Hidden LMS 적응 필터링 알고리즘을 이용한 경쟁학습 화자검증

論 文
51D-2-5

Speaker Verification Using Hidden LMS Adaptive Filtering Algorithm and Competitive Learning Neural Network

趙 成 元* · 金 載 敏**
(Cho, Seongwon · Kim, Jaemin)

Abstract - Speaker verification can be classified in two categories, text-dependent speaker verification and text-independent speaker verification. In this paper, we discuss text-dependent speaker verification. Text-dependent speaker verification system determines whether the sound characteristics of the speaker are equal to those of the specific person or not. In this paper we obtain the speaker data using a sound card in various noisy conditions, apply a new Hidden LMS (Least Mean Square) adaptive algorithm to it, and extract LPC (Linear Predictive Coding)-cepstrum coefficients as feature vectors. Finally, we use a competitive learning neural network for speaker verification. The proposed hidden LMS adaptive filter using a neural network reduces noise and enhances features in various noisy conditions. We construct a separate neural network for each speaker, which makes it unnecessary to train the whole network for a new added speaker and makes the system expansion easy. We experimentally prove that the proposed method improves the speaker verification performance.

Key Words :speaker verification, Hidden LMS adaptive algorithm, neural network

1. 서 론

현대 사회의 정보화가 빠르게 진행되면서 많은 정보들이 집중되고 이것을 관리하는데 보안의 문제가 심각하게 대두되고 있다. 이를 위해서 다양한 방법의 출입관리 시스템이 개발되고 있다. 카드나 열쇠 등은 분실할 경우에 타인에 의해서 악용되어 큰 불이익을 초래할 수 있다는 단점이 있어서 고도의 보안이 필요한 곳에 대한 충분한 해결책이 되지 못한다는 문제가 있다. 이에 대한 해결책으로서 사람 개개인이 가지는 고유한 특징을 이용하는 보안 방법이 많이 연구되어 왔다. 이를 위해서는 개개인의 특징으로 어떤 것을 사용할 것인가와 인식을 위해 어떤 방법을 이용할 것인가, 그리고 그에 따른 비용, 신뢰성 및 사용자에게 친근감을 줄 수 있는가 등을 고려하여야 한다. 이중 최근에 들어서 많이 연구되고 있고, 실현이 가능한 방법으로는 개인의 지문, 음성, 얼굴, 사인, 홍채 등이 있다[1]. 물론 아직까지는 개개인의 특징을 이용하는 방법이 독자적으로 기존의 보안 문제를 해결할 수는 없지만 기존의 방법과 병행하여 쓰인다면 좀 더 높은 신뢰성을 보장하는 보안방법으로서 이용될 수 있을 것으로 보이며, 최근 이에 대해 많은 연구가 이루어지고 있다. 본 논문에서는 위에서 기술한 방법들 중에서 사람마다 성도(vocal apparatus)의 구조가 다른데서 기인한 개인별 음성 파형의 물리적인 특징을 이용한 화자검증 방법에 대해서 논의한다.

화자검증시스템은 검증하고자 하는 화자에 대해서 주장하는 신원과 음성을 받아들이고 주장하는 신원에 대해서 음성 데이터를 적절히 비교하여 해당화자가 맞는지를 검증하는 것이다. 현재까지 연구되어온 화자검증 기법에는 패턴정합(pattern matching)방법과 통계적 성질을 이용한 분류방법이 있었다[2]. 패턴 정합의 경우에는, 먼저 각 화자를 대표할 수 있는 기준패턴을 미리 작성한 다음, 시험패턴과 기준패턴 사이의 유사도를 측정하여 시험패턴의 유사도에 따라서 시험패턴의 신원을 확인한다. 통계적 성질을 이용한 분류방법은 각 화자에서 추출한 요소들을 오랜 시간 동안 관찰하여 통계량을 구한 후 이것으로 화자의 신원을 확인하는 방법이다. 패턴정합 방법이 통계적 분류 방법보다 계산량은 많지만 인식률이 높기 때문에 더 많은 연구가 진행되어 왔다.

이러한 화자검증의 인식률을 향상시키기 위한 노력으로는 크게 두 가지 관점에서 연구가 진행되어왔다. 첫째는 각 화자의 음성으로부터 추출해 내는 특징에 대해 좀 더 화자의 특징을 잘 나타낼 수 있는 정보를 추출하려는 연구[3]와, 둘째로는 추출된 특징정보를 이용하여 적절한 판단을 내리는 인식부에 대한 연구가 진행되어 왔다.

본 논문에서는 특징 추출의 성능을 향상시키기 위해 전처리 과정에서 LMS 필터링 알고리즘에 신경망 구조를 적용시켜 제안한 Hidden LMS 적응필터링 알고리즘을 사용하고, 화자의 특징으로는 가장 나은 성능을 보인다고 알려져 있는 cepstrum계수를 추출하여 사용한다. 화자검증의 판단을 내리기 위한 인식부로는 비지도 학습 신경회로망의 대표적인 알고리즘인 SCL (simple competitive learning)의 학습방정식을

* 正 會 員 : 弘益大 電氣電子工學部 副敎授 · 工博
** 正 會 員 : 弘益大 電氣電子工學部 助敎授 · 工博
接受日字 : 2001年 12月 15日
最終完了 : 2002年 1月 12日

이용하는데, 전체 신경회로망의 구조는 검증하고자 하는 화자마다 하나씩의 신경회로망을 가지는 구조를 사용하고, 실험을 통해 화자검증에 대해서 본 논문에서 제안하는 방법의 타당성을 보인다.

2. 기존관련연구

화자인식의 인식률을 향상시키기 위한 노력으로 각 화자의 음성을 입력받는 것에서부터, 특징을 추출하기까지의 과정인 전처리 과정을 통해서 좀 더 화자의 특징을 잘 나타낼 수 있는 정보를 추출하려는 연구가 진행되어왔다. 특히, 잡음이 있을 때 화자의 발음이 조용한 환경에서 발음한 것과 다르고 음성인식 시스템의 입력으로 잡음과 음성이 동시에 들어가기 때문에 높은 성능의 시스템을 구현하기 어려울 수밖에 없다. 음성인식 시스템의 전처리 과정으로서 사용될 수 있는 잡음제거 방법으로는 스펙트럼 차감법, 자기상관 차감법, 적응 잡음 제거법, 음향빔 형성법이 있다. 스펙트럼 차감법은 주파수 영역에서 측정된 오차신호의 스펙트럼을 소거함으로써 음성신호의 스펙트럼을 복원하여 음질을 향상시킬 수 있는 방법이다. 이 방법은 잡음제거방법으로 널리 알려져 있고, 성능도 우수하지만, 잡음의 스펙트럼형태를 미리 알고 있거나 잡음의 스펙트럼을 추정하기에 충분한 묵음구간이 주어져야 하고, 잡음이 최소한 부분적으로 정적인 특성을 가져야 하며 통계적 특성이 서서히 변화하는 환경에서는 음성이 존재하는 구간과 잡음만이 존재하는 구간을 검출할 수 있는 방법이 필요 하는 등 사전 정보를 확보해야 하는 어려움이 있다. 자기상관 차감법은 잡음이 섞인 신호의 자기 상관행렬에서 추정된 잡음신호의 자기 상관행렬을 빼줌으로써 잡음을 제거하는 방법이다. 이 방법은 자기 상관행렬을 이용하기 때문에 이 행렬로부터 쉽게 선형예측 계수들을 구할 수 있으므로 음성 인식 시스템 등에 쉽게 적용 할 수 있지만, 잡음추정에 대한 적절한 방법이 없이는 뚜렷한 성능을 보이기 힘들다. 적응 잡음제거법은 잡음 경로를 통해서 들어온 신호를 적응필터를 통해서 추정하여 원 음성신호에 내재된 잡음신호를 제거하는 방법이다. 적응 필터의 형태는 자동적으로 필터 인자들을 조절하는 능력이 있으므로, 신호와 잡음에 대한 아무런 사전 정보가 필요치 않다. 특히, 입력 신호의 특성을 모르거나 그 신호가 시변 신호일 때 가장 유용한 기술이 된다[4][5]. 음향빔 형성법은 특정방향의 신호대잡음비가 높은 잡음환경에서 신호의 방향정보를 이용하여 잡음을 제거하는 방법이다. 화자인식을 위해 음성신호로부터 추출되는 특징으로는 음성의 피치, short-time energy, PARCOR 계수, LPC계수, LPC계수에서 유도되는 Cepstrum계수 등이 사용되며, 이러한 요소(parameter)는 각 화자에 따라서 성도의 구조가 다르므로 화자의 특징으로서 의미를 지닌다고 알려져 있다. 기존의 연구결과에 의하면 LPC-Cepstrum계수가 화자인식에 가장 뛰어난 성능을 보인다고 알려져 있다[6][7]. 본 논문에서는 전처리 과정으로 적응 잡음제거법과 LPC-Cepstrum계수를 사용하여 심각한 잡음환경하에서 검증하고자 하는 각 화자의 음성 특징을 추출하고, 인식부로는 미리 기준화자의 음성특징으로부터 기준패턴을 작성하고, 이후 검증하고자 하는 화자의 음성과 신원을 입력으로 받아들여서 해당되는 화자의 기준 패턴과의 유클리디안 거리를 구하여 정해진 임계치 값에 의해서 판단을 내리는 방법을 사용했다[8][9].

3. 잡음제거

화자인식기는 사용 환경상의 소음 문제가 큰 약점으로 작용하게 된다. 따라서 적절한 필터를 적용시킴으로써 화자인식기의 성능을 향상시킬 수 있다.

잡음제거의 형태에는 고정 형태와 적응 형태가 있는데, 고정 형태는 사전에 신호와 잡음의 특성에 대한 정보를 알고있어야 되는 반면, 적응 형태는 자동적으로 필터 인자들을 조절하는 능력이 있으므로, 신호와 잡음에 대한 아무런 사전 정보가 필요치 않다. 특히, 입력 신호의 특성을 모르거나 그 신호가 시변 신호일 때 가장 유용한 기술이 된다. LMS 알고리즘은 필터 인자를 찾아내는데 이용되며 그 식의 간편함과 구현의 용이함으로 가장 널리 이용되어 왔다.

본 논문에서는 적응 잡음제거 방법에서 적응필터의 인자들을 조절하는데 Hidden LMS 알고리즘을 제안하여 사용하였다.

3.1 LMS 적응 필터

기존의 LMS 적응 필터 구조는 그림 3.1과 같고, 그 중 적응 시스템의 주요한 요소인 적응 선형 합성기를 그림 3.2과 같이 나타냈다 [4].

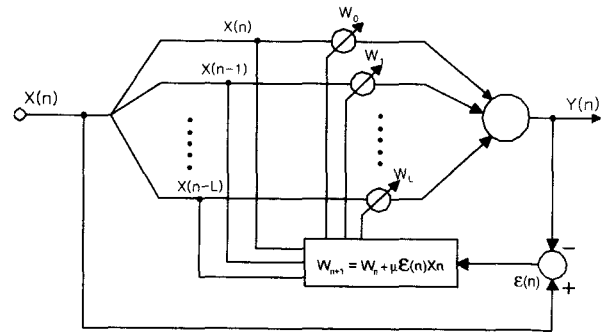


그림 3.1 LMS 적응 필터 구조

Fig. 3.1 Structure of LMS adaptive filter

입력 신호는 X(n)의 지연된 신호로 다음과 같이 벡터 X를 정의한다.

$$\mathbf{X} = [X_0, X_1, X_2, \cdots\cdots, X_L]^T \qquad (식3.1)$$

가중 계수(weight) $w_0, w_1, w_2, \cdots\cdots, w_L$는 가변 조정이 가능하고, 가중 계수 벡터와 출력은 각각 (식3.2)와 (식3.3)과 같다.

$$W = [w_0, w_1, w_2, \cdots\cdots, w_L]^T \qquad (식3.2)$$

$$y = X^T W = W^T X \qquad (식3.3)$$

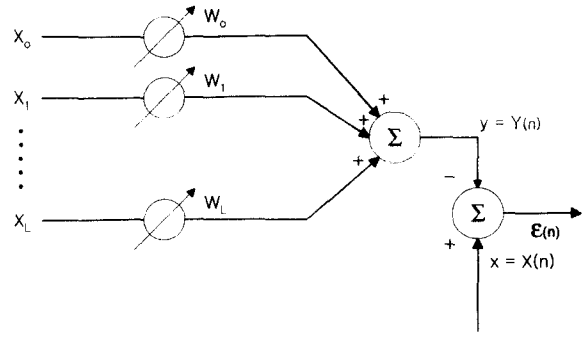


그림 3.2 적응 선형 합성기
Fig. 3.2 Adaptive linear combinator

오차 ε은 입력 x와 출력 y와의 차로서 정의된다.

$$\varepsilon = x - X^T W = x - W^T X \qquad (식3.4)$$

앞에서 설명한 바와 같이 $E[\varepsilon^2]$의 값을 최소로 만드는 W의 값을 추정하는 알고리즘이 필요하며, 대표적인 것이 LMS 알고리즘이다. 적응 알고리즘의 목적은 (식3.4)의 평균 제곱 오차(Mean Square Error : MSE)를 최소화하기 위한 가중계수를 조정하는 것이다.

(식3.4)에 제곱을 취하여 전개하면 (식3.5)와 같이 나타낼 수 있다.

$$\varepsilon^2 = x^2 + W^T X X^T W - 2xX^T W \qquad (식3.5)$$

여기서 MSE가 최소가 되는 W를 구하기만 하면 되므로 직접 MSE를 구할 필요는 없다. 따라서, (식3.5)를 가중계수 벡터 W 성분의 2차 함수로 나타냈을 때, MSE가 최소가 되는 가중계수는 그림 3.3과 같이 ε^2 곡선의 접점 기울기가 0이 되도록 조절함으로써 구할 수 있다.

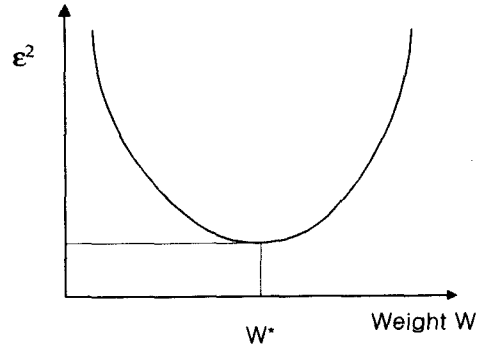


그림 3.3 W의 2차 함수 성능곡선
Fig. 3.3 The performance curve of W

LMS 알고리즘에서의 차기 가중계수 W_{n+1}은 다음과 같다.

$$W_{n+1} = W_n - \mu \nabla E[\varepsilon^2_n] \qquad (식3.6)$$

여기서 μ는 안정도와 수렴률을 결정하는 요소로서 다음과 같은 조건을 만족해야 한다.

$$0 < \mu < \frac{1}{\lambda_{max}} \qquad (식3.7)$$

여기서 λmax는 XX^T의 최대 고유치(eigenvalue)이다. 실제 기울기 $\nabla E[\varepsilon^2]$는 추정된 기울기 $\hat{\nabla}$로 근사화시킬 수 있다.

$$\hat{\nabla} = \begin{pmatrix} \frac{\partial \varepsilon^2}{\partial w_0} \\ \vdots \\ \frac{\partial \varepsilon^2}{\partial w_L} \end{pmatrix} = 2\varepsilon \begin{pmatrix} \frac{\partial \varepsilon}{\partial w_0} \\ \vdots \\ \frac{\partial \varepsilon}{\partial w_L} \end{pmatrix} = -2\varepsilon X \qquad (식3.8)$$

따라서 Widrow-Hoff LMS 알고리즘은 다음과 같다.

$$W_{n+1} = W_n - \mu \nabla$$
$$= W_n + 2\mu \varepsilon X \qquad (식3.9)$$

이러한 LMS 알고리즘은 단순성과 효율성 때문에 많이 사용된다.[10]

3.2 Hidden LMS 적응 필터

그림 3.1과 같은 기존의 LMS 필터구조에 Backpropagation 신경회로망과 같은 Hidden Layer 구조를 도입하여 제안한 구조가 그림 3.4의 Hidden LMS 적응 필터 구조이다. 제안한 Hidden LMS 적응 필터는 크게 필터 중간에 위치해 있는 Hidden part와 출력부 측에 위치한 Out part로 나눠지게 된다.

그림 3.4에서의 Out part의 가중 계수 계산은 아래와 같이 유도 할 수 있다. $E[\varepsilon(n)^2]$는 기울기 계산의 편의를 위해 (식3.10)과 같이 정의하여 계산한다.

$$E[\varepsilon(n)^2] \simeq \frac{1}{2}(X(n) - W_y{}^T H_n)^2 \qquad (식3.10)$$

앞에서 설명했던 바와 같이 가중 계수 식은 (식3.11)과 같으며, Out part의 출력은 (식3.12)와 같다.

$$W_{n+1} = W_n - \mu \nabla E[\varepsilon(n)^2] \qquad (식3.11)$$
$$Y(n) = W_y{}^T H_n \qquad (식3.12)$$

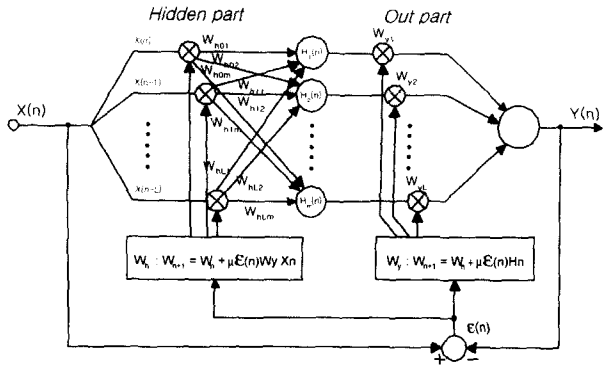


그림 3.4 Hidden LMS 적응 필터 구조
Fig. 3.4 Structure of Hidden LMS adpative filter

$E[\varepsilon(n)^2]$을 Wy에 대해 편미분을 하면 다음과 같이 쓸 수 있다.

$$\nabla E[\varepsilon(n)^2] = (W_y{}^T H_n - X(n)) \cdot H_n \qquad (식3.13)$$

따라서 Out part의 계수 Wy는 (식3.14)에 의해 갱신하게 된다.

$$W_y : W_{n+1} = W_n + \mu(X(n) - W_y^T H_n) \cdot H_n$$
$$= W_n + \mu\varepsilon(n) \cdot H_n \qquad (식3.14)$$

Hidden part의 가중 계수 계산은 다음과 같이 유도할 수 있는데, Hidden part의 Node 출력은 (식3.15)와 같다.

$$H_L(n) = W_h^T X_n \qquad (식3.15)$$

$E[\varepsilon(n)^2]$을 Wh에 대해 편미분을 하기 위해 다음과 같은 과정을 거치게 된다.

$$\nabla E[\varepsilon(n)^2] = \frac{\partial E[\varepsilon(n)^2]}{\partial H_n} \cdot \frac{\partial H_n}{\partial W_h^T} \qquad (식3.16)$$

$$\frac{\partial E[\varepsilon(n)^2]}{\partial H_n} = (W_y^T H_n - X(n)) W_y \qquad (식3.17)$$

$$\frac{\partial H_n}{\partial W_h^T} = \frac{\partial W_h^T X_n}{\partial W_h^T} = X_n^T \qquad (식3.18)$$

따라서 Hidden part의 가중 계수 식은 다음과 같이 될 수 있다.

$$\nabla E[\varepsilon(n)^2] = (W_y{}^T H_n - X(n)) W_y \cdot X_n^T \qquad (식3.19)$$

$$W_h : W_{n+1} = W_n + \mu(X(n) - W_y^T H_n) W_y \cdot X_n^T \qquad (식3.20)$$
$$= W_n + \mu\varepsilon(n) W_y \cdot X_n{}^T$$

여기서 X(n)은 입력신호이며, L은 필터 계수의 차수이다. Hidden Part와 Out Part의 필터 계수 W는 Widrow Hoff의 알고리즘을 의하여 위의 식에 의해 한 sample씩 기준으로 갱신된다.

4. 음성신호 전처리 과정

4.1 음성부분 추출

사운드카드를 이용하여 획득한 음성데이터를 30ms씩의 프레임으로 나누었으며(한 프레임당 300샘플), 20ms씩 중첩시키고 각 프레임에서는 Hamming윈도우를 씌어줌으로서 각 프레임의 중간 부분을 강조하면서 처음과 끝 부분에서의 불연속성을 최소화하였다.

묵음 부분을 제거하고 실제 발음이 있는 부분을 뽑아내는 끝점 추출방법으로는 평균에너지법과 영교차율(zero crossing rate)을 이용하였다[11].

유성음의 경우에는 무성음이나 noise에 비해서 큰 에너지 성분을 갖기 때문에 유성음 부분을 추출해내는데 평균 에너지법이 이용된다. 평균에너지를 구하는 식은 (식4.1)과 같다.

$$M_n = \sum_{m=n-N+1}^{n} |x(m)| w(n-m) \qquad (식 4.1)$$

$$w(n) = \begin{cases} 1 & (0 \le n \le N-1) \\ 0 & otherwise \end{cases}$$

여기서 N은 한 프레임에서의 샘플 수를 의미하며, $w(n)$은 해당 구간을 정해주는 윈도우 함수이다. 영교차는 음성신호에서 두 개의 연속되는 샘플의 부호가 다른 경우가 생기는 비율, 즉 샘플링된 음성신호의 값이 얼마나 빈번하게 영을 지나치는가를 나타내는 척도가 된다. 발음이 되는 부분, 즉 음성으로 간주되는 부분은 유성음만을 의미하지는 않는다. 무성음 부분은 에너지가 작기 때문에 평균 에너지 법만으로는 추출할 수가 없다. 따라서 무성음의 경우 유성음과 일반 잡음에 비해 주파수 성분이 크다는 점을 고려하여 분리 해 낼 수가 있다. 음성신호의 주파수가 큰 경우에는 영교차율이 크고, 작은 경우에는 영교차율이 작아진다. 그러므로 무성음인 경우에는 유성음에 비해서 영교차율이 크기 때문에 무성음 부분을 추출해내는데 이용되며, 영교차율을 구하는 식은 (식4.2)와 같다.

$$Z_n = \sum_{m=n-N+1}^{n} |sgn[x(m)] - sgn[x(m-1)]| w(n-m)$$

$$sgn[x(n)] = \begin{cases} 1 & (x(n) \ge 0) \\ -1 & (x(n) < 0) \end{cases} \qquad (식4.2)$$

$$w(n) = \begin{cases} \frac{1}{2N} & (0 \le n \le N-1) \\ 0 & otherwise \end{cases}$$

측정된 영교차율이 이미 결정된 임계치보다 크면 해당 부분에 무성음이 존재한다고 볼 수 있다. 따라서 음성 부분의

추출을 위해서는 먼저 평균 에너지 법으로 유성음 부분을 찾은 후 유성음 부분의 앞쪽으로 마찰음이 있는지 영교차율을 통해서 찾아내어 음성 부분을 추출한다. 그림 4.1은 실제 음성 데이타에 대해서 끝점을 추출한 것이다.

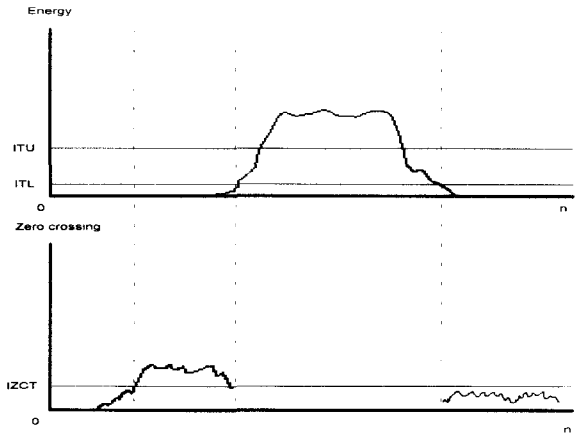


그림 4.1 음성부분 추출

Fig. 4.1 Extraction of speech area

4.2 음성신호의 특징추출

실제로 발음있는 부분만을 추출한 후 autocorrelation analysis와 LPC analysis를 거쳐서 LPC계수를 구하고, 이를 cepstrum계수로 변환하였다. LPC모델의 기본 개념은 시간 n에서의 음성샘플 s(n)이 과거 p개의 음성샘플의 선형조합으로 근사화 될 수 있다는 것으로, 이를 수식적으로 표현하면 (식4.3)과 같다. 단, 계수 $a_1, a_2, \ldots, a_p$는 분석할 프레임에 대해서 일정하다고 가정한다.

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \cdots + a_p s(n-p) \quad (식4.3)$$

(식4.3)을 항등식으로 나타내면 (식4.4)와 같으며, 여기서 u(n)은 normalized excitation이고, G는 excitation의 이득이다.

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \quad (식4.4)$$

LPC모델에 기반한 s(n)과 u(n)의 관계식과 과거음성 샘플들의 선형조합인 예측신호 s(n)과의 예측오차 e(n)은 (식4.5)가 된다.

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \quad (식4.5)$$

이상에서 LPC분석의 기본문제는 (식4.5)의 e(n)을 최소화하는 예측계수 $\{a_k\}$를 정하는 것이며, 시간에 따라 변하는 음성신호의 스펙트럼 특성상, 일정구간, 즉 프레임에 대해서 MSE를 최소화하는 값을 갖도록 해야 한다. 이렇게 구한 LPC계수는 실험적으로 음성인식 및 화자인식을 위한 특징으로서 좀 더 강인하고 신뢰성 있는 것으로 알려져 있는 cepstrum계수 cm으로 변환한다.

$$c_0 = \ln \sigma^2 \quad (식4.6)$$

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \quad 1 \le m \le p \quad (식4.7)$$

$$c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \quad m > p \quad (식4.8)$$

4.3 학습과정

본 논문에서 각 화자별 음소의 특징을 클러스터링하는데 쓰이는 신경회로망의 구조는 그림 4.2와 같이 화자별로 network를 갖는 구조로서, 전체 화자가 M명이라면 총 M개의 신경망이 필요하다. 각 화자별 신경망을 학습시키는 방법은 해당 신경망에 대해서 외부화자의 데이터를 제외하고 기준화자의 학습데이터만을 가지고 학습시키게 되어 결국 신경망이 학습되고 나면 해당 화자별 음속의 특징이 학습된다고 할 수 있다. 학습 알고리즘은 비지도 학습 신경회로망의 대표적인 알고리즘인 SCL을 사용하였고[12][13], 학습과정은 표 4.1과 같다.

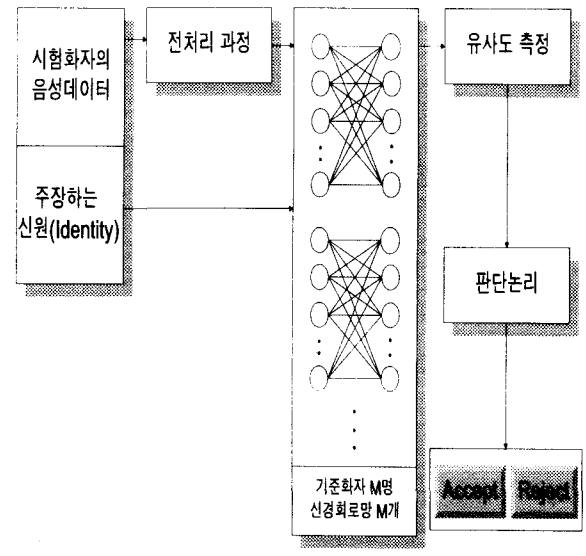


그림 4.2 화자검증 시스템 전체구조

Fig. 4.2 The overal structure of Speaker verification system

step1. 학습과정에 필요한 변수 및 가중치 벡터를 초기화한다. i=j=1.

step2. 해당화자에 대한 i 번째 학습데이터(M*N패턴)를 읽어 들인다.

step3. j번째 프레임의 cepstrum계수를 입력으로 승자뉴런을 찾는다.

step4. 승자가 된 출력 뉴런에 대해서 다음의 학습방정식에 의해 가중치벡터를 갱신한다.

$$w_i(t+1)=\begin{cases} w_i(t)+\alpha(t)\cdot[x(t)-w_i(t)], & \text{if } i=c \\ w_i(t) & \text{if } i\neq c \end{cases}$$

step5. 음성이 총 N개 프레임으로 구성되어 있다면, 1번 프레임에서 N 프레임까지 j를 1씩 증가시키면서 step3~step4를 반복한다.

step6. 해당화자의 학습데이터가 M개라고 하면, 1번째 데이터에서 M번째 데이터까지 i를 1씩 증가시키면서 step2~step5를 반복한다.

step7. 학습의 총 반복횟수(epoch)만큼 step2~step6을 반복한다. i=j=1.

step8. 학습이 끝난 후 가중치벡터를 저장한다.

표 4.2
Table 4.2

위와 같은 과정을 통해서 한 명의 화자에 대한 신경회로망이 학습된다. 본 연구에서는 각 화자가 하나의 신경회로망을 갖는 구조이므로 검증하고자 하는 화자의 수만큼 위의 과정을 반복하여 각 화자별로 가중치벡터를 저장한다. 이후 화자검증의 과정에서 주장되는 화자의 신원에 해당하는 신경회로망의 가중치벡터를 읽어들여서 검증과정을 수행하게 된다.

본 논문에서 화자 검증 과정은 표4.2와 같다.

step1. 검증을 원하는 시험화자의 음성과 주장하는 신원을 받아들인다.

step2. 주장된 신원에 해당하는 화자의 신경회로망에 해당하는 가중치벡터와 임계치값을 읽어들인다.

step3. 전처리 과정을 통해서 시험화자의 음성데이터로부터 음성 특징벡터를 추출한다.

step4. 시험화자의 음성데이터가 총 N프레임으로 구성된다고 할 때, 각 프레임별로 승자뉴런을 찾는다.

$$\|\boldsymbol{x}-\boldsymbol{w}^*\|=\min_i\{\|\boldsymbol{x}-\boldsymbol{w}_i\|\}$$

step5. 각 프레임에 대해서 승자뉴런과의 거리를 구한다.

$$d_i=\sum_{m=1}^{p}(C_m-W_m^*)^2 \qquad i: \text{프레임번호}$$

step6. 전체 프레임에 대해서 거리를 합하고, 총 프레임수로 나누어주어 패턴의 거리를 구한다.

$$D=\frac{1}{N}\sum_{i=1}^{N}d_i \qquad D: \text{패턴과의 거리}$$

step7. 이렇게 구한 패턴간의 거리 D와 임계치값을 비교하여 유사도에 따라 적절한 판단을 내린다.

$$Decision=\begin{cases} Accept & \text{, if } Th\leq D \\ Reject & \text{, if } Th>D \end{cases}$$

Th : 임계치값

표 4.2
Table 4.2

5. 실험결과

본 논문에서는 텍스트 종속 화자검증 실험을 위해 "제어실험실"이란 발음을 총 10명의 화자로부터 20개씩 수집하였다. 음성데이터는 사운드카드를 통해서 11.025kHz(16 bit)로 얻었으며, 각 음성은 한 프레임에 300샘플(200샘플씩 중첩)씩으로 나누었다. 여기서 사용된 Hidden LMS 적응 필터 계수의 차수와 Hidden node의 수는 각각 5 씩이며, 앞에서 기술한 전처리 단계를 거쳐서 10차의 cepstrum계수를 추출하여 신경회로망의 입력으로 사용하였다. 실험에 사용된 잡음은 mean이 0이고, variance가 1인 random noise가 사용되었다.

음성신호의 전처리 과정과 화자검증 과정을 도식적으로 나타내면 각각 그림 5.1과 그림 5.2와 같다.

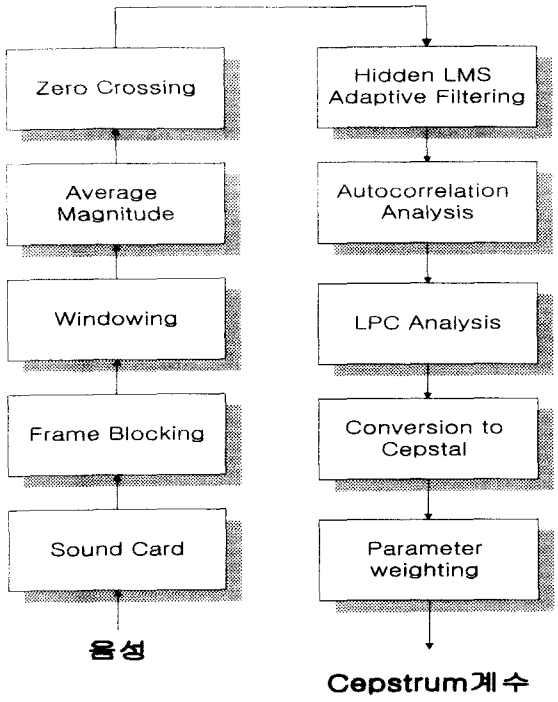


그림 5.1 전처리과정 블록도

Fig. 5.1 The block diagram of the preprocessing procedure

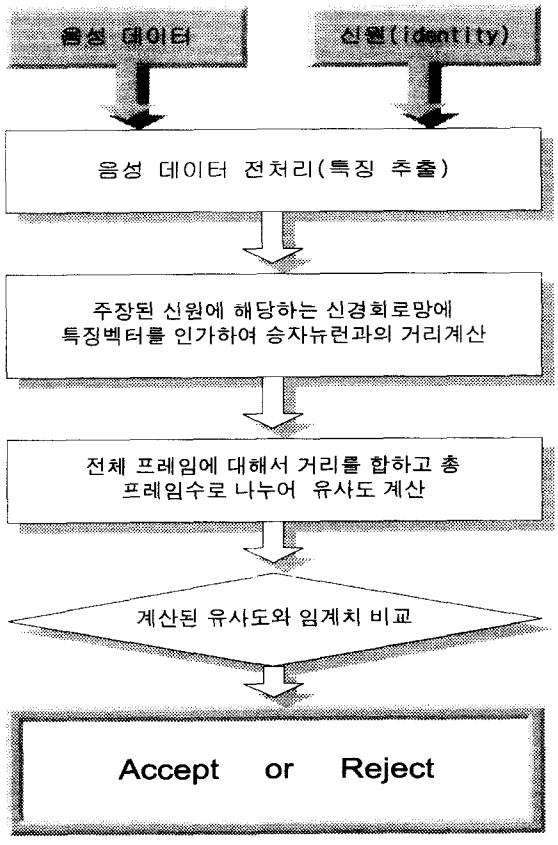


그림 5.2 화자검증 단계

Fig. 5.2 Steps for Speaker verification

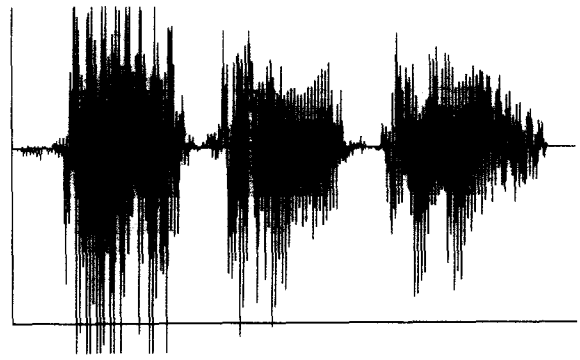


그림 5.3 음성 신호

Fig. 5.3 Speech signal

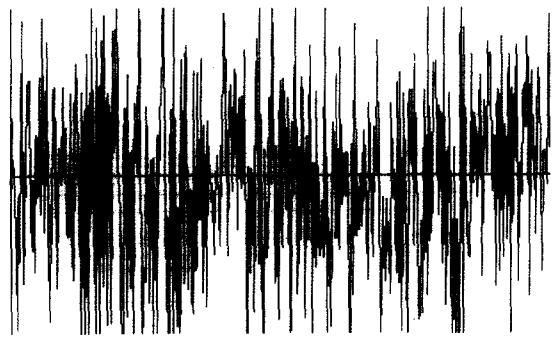


그림 5.4 잡음이 섞인 음성 신호

Fig. 5.4 Speech signal with a noise

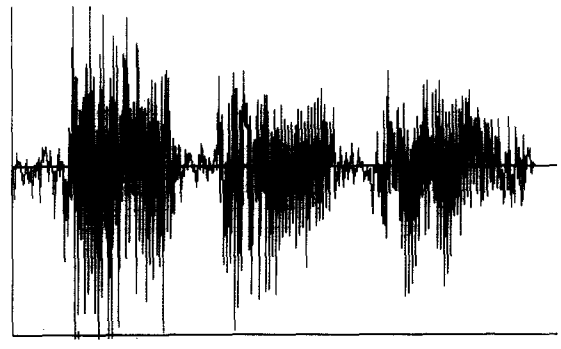


그림 5.5 Hidden LMS 적응 필터를 사용하여 잡음을 제거한 음성신호

Fig. 5.5 Speech signal when a noise is removed by Hidden LMS adaptive filter.

그림 5.3은 본 논문의 화자검증 실험에서 사용된 "제어실험실"이라는 발음에 대한 음성신호의 파형이고, 그림 5.4는 random noise를 삽입시킨 후의 음성신호를 나타낸 것이며, 그림 5.5는 Hidden LMS 필터를 이용하여 잡음을 제거시킨 음성신호의 파형을 나타낸 것이다.

다음의 실험결과에서 FR에 해당하는 것은 본인의 신원을 주장하였을 때 잘못하여 거부하는 경우이고, FA는 본인 이외 9명의 데이터(각 20개)를 입력으로 하였을 때 잘못하여 accept하는 경우를 뜻한다.

표 5.1은 잡음이 포함된 후에 기존의 화자 검증 방법으로 실험한 결과를 나타낸 것이며, 표 5.2와 표 5.3은 각각 잡음이 포함된 신호에 대해 LMS 적응 알고리즘과 Hidden LMS 적응 알고리즘을 이용하여 잡음을 제거시킨 결과 값을 나타내었다.

표 5.1 잡음을 포함시킨 후의 화자검증 결과 (No Filter)

Table 5.1 Results of Speaker Verification when a noise is added.

화자	FR	FA
A	9/20	72/180
B	7/20	39/180
C	13/20	56/180
D	10/20	61/180
E	10/20	14/180
F	9/20	53/180
G	7/20	11/180
H	8/20	33/180
I	15/20	13/180
J	9/20	28/180
결과	97/200	380/1800

표 5.2 LMS 필터를 적용시킨 경우 화자검증 결과

Table 5.2 Results of Speaker Verification with LMS filter

화자	FR	FA
A	0/20	62/180
B	3/20	26/180
C	4/20	41/180
D	1/20	42/180
E	9/20	0/180
F	5/20	49/180
G	2/20	2/180
H	3/20	24/180
I	8/20	7/180
J	3/20	13/180
결과	38/200	266/1800

표 5.3 Hidden LMS필터를 적용시킨 경우 화자검증 결과

Table 5.3 Results of Speaker Verification with Hidden LMS filter

화자	FR	FA
A	2/20	24/180
B	0/20	26/180
C	3/20	29/180
D	1/20	6/180
E	0/20	9/180
F	1/20	78/180
G	2/20	2/180
H	2/20	7/180
I	4/20	2/180
J	1/20	10/180
결과	18/200	193/1800

표 5.1을 살펴보았을 때, 잡음이 포함된 경우의 기존 화자 검증 방법으로는 거의 인식이 불가능했다. 표 5.4는 잡음을 포함시킨 음성신호에 대해 각각 인식률을 비교한 것이다.

표 5.4에서 보는 바와 같이 필터를 적용하였을 때의 결과 값이 필터를 적용하지 않은 경우의 결과 값보다 우수한 성능을 나타내고 있음을 알 수 있었고, 그 중 기존의 LMS 적응 필터에 비해 Hidden LMS 적응 필터가 월등한 성능을 나타내고 있다.

표 5.4 인식률 비교

Table 5.4 Comparison of Recognition rate

	인 식 률	
	본인 Accept 비율	타인 Reject 비율
기존의 화자 검증	51.5%	78.8%
LMS 적응 필터적용 화자검증	81%	85.2%
Hidden LMS 적응 필터적용 화자검증	91%	89.3%

6. 결 론

컴퓨터 시스템이 급속도로 발달하면서 사람의 음성신호를 처리하는 분야에 대한 연구가 많이 이루어졌으며, 인간의 음성을 실생활에서 유용하게 이용하려는 많은 응용사례 연구들이 있었다.

본 논문에서는 심각한 잡음 환경 하에서 화자검증 시스템의 성능을 향상시키기 위한 연구결과를 기술하였다. 잡음환경에서의 일반 화자검증시스템은 실험한 결과 값에서 확인했듯이 올바른 사용자가 시스템에 접근을 시도하려 하여도 접근할 수 없는 최악의 인식률을 보이게 된다. 본 연구에서는 전처리 과정에서 Hidden LMS 적응 필터링 알고리즘을 제안하여 잡음을 제거하고, 특징을 강화시키는데 사용하여, 잡음환경에서 인식률의 저하를 최소화시켜 화자 검증 시스템에서의 오동작 비율을 줄이고, 사용자의 불편함을 최소화시킴으로써 잡음 환경 하에서의 화자검증을 이용한 보안 시스템에의 적용 가능성을 확인할 수 있었다.

참 고 문 헌

[1] Benjamin Miller, "Vital Signs of Identity", IEEE Spectrum Special Report pp.22-30

[2] L.B.Rabiner and R.W.Schafer, Digital Processing of Speech Signals, Prentice Hall, New Jersey, 1978

[3] 강문기, "일반화된 성도모델에 기반을 둔 선형예측기법", 석사학위논문, 서울대학교 공과대학, 1988

[4] C.F.N. Cowan & P.M. Grant "Addative Filters", Prentice-Hal, New Jersey, 1985

[5] 이인홍, "잡음이 섞인 신호를 개선하기 위한 LMS 적응 필터링에 관한 연구", 석사학위논문, 광운대학교 대학원, 1987

[6] S.Furui, "Cepstral analysis technique for automatic speaker verification", IEEE Trans. on ASSP, vol. 29, pp.254-272

[7] F.K.Soong, A.Aron E. Rosenberg, "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition", IEEE Trans. on ASSP, vol. 36, pp.871-879

[8] D.K.Burton, "Text-Dependent Speaker Verification Using Vector Quantization Source Coding", IEEE Trans. on ASSP vol. 35, pp.133-143

[9] 한국과학재단, "음성처리 및 인식의 기초연구", 1989

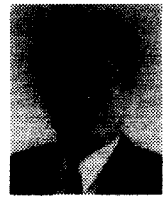
[10] R.W.Hamming, "Digital Filters", Prentice-Hall, New Jersey, 1989

[11] L.Rabiner, Biing-Hwang Juang, "Fundamentals of Speech Recognition", Prentice-Hall, New Jersey, 1993

[12] Laurene Fausett, "Fundamentals of Neural Network", Prentice Hall, 1994

[13] Philip D. Wasserman, "Neural Computing", Van Nostrand Reinhold, 1989

저 자 소 개



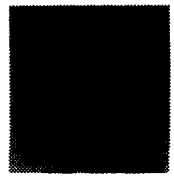
조 성 원(趙 成 元)

홍익대학교 전자전기공학부 부교수

1982년 서울대학교 공학사

1987년 미국 Purdue대학교 공학 석사

1992년 미국 Purdue대학교 공학 박사



김 재 민(金 載 敏)

홍익대학교 전자전기공학부 조교수

1984년 서울대학교 공학사

1986년 서울대학교 공학 석사

1994년 Rensselaer Polytechnic Institute박사