論文2002-39SP-3-10

학습속도 개선과 학습데이터 축소를 통한 MLP 기반 화자증명 시스템의 등록속도 향상방법

(An Improvement of the MLP Based Speaker Verification System through Improving the Learning Speed and Reducing the Learning Data)

李 百 永 *, 李 泰 承 *, 黃 秉 元 *

(Baeg Yeong Rhee, Tae Seung Lee, and Byong Won Hwang)

요 약

MLP(multilayer perceptron)는 다른 패턴인식 방법에 비해 몇 가지 유리한 이점을 지니고 있어 화자증명 시스템의 화자학습 및 인식 방법으로서 사용이 기대된다. 그러나 MLP의 학습은 학습에 이용되는 EBP(error backpropagation) 알고리즘의 저속 때문에 상당한 시간을 소요한다. 이 점은 화자증명 시스템에서 높은 화자인식률을 달성하기 위해서는 많은 배경화자가 필요하다는 점과 맞물려 시스템에 화자를 등록하기 위해 많은 시간이 걸린다는 문제를 낳는다. 화자증명 시스템은 화자 등록후 곧바로 증명 서비스를 제공해야 하기 때문에 이 문제를 해결해야 한다. 본 논문에서는 이 문제를 해결하기 위해 EBP의 학습속도를 개선하는 방법과, 기존의 화자증명 방법에서 화자군집 방법을 도입한 배경화자 축소방법을 사용하여 MLP 기반 화자증명 시스템에서 화자등록에 필요한 시간의 단축을 시도한다.

Abstract

The multilayer perceptron (MLP) has several advantages against other pattern recognition methods, and is expected to be used as the learning and recognizing speakers of speaker verification system. But because of the low learning speed of the error backpropagation (EBP) algorithm that is used for the MLP learning, the MLP learning requires considerable time. Because the speaker verification system must provide verification services just after a speaker's enrollment, it is required to solve the problem. So, this paper tries to make short of time required to enroll speakers with the MLP based speaker verification system, using the method of improving the EBP learning speed and the method of reducing background speakers which adopts the cohort speakers method from the existing speaker verification.

Keyword : speaker verification, multilayer perceptron, error backpropagation, cohort speakers, pattern recognition

Ⅰ. 서 론

1.1. MLP 화자증명 시스템

컴퓨터, PDA, 휴대폰 등 각종 정보기기가 생활저변에서 활용되는 시대를 맞이하여 개인정보를 보호하는 수단이 점차 중요하게 인식되고 있다. 이 중 생체인식

* 正會員, 韓國航空大學敎 航空電子工學科
(Dept. of Avionics, Hankuk Aviation Univ.)
接受日字:2002年2月25日, 수정완료일:2002年4月2日

은 사람마다 고유한 특징을 이용하여 이러한 정보기기에 대한 접근을 통제하는 수단으로, 보안정보 관리의 편의성이 뛰어나고 보안능력 자체도 탁월하다. 주요한 생체특징으로 지문, 홍체, 얼굴모양, 음성 등이 있으나, 이 중 음성은 우리가 가장 쉽게 접할 수 있고 음성처리에 필요한 비용이 다른 특징에 비해 상대적으로 저렴하다는 이점 때문에 이를 이용한 연구가 진행되고 있다[1].

음성을 이용한 생체인식 방법을 화자인식이라고 한다. 화자인식은 시스템에 여러 화자를 등록하고 이들 중 현재 음성과 일치하는 화자를 선택하는 화자식별과, 시스템에 미리 등록해 둔 특정 화자의 신원을 선택한 뒤 현재 음성이 그 화자와 일치하는지의 여부를 판별하는 화자증명으로 나뉜다. 이 중 화자증명은 불특정 다수를 대상으로 신원을 확인하고, 이러한 화자증명 모듈을 다수 결합하여 화자식별 시스템을 구성할 수 있기 때문에 화자증명 시스템의 연구가 더 활발하게 이루어지고 있다[2].

특정화자의 음성을 학습하고 입력된 음성을 이렇게 학습된 음성과 비교하여 신원을 판별하기 위해 다양한 인식방법이 사용된다. 이들 중 MLP(multilayer perceptron)는 다음과 같은 이점을 갖고 있어 다양한 인식문제에서 사용될 수 있다[3-6].

- 논파라메트릭(nonparametric) 방식이기 때문에 문제에서 가정해야 하는 하부확률분포가 필요없다.
- 학습되는 각 모델 사이의 차이를 최대한 구별하는 거부학습능력이 있기 때문에 인식오류 가능성을 최소화한다.
- 학습모델별로 +1, 0(또는 -1)의 학습목표치를 사용할 때 LDA(linear discriminant analysis)와 유사한 특징공간 변환능력을 갖는다.

1.2. MLP 학습속도 문제와 화자증명 시스템의 실시간 성능요구

일반적으로 MLP의 학습에는 상당한 시간이 소요된다. 이는 MLP의 학습에 사용하는 EBP(error back-propagation) 알고리즘에 주요한 원인을 둔다. EBP 알고리즘은 최대 기울기 감소 방법을 바탕으로 한 것으로, MLP의 현재출력과 목표출력 사이의 오류를 출력층에서 은닉층으로 역방향으로 전파하면서 가중치를 조정하는 방법으로 최종적인 목표치를 달성한다[4]. EBP 학습이 느린 까닭은 최대 기울기 감소 방법이 현재의 가중치에 대한 지역적인 정보만 사용하는 것에서 연유한다.

한편, 학습 데이터의 크기도 MLP의 학습속도에 영향을 줄 수 있다. 화자증명을 위해서는 신원점수를 측정할 의뢰화자와 비교를 위한 배경화자가 필요하며, 이 중 의뢰화자는 다시 실제화자와 사칭화자로 나뉜다. 높은 화자증명 인식률을 위해서는 의뢰화자를 되도록 특성이 유사한 배경화자와 비교하여 엄밀한 판별이 이뤄지도록 해야 한다. 그러나 의뢰화자의 특성을 미리 알 수 없기 때문에 가능한 한 많은 배경화자를 준비하여 어떤 화자가 신원확인을 의뢰하더라도 정확한 판별이 이뤄질 수 있도록 한다[7]. 하지만, 높은 인식률을 달성하기 위해 배경화자의 수를 늘리면 학습시간이 긴 MLP 화자증명 시스템의 등록시간도 함께 길어진다는 점은 자명하다.

화자증명 시스템은 화자의 등록과 동시에 신원증명 서비스가 가능해야 할 뿐 아니라 빈번히 이용되는 정보기기 사용의 편의성을 고려하여 신속한 증명 서비스가 이루어져야 한다. 앞서 설명한 두 요인으로 인해 MLP를 이용한 시스템의 화자등록 시간은 긴 편이지만, 구별함수에 기반한 MLP의 빠른 동작특성 때문에 화자증명 시간은 보편적인 정보기기의 처리능력을 기준으로 볼 때 만족할만한 수준이다. 이에 따라 MLP에 기반한 화자증명 시스템의 실시간 성능을 보장하기 위해 등록속도 개선에 노력이 집중되어야 한다.

1.3. 기존의 MLP 학습속도 개선노력

MLP의 학습속도를 개선하려는 시도는 크게 두 방향으로 이루어졌다. 첫 번째 방향은 경험과 실험결과를 활용한 것으로, 출력치가 목표치에서 멀 경우에는 학습률을 크게 하고 가까울 경우에는 작게 하는 것이다. 이것은 다시 가중치 벡터 전체에 일괄적으로 영향을 미치는 전역 학습률을 변경하는 방법[8]과 각 가중치마다 최대 기울기의 변화에 따라서 지역 학습률을 변경하는 방법[9]으로 나뉜다. 두 번째 방향은 최적화 이론을 활용한 것으로, 가중치에 대한 2차 미분정보를 사용한다. 이러한 부류로는 모멘텀(momentum)을 사용하여 이전의 학습추세를 현재 갱신에 반영하거나[8], Newton의 최적화 이론[10] 또는 이를 변형한 알고리즘[11 12]을 이용하여 목표치로 가장 빠르게 수렴할 수 있는 가중치 벡터 갱신치를 계산하는 방법이 있다.

가중치 갱신은 두 가지 방식으로 이루어진다. 하나는 모든 학습데이터를 제시한 후 그에 따른 변경치들의

평균을 적용하는 방법이고, 다른 하나는 학습데이터를 하나씩 제시할 때마다 변경치를 적용하는 방법이다. 전자를 온라인(또는 확률적) 방식이라고 하고, 후자를 오프라인(또는 일괄적) 방식이라고 부른다. 두 방식 모두 모든 학습데이터가 제시되는 한 주기를 에폭(epoch)이라 하고, 에폭마다 MLP 목표치와 출력치 사이의 차이를 검사하여 학습의 속행여부를 결정한다.

화자증명을 포함한 패턴인식에서 MLP를 사용할 경우 오프라인 학습보다 온라인 학습이 더 빠른 속도로 이루어지는데, 그 원인으로 아래와 같은 이유를 찾을 수 있다[13].

- 모델 내의 모든 패턴이 서로에 대해 상당한 중복성을 내포하므로 모든 패턴이 EBP의 최대 기울기 계산에 기여한다. 이 때문에 모델에 포함된 패턴수가 많을수록 에폭 단위의 학습속도가 빨라진다.
- 모델 내 패턴들의 최대 기울기가 90° 이내일 경우 오류를 최소화하는 방향으로 학습이 진행된다.
- 로컬 미니마에 빠질 가능성을 크게 줄인다. 이러한 특성은 모델 내의 모든 패턴마다 가중치 벡터의 갱신이 이루어질 때 중심위치에서 상대적으로 멀리 떨어진 패턴에 의해 전체 진행방향과 다른 임의적 진동이 발생하기 때문이다.

패턴인식의 여러 연구에서 온라인 방식의 EBP 학습이 위에서 소개한 오프라인 계열의 여러 방법보다도 빠르게 이루어진다는 결과가 보고되었다[11, 13-14].

1.4. 기존의 화자증명 시스템의 등록속도 향상노력

파라메트릭 방식의 기존 화자증명 시스템에서는 화자등록시 등록화자와 유사한 제한된 수의 배경화자를 선택한 뒤, 의뢰화자의 신원판별시 이 군집의 화자정보를 이용하여 의뢰화자의 신원점수를 계산하는 방법이 도입되었다[7, 15].

Higgins 등[16]은 신뢰할만한 신원점수를 얻기 위해 의뢰화자의 유사도 점수를 의뢰화자와 가장 유사한 배경화자의 점수로 평준화하는 방법을 다음과 같이 제안하였다.

$$L(\mathbf{X}) = \frac{p(\mathbf{X} \mid S_C)}{\max_{S_{BG}, S_{BG} \neq S_C} p(\mathbf{X} \mid S_{BG})} \quad (1)$$

여기서, X는 의뢰화자의 음성열이고, S_c는 의뢰화자가 주장하는 신원을, S_{BG}는 의뢰화자에 가장 근접한 배경화자를 나타낸다.

그러나 이 점수는 (a) 의뢰화자와 가장 근접한 배경화자를 찾기 위해 모든 배경화자를 탐색하므로 계산량이 배경화자의 규모와 비례하고, (b) 배경화자 집단에서 의뢰화자의 위치에 따라(즉, 의뢰화자 주위의 배경화자 분포밀도에 따라) 식(1)의 분모항의 값이 크게 달라지므로 안정적인 점수를 얻기 힘들다는 단점을 갖는다.

이 문제를 해결하기 위해 화자를 시스템에 등록할 때 일정한 수의 근접 배경화자를 선택한 뒤, 의뢰화자 점수를 계산할 때 이들의 유사도 점수를 합한 값을 평준화 항으로 사용하는 화자군집 방법이 제안되었다[7].

$$L(\mathbf{X}) = \frac{p(\mathbf{X} \mid S_C)}{\sum_{S_{Cohort}, S_{Cohort} \neq S_C} p(\mathbf{X} \mid S_{Cohort})} \quad (2)$$

여기서, S_{Cohort}는 등록화자와 근접한 화자군집을 뜻한다. S_{Cohort}을 이루는 배경화자의 수가 많을수록 증명점수의 신뢰성이 높아진다[7].

처리속도의 측면에서 봤을 때 화자군집 방법은 식(1)의 신원증명에 요구되는 처리시간을 단축하기 위해 근접 배경화자 탐색처리를 화자등록 단계로 분산시킨 것으로 간주할 수 있다. 화자증명 시스템의 등록과 증명 단계 중에서 등록단계가 증명단계보다 비교적 실시간에 대한 엄격도가 낮으므로 이러한 계산량 분산은 전반적인 실시간 성능을 높이는 효과를 갖는다.

1.5. 등록속도 개선가능성

패턴인식에서 EBP 알고리즘으로 학습되는 MLP는 3단계의 학습단계를 밟는다. 즉, (a) 모델의 중심위치를 학습하고, (b) 모델의 분산을 학습한 다음, (c) 모델분포의 윤곽을 학습한다. 후자 쪽으로 진행해 갈수록 학습에 기여하는 패턴영역이 외곽으로 한정되고 패턴마다 효율적인 학습률이 달라질 뿐 아니라, 학습에 기여하지 않는 패턴이 계속해서 계산과정에 포함되는 비효율성이 발생한다. 기존의 온라인 EBP 알고리즘이 이러한 변화를 수용할 수 있다면 더욱 빠른 학습이 가능할 것이고, 이는 곧 MLP 기반의 화자증명 시스템의 등록속도 향상으로 이어질 수 있다.

MLP의 학습속도를 증가시키는 방법 외에 MLP의 학습에 필요한 데이터량을 줄이는 방법으로도 화자증명 시스템의 등록속도를 향상시킬 수 있다. 이를 위해 등

록단계에 기존의 화자군집 방법을 도입하여 등록화자와 비슷한 소수의 배경화자를 선택함으로써 등록시 발생하는 데이터량을 줄인다.

본 논문에서는 이 두 아이디어의 구체적인 실현방법을 제시하고 지속음을 화자의 인식단위로 하는 시스템에 이 방법을 적용한다. 이를 위한 이후 논문의 구성은 다음과 같다. 2장과 3장에서 두 아이디어의 배경을 다루고, 4장과 5장에서 각 아이디어의 구체적인 실현방법을 설명한다. 6장에서 본 연구에서 구현한 화자증명 시스템을 설명한 다음, 7장에서 음성 데이터베이스를 이용한 실험을 통해 두 방법의 효과를 입증한다. 그리고 마지막으로 본 논문의 성과를 8장에서 정리한다.

Ⅱ. 척력개념에 의한 EBP 학습해석과 온라인 EBP의 학습속도 개선 가능성

EBP 알고리즘에서 가중치의 현재 값에 대해 목표치에 가장 빠르게 접근할 수 있는 변위는 아래와 같이 계산된다.

$$\frac{\partial e_p}{\partial w_{ij}} = \frac{\partial e_p}{\partial s_i}\frac{\partial s_i}{\partial n_i}\frac{\partial n_i}{\partial w_{ij}} \tag{3}$$

여기서, e_p는 패턴 p에 대한 출력뉴런층의 오류측정함수이고, w_{ij}는 j번째 뉴런과 i번째 뉴런 사이의 연결 가중치, s_i는 i번째 뉴런의 동작치, n_i는 i번째 뉴런에 대한 가중된 입력의 총합을 나타낸다. 이 식에서 계산된 변화량을 이전의 가중치 벡터에 적용하면 목표치에 더욱 가까운 값을 도출할 수 있으며, 이를 아래의 방법으로 실현한다.

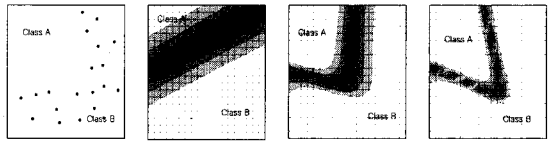
$$w_{ij}(t+1) = w_{ij}(t) - \eta\frac{\partial e(t)}{\partial w_{ij}(t)} \tag{4}$$

여기서, t는 가중치 벡터의 특정 상태 시각을 나타내며, $e(t)$는 시각 t에서 학습패턴(들)의 목표치에 대한 오류치이고, η는 적용할 변화량의 비율을 결정하는 학습률이다.

패턴인식을 위한 EBP 학습에서 각 모델은 식(4)에 의해 반복적인 방식으로 학습되므로 다음과 같은 과정을 거친다.

(1) 중심위치학습
(2) 영역분산학습
(3) 영역윤곽학습

2모델 분류문제에서 이 과정을 그림 1에서 보여준다.



(a) (b) (c) (d)

그림 1. (a) 2모델 데이터 분포, (b) 중심위치학습, (c) 분산학습, (d) 윤곽학습 EBP 학습의 3단계
Fig. 1. (a) data distribution in two models, (b) learning model center locations, (c) learning model area variations, and (d) learning model contours. The three phases in EBP learning.

이 그림의 (b)~(d)에서 검은색 띠는 두 모델의 결정경계선을 나타내며 색이 짙을수록 경계가 뚜렷함을 의미한다. 여기서 경계선이 (b)에서는 두 모델의 중심을 가로지르는 모습을 보여주고, (c)에서는 두 모델의 영역분산정도를 구분하고 있으며, (d)에서는 세부적인 윤곽을 표현하고 있음을 알 수 있다.

한편, 식(3)의 e_p는 아래와 같이 표현된다.

$$e_p(t) = \frac{1}{2}\sum_{o=1}^{N} e_{o,p}^2(t) \tag{5}$$

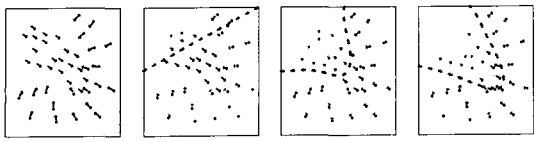
여기서, $e_{o,p}$는 현재 패턴에 대한 개별 출력뉴런의 오류치이고, N은 출력뉴런의 개수이다. $e_{o,p}$는 다시 다음과 같이 표현된다.

$$e_{o,p}(t) = d_{o,p}(t) - y_{o,p}(t) \tag{6}$$

여기서, $d_{o,p}$는 학습 목표치이고, $y_{o,p}$는 현재 출력치이다.

그림 1의 각 단계에서 식(6)에 의해 각 패턴이 나타내는 오류치는 결정 경계선을 움직이는 척력으로 생각할 수 있으며, 이는 곧 각 단계에서 개별 패턴이 학습에 기여하는 정도로 간주할 수 있다. 그림 2에서 그림 1을 패턴별 척력 개념으로 묘사한 모습을 보여준다.

이 그림에서 점선은 결정 경계선을 나타내고, 화살표는 패턴별 척력을 나타낸다. (a)에 해당하는 미학습 상



(a) (b) (c) (d)

그림 2. EBP 학습의 패턴별 척력

Fig. 2. The repulsive force of each pattern in EBP learning.

태에서는 모든 패턴이 강한 척력을 나타내어 중심위치를 학습할 수 있게 하며, (b) 이후부터는 결정 경계선 인근과 경계선을 넘어 상대방 모델 영역에 위치한 패턴이 강한 척력을 발생시켜 경계선의 형상과 위치를 변화시킨다.

패턴인식에서 온라인 방식이 오프라인 방식에 비해 빠른 학습을 달성하는 원인을 패턴별 척력의 적용방식에서 찾을 수 있다. 오프라인 방식에서는 모든 패턴별 척력벡터의 합벡터를 적용하지만, 온라인 방식에서는 이들 척력벡터를 개별적으로 적용한다. 따라서 온라인 방식이 결정 경계선의 모양을 변화시킬 기회를 더 많이 갖게 되며, 모델의 영역이 복잡한 형상을 띠는 경우 그 효과는 더욱 크다.

이처럼 온라인 방식이 오프라인 방식에 비해 학습속도 면에서 이점을 갖고 있으나, 아직도 최적의 속도를 달성하기 위한 여지가 남아있다.

온라인 EBP에서 학습단계에 따른 패턴의 척력분포를 보다 적극적으로 활용하면 학습시간을 더욱 단축할 수 있다.

그림 2에서 모델의 결정 경계선이 아직 학습되지 않은 패턴의 큰 척력에 의해 변형되는 것을 볼 수 있다. 이 척력이 결정 경계선에 미치는 영향력은 식(4)의 학습률 η에 의해 조정된다. 즉, 학습이 덜된 패턴에 대해 η가 클수록 그 패턴의 인근 결정 경계선은 급격히 이동하게 된다. 그러나 학습이 진행되면서 결정 경계선이 학습목표 주변에 도달하면, 그 부근의 패턴에 대한 η가 점차 작아져야만 결정 경계선이 학습목표에 수렴할 가능성이 높아진다. 또한, 여러 모델 사이의 중첩된 영역은 완벽한 학습이 불가능하므로, 결정 경계선의 불필요한 진동을 막기위해 이 영역의 패턴에 대해서는 η가 작아야 한다.

이를 위해 기존의 온라인 EBP 알고리즘에서와 같이 η를 학습기간 동안 고정하는 대신, η가 상황에 맞춰 변경될 수 있어야 한다. 이와 같은 가변 η는 이미 오프라인 EBP에서 시도된 바 있지만[8, 18], 온라인 EBP에서 이를 적용하면 온라인 방식의 데이터 중복성 활용 이점을 더욱 향상시킬 수 있다.

한편, 기존의 온라인 EBP 알고리즘에서는 패턴이 이미 학습되었더라도 그 패턴이 계속해서 학습계산에 참여한다. 일정한 학습성과의 기준을 설정하고 패턴이 그 기준을 만족하면 학습계산에서 그 패턴을 제외하더라도 전체 학습진행에는 거의 영향을 주지 않으면서 학습시간을 단축할 수 있다.

Ⅲ. 화자군집 방법의 MLP 적용

식(2)에서 S_{Cohort}의 배경화자 수가 무한대라고 가정한다면 Bayes 법칙에 따른 의뢰화자의 주장화자에 대한 사후확률에서 화자들의 사전확률이 고정됨에 따라 식(2)를 사후확률로 볼 수 있다.

$$P(S_C \mid \mathbf{X}) = \frac{p(\mathbf{X} \mid S_C)}{\sum_{\infty} p(\mathbf{X} \mid S)} \cong L(\mathbf{X}) = \frac{p(\mathbf{X} \mid S_C)}{\sum_{S_{Cohort}} p(\mathbf{X} \mid S_{Cohort})} \quad (7)$$

Gish는 MLP가 충분한 학습용량을 가졌다면 다수모델을 분류하는 패턴인식에서 각 모델의 사후확률을 학습할 수 있다는 사실을 증명했다[17]. 이에 따라 2개의 학습모델로서 등록화자와 충분한 수의 배경화자가 주어진다면 MLP의 학습은 식(7)으로 표현되는 화자군집 방법을 근사화한다.

그러나, 화자군집을 이루는 배경화자의 수가 일정한 수준 이상이면 인식률 향상에는 거의 기여하지 못하면서 계산량은 선형적으로 증가하게 된다. 따라서 실험을 통해 만족할만한 인식률 수준에서 가장 적은 배경화자를 선택하는 것이 유리하다.

Ⅳ. 패턴별 가변학습률 및 학습생략 방법

2절에서 제기한 온라인 방식 EBP 알고리즘의 학습속도 향상 아이디어를 실현하기 위해 본 논문에서는 두 가지 방법을 제안한다.

첫 번째 방법은 학습패턴별로 학습률을 가변하는 방법이다.

학습률은 모델의 학습진행에 따라 큰 값에서 작은 값으로의 변화가 필요하다. 식(5)에서 현재 패턴의 모델과 대응하는 출력뉴런의 $e^{2}_{o,p}(t)$는 현재패턴의 학습

상태를 알 수 있는 수치적 측정수단을 제공한다. 즉, 패턴이 충분히 학습되지 않았을 경우 큰 값을 나타내고, 충분히 학습되었을 경우 작은 값을 나타낸다. 따라서 학습 초기에는 큰 값이, 그 뒤 학습이 완결되어감에 따라 작은 값이 구해진다.

온라인 학습방식에서 보편적으로 선택되는 학습률의 범위는 1~0.0001이다[18]. 이 범위에서 상한을 넘어서면 내부변수의 값이 발산하기 쉽고, 하한을 넘어서면 학습이 불필요하게 길어진다. 그러나 문제에 따라 적절한 값의 범위가 다르므로 이 범위는 실험을 통해 결정해야 한다. 식(5)의 $e^{2}_{o,p}(t)$를 학습률로 사용하면 하한에 대해서는 걱정할 필요가 없으므로 적절한 상한을 알아내기만 하면 된다. 그런 다음 0에서부터 이렇게 알아낸 상한까지 $e^{2}_{o,p}(t)$의 범위를 제한한다.

$$y(x) = \frac{2 \cdot V_{UL}}{1 + e^{-2x}} - V_{UL} \tag{8}$$

여기서, V_{UL}은 실험을 통해 알아내는 학습률의 상한이고, x는 $e^{2}_{o,p}(t)$의 출력치이며, y는 제한된 값이다.

그러나 식(8)는 현재패턴에 한정된 오류정보만을 사용한다. 만약 학습모델 자체가 필연적인 오류를 크게 내포하는 경우, 즉 여러 모델 사이에 모델 분포영역이 겹치는 부분이 클 경우에는 그 부분의 패턴에 의한 가현재 학습상황과 대체적으로 무관하게 큰 값을 나타냄으로써 전체 학습을 방해하는 결과를 초래할 수 있다. 이런 경우에 대처하기 위해 식(8)의 값을 한 단계 이전 에폭의 평균오류로 제한한다.

$$y'(x) = \begin{cases} V_{ASEE} \cdot V_{UL} & if\ y(x) \geq V_{ASEE} \cdot V_{UL} \\ y(x) & otherwise \end{cases} \tag{9}$$

여기서, V_{ASEE}는 아래와 같이 정의되는 평균오류제곱 에너지를 나타낸다.

$$V_{ASEE} = \frac{1}{NM} \sum_{p=1}^{M} e_p(t-1) \tag{10}$$

여기서, M은 1에폭 동안 사용된 패턴의 수이다.

두 번째 방법은 패턴의 학습을 생략하는 방법이다.

온라인 방식 EBP에서 이뤄지는 주요 계산은 패턴의 오류계산, 오류 역전파, 가중치 갱신이다. 현재 학습단계에서 현재 패턴의 기여도는 $e^{2}_{o,p}(t)$를 통해 알 수 있으므로, $e^{2}_{o,p}(t)$의 값이 학습 전에 설정하는 식(10)의 최종목표 오류치보다 작은 경우 현재 패턴이 학습에 기여하는 바가 적다고 판단하여 오류 역전파와 내부변수 갱신 계산과정을 생략할 수 있다. 만일 이후 다른 패턴의 학습으로 현재 패턴의 기여도가 높아진다면 이 상태는 $e^{2}_{o,p}(t)$에 의해 발견되기 때문에 다시 학습에 참여할 수 있게 된다.

본 논문에서는 이 두 방법을 가리켜 패턴별 가변학습률 및 학습생략(Changing rate and Omitting patterns in Instant Learning) 방법이라고 부른다.

V. 인접화자정보를 이용한 MLP 화자등록 속도향상 방법

3절에서 제기한 대로 MLP 학습에 필요한 배경화자의 수를 줄이기 위해 식(2)와 같이 등록하는 화자에 인접한 제한된 수의 배경화자를 선택하되, 화자증명 시스템의 인식률을 저하시키지 않는 수준에서 최소수를 선택한다.

이 방법을 위해서 등록화자가 각 배경화자와 얼마나 근접하는 지를 평가하는 MLP와, 이 MLP를 통해 선택된 배경화자를 이용하여 등록화자의 특성을 학습하는 MLP가 필요하다. 본 논문에서는 전자를 MLP-I, 후자를 MLP-II라고 한다.

MLP-I은 배경화자의 수와 동일한 개수의 출력뉴런을 가지며, 사전에 배경화자의 데이터를 이용하여 각각을 분류할 수 있도록 학습된다. 이 때 은닉뉴런이 많으면 배경화자의 근접도 평가만으로도 많은 시간이 소요될 수 있으므로, 각 배경화자 데이터를 80% 이상의 정확도로 분류할 수 있을 정도로만 은닉뉴런 개수를 선택한다.

MLP-II는 MLP-I을 통해 선택된 배경화자의 데이터를 이용하여 등록화자를 학습한다. 단, MLP의 경우 화자군집 방법과 달리 학습모델을 등록화자와 배경화자의 2개로 설정하므로 MLP-I으로 선택된 배경화자들의 패턴을 한 모델로 통합한다.

MLP-I에서 선택될 배경화자의 수는 실험을 통해 결정한다. 이 수는 등록화자의 분포범위를 선택된 배경화자들이 둘러쌀 수 있는 수준 이상이어야 하며, 인식률의 하락을 방지하기 위해서는 배경화자가 등록화자를 밀도있게 둘러싸야 한다. 배경화자의 밀도에 따른 등록화자의 학습양상을 그림 3에서 설명하고 있다. 이 그림에서 점선은 등록화자와 배경화자의 영역을 결정하는

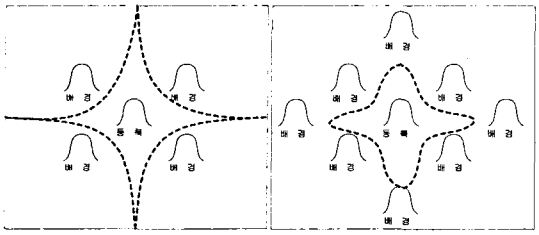


그림 3. (a) 저밀도 배경화자와 (b) 고밀도 배경화자의 경우에서 MLP의 등록화자학습

Fig. 3. The MLP learnings of the enrolling speaker for (a) the sparse background speakers and (b) the dense background speakers.

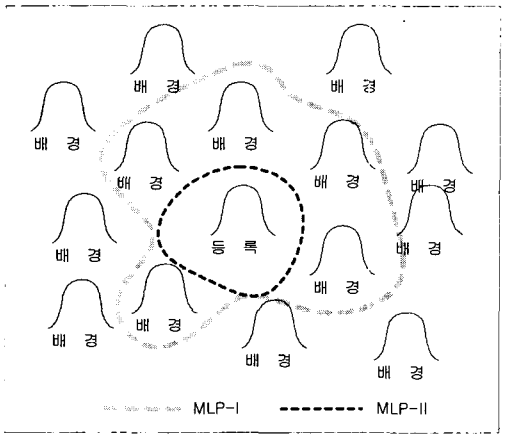


그림 4. MLP-I의 배경화자군집 선택과 MLP-II의 등록화자학습

Fig. 4. The MLP-I used to select background speakers and the MLP-II used to learn the enrolling speaker with the background speakers.

경계선을 나타낸다.

그림 4는 MLP-I과 MLP-II의 동작원리를 설명한다.

VI. 지속음 MLP 기반 화자증명 시스템

본 연구에서 구현한 화자증명 시스템은 입력음성에서 고립단어를 추출하고, 이 고립단어에서 한국어 지속음(/a/, /e/, /ə/, /o/, /u/, /ɨ/, /i/, /l/, 비음)을 인식한 다음, 각 지속음별로 MLP-I과 MLP-II를 이용하여 화자를 학습하고 증명점수를 계산한다. 이 시스템에서 이뤄지는 처리를 그림 5에서 보여주며, 각 처리의 설명은 아래와 같다.

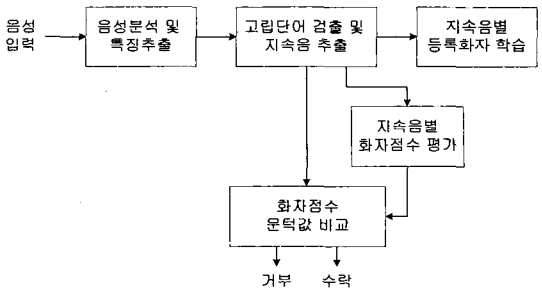


그림 5. 지속음 MLP 기반 화자증명 시스템 처리 흐름도

Fig. 5. The procedures of the continuant-and MLP-based speaker verification system.

(1) 음성분석 및 특징추출

- 16bit 16kHz로 샘플링된 등록화자의 입력음성을 20ms 오버랩시킨 30ms 길이의 프레임으로 나눈다.
- 각 프레임에 대해 16차 Mel 간격 필터뱅크(filter bank)[19]를 추출하여 고립단어 및 지속음 검출에 사용한다. 필터뱅크 계수는 전체 스펙트럼 포락에 미치는 성량의 영향을 제거하기 위해 1kHz까지의 계수를 평균하여 이 값을 모든 계수에서 차감한다. 그리고, MLP의 효과적인 학습을 위해 다시 모든 계수의 평균이 0이 되도록 조정한다.
- 각 프레임에 대해 50차의 0~3kHz 대역 균등간격 Mel 필터뱅크를 추출하여 화자증명에 사용한다. 이 음성특징은 2차 포만트(formant)에 더 많은 화자정보가 집중된다는 연구결과[20]에 의한 것이다. 필터뱅크 계수는 전체 스펙트럼 포락에 미치는 성량의 영향을 제거하기 위해 1kHz까지의 계수를 평균하여 이 값을 모든 계수에서 차감한다. 그리고, MLP의 효과적인 학습을 위해 다시 모든 계수의 평균이 0이 되도록 조정한다.

(2) 고립단어 검출 및 지속음 검출

- 각 지속음과 묵음을 화자독립 방식으로 검출하도록 학습된 MLP를 사용하여 고립단어와 고립단어 내의 지속음을 검출한다.

(3) 지속음별 등록화자 학습

(3-1) 기본 온라인 EBP 또는 COIL 적용의 경우

- 등록화자와 배경화자 데이터를 이용하여 MLP를 학습한다.

(3-2) MLP-I 및 MLP-II 적용의 경우

- 지속음별로 전체 고립단어의 각 지속음을 MLP-I에 입력한 뒤, 출력뉴런의 수치를 평균하고, 이 평균치

가 높은 순서로 n명의 배경화자를 선택한다.

- 지속음별로 선택된 n명의 배경화자 데이터를 이용하여 MLP-II에 등록화자를 학습시킨다.
- 지속음별로 MLP-II 학습에 사용되는 패턴수는 등록화자 당 10개씩이다.

(4) 지속음별 화자점수 평가

(4-1) 기본 온라인 EBP 또는 COIL 적용의 경우

- 지속음별로 전체 고립단어의 각 지속음을 MLP 에 입력한 뒤, 출력뉴런의 수치를 평균한다.

(4-2) MLP-I 및 MLP-II 적용의 경우

- 지속음별로 전체 고립단어의 각 지속음을 MLP-I에 입력한 뒤, 출력뉴런의 수치를 평균하고, 이 평균치가 높은 순서로 n명의 배경화자를 선택한다.
- 선택된 배경화자 가운데 (3)의 배경화자가 1명 이상 포함된 모든 지속음에 대해 MLP-II의 출력치를 평균한다.
- (3)의 배경화자가 1명 이상 포함된 지속음이 전무한 경우 의뢰화자를 거부한다.

(5) 등록어 및 화자점수 문턱값 비교

- (4)의 평균치와 사전 설정한 문턱값을 비교하여 최종적인 거부/수락을 결정한다.

이 화자증명 시스템에서는 지속음을 화자인식단위로 사용하기 때문에 하부확률분포가 단변량(univariate)의 형태를 띤다[21]. 따라서 화자학습에 관련한 MLP, MLP-I, MLP-II는 모두 1개의 은닉계층이 포함된 2층 구조만으로 충분하다[22-23]. 이 중 MLP와 MLP-II는 학습하는 모델이 2개뿐 이므로 1개의 출력뉴런과 2개의 은닉뉴런이면 충분히 이들을 학습할 수 있다.

Ⅶ. 실 험

7.1. 음성 데이터베이스

이 실험에 사용할 데이터는 한국인 남녀 40명의 4연숫자 발성을 녹음한 것이다. 여기서 4연숫자라 함은 아라비아 숫자 0~9에 해당하는 /goN/, /il/, /i/, /sam/, /sa/, /o/, /yug/, /cil/, /pal/, /gu/ 음을 연속해서 4자리를 발성한 것을 의미한다. 각 화자가 총 35개의 서로 다른 숫자음 배열을 4회씩 발성하였고, 발성은 16kHz 주기의 16bit 크기로 녹음되었다. 4회 발성 중 3회를 각 화자의 등록음성으로 사용하고 나머지 1회를 증명시험 음성으로 사용한다.

등록화자 학습시 필요한 배경화자로는 위의 40명 외의 남녀 29명을 사용한다.

7.2. MLP 학습설정

실험에서 화자등록에 사용되는 MLP(MLP-I/II 방법의 경우 MLP-II)의 각종 파라미터[4]는 다음과 같다. 입력 데이터는 -1.0~+1.0으로 평준화되며, 학습속도 향상을 위해 각 출력뉴런의 목표치는 등록화자에 +0.9, 배경화자에 -0.9를 지정한다. 모든 내부변수는 학습 전 -0.5~+0.5의 임의수치로 초기화된다. 학습시 두 모델의 음성패턴은 교대로 MLP에 제시되는데, 거의 대부분의 경우에 있어서 두 모델의 학습패턴수가 일치하지 않으므로 많은 쪽 패턴이 모두 제시될 때까지 적은 쪽 패턴을 반복해서 제시하여 1에폭을 채운다. 최대 학습에폭은 로컬 미니마에 빠지는 경우를 고려하여 1000회로 제한하고, 학습목표는 식(10)의 계산치가 0.01 이하가 되는 동시에 조기 학습중지를 막기위해 이 값의 변화율(average squared error energy rate)이 0.01 이하가 되는 것으로 한다.

7.3. 실험결과

실험화자 40명을 한 명씩 차례로 등록화자와 실제화자로 사용하고 이를 제외한 나머지 39명을 사칭화자로 사용한다. 결과적으로 화자당 35회의 실제화자 시도와 1,560회의 사칭화자 시도를 평가하게 되고, 실험 전체로 봤을 때는 1,400회의 실제화자 시도와 54,600회의 사칭화자 시도를 평가하게 된다.

실험은 AMD 1.4GHz급 컴퓨터에서 실시하였으며, 실험결과에서 오류율은 동일오류율(Equal Error Rate)을 의미하고, 학습패턴수는 1명의 화자를 등록하기 위해 학습된 총 패턴수를 나타내며, 학습시간은 이 패턴들을 학습하는 데 걸린 실제시간을 가리킨다. 오류율, 학습패턴수, 학습시간은 전체 증명시도에 대한 평균치이다.

먼저, 기본 온라인 EBP 알고리즘만을 사용하여 학습률을 변경해 가며 최적의 학습률을 찾는 실험결과를 그림 6에 보였다.

이 결과에서 최고 학습속도는 학습률 0.5에서 달성되었으나 최저 오류율을 고려할 때 학습률 1.0을 선택하는 것이 타당하다.

다음으로 이 학습률에서 온라인 EBP 대신 COIL을 적용했을 때의 실험결과를 그림 7에 보였다. 여기서 온라인 EBP를 적용한 실험을 OnEBP로, 가변학습률만

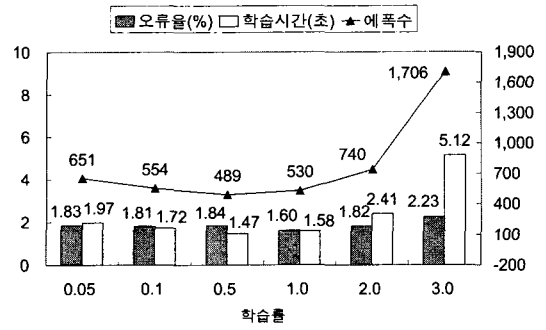


그림 6. 온라인 EBP 알고리즘의 실험결과
Fig. 6. The experiment result of the online EBP algorithm.

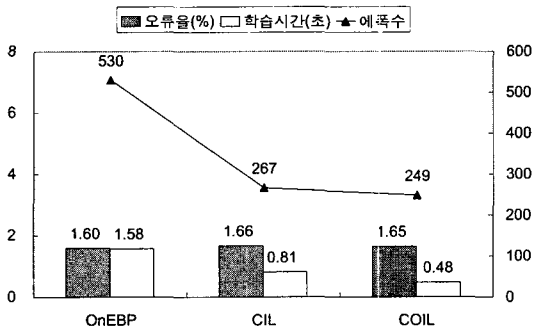


그림 7. COIL 적용 실험결과
Fig. 7. The experiment result of applying the COIL.

적용한 실험을 CIL로, 가변학습률과 패턴생략을 모두 적용한 실험을 COIL로 표기한다.

이 결과에서 CIL과 COIL의 학습률 상한은 1.0이며, 이 수치 역시 온라인 EBP와 마찬가지로 학습시간과 오류율을 함께 고려하여 최적의 수치를 선택한 것이다. 각각의 경우 학습시간이 95.1%, 229.2% 향상되었다.

그림 8에서는 온라인 EBP를 사용하면서 MLP-I/II를

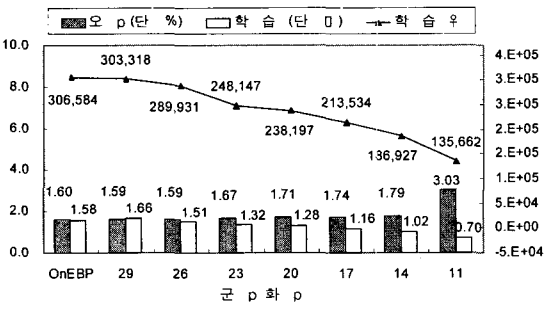


그림 8. MLP-I/II를 이용한 실험결과
Fig. 8. The experiment result of applying the MLP-I/II.

적용했을 때의 실험결과를 보인다. 이 결과에서는 실제로 학습에 투입되는 패턴수가 줄어드는 것을 확인해야 하므로 에폭수 대신 총학습기간 동안 학습에 사용된 패턴수를 기록한다.

이 결과에서 오류율이 높아지지 않는 수준에서 군집내 최소 화자수는 26명이고 0.19% 상승시 화자수는 14명이며, 각각의 경우에서 학습시간은 온라인 EBP에 비해 4.6%, 54.9% 향상되었다. 또한, 군집내 화자수가 배경화자 수와 같을 때(즉, 학습데이터 감축 효과가 없을 때) 화자군집 방법을 적용하지 않은 경우보다 5% 정도의 학습시간 지연이 관측되었는데, 이는 MLP-I의 적용에 따른 계산량의 증대 때문이다.

마지막으로 COIL과 MLP-I/II를 모두 이용한 실험결과를 그림 9에서 보인다.

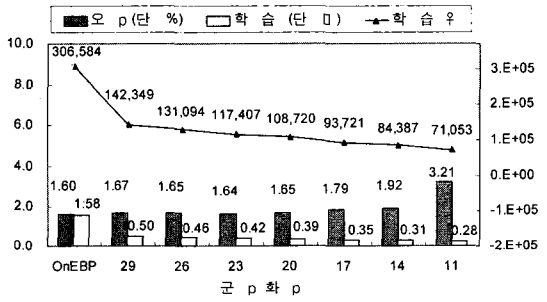


그림 9. COIL 및 MLP-I/II 동시 적용시 실험결과
Fig. 9. The experiment result of applying both the COIL and MLP-I/II.

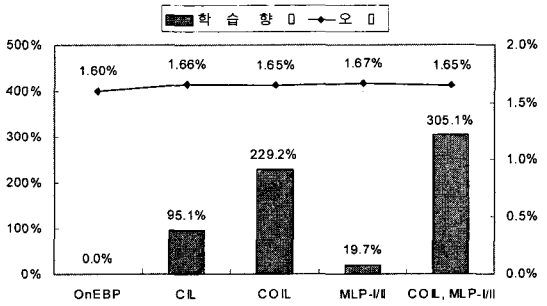


그림 10. 각 실험결과의 향상률 추이
Fig. 10. The transition of the experiment results.

이 결과에서 오류율이 0.05% 상승하는 수준에서 군집내 최소 화자수는 20명이고 0.19% 상승시 화자수는 17명이며, 각각의 경우에서 학습시간은 온라인 EBP에 비해 305.1%, 351.4% 향상되었다.

오류율이 1.6% 수준일 때 각 실험결과의 향상률을

그림 10에서 정리한다.

이 결과에서 볼 수 있듯이, COIL 방법과 MLP-I/II 방법이 동시에 적용될 경우 서로 시너지 효과를 발휘하여 개별적으로 적용될 때보다 높은 학습시간 향상률을 기록한다.

Ⅷ. 결 론

다른 패턴인식 방법에 비해 몇 가지 이점을 제공하는 MLP(multilayer perceptron)는 화자증명 시스템의 화자학습 및 인식 방법으로도 유망하다. 그러나 MLP의 학습은 학습에 이용되는 EBP(error backpropagation) 알고리즘의 저속 때문에 상당한 시간을 소요한다. 이 점은 화자증명 시스템에서 높은 화자인식률을 달성하기 위해서는 많은 배경화자가 필요하다는 점과 맞물려 시스템에 화자를 등록하기 위해 많은 시간이 든다는 문제를 낳는다. 이 점은 화자 등록후 곧바로 증명 서비스를 제공해야 하는 화자증명 시스템의 실시간 요구를 만족시키지 못한다. 본 논문에서는 이 문제를 해결하기 위해 EBP의 학습속도를 개선하는 COIL 방법과, 기존의 화자증명에서 화자군집 방법을 도입한 MLP-I/II 방법을 사용하여 MLP 기반 화자증명 시스템에서 화자등록에 필요한 시간의 단축을 시도하였다. 지속음을 화자인식 단위로 하는 MLP 기반의 화자증명 시스템에서 이 두 방법을 적용한 결과 각각 등록시간 향상을 확인하였고, 두 방법을 동시적용했을 때 기존의 온라인 EBP 학습 알고리즘을 사용했을 때보다 대략 4배 빠른 등록속도가 기록되었다.

한편, 실시된 실험결과에 의하면 COIL 방법보다 MLP-I/II의 속도향상폭이 작은 것으로 나타났는데, 이것은 실험의 화자증명 시스템에 사용된 배경화자의 수가 적어 MLP-I/II의 배경화자 감축효과가 작게 나타난 것으로 추측된다. 추후 보다 많은 배경화자를 적용하면 인식률의 향상과 함께 MLP-I/II의 더 큰 효과를 기대해 볼 수 있을 것이다.

참 고 문 헌

[1] Q. Li et al., "Recent Advancements in Automatic Speaker Authentication," IEEE Robotics & Automation Magazine, Vol. 6, pp. 24-34, Mar 1999.

[2] S. Furui, "An Overview of Speaker Recognition Technology," Automatic Speech and Speaker Recognition, Kluwer Academic Publishers, 1996.

[3] N. Morgan and H. Bourlard, "Hybrid Connectionist Models for Continuous Speech Recognition," Automatic Speech and Speaker Recognition, Kluwer Academic Publishers, 1996.

[4] S. Haykin, Neural Networks, Prentice Hall, 1999.

[5] Y. Bennani and P. Gallinari, "A Modular Connectionist Architecture for Text-Independent Talker Identification," International Joint Conference on Neural Networks, Vol. 2, pp. 857-860, Seattle, USA, 1991.

[6] N. Fakotakis and J. Sirigos, "A High Performance Text Independent Speaker Recognition System Based on Vowel Spotting and Neural Nets," IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 661-664, Atlanta, USA, 1996.

[7] A. E. Rosenberg, and S. Parthasarathy, "Speaker Background Models for Connected Digit Password Speaker Verification," IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 81-84, Atlanta, USA, 1996.

[8] H. Demuth and M. Beale, Neural Network Toolbox, The MathWorks, 2001.

[9] M. Riedmiller and H. Braun, "A Direct Adaptive Method for Faster Backpropagation Learning : The RPROP Algorithm," IEEE International Conference on Neural Networks, pp. 586-591, Vol. 1, San Francisco, USA, 1993.

[10] R. Fletcher, Practical Methods of Optimization, Wiley, 1987.

[11] M. Moller, "Supervised Learning on Large Redundant Training Sets," Proceedings of the 1992 IEEE-SP Workshop Neural Networks for Signal Processing, pp. 79-89, Helsingoer, Denmark, 1992.

[12] S. Becker and Y. LeCun, "Improving the Convergence of Back-Propagation Learning

with Second-Order Methods," Proceedings of the 1988 Connectionist Models Summer School, pp. 29-37, 1988.

[13] Y. Bengio, Neural Networks for Speech and Sequence Recognition, International Thomson Computer Press, 1995.

[14] Y. LeCun, "Generalization and Network Design Strategies," Technical Report CRG-TR-89-4, Department of Computer Science, University of Toronto, 1989.

[15] T. Matsui and S. Furui, "Likelihood Normalization for Speaker Verification Using a Phoneme-and Speaker-Independent Model," Speech Communication, Vol. 17, pp. 109-116, Aug 1995.

[16] A. L. Higgins et al., "Speaker Verification Using Randomized Phrase Prompting," Digital Signal Processing, Vol. 1, pp. 89-106, 1991.

[17] H. Gish, "A Probabilistic Approach to the Understanding and Training of Neural Network Classifiers," IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 3, pp. 1361-1364, Albuquerque, USA, 1990.

[18] D. R. Wilson and T. R. Martinez, "The Need for Small Learning Rates on Large Problems," International Joint Conference on Neural Networks, Vol. 1, pp. 115-119, Washington, USA, 2001.

[19] C. Becchetti and L. P. Ricotti, Speech Recognition, John Wiley & Sons, 1999.

[20] P. Cristea and Z. Valsan, "New Cepstrum Frequency Scale for Neural Network Speaker Verification," IEEE International Conference on Electronics, Circuits and Systems, Vol. 3, pp. 1573-1576, Pafos, Cyprus, 1999.

[21] M. Savic and J. Sorensen, "Phoneme Based Speaker Verification," IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 165-168, San Francisco, USA, 1992.

[22] R. P. Lippmann, "An Introduction to Computing with Neural Nets," IEEE Acoustics, Speech, and Signal Processing Magazine, Vol. 4, pp. 4-22, Apr 1987.

[23] D. P. Delacretaz and J. Hennebert, "Text-Prompted Speaker Verification Experiments with Phoneme Specific MLPs," IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 777 - 780, Seattle, USA, 1998.

저 자 소 개

李 百 永(正會員)

1969년 : 한국항공대학교 항공전자공학과 학사. 1982년 : 연세대학교 전자공학과 석사. 1999년 9월~현재 : 한국항공대학교 항공전자공학과 박사과정 재학. 1974년 3월~1990년 9월 : 롯데파이오니아㈜ 개발부장. 1990년 10월~1998년 12월 : 현대전자산업㈜ 전장연구소장(상무).

李 泰 承(正會員)

1997년 : 한국항공대학교 항공전자공학과 학사. 2000년 : 한국항공대학교 항공전자공학과 석사. 2000년 3월~현재 : 한국항공대학교 항공전자공학과 박사과정 재학. 2000년 3월~현재 : 화음소㈜ 재직.

黃 秉 元(正會員)

1972년 : 한국항공대학교 항공전자공학과 학사. 1981년 : 도쿄대학교 전자공학과 석사. 1984년 : 도쿄대학교 전자공학과 박사. 1973년~1984년 : 여수대학 교수. 1984년~1985년 : 국방과학연구소 연구원. 1985년~현재 : 한국항공대학교 전자정보통신컴퓨터공학부 교수.