성의 왜곡이 클 경우에는 정확한 채널왜곡의 추정이 어렵다는 문제점이 있다. 다음절에서는 채널 왜곡의 보상과 더불어 음성에서 공통된 화자특성을 증가시키고, 채널 왜곡은 억제하는 독립성분분석을 이용한 강인한 화자 특징 추출 방법을 제안한다.

3. 독립성분분석을 이용한 강인한 화자식별

본 절에서는 전화음성의 채널 변이에 강인한 공간으로 특징 파라미터를 변환하는 방법으로 독립성분분석을 이용한, 제안된 강인한 화자식별 방법인 '독립성분분석을 이용한 켑스트럼 벡터 변환(CVTICA; cepstrum vector transformation using independent component analysis)' 에 대해서 설명한다. 이는 기존의 채널 왜곡 보상 방법들을 일반화한 것으로, 채널 왜곡을 억제함과 더불어 벡터 공간을 구분이 보다 명확하도록 변환한다. 3.1절에서는 독립성분분석 방법에 대해 소개하며, 3.2절에서는 제안한 방법의 개념 및 왜곡 가정에 대해 설명한다. 그리고, 3.3절에서는 제안한 방법의 과정을 자세히 소개한다.

3.1 독립성분분석

독립성분분석이란, 특징이 상이한 둘 이상의 신호들이 선형적으로 혼합되어 구성된 확률변수(random variable)들을 통계적인 방법에 따라 서로 독립적인 신호들로 분리하는 것을 통칭한다. 임의의 신호들은 몇 개의 신호의 가중치가 곱해진 혼합으로 가정되고, 정보이론에 기반하여 혼합된 여러 신호들간의 통계적인 의존성을 정의한다. 그리고, 신경회로망에 사용되는 학습 방법에 따라 신호들간의 의존성이 최소가 되는 가중치를 추정하고 이를 곱하면 서로 독립적인 신호들을 얻을 수 있다[6,7,8].

먼저 q개의 변수 $x_1, x_2, \ldots, x_q$가 p개의 확률적으로 독립적인 변수 $s_1, s_2, \ldots, s_p$의 선형결합으로 이루어졌다고 가정한다. q는 p보다 작지 않으며 평균은 0인 확률변수들이다. 이 확률변수들은 두 개의 확률벡터 $\mathbf{x} = [x_1\ x_2\ \ldots\ x_q]^T$와 $\mathbf{s} = [s_1\ s_2\ \ldots\ s_p]^T$를 구성하며, $\mathbf{x}$는 $\mathbf{s}$와 혼합행렬 $\mathbf{A}$의 곱으로 표현된다. 이때 약간의 오류와 상수편차가 가산됨을 가정한다.

$$\mathbf{x} = \mathbf{As} + \mathbf{e} + \mathbf{b} \qquad (6)$$

독립성분분석의 목적은 원래의 벡터 $\mathbf{s}$와 선형혼합행렬 $\mathbf{A}$의 추정치를 구하는 것이다. $\mathbf{y}$를 $\mathbf{s}$의 추정치, $\mathbf{W}$를 $\mathbf{A}^{-1}$의 추정치로 표현하면 식 6은 다음과 같이 바꾸어 쓸 수 있다.

$$\mathbf{s} \cong \mathbf{y} = \mathbf{W}(\mathbf{x} - \mathbf{b}) \qquad (7)$$

혼합행렬 $\mathbf{A}$대신 관찰신호를 독립신호로 변환하는 $\mathbf{W}$를 구함으로써 독립성분분석의 해를 얻을 수 있다. 통계적인 독립성은 추정된 독립성분들의 결합 엔트로피와 각각의 엔트로피의 차로 계산할 수 있으며, 이를 상호정보(mutual information)로 정의한다. 상호정보는 추정성분들과 같은 공분산과 평균을 가지는 정규분포로 추정한 음정규화 엔트로피(negentropy; negative normalized entropy)로 계산할 수 있다[6].

$$
\begin{aligned}
J(\mathbf{y}) &= S(\mathbf{V}_y\nu) - S(\mathbf{y}) \\
I_m(\mathbf{y}) &= \sum_{i=1}^{p} S(y_i) - S(\mathbf{y}) \\
&= J(\mathbf{y}) - \sum_{i=1}^{p} J(y_i) + \frac{1}{2}\log\frac{\prod v_{ii}}{\det \mathbf{V}_y}
\end{aligned} \tag{8}
$$

$\mathbf{V}_y$는 $\mathbf{y}$의 공분산 행렬이며, $S(\cdot)$은 Shannon의 정의에 따른 엔트로피이다. 이렇게 정의된 상호정보를 최소화하는 가중치들은 각 성분들의 결합확률을 최소화할 수 있으며, 이때의 분리행렬 $\mathbf{W}$는 성분들간의 독립성을 최대화하는 계수들이다. 음정규화 엔트로피는 비선형의 고차 누적분포 함수에 의해 추정할 수 있으며 다음의 대비함수(contrast function) G_k로 계산할 수 있다[9,10].

$$
\begin{aligned}
J_{G_k}(\mathbf{y}) &= E[G_k(\mathbf{y})] - E[G_K(\mathbf{V}_y\nu)] \\
G_1(u) &= \log\cosh(u) \\
G_2(u) &= -\exp(-\tfrac{1}{2}u^2) \quad \text{(Gaussian)} \\
G_3(u) &= \tfrac{1}{4}u^4 \quad \text{(kurtosis)}
\end{aligned} \tag{9}
$$

독립성분의 분리행렬은 상호정보를 최소함으로써 구할 수 있고, 분리된 독립성분들은 대비함수에 의해 특징지어진다.

3.2 켑스트럼의 선형혼합

제안된 방법에서는 독립성분분석을 화자의 특징 벡터인 켑스트럼 파라미터에 적용하여, 채널 왜곡에 강인한 특징을 추출한다. 즉, 공통되는 신호는 강조하고 상이한 신호는 억제하는 독립성분분석의 특징을 이용하여, 일반적인 전화음성에서 화자특성은 강조하고 채널 왜곡은 억제하는 선형변환을 추정한다. 새로운 공간의 켑스트럼 벡터를 학습과 식별에 사용하면, 화자들의 구분이 뚜렷하고, 채널에 의한 화자 모델의 왜곡을 최소화할 수 있다.

2장에서 설명한 기존의 채널 보상방법들의 가정을 다시 정리해 보면, 첫번째로 매 통화시 변경되는 연결 경로에 따라 채널의 왜곡되는 함수의 특성이 바뀐다는 것이다. 화자식별기는 채널에 대해 정확한 정보를 얻을 수 없으므로, 정확한 채널 왜곡함수들을 계산해 내는 것은 실제로 불가능하다. 두번째로, 연결된 한 통화 내에서는 그 특성이 거의 일정하다는 것이다. 채널 왜곡은 전체 통화에서 일정하며, 음성의 변화에 비해 매우 느리게 변한다는 것이다. 이 역시 전화음성에 몇 가지의 서로 다른 채널 특성 함수들이 포함되어 있는지 알 수 없으며, 이와 더불어 존재하는 음성의 특징을 나타내는 전달함수들 간에도 서로 구분이 명확하지 않다. 제안된 방법에서는 그 숫자를 임의의 고정된 숫자로 가정한다. 즉, 전화음성의 켑스트럼을 나타내는 식 2는 다음과 같이 임의의 숫자의 특성 함수들의 합으로 표현할 수 있다.

$$
\begin{aligned}
\mathbf{c}[n] &= \mathbf{h}[n] + \mathbf{m}[n] + \sum_{i=1}^{p} \mathbf{t}_i[n] \\
&= \sum_{i=1}^{p} \mathbf{f}_i[n]
\end{aligned} \tag{10}
$$

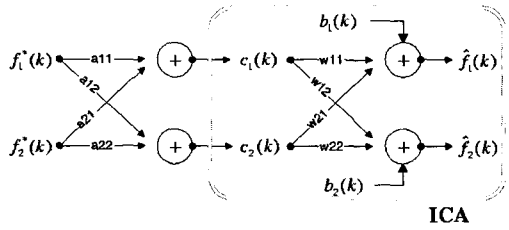
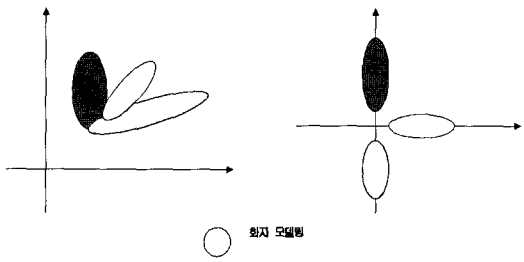


그림 4 입력 켑스트럼의 혼합과정

식 10에 따라 켑스트럼은 여러 가지 음성의 전달함수와 채널 왜곡 함수들의 합으로 표현된다. 그 결과인 켑스트럼은 p개의 다른 성분들로 구성된다. 이렇게 서로 특징이 통계적으로 구분되는 함수들을 벡터 공간에서 p개의 함수들의 선형 혼합으로 가정한다.

가정 1.
p차 켑스트럼은 성도의 특성을 표현하는 p개의 통계적으로 독립적인 전달함수들의 선형혼합이다.



(a) 원래의 켑스트럼 공간 (b) 변환된 켑스트럼 공간

그림 5 독립성분분석에 의한 영역변환 효과: (a) 채널 왜곡과 공간의 겹침으로 인한 잘못된 화자 모델링 (b) 각 성분의 분리로 뚜렷하게 구분된 화자 모델링

실제로는 이보다 많거나 적을 수 있으며, 그 구분 또한 명확하지 않다. 가정 1에 따라 켑스트럼은 식 6의 선형 혼합 모델의 하나로 생각할 수 있다. 식별을 위해서는 음성의 특징을 나타내는 성도 전달함수 $\mathbf{h}[n]$만이 중요하며 채널 왜곡 $\mathbf{t}[n]$은 억제시켜야 한다. 켑스트럼

에 포함되는 채널 전달함수들은 한 통화 내에서는 고정적이지만 서로 다른 통화 사이에서는 연관관계가 없으며, 한 통화 내에서도 조금씩 느리게 변하는 특성을 가지고 있다. 충분히 많은 환경에서 수집된 전화음성의 켑스트럼에 포함된 채널 전달함수들의 평균을, 한 통화에 대해서 0으로 정규화하면 가우시안 분포를 따르는 난수 잡음(Gaussian random noise)으로 간주할 수 있다.

가정 2.
전화음성의 켑스트럼 벡터에 포함된 채널 왜곡의 동적인 변이는 가우시안 분포를 따르는 잡음의 성질을 가진다.

$$\begin{aligned} \mathbf{c}[n] &= \mathbf{A}\mathbf{s}[n] + (\mathbf{t}_o[n] - \bar{\mathbf{t}}_o) + \mathbf{b} \\ &= \mathbf{A}\mathbf{s}[n] + \mathbf{t}[n] + \mathbf{b}, \\ &\text{s.t. } \mathbf{t}[n] \sim N(\mathbf{0}, \Sigma_t) \end{aligned} \tag{11}$$

식 6의 $\mathbf{e}[n]$은 평균이 0으로 정규화된 채널 왜곡함수 $\mathbf{t}[n]$이고, 이는 가우시안 잡음으로 가정된다. $\mathbf{c}$에 포함된 음성의 전달함수 $\mathbf{h}$는 서로 다른 통화에서도 공통적인 특징을 가지고 있다. 다양한 환경의 전화음성의 켑스트럼 벡터들이 독립성분분석 과정을 거치면 공통적인 음성의 특징함수들은 공간상에서 뚜렷하게 분리되고, 채널 함수들은 첨가되는 잡음으로 간주되어 억제되며, 그림 5와 같은 환경변이에 강인한 새로운 켑스트럼 공간을 얻을 수 있다. 독립성분분석은 원점을 중심으로 변환되는 선형변환 행렬을 추정하기 때문에 입력 켑스트럼 공간에 상수 채널 왜곡이 존재하는 경우, 최적의 선형분리 행렬을 얻기 어렵다. 먼저 켑스트럼 공간의 평균은 0으로 정규화한 후 그 선형변환 행렬을 추정한다.

3.3 변환과정

가정 2를 만족시키기 위해서는 일반적인 왜곡의 음성에서 혼합행렬을 구해야 한다. 그 변환의 역변환을 통해 강인한 공간의 켑스트럼을 얻은 후, 이를 식 11의 $\mathbf{s}[n]$으로 가정한다. 실제로 최적의 혼합행렬 $\mathbf{A}$를 구할 수 없으므로, 독립성분분석을 통해 그의 역변환 행렬 $\mathbf{W}_A$를 추정한다. 추정된 행렬은 혼합행렬의 역행렬 $\mathbf{A}^{-1}$이 아니라, 각 성분들의 순서가 섞이고 크기가 다른 행렬이 추정된다.

$$\begin{aligned} \mathbf{c}_I[n] &= \mathbf{W}_A(\mathbf{c}[n] - \mathbf{b}) \\ \mathbf{W}_A &\approx \mathbf{PDA}^{-1} \end{aligned} \tag{12}$$

$\mathbf{P}$는 행들의 순서를 바꾸는 교환행렬이고 $\mathbf{D}$는 행들의 크기를 맞추는 대각행렬이다. 독립성분분석과정에서 채널의 특징은 평균이 0인 가우시안 잡음 $\mathbf{t}[n]$으로 간주되어 억제되고, 구분이 모호한 음성의 전달함수들은 서로 독립적인 성분들끼리 분리되기 때문에 벡터 공간에서 뚜렷이 구분되는 형태로 얻어지게 된다. 이렇게 일반적인 전화음성에서 독립성분 분리 행렬로 얻은 선형변환 켑스트럼 $\mathbf{c}_I[n]$은 채널 변이에 강인하고 벡터공간의 분별력이 높은 특징으로 얻을 수 있다. 그러면, 새로운 켑스트럼을 확률 모델에서 특징벡터로 사용하기 위해서는 새로운 영역에서 두 벡터간의 거리를 정의해야 한다. $\mathbf{W}_A$에 의해 선형변환된 새로운 켑스트럼 벡터 영역에서의 거리와 변환되기 전의 켑스트럼의 거리와의 관계는 식 13에서 얻을 수 있다.

$$\begin{aligned} \mathbf{c}_{Ii} &= \mathbf{W}_A\,\mathbf{c}_i, \quad \mathbf{c}_{Ij} = \mathbf{W}_A\,\mathbf{c}_j, \quad \Delta\mathbf{c}_I = \mathbf{W}_A\Delta\mathbf{c} \\ |\mathbf{c}_{Ii} - \mathbf{c}_{Ij}|_e &= \{\mathbf{W}_A(\mathbf{c}_i - \mathbf{c}_j)\}^2 \\ &= \Delta\mathbf{c}^T\,\mathbf{W}_A{}^T\,\mathbf{W}_A\Delta\mathbf{c} \end{aligned} \tag{13}$$

위의 식에 따라 두 켑스트럼 벡터 사이의 거리를 계산할 때 행렬 $\mathbf{W}_A{}^T\,\mathbf{W}_A$를 두 벡터의 사이에 곱하는 Mahalanobis 거리와 비슷한 형태의 거리 척도를 얻을 수 있다. Mahalanobis 거리는 $|\Delta\mathbf{x}| = \Delta\mathbf{x}^T\,\mathbf{C}^{-1}\Delta\mathbf{x}$로 정의되며, 이 때의 행렬 $\mathbf{C}$는 공분산 행렬로 주로 대각행렬로 근사된다.

이러한 과정들을 그림 6의 블럭도와 같이 3단계로 나누어진다. 먼저, 변환추정 과정에서는 모든 화자들의 학습자료들의 특징을 반영하는 독립성분 변환행렬 $\mathbf{W}_A$를 계산한다. 독립성분분석 알고리즘은 Hyvaerinen [9]에 의해 제안된 음정규화 엔트로피 최대화 알고리즘을 사용한다. 두 번째로, 화자모델의 학습과정에서는 그 화자의 음성을 $\mathbf{W}_A$를 이용하여 변환하고, 그 영역에서의 모델을 구축한다. 마지막으로, 화자식별에서는 임의의 입력음성을 같은 영역으로 변환하고, 각 화자모델에 대한 출현확률을 계산한다.

3.4 화자식별 시스템의 구축

그림 7은 구축된 화자식별 시스템의 전처리 부분을 나타낸다. 독립 켑스트럼으로의 변환 행렬 $\mathbf{W}_A$는 임의의 입력 켑스트럼에 대한 변이를 흡수하기 위해 전체 화자들의 학습자료들을 독립성분 행렬의 추정에 사용한다. 이렇게 변환된 독립 켑스트럼은 학습과 인식에 모두 동일한 방식으로 사용한다. 화자의 분포 모델링을 위해서는 연속형 HMM의 특별한 형태인 HMVQM (hidden Markov VQ-codebook model)을 사용한다[11]. HMM의 관측확률을 정규분포의 가중 합으로 추정하는 연속

형 HMM과는 달리, 이 모델에서는 식 3.9와 같이 최소 벡터양자화 거리로 정의한다. 문장독립형 화자식별 시스템을 구축하기 위해, HMM의 연결구조는 그림 7과 같이 모든 상태 쌍간에 연결이 존재하는 완전연결구조(ergodic topology) [12]를 사용한다. 학습된 HMM에서 가능한 상태천이는 학습자료에서 발생하는 모든 음소열을 확률적으로 사상할 수 있다. 이로써 상태천이는 대분류 음소열을 사상하게 된다.

$$b_i(\mathbf{x}) = \exp(-\min\|\mathbf{x} - \mu_{ik}\|^2) \qquad (14)$$

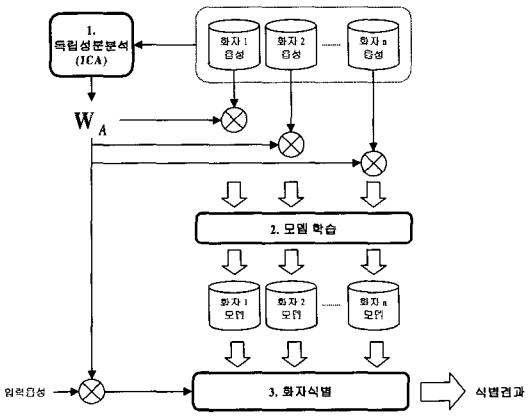


그림 6 CVTICA: 독립성분분석에 의한 전체 화자식별 과정

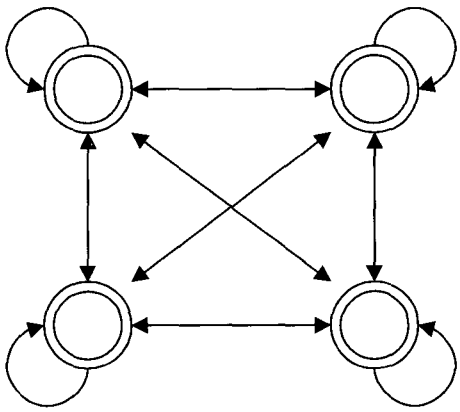


그림 7 완전연결구조(ergodic) HMM

4. 실험결과

실험에 사용된 자료는 장거리 전화음성 데이타 베이스인 SPIDRE (SPeaker IDentification REsearch corpus) [13]이며 특징 추출 단계에서는 인간의 청각특성을 반영한 13차 멜 단위 켑스트럼 벡터를 사용하였다. 화자 모델링에 사용된 HMVQM에 상태 수를 1,2,4,8로 변이하면서 비교 실험하였다. 학습음성의 길이는 30초, 인식음성의 단위는 10초로 하였다. 실험은 네 가지로 나누어 진행하였다. 먼저, 채널 정규화 방법을 적용하지 않은 기본 시스템과, 기존의 채널 정규화 방법인 켑스트럼 평균 차감법(CMS)과 신호편차 제거법(SBR), 그리고 제안된 방법인 특징 파라미터 변환(CVTICA)을 사용하였다. ICA 행렬은 각 화자의 학습에 사용한 84(42명×2)개 음성으로부터 약 10초씩 임의로 선택하여 켑스트럼의 분석차수와 같은 13차로 구하였다. 또한 채널 정규화 방법의 효과를 보기 위해 학습음성의 채널조건과 실험음성의 채널조건이 동일한 경우와, 두 조건이 상이한 경우로 나누어 실험하였다. 채널조건의 분류는 SPIDRE 음성자료에 기재되어 있는 채널 번호를 따랐다

4.1 기존의 채널 정규화 방법

먼저 HMVQM을 이용한 기본 화자식별 시스템(BASE)을 구현하고, CMS, SBR과 그 화자식별 결과를 비교한다. 실험 결과는 표 1과 표 2에 HMVQM의 상태수에 따라 보였다. 학습환경과 인식환경이 같은 동일채널의 경우, 정규화를 수행하지 않은 BASE가 모든 상태수에서 채널 보상 방법을 적용한 CMS, SBR보다 좋은 성능을 얻었다. 이상적인 경우, 채널 보상 방법을 적용하면 음성의 정적인 정보들이 손실되기 때문에 BASE보다 높은 성능을 얻을 수 없다. 하지만, 상태수가 증가하면 모델의 정확도가 향상되기 때문에, CMS와 SBR의

표 1 기존의 채널정규화 방법(동일채널)

상태수	1	2	4	8
BASE	85.7%	87.3%	87.3%	86.5%
CMS	75.4%	80.2%	83.3%	84.9%
SBR	80.9%	83.3%	84.9%	86.5%

표 2 기존의 채널정규화 방법(상이채널)

상태수	1	2	4	8
BASE	28.6%	34.9%	33.3%	33.3%
CMS	43.6%	48.4%	53.9%	55.6%
SBR	37.3%	47.6%	53.2%	51.6%

성능은 BASE에 근접한 결과를 얻을 수 있었다. 상이채널의 경우, BASE는 학습환경과 실험환경의 불일치로 인하여 인식률이 현저하게 떨어졌다. CMS와 SBR은 환경의 차이를 줄여주므로 10%에서 20%까지 BASE보다 성능향상을 보였다. 하지만, 동일채널의 경우보다 30%~50%의 성능저하가 발생하며, 이는 왜곡의 정도가 보상이 불가능할 정도로 심한 것과, 채널 추정의 부정확성에서 기인된다.

4.2 CVTICA: 독립성분분석 적용

표 3과 표 4는 동일채널과 상이채널 조건에서의 CVTICA의 실험결과를 나타낸다. 세 가지 대비함수들의 적용결과들은 동일채널에서는 큰 차이를 보이지 않았으나, tanh와 u^3는 거의 같은 성능을 보였으며 Gaussian은 대체적으로 이들보다 좋은 성능을 보였다. 상이채널에서는 u^3가 다른 두 가지보다 대부분 1~5% 정도의 높은 인식률을 보였다. 따라서, 전체적으로 볼 때 u^3가 안정적이고 분별력이 높은 벡터공간을 추정한다고 결론 내릴 수 있다.

표 3 CVTICA 실험결과(동일채널)

상태수	1	2	4	8
Gaus	81.8%	86.5%	88.9%	89.7%
tanh	83.3%	89.7%	88.1%	89.7%
u^3	82.5%	87.3%	88.9%	92.7%

표 4 CVTICA 실험결과(상이채널)

상태수	1	2	4	8
Gaus	54.8%	57.9%	65.9%	62.7%
tanh	56.4%	59.5%	61.1%	61.1%
u^3	59.5%	57.9%	64.3%	65.9%

4.3 실험결과 종합

표 5와 표 6에서 기존의 방법들과 CVTICA의 실험결과들을 비교하여 나타내었다. CVTICA는 표 3과 표 4에서 가장 좋은 결과를 보였던 u^3를 대비함수로 사용한 결과들이다. CVTICA는 동일채널, 상이채널 조건에 관계없이 대체로 기존의 정규화 방법들보다 높은 성능향상을 보였다. 동일채널 조건에서는 채널 정규화를 수행하지 않은 BASE와 거의 같은 수치의 결과가 나타났다. 하지만, BASE는 상이채널에서 현저하게 그 성능이 저하되었으나, CVTICA는 오히려 상이채널조건에서 기존의 정규화 방법들보다도 높은 수치를 보였다. 보다 자세히 살펴보면, 상이채널의 경우 기존의 CMS, SBR보다 상태수 8에서 11~15% 정도의 인식률 향상을 보였고 나머지에서도 10%이상의 성능향상을 보였다. 따라서, 본 논문에서 제안된 CVTICA 방법이 기존의 채널 정규화 방법들보다 채널 변이에 강인함을 알 수 있다. 동일채널의 경우도 상이채널의 경우보다 향상 폭은 적지만 거의 모두 5%이상씩 인식률이 향상되었다. 특히, CMS나 SBR의 경우 정규화의 영향으로 기본 시스템보다 성능이 떨어지지만 CVTICA는 오히려 더 좋은 성능을 보였다. 즉, CMS나 SBR의 경우 정규화에 의해 화자정보의 손실이 크지만 CVTICA의 경우에는 그 정보 손실이 크지 않으며 오히려 켑스트럼의 각 차수간의 구분을 더 크게 하였기 때문에 정규화에 의한 영향을 감소시킬 수 있었다.

표 5 전체 실험결과(동일채널)

상태수	1	2	4	8
BASE	85.7%	87.3%	87.3%	86.5%
CMS	75.4%	80.2%	83.3%	84.9%
SBR	80.9%	83.3%	84.9%	86.5%
CVTICA	82.5%	87.3%	88.9%	92.7%

표 6 전체 실험결과(상이채널)

상태수	1	2	4	8
BASE	28.6%	34.9%	33.3%	33.3%
CMS	43.6%	48.4%	53.9%	55.6%
SBR	37.3%	47.6%	53.2%	51.6%
CVTICA	59.5%	57.9%	64.3%	65.9%

5. 결 론

본 논문에서는 전화 음성과 같이 왜곡이 심한 음성에서 화자 식별기의 성능 향상을 위한 방법을 제안하였다. 기존의 채널 보상 방법들은 채널 왜곡이 한 통과에 대해서 고정적이라고 가정하고 추정하기 때문에 전체 음성에서 정적인 화자 정보의 손실을 가져올 수 있으며,

시간에 따라 변하는 채널 왜곡의 해결은 불가능하다. 제안된 방법은 화자의 특징 공간인 켑스트럼 벡터를 독립성분분석을 통해 변환된 켑스트럼으로 학습과 인식을 수행하였다.

제안된 방법의 유효성을 검증하기 위해서 문장 독립형 화자인식 연구에 널리 쓰이고 있는 실제 장거리 전화음성 자료인 SPIDRE로 실험을 수행하였다. 채널 보상의 효과를 검증하기 위해서 학습 자료와 실험 자료의 채널이 동일한 경우와 상이한 경우로 분류하여 기존의 방법들과 그 성능을 비교하여 보았다. 전체적으로 제안된 방법은 기존의 채널 보상 방법들에 비해 모든 시스템 환경에서 9%에서 15%까지 인식률이 향상되었다. 기존의 방법들은 동일 채널의 경우, 채널 보상을 수행하지 않은 기본 시스템보다 인식률이 많게는 10%까지 떨어진 반면, 제안된 방법은 그 성능이 저하되지 않았으며, 공간 분별 능력의 향상으로 오히려 6% 정도 인식률이 향상되는 경우도 보였다. 상이 채널의 경우 기존의 채널 보상 방법들과 제안된 방법 모두, 채널 보상을 수행하지 않은 기본 시스템보다 현저하게 인식률이 향상되었다. 이는 제안된 방법이 기존의 방법들을 일반화하면서 보다 분별력이 높은 공간으로 변환함으로써, 순간적인 채널 왜곡에 의한 잘못된 화자 모델화를 방지하였기 때문에, 채널 환경 변화에 대해 보다 우수한 강인성을 가지기 때문이다.

구현된 시스템에서 사용한 방법은 일반적인 특징 공간에서의 변환을 다루고 있기 때문에, 화자 인식기뿐만이 아니라 다른 음성 인식 분야에서 적용이 가능하다. 하지만, 통계적인 방법에 의해 변환공간을 추정하기 때문에 입력음성의 길이가 짧고, 기본 인식기의 성능이 좋은 시스템의 경우에는 오히려 성능을 저하시킬 가능성이 있다. 앞으로 화자 인식만이 아니라 일반적인 전화음성 인식기에 적용할 수 있는 방법과, 벡터 공간에서의 분별능력을 좀 더 높이기 위해 독립성분 추정방법의 개선과 함께 독립성분분석 자체를 화자 모델링에 활용하는 방법에 대한 연구가 진행되어야 할 것이다.

참 고 문 헌

[1] J. P. Campbell, Jr., "Speaker recognition: a tutorial," Proceedings of the IEEE, vol.85, pp.1436-1462, 1997.

[2] J. de Veth and H. Bourlard, "Comparison of Hidden Markov Model techniques for automatic speaker verification in real-world conditions," Speech Communications, vol.17, pp.81-90, 1995.

[3] A. E. Rogenberg, C.-H. Lee, and F. K. Soong, "Cepstral channel normalization techniques for HMM based speaker verification," in Proceedings of ICSLP, Yokohama, pp.1835-1838, 1994.

[4] M. G. Rahim and B.-H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," IEEE Trans. on Speech and Audio Processing, vol.4, pp.16-30, 1996.

[5] R. J. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition: a feature-based approach," IEEE signal processing magazaine, pp.58-71, 1996.

[6] P. Comon, "Independent component analysis, A new concept?" Signal Processing, vol.36, pp.287-314, 1994.

[7] K. J. Pope and R. E. Bogner, "Blind signal separation: linear, instaneous combinations," Digital signal processing, pp.5-16, 1996.

[8] T.-W. Lee, A. Ziehe, R. Orglmeister, and T. Sejnowski, "Combining time-delayed decorrelation and ICA: towards solving the cocktail party problem," in Proceedings of ICASSP, pp.1249-1252, 1998.

[9] A. Hyvaerinen, "A family of fixed-point algorithms for independent component analysis," in Proceedings of ICASSP, pp.3917-3920, 1997.

[10] A. Hyvaerinen, "Independent component analysis by minimization of mutual information," Technical Report A46, Helsinki University of Technology, 1997.

[11] S.-J. Yun and Y.-H. Oh, "Performance improvement of speaker recognition system for small training data," in Proceedings of ICSLP, Yokohama, pp.1863-1866, 1994.

[12] X. D. Huang, Y. Ariki, and M. A. Jack, "Hidden Markov models for speech recognition," Redwood Press Limited, 1990.

[13] J. Godfrey, D. Graff, and A. Martin, "Public databases for speaker recognition and verification," ESCA Workshop on Automatic Speaker Recognition Identification and Verification, pp.39-42, 1994.

[14] C.-C. T. Chen, C.-T. Chen, and C.-M. Tsai, "Hard-limited Karhunen-Loeve transform for text independent speaker recognition," Electronics Letters, vol.33, pp.2014-2016, 1997.

[15] J.-H. Kim, G.-J. Jang, S.-J. Yun, and Y.-H. Oh, "Candidate selection based on significance testing and its use in normalisation and scoring," in Proceedings of ICSLP, pp.141-144, 1998.

장 길 진

1997년 한국과학기술원 전산학과 학사. 1999년 한국과학기술원 전산학과 석사. 1999년 ~ 현재 한국과학기술원 전산학과 박사과정 재학중. 관심분야는 음성인식, 화자인식, 잡음처리.

오 영 환

1972년 2월 서울대학교 공과대학 전자공학과. 1974년 2월 서울대학교 교육대학원 공업교육학과(석사). 1980년 3월 Toko Institute of Technology 정보공학전공(박사). 1981년 4월 ~ 1985년 6월 충북대학교 공과대학 전산학과 조교수. 1983년 12월 ~ 1984년 11월 University of California (Davis) 연구교수. 1985년 7월 ~ 현재 한국과학기술원 전산학과 교수로 재직중. 관심분야는 음성인식, 음성합성, 음성코딩, 화자인식, 대화관리, 신경회로망, 전문가 시스템