1998년도 한국음향학회 학술발표대회 논문집 제 17 권 제 2(s) 호

독립성분분석을 이용한 강인한 화자인식

장 길 진○ 윤 성 신 오 영 환

한국과학기술원 전산학과

Robust Speaker Recognition using Independent Component Analysis

Gil-Jin Jang○ Seong-Jin Yun Yung-Hwan Oh

Department of Computer Science

Korea Advanced Institute of Science and Technology

{jangbal,sjyun,yhoh}@bulsai.kaist.ac.kr

요 약

독립성분분석(ICA: Independent Component Analysis)이란 특징이 상이한 둘 이상의 신호들이 선형적으로 결합되어 있을 때 이를 효과적으로 분리하는 방법들을 통칭하며 잡음제거, 음질개선 및 신호처리 분야에서 많이 활용되고 있다. 본 논문에서는 전화음성 화자인식 시스템의 성능향상을 위해 독립성분분석을 이용하는 방법을 제안한다. 먼저 화자가 발성한 음성신호의 켑스트럼 계수를 여러 채널 함수들의 선형적인 합으로 가정하고, 독립성분분석을 이용하여 연속 새로운 켑스트럼 벡터를 학습과 인식에 사용하였다. 실험자료는 전화음성 화자식별기의 성능평가에 널리 쓰이고 있는 SPIDRE를* 사용하였고 ergodic 은닉 마코프 모델을 이용하여 본질 독립 화자식별 시스템을 구성하였다. 학습음성의 특징과 실험음성의 특징이 다른 조건에서 기존의 채널 정규화 방법들에 비해 10~15%이상 인식률이 향상되었다.

1 서론

전화음성은 대역폭의 제한, 핸드셋의 특성, 채널의 특성 변이, 배경잡음 등으로 인해 고품질 음성의 경우보다 훨씬 인식하기가 어렵다. 전화음성의 여러 가지 잡음 요인들 중에서도 학습환경과 실험환경의 채널 주파수 특성의 차이가 인식률을 가장 크게 떨어뜨리는 요인이다. 채널의 컨볼루션 잡음을 처리하는 방법들이 많이 연구되어 왔으나 채널의 특성을 제거하는 과정에서 화자의 특성까지도 손실되기 때문에 화자인식에 있어서는 좋은 성능들을 얻지 못했다[2].

본 연구에서는 독립성분분석을 이용하여 입력음성의 특징벡터에서 채널특성을 제거하고, 화자의 특징공간의 분별력을 향상시키는 효과적인 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 전화음성의 화자인식에서 해결해야 할 전화음성의 왜곡과정에 대하여 살펴보고, 기존에 연구되어 있는 채널 왜곡 추정 및 정규화 방법들에 대해 간단히 설명한다. 3장에서는 본 연구에서 사용된 독립성분분석의 정의와 그 방법을 기술한다. 4장에서는 기존의 방법들과 독립성분분석에 의한 제안된 방법을 비교실험하고, 5장에서 결론을 맺는다.

*SPeaker IDentification REsearch corpus. NIST Speech Discs 18-1.1 and 18-2.1. LDC, 1994.

2 기존의 채널왜곡 보상 방법

2.1 전화음성의 특징

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

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

전화음성의 스펙트럼은 그림 1과 같은 과정을 거쳐 식 1과 같이 왜곡된다. $G(\omega)$, $H(\omega)$, $M(\omega)$, $T_i(\omega)$는 각각 성대의 기본진동, 성도의 특성, 채집 마이크의 특성, 전송선의 특징을 나타내는 전달함수들이다. 켑스트럼 영역에서는 이러한 필터함수들의 곱이 선형적인 합으로 나타나게 된다.

$$\begin{aligned} \mathbf{c}[n] &= FFT^{-1}\big(\log S(\omega)\big) \\ &= FFT^{-1}\big(\log H(\omega)M(\omega)\textstyle\prod_{i=1}^{r} T_i(\omega)\big) \\ &= \mathbf{h}[n] + \mathbf{m}[n] + \textstyle\sum_{i=1}^{r} \mathbf{t}_i[n] \qquad (2) \\ &= \textstyle\sum_{i=1}^{N} \mathbf{f}_i[n] \qquad (3) \end{aligned}$$

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

전화망의 컨볼루션 왜곡을 나타내는 전달함수 $\mathbf{t}_i[n]$은 한번 연결된 통화에 대해서는 거의 불변하며 매 통화마다 바뀐다고

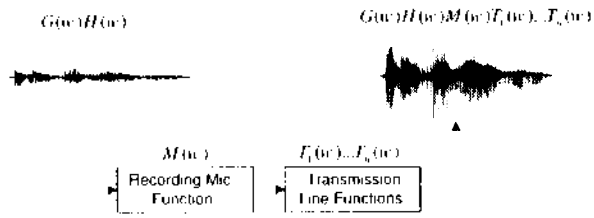

그림 1: 전화 음성의 왜곡 과정.

알려져 있다[3]. 따라서, 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 일치시켜주는 방법으로 크게 분류된다. 전자의 대표적인 방법으로는 켑스트럼 평균 차감법(CMS: cepstral mean subtraction)[3], 후자로는 최대우도 추정법에 의한 신호 편차 제거(SBR: signal bias removal)[4] 등이 있다.

이외 같은 상수 채널편차 제거법 이외에도 켑스트럼 영역을 선형변환(affine transform)시켜 채널 조건에 강인한 새로운 영역의 켑스트럼을 인식기에 사용하는 방법이 있다[5]. 이 방법은 기존의 채널차감법들을 일반화시킨 형태로 켑스트럼 벡터가 전송선을 통과할 때 선형 변환의 형태로 왜곡된다고 가정하며, 그의 보상과 더불어 벡터 영역에서 분별력이 높아지도록 각 성분에 가중치를 부여한다.

이러한 방법들은 화자의 모델링 방법은 고려하지 않고 전처리 단계에서 채널의 특성을 추정하여 특징 파라미터만을 변환시킨다. 그러므로 주변환경에 대한 정보가 불필요한 장점이 있으나 채널특성을 제거하는 과정에서 음성의 특징까지 손실되기 때문에 어느 정도 한계가 있다.

2.2 켑스트럼 평균 차감법(CMS)

채널의 특성이 한 발화 내에서는 거의 변하지 않는 특성을 이용하여 전체 켑스트럼의 평균을 채널 특성이라고 가정하고 차감하여 학습과 인식에 사용한다.

$$\hat{\mathbf{c}}[n] = \mathbf{c}[n] - \frac{1}{T}\sum_{t=1}^{T}\mathbf{c}[t] \tag{4}$$

계산량이 많지 않은 장점이 있으나 입력음성의 길이가 짧을 경우 추정값의 신뢰도가 떨어지며 채널의 영향이 크지 않을 때 오히려 성능저하를 유발한다.

2.3 신호 편차 제거(SBR)

최대우도 추정법(maximum likelihood estimation)에 의해 인식기의 학습음성과 가장 특징 벡터의 분포가 유사하도록 입력 벡터에서 상수 채널 편차를 추정하고 이를 차감한다. 실제로 전화음성 이산 HMM 인식기에 적용되어 높은 성능 향상을 보였다[4].

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

$$\mathbf{b}^* = \arg\min_{\mathbf{b}} \frac{1}{T}\sum_{t=1}^{T}\left\|(\mathbf{c}[t]-\mathbf{b}) - z(\mathbf{c}[t]-\mathbf{b})\right\|, \tag{5}$$

최대우도 추정법에 의해 반복식으로 채널왜곡 추정값 $\hat{\mathbf{b}}$를 찾는다.

$$\begin{aligned}
\mathbf{b}_{i+1} &= \frac{1}{T}\sum_{t=1}^{T}\left(\hat{\mathbf{c}}_i[t] - z(\hat{\mathbf{c}}_i[t])\right) \\
\hat{\mathbf{c}}_{i+1}[n] &= \hat{\mathbf{c}}_i[n] - \mathbf{b}_{i+1} \\
\hat{\mathbf{b}} &= \mathbf{b}_k, \quad \text{when } \|\mathbf{b}_k - \mathbf{b}_{k-1}\| < \epsilon
\end{aligned} \tag{6}$$

학습자료에 가장 적합한 채널함수를 추정하므로 입력음성의 길이에 의한 영향이 적다. 하지만, 채널왜곡을 입력음성으로부터 반복적인 방법으로 추정하기 때문에 입력음성의 왜곡이 클 경우 정확한 채널왜곡의 추정이 어렵다.

2.4 켑스트럼 영역의 선형변환

켑스트럼 벡터 영역에서 선형변환된 새로운 켑스트럼을 학습과 인식에 사용한다. 그때의 거리는 선형변환 행렬 $\mathbf{W}$에 의해 정의된다.

$$\mathbf{c}_i' = \mathbf{W}\mathbf{c}_i,\ \mathbf{c}_j' = \mathbf{W}\mathbf{c}_j,\ \Delta\mathbf{c}' = \mathbf{W}\Delta\mathbf{c}$$

$$\begin{aligned}
\left\|\mathbf{c}_i' - \mathbf{c}_j'\right\|_{\mathbf{W}} &= \left\{\mathbf{W}(\mathbf{c}_i - \mathbf{c}_j)\right\}^2 \\
&= \Delta\mathbf{c}^T\mathbf{W}^T\mathbf{W}\Delta\mathbf{c}
\end{aligned} \tag{7}$$

Zhang[5]은 음성의 포만트를 나타내는 선형예측 계수의 극점은 채널 변화에 강인하다는 특성을 이용하여 선형예측 계수에서 얻은 켑스트럼 벡터의 선형변환 행렬을 구하였다. 그러나 일반적인 전화음성 왜곡에 대해 강인함을 가지는 선형변환 행렬을 찾는 효과적인 방법이 알려져 있지 않다.

3 독립성분분석

독립성분분석(ICA)이란, 선형적으로 혼합된 둘 이상의 신호들을 서로 독립적인 신호들로 분리하는 방법을 통칭한다[8]. 선형적으로 혼합된 신호를 분리하는 다른 방법으로는 주성분분석(PCA: principal component analysis)이 있는데 ICA 사이와 구분되는 것은 신호들간의 상관도(correlation)뿐만이 아니라 의존성(dependence)까지 최소가 되도록 신호들을 분리한다는 점이다. 따라서, ICA는 PCA를 일반화시킨 방법으로 볼 수 있다. ICA는 음원 제거 및 분리, 음질 개선 등의 분야에 응용되었으며, 특히 칵테일 파티 문제(cocktail party problem)의 해결에 이용되어 좋은 결과를 얻었다[6].

3.1 정의

n개의 독립적인 신호 $s_1[t], s_2[t], \ldots, s_n[t]$ 들이 선형적으로 혼합된 m개의 신호 $x_1[t], x_2[t], \ldots, x_m[t]$가 있을 때 독립성분분석의 목표는 혼합된 m개의 신호들을 가지고 n개의 원신호들을 손실없이 복원하는 것이다. 이때 원신호에 대하여 주어지는 가정은 서로 통계적으로 독립(statistically independent)이라는 사실만이 주어진다. 하지만, 원신호를 완전히 복원하는 것은 불가능하며 단지 통계적으로 독립인 n개의 신호들만을 생성해 낼 수 있다.

원신호에의 혼합은 선형 시불변(linear time invariant) 시스템에 의해 이루어지며 혼합된 신호에는 임의의 가산잡음이

된다고 가정한다. 그러면 이 시스템은 다음과 같이 행렬과 벡터의 식으로 표현할 수 있다.

$$\mathbf{x}[n] = \mathbf{A}\mathbf{s}[n] + \mathbf{e}[n] + \mathbf{b} \quad (8)$$

$$\mathbf{y}[n] = \mathbf{W}(\mathbf{x}[n] - \mathbf{b}) \quad (9)$$

$\mathbf{s}$는 혼합되기 전의 원신호이고, $\mathbf{x}$는 혼합된 신호이다. 잡음 $\mathbf{e}$와 $\mathbf{b}$는 각각 가우시안 백색 잡음과 전체 신호에 걸쳐진 상수 벡터이다. 독립성분분석은 관찰된 신호들로부터 n개 원신호의 복원해내는 선형 시불변 분리 시스템을 나타내는 행렬 $\mathbf{W}$를 찾아내는 과정이라 할 수 있다. 또한, 결과로 얻은 독립신호 $\mathbf{y}$는 각각의 성분들이 독립적이어야 한다. 즉, $p(y_i[t], y_j[t]) = p(y_i[t])p(y_j[t])$, $\forall i \neq j$를 만족하도록 하는 행렬을 찾는다.

3.2 정보이론에 기반한 독립성분분석

확률 밀도함수 $f(\cdot)$로 따르는 벡터 $\mathbf{y}$의 엔트로피는 다음과 같이 정의된다

$$H(\mathbf{y}) = -\int f(\mathbf{y}) \log f(\mathbf{y}) d\mathbf{y} \quad (10)$$

엔트로피는 확률변수에 존재하는 정보의 양을 의미한다. 즉, $H(\mathbf{y})$는 $\mathbf{y}$의 분포를 기술하는데 필요한 정보의 양을 의미한다.

각 변수들의 엔트로피로부터 전체 변수들간의 의존성을 나타내는 적도인 상호정보를 정의할 수 있다. 벡터 $\mathbf{y}$의 각 성분들간의 상호정보(mutual information)의 양 I는 다음과 같다

$$I(y_1, y_2, \ldots, y_m) = \sum_{i=1}^{m} H(y_i) - H(\mathbf{y}) \quad (11)$$

상호정보는 확률변수간의 관계를 표현할 수 있는 척도이다. 모든 변수들이 통계적으로 관계가 없을 때 개개의 엔트로피의 합과 전체 엔트로피의 합이 같아지므로, 상호정보의 값은 0이 된다. 즉, 확률 벡터의 상호정보가 작아질수록 각 성분들의 독립성은 커지게 되므로 $I(\mathbf{y})$가 주어진 자료에 대하여 최소가 되도록 변환 행렬 $\mathbf{W}$를 구하면 최적의 변환행렬의 독립성분을 얻을 수 있다[8].

3.3 독립성분분석에 의한 벡터공간 변환

음성의 특징을 나타내는 전달함수들은 서로 구분이 명확하지 않으며 그 수가 고정적이지 않다. 또한, 캡스트럼에 포함되는 채널 전달함수들은 한 동회 내에서는 고정적이지만 서로 다른 동회 사이에서는 연관관계가 없다. 따라서, 다음과 같이 두가지 가정을 할 수 있다

가정 1 p차 캡스트럼은 성도와 특성을 표현하는 p개의 통계적으로 독립적인 전달함수들의 선형결합이다.

가정 2 캡스트럼에 포함된 채널 특성들은 가우시안 분포를 따르는 잡음의 성질을 지닌다.

위의 가정과 식 2, 3에 따라 캡스트럼을 분석 차수와 같은 p개의 특징 함수들의 선형적인 합으로 표현한다. 채널 간섭함수는 식 8에서 가우시안 분포를 따르는 잡음으로 간주한다. 그리고, 여러 가지 채널 특성이 포함된 일반적인 음화음성의 캡스트럼으로부터 독립성분분석을 통해 새로운 p차 독립 캡스트럼 영역으로 변환하는 행렬 $\mathbf{W}_A$를 찾는다.

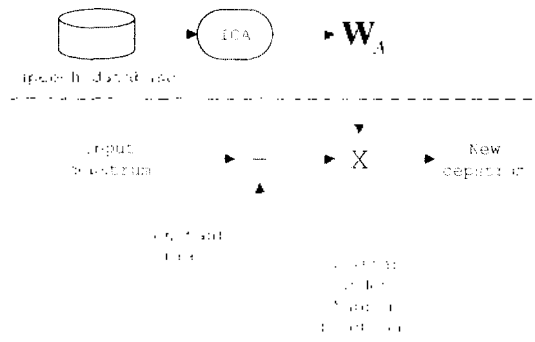

그림 2: ICA에 의한 캡스트럼 변환

$$\mathbf{c}[n] = \mathbf{A}\mathbf{c}_I[n] + \mathbf{t}[n] + \mathbf{b}$$

$$\hat{\mathbf{c}}_I[n] = \mathbf{W}_A(\mathbf{c}[n] - \mathbf{b}) \quad (12)$$

$\mathbf{W}_A$에 의해 선형변환된 새로운 캡스트럼 영역에서의 처리 식 7에서 얻을 수 있다. 독립성분분석과정에서 채널의 특징은 평균이 0인 가우시안 잡음 $\mathbf{t}[n]$으로 간주되어 억제되고, 구분이 모호한 음성의 전달함수들은 서로 독립적인 성분들에 더 분리된다. 따라서, 일반적인 전화음성에서 독립성분 분석 행렬로 얻은 선형변환 캡스트럼 $\hat{\mathbf{c}}_I[n]$은 채널 면이에 강인하고 벡터공간의 분별력이 높아진다.

2장에서 소개한 CMS, SBR과 새로운 캡스트럼 $\hat{\mathbf{c}}_I[n]$은 2.1절의 선형변환 방법의 특별한 형태들이며, 변환 행렬들과 상수벡터 추정방법들에 의해 구분된다. 기존의 채널 정규화 방법들은 음성신호의 특징을 이용하여 채널신호를 억제하고 음성신호를 강조한다. 본 장에서 제안한 방법은 통계적인 관점에서 벡터 공간의 분별력을 높이고 채널신호를 억제하였다.

4 결 과

4.1 실험환경

실험에 사용된 자료는 상업적 전화음성 데이터 베이스인 SPIDRE의 특성 추출 단계에서는 인간의 청각특성을 반영한 13차 멜단위 캡스트럼 벡터를 사용하였다. 인식시스템은 상태마다 32개씩 가지고 각 혼합가우시안을 중력확률로 계산하는 HMVQM[1]으로 구현하였으며, 음장자유형을 지향하여 에르고딕 위상구조(ergodic topology)로 구성하였다. 또한, 상태수에 따른 성능의 변화를 알아보기 위해 1,2,4,8 네가지로 나누어 비교실험하였다. HMVQM의 학습은 30초, 인식은 10초 단위로 하였다.

실험은 네가지로 나누어 진행하였다. 먼저, 채널 정규화 방법을 적용하지 않은 기본 시스템과, 2절에서 소개한 기존의 채널 정규화 방법인 CMS와 SBR, 3.3절에서 제안한 ICA를 이용한 특징 파라미터 변환방법을 사용하여 비교실험하였다. ICA 행렬은 각 화자의 학습에 사용한 84(42명 × 2)개 음성으로부터 약 10초의 입력로 산출하여 캡스트럼의 분석차수와 같은 13차로 구하였다. 구하는 방법은 3.2절의 상호정보를 최소화하는 Hyvärinen의 고정점수정 알고리즘[7]을 따랐다.

표 1: 동일 채널 조건 결과

실험	상태수 1	상태수 2	상태수 4	상태수 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%
ICA	83.3%	89.7%	88.9%	90.5%

표 2: 상이 채널 조건 결과

실험	상태수 1	상태수 2	상태수 4	상태수 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%
ICA	56.4%	59.5%	62.7%	66.7%

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.

4.2 실험결과

표 1과 표 2는 각각 학습과 인식환경을 같게 했을 때와 다르게 했을 때의 결과들을 나타낸다. 거의 모든 경우에 있어서 ICA방법이 기존의 채널 정규화 방법들에 비하여 좋은 성능을 보였다. 상이채널의 경우 기존의 CMS, SBR보다 상태수 8에서 11~15% 정도의 인식률 향상을 보였고 나머지에서도 10%이상의 성능향상을 보였다. 따라서, 본 논문에서 제안된 ICA방법이 기존의 채널 정규화 방법들보다 채널 변이에 강인함을 알 수 있다.

동일채널의 경우도 상이채널의 경우보다 향상폭은 적지만 거의 모두 5%이상씩 인식률이 향상되었다. 특히, CMS나 SBR의 경우 정규화의 영향으로 기본 시스템보다 성능이 떨어지지만 ICA는 오히려 더 좋은 성능을 보였다. 즉, CMS나 SBR의 경우 정규화에 의해 화자정보의 손실이 크지만 ICA의 경우에는 그 정보손실이 크지 않으며 오히려 캡스트럼의 각 차수간의 구분을 더 크게 하였기 때문에 정규화에 의한 영향을 감소시킬 수 있었다.

5 결론

본 연구에서는 전화음성하에서 강인한 화자식별기의 구현을 위해 독립성분분석을 특징 파라미터에 적용할 수 있는 방법을 고안하였다. 또한, 기존의 채널 정규화 방법들에 비해 10~15% 정도 인식률이 향상됨으로 해서 제안한 방법이 채널환경 변화에 대해 보다 우수한 강인성을 가짐을 보였다. 현재 벡터 공간에서의 분별능력을 좀 더 높이는 방법과 독립성분분석 자체를 화자 모델링에 활용하는 방법에 대한 연구를 진행중이다.

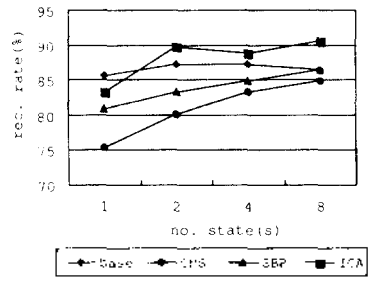

그림 3: 동일 채널 조건

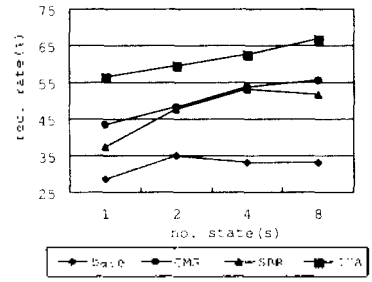

그림 4: 상이 채널 조건

참고 문헌

[1] 윤성진, 적은 학습자료 환경하에서 화자인식 시스템의 성능향상에 관한 연구, 한국과학기술원 전산학과 석사학위논문, 1994.

[2] D.A. Reynolds et al., "The effects of telephone transmission degradations on speaker recognition performance," *Proceedings of ICASSP*, pp. 329-332, 1995.

[3] A.E. Rogenberg, C.-H. Lee, and F.K. Soong, "Cepstral channel normalization techniques for HMM based speaker verification," *Proc. of ICSLP* , pp. 1835-1838, Yokohama, 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, No. 1, pp. 16-30, Jan. 1996.

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

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

[7] A.Hyvärinen, "A family of fixed-point algorithms for independent component analysis," *Proceedings of ICASSP*, pp. 3917-3920, 1997.

[8] A.Hyvärinen, *Independent component analysis by minimization of mutual information*, Technical Report, Helsinki University of Technology, 1997.