차량 보안을 위한 어구독립 화자증명의 등록시간 단축에 관한 연구

A Study on the Fast Enrollment of Text-Independent Speaker Verification for Vehicle Security

이태승*, 최호진*

Tae-Seung Lee* and Ho-Jin Choi*

요 약

음성은 차량 운행시 여러 가지 조작으로 분주한 운전자에게 간편한 장비 입력 및 조작을 위한 수단으로 유용한 특성을 지니고 있다. 본 논문에서는 이런 음성의 특성을 이용하여 차량 도난이나 중요한 온라인 서비스 접근시 신원을 증명할 수 있는 화자증명 방식을 제안한다. 이 방식에서는 음성의 언어정보를 이용하는 지속음 인식 기법과 함께 확률적 인식 방식에 비해 몇 가지 이점을 갖는 MLP(multi-layer perceptron)를 사용한다. 하지만 MLP를 사용하는 인식 기법은 학습에 많은 계산량을 요구하므로 실시간으로 화자를 등록해야 하는 화자증명에서는 적용하기가 쉽지 않다. 이 문제를 해결하기 위해 본 논문에서는 기존의 화자점수 평준화 방법에서 화자군집 모델 기법을 도입하여 배경화자를 사전에 여러 개의 작은 화자군집으로 분리하는 방법을 제안한다. 등록화자를 이렇게 나눈 화자군집 중 하나로 분류한 뒤 해당 화자군집에 대해서만 등록 학습과정을 거치는 방법으로 계산량을 큰 폭으로 줄일 수 있다.

Abstract

Speech has a good characteristics of which car drivers busy to concern with miscellaneous operation can make use in convenient handling and manipulating of devices. By utilizing this, this works proposes a speaker verification method for protecting cars from being stolen and identifying a person trying to access critical on-line services. In this, continuant phonemes recognition which uses language information of speech and MLP(multi-layer perceptron) which has some advantages against previous stochastic methods are adopted. The recognition method, though, involves huge computation amount for learning, so it is somewhat difficult to adopt this in speaker verification application in which speakers should enroll themselves at real time. To relieve this problem, this works presents a solution that introduces speaker cohort models from speaker verification score normalization technique established before, dividing background speakers into small cohorts in advance. As a result, this enables computation burden to be reduced through classifying the enrolling speaker into one of those cohorts and going through enrollment for only that cohort.

* 한국항공대학교 항공전자공학과(Dept. of Avionics Eng., Hankuk Aviation Univ.)

· 논문번호 : 2001-1-1
· 접수일자 : 2000년 12월 5일

Ⅰ. 서 론

차량 내부의 GPS(global positioning system), 무선 전화 및 인터넷, 전자 통제장치 등과 같은 새로운 기술이 승차 시간을 조절하고 관리하는 방식을 향상시킬 목적으로 개발, 보급되고 있다. 이런 장비에서 발생하는 막대한 양의 정보를 관리하고 운전자가 적절히 반응할 수 있을 정도로 업무량을 유지하기 위해서는 최근 몇 년 동안 개발해 온 몇 가지 인터페이스의 장점을 취합하여 제공하는 -경량의 음성 능력을 갖춘 장비나 터치 스크린 같은- 장비를 기반으로 하는 해결책이 모색되어야 한다.

그런 해결책 가운데 음성은 운전자의 손과 눈을 운전 기능에 전적으로 할당하면서도 정보를 검색하고 필요한 명령과 메시지를 정보 장비에 전달할 수 있다는 장점 때문에 다른 인터페이스보다 우수한 방식으로 받아들여지고 있다. 음성을 활용할 수 있는 차내 기능으로는 크게 세 가지를 들 수 있다[1]. 첫째는 주행 경로 지시이다. 일부 고급 차량과 대여 차량에 현재 위치를 알려주고 특정 목적지까지의 경로를 배정하며 음성 안내까지 해주는 GPS 시스템이 장비되고 있다. 둘째는 정보 검색 기능이다. 이 경우 정보 검색이라 함은 교통량 정보나 휴게소, ATM의 위치, e-mail 메시지 같은 정보를 알려주는 지능적인 메커니즘을 가리킨다. 세째는 안전과 보안 기능이다. General Motors에서는 자동차 사고나 도난을 당했을 때 GPS와 휴대 전화를 이용하여 서비스 센터에 현재의 정확한 위치를 알려주는 OnStar 시스템을 개발하여 보급하고 있다[2].

본 논문에서는 이 중 세 번째의 안전과 보안에 사용할 수 있는 음성 기능으로 화자증명(speaker verification)을 제안한다. 화자증명은 음성파형에서 얻어진 정보를 바탕으로 현재 발성하고 있는 화자를 자동으로 인식하는 과정을 말한다. 이 기술을 사용하면 도용 가능성이 높은 기존의 방식 대신 개인의 생체적 특성을 반영하는 음성을 통해 다양한 서비스의 접근을 통제하는 개인 접근 시스템에서의 신원 확인이 가능하다. 차량에 적용할 수 있는 화자증명의 응용 예로, OnStar 시스템에서 차량의 문 잠금을 해제하는 방법으로 비밀번호나 암호문을 음성으로 입력할 때 화자의 미리 저장된 음성과 비교를 한다거나, 주행 중 헤드셋이 달린 휴대폰을 사용하여 온라인으로 은행 업무를 볼 때 화자를 증명하는 방식[3]을 들 수 있다.

인간의 음성은 언어정보를 기준으로 그 범위 내에서 화자간 차이를 보인다. 따라서 먼저 언어정보에 따른 분류를 수행한 뒤 다시 그 범위 내에서 화자증명을 처리하는 방법이 유리하다. 화자간 차이를 드러내는 성능을 여러 음소범주에 대해 조사한 기존 연구결과[4][5]에 의거하여 본 연구에서는 지속적인 음성부분이 화자간 인식에 많은 기여를 한다는 가정을 설정하고 각 지속음별로 화자증명을 수행하도록 한다. 이 방법은 구현 시스템을 어구독립(text-independent) 방식[6]으로 만들지만, 차후 음성인식 기술을 추가함으로써 어구지시(text-prompted) 방식[6]으로의 개선이 가능하다.

화자증명을 수행하는 인식기술로는 최근 음성인식 분야에서 활발히 연구되고 있는 MLP(multi-layer perceptron)[7]를 사용한다. MLP는 기존의 확률적 방법에 비하여 경쟁학습을 통한 높은 인식률과 확률분포에 대한 사전지식이 필요 없다는 점 때문에 관심을 모으고 있다.

하지만 MLP의 학습에는 많은 계산이 필요하다. 이 점은 오프라인에서 인식대상을 학습하는 음성인식에서는 그다지 큰 문제가 되지 않지만, 차재 장비와 같이 한정된 계산능력을 가진 장비에서 화자를 반드시 온라인에서 등록해야 하는 화자인식에서는 그렇지 못하다. 이에 따라 본 연구에서는 합리적인 계산량 수준에서 화자등록을 처리할 수 있도록 기존의 화자점수 평준화 방법[6]에서 고안된 화자군집 모델을 도입한다. 이 모델은 등록화자와 유사한 배경화자들로 군집을 형성하여 이들에 대해서만 평준화-MLP의 경우 학습-를 수행하게 해주어 군집의 개수에 반비례하게 계산량을 줄인다.

본 논문의 구성은 다음과 같다. Ⅰ장 서론에 이어 Ⅱ장에서 화자증명에 이용되는 음성의 화자간 차이를 설명하고, Ⅲ장에서 화자점수 평준화 방식을 소개하며, Ⅳ장에서 확률적 화자군집 모델을 MLP로 구현하는 방법을 제안한다. Ⅴ장에서는 음성 데이터베이스를 이용하여 세 가지 방법으로 수행한 실험 결과

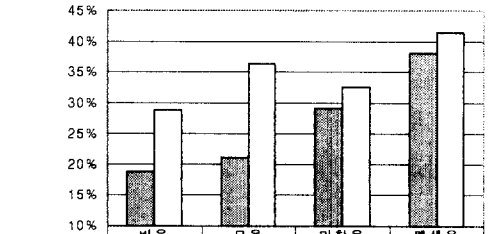
를 제시하고, Ⅵ장에서 본 연구의 결과를 정리한다.

Ⅱ. 음성의 화자간 차이 모델링

화자인식의 기본원리는 화자 사이에 존재하는 음향적 특성의 편차를 이용하여 각 화자를 구별하는 것이다. 화자의 음향적 특성은 크게 정적인 특성과 동적인 특성으로 구분할 수 있다. 정적 특성은 화자의 연령이나 성별 등에 따른 구강(oral cavity), 비강(nasal cavity), 성대(vocal cord)에서 만들어지는 음성에 의해 형성되는데, 기본주파수(fundamental frequency)와 포만트(formant)가 대표적인 정적 특성이다. 이에 비해 동적 특성은 화자의 억양, 강세, 빠르기 등에 의해 형성된다. 일반적으로 성대모사를 통해 다른 화자를 흉내낼 때 동적 특성을 비슷하게 만드는 것이므로 각 화자를 고유하게 구별하는 특징은 정적 특성이라고 할 수 있고, 이에 따라 정적 특성이 동적 특성보다 보안성 면에서 더 뛰어나다고 할 수 있다.

음성 신호가 가지는 정보는 발성하는 문장에 대한 언어정보와 화자정보를 동시에 포함하고 있다. 일반적으로 사용되는 단구간 스펙트럼과 같은 음향파라미터의 경우, 개인차가 음운특성에 따른 영향을 넘어설 정도로 크지 않아 음성으로부터 언어정보를 제거한 화자정보만을 추출하기는 어렵다. 따라서, 화자인식을 위해서는 언어정보를 완전히 제거하기보다는 화자특성을 언어에 무관한 것과 언어 종속적인 것으로 구분하여 이 둘을 동시에 사용하는 것이 분석이나 인식률 측면에서 바람직하다. 현재, 이러한 점을 인식 시스템에 반영하기 위해 여러 개의 문장을 발성하거나 음성인식을 이용하여 인식된 음소단위로 화자를 인식하는 방법 등이 고안되고 있다.

음성을 이루는 음운 부분에 따라서는 다른 부분에 비해 화자인식에 더 유용하다는 사실이 밝혀져 있다. 따라서 화자인식에 어떤 음소를 사용하는 것이 가장 좋은지 명확히 알 수만 있다면 인식성능을 향상시킬 수 있음이 분명하다. 예를 들어 양호한 음소의 비중이 높은 비밀번호를 선택한다면 이러한 특성의 효과를 기대할 수 있다. 또는, 화자인식에 유리한 음소를 자동으로 식별하는 전단(front end) 인식기를 사용할 수 있다면 최종 인식 단계에서 적절한 가중치를 부여



	비음	모음	마찰음	폐쇄음
Eatock	18.8%	21.1%	29.2%	36.1%
Delacretaz	28.8%	36.3%	32.6%	41.4%

그림 1. 음소 범주별 오류율

Fig., 1. Verifying error rate for each phonemic category.

함으로써 인식률 향상을 꾀할 수 있을 것이다.

Eatock 등은 영어의 각 음소와 이들이 이루는 범주에 따른 화자간 인식능력의 차이를 연구하였다. [4]. 이들에 따르면 비음과 모음이 가장 뛰어난 성능을 보이고 그 뒤를 마찰음과 폐쇄음이 잇는다. 이 사실은 Delacrtaz 등의 유사한 연구에 의해 뒷받침되고 있고[5], 그림 1에서 이들의 연구결과를 간단히 요약하였다.

본 연구에서는 언어정보 분포 내의 화자정보 분포를 이용하며, 음소범주에 따라 화자인식 능력이 달라진다는 결과에서 음이 지속되는 부분이 많은 범주일수록 화자인식 성능이 좋다고 가정한다. 이것은 빠르게 발성할 때보다 느리게 발성할 때 다른 화자를 인식할 수 있는 우리의 일상적인 청각인지 경험을 통해서도 추측할 수 있다. 이에 따라 본 연구에서는 지속적인 부분이 비교적 많은 비음, 모음, 마찰음에서 지속부분(이후 지속음)을 채취하여 화자간 인식의 기본단위로 사용한다. 이 방식이 갖는 장점으로 이후 음소범주별 가중치 설정이 가능하고 어구요구 화자증명으로의 개선이 용이하다는 점을 들 수 있다.

Ⅲ. 화자점수 평준화 방법

3-1 폐쇄화자형과 개방화자형 인식

폐쇄화자형(close set) 화자인식[6]에서는 아래의 식에서 의미하듯이 의뢰화자의 신원을 등록된 여러 화자 가운데 가장 높은 유사도 값을 갖는 화자를 선택한다.

$$\arg\max_{\forall i} P(S_i \mid X) = \frac{\arg\max_{\forall i} P(X \mid S_i)P(S_i)}{P(X)} \quad (1)$$

여기서, X는 의뢰화자의 음성신호이고, S_i는 유사도(likelihood)를 계산하는 화자를 나타낸다. 화자 증명에서는 신원을 주장하는 화자가 가장 높은 유사도 값을 갖는 경우 증명된 것으로 처리한다.

개방화자형(open set) 화자인식[6]에서는 이와 달리 의뢰화자의 참조 모델이 시스템에 없을 수도 있기 때문에 위의 방식으로 처리할 수 없다. 그 대신 전체 화자를 대표하는 한정된 배경화자를 설정하고 이들 가운데 주장화자를 선택한 뒤 의뢰화자, 주장화자, 배경화자 사이의 거리(유사도 차이)를 가늠하는 방법으로 화자를 인식한다. 이 방법을 화자점수 평준화 방법이라고 한다.

3-2 화자점수 평준화 방법

Higgins 등은 유사도 비율을 이용하는 거리(유사도) 값에 대한 평준화 방법을 제안했다[8]. 이 유사도 비율은 주장화자가 주어졌을 때의 관측된 발성열의 조건확률과 화자가 사칭자일 때 관측되는 조건확률의 비율로 정의된다. 이 비율의 수학적 표현은 다음과 같다.

$$\log L(X) = \log p(X \mid S = S_C) - \log p(X \mid S \neq S_C) \quad (2)$$

여기서, X는 의뢰화자의 음성열이고, S_C는 주장화자이다. 일반적으로 log L이 양일 때 실제화자를 나타내고 반대로 음일 때 사칭화자를 나타낸다. 식 (2)의 오른쪽의 두 번째 항목을 평준화 항이라고 한다.

참조화자 집단이 모든 화자를 대표한다고 가정할 때 주장화자 외의 모든 화자에 대해 X 지점의 확률밀도는 최근접 참조화자의 확률밀도로 좌우될 수 있다. 그러므로 다음과 같은 결정 공식을 유도할 수 있다.

$$\log L(X) = \log p(X \mid S = S_C) - \max_{S \in Ref, S \neq S_C} \log p(X \mid S \neq S_C) \quad (3)$$

이 식은 유사도 비율 평준화 방법이 Bayes의 최적점수 개념을 근사화한다는 것을 보여준다. 그러나 이 결정식은 두 가지 이유로 비현실적이다. 첫 번째 이유는 최근접 화자를 선택하기 위해서는 모든 참조화자에 대해 계산을 수행해야 하므로 과도한 계산이 요구될 수 있다는 점이고, 두 번째는 참조화자 내에서 최근접 화자가 얼마나 근접해 있는지에 따라 최대 조건확률 값이 화자마다 달라지기 쉽다는 점이다.

이에 대한 해결책으로 식(2)의 평준화 항을 계산하기 위한 화자군집(cohort)을 선택한다. Higgins 등은 주장화자 인근의 화자들을 대표하는 화자군집의 사용을 다음과 같이 제안했다.

$$\log L(X) = \log p(X \mid S = S_C) - \log \sum_{S \in Cohor, S \neq S_C} p(X \mid S) \quad (4)$$

반복적인 실험을 통해 이 평준화 방법으로 주장화자 모델만 사용하는 점수계산 방법에 비해 화자 구별력이 향상되고 화자종속 또는 어구종속 문턱값을 사용하지 않아도 된다는 것이 밝혀졌다[8].

Ⅳ. MLP 화자점수 평준화 방법

MLP는 음성인식 분야에서 이미 MLP가 가진 여러 가지 장점을 인정받아 단독[9]으로 또는 HMM (hidden morkov model)과의 혼합형태[10]로 채택되고 있다. MLP의 장점을 열거하면 아래와 같다.

- 유사도 비교 방식에 비해 경쟁 집단의 거부 학습이 가능하다.
- 입력 특징의 통계적 분포에 대한 사전 지식이 필요없다.
- 고도의 병렬성과 규칙성을 가지고 있어 고성능 하드웨어 구현이 용이하다.

이 같은 장점은 화자인식에서도 그대로 적용될 수 있으므로[11]~[13], 본 연구에서는 MLP의 위와 같은 장점과 화자점수 평준화 방법을 결합하고 MLP를 화자인식에 사용할 때 등록에 걸리는 학습시간을 단축하기 위해 사전정의 화자군집 모델을 MLP에 도입하였다.

4-1 MLP 개요

MLP는 입력계층과 출력계층 사이에 임의 개수(≥0)의 은닉계층이 자리하는 구조로 되어 있다. 입력계층을 제외한 각 계층은 하나 이상의 노드(뉴런)로 구성되는데, 이는 동작함수로 시그모이드(sigmoid) 함수를 사용하는 계산 단위이다. Oglesby 등의 연구에서 화자인식 응용의 경우 둘 이상의 은닉계층을 사용하더라도 성능에는 향상이 없다[14]는 것이 밝혀졌으므로 본 연구에서는 은닉계층이 하나만 있는 2층 MLP를 사용한다.

MLP의 사용은 대개 여러 군집의 학습 데이터를 사용하여 훈련시킨 뒤 시험 데이터를 입력하여 학습된 군집 중 하나로 분류하는 방식으로 이루어진다. 군집 지정은 목표 벡터 $d(x_n)$을 출력 노드에 배당함으로써 이루어지고, 화자인식에서 각 군집은 개별 화자가 된다. 일반적으로 d의 값으로는 0/1 또는 1/1이 지정된다.

훈련에 사용되는 오류 측정 기준은 다음과 같다.

$$E = \sum_{n=1}^{N} \| g(x_n) - d(x_n) \|^2 \qquad (5)$$

여기서 $g()$는 입력 데이터 벡터 x_n에 대해 MLP에서 동작하는 비선형 벡터 함수이고, N은 출력노드의 총수를 나타낸다. MLP의 파라미터(가중치 벡터)는 식(5)의 오류를 최소화하기 위한 기울기 감소 과정을 통해 반복적으로 갱신된다[15]. 가중치의 갱신은 훈련 과정 동안 입력이 주어질 때마다 이루어진다.

4-2 광역모델 MLP 화자증명

화자증명에서 MLP를 이용하기 위해서는 먼저 등록화자와 배경화자의 데이터를 구분하고 MLP가 이 둘을 구분하도록 학습시켜야 한다. 이 경우 하나의 출력노드를 사용하여 등록화자의 데이터에 대해서는 긍정 목표치(1)를 부여하고 배경화자에 대해서는 부정 목표치(0 또는 -1)를 부여한다. 학습 후 의뢰화자의 데이터가 입력될 때 출력노드의 값이 이 두 목표치의 중간보다 크면 수락된 것으로, 작으면 거부된 것으로 판정한다. 배경화자의 수와 각 화자의 발성 데이터가 많을수록 화자증명 성능이 좋게 나온다. 이와 같이 전체 화자를 대표하는 배경화자가 모두 학습에 사용되기 때문에 본 연구에서는 이 방식을 광역모델이라고 정의한다.

그러나, 화자를 등록할 때는 여건상 충분한 화자 데이터를 얻을 수 없다. 충분한 화자 데이터가 확보되지 않으면 등록화자와 배경화자 사이의 정확한 경계를 찾기 힘들다. 이에 대한 대안으로 등록 전에 배경화자 집단 내의 충분한 데이터를 이용하여 오인수락률과 오인 거부율이 같아지는 EER(equal error rate)을 달성하는 문턱값을 결정한 뒤 등록 후 증명에 이를 사용하는 방법을 생각할 수 있다. 즉, 배경화자 각각에 대해 한 번씩 등록화자로 지정하고 제한된 데이터로 MLP를 학습시킨 후 남은 데이터를 통해 EER을 달성하는 문턱값을 결정하는 것이다. 이 때 등록화자로 지정된 배경화자마다 문턱값이 다를 것인데, 통일된 문턱값을 얻기 위해 평균을 취하는 등 여러 방법을 사용할 수 있다.

MLP 학습은 이와 같이 문턱값을 결정하는 과정과 화자를 등록하는 과정으로 나뉘므로 본 연구에서는 전자를 오프라인 학습이라고 하고 후자를 온라인 학습이라고 지시한다.

학습이 끝나면 의뢰화자의 데이터를 검증할 수 있게 된다. 본 연구에서 화자인식을 위한 기본 모델로 지속음을 사용하므로 각 지속음마다 MLP가 할당되게 된다. 검증과정에서는 의뢰화자의 지속음을 해당 MLP에 입력하여 전체 입력 프레임에서 문턱값을 넘는 프레임의 비율이 50% 이상일 때 등록화

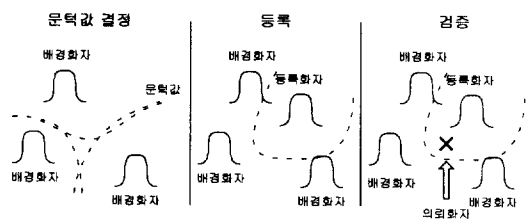


그림 2. 광역모델 MLP 화자증명 과정

Fig. 2. Speaker verifying scheme for global speaker MLP modeling.

자로 수락하고 이하일 때 거부한다. 이 과정을 그림 2에서 설명하고 있다.

MLP를 이용한 화자증명은 등록화자와 배경화자 사이의 차이를 학습한다는 점에서 Ⅲ장의 화자점수 평준화 방법과 유사한 효과를 갖는다.

4-3 지역모델 MLP 화자증명

전술한 바와 같이 광역모델에서 충분한 인식률을 달성하기 위해서는 전체 화자를 충분히 대표할 수 있는 화자가 필요하다. 하지만 충분한 수의 화자 데이터를 사용할 경우 등록 시간이 너무 길어지며 실용 조건에서는 이런 장시간의 등록이 허용되지 않는다. 더군다나 본 연구에서는 지속음을 인식모델로 사용하기 때문에 사용하는 지속음이 많아질수록 등록 학습이 필요한 MLP가 증가하여 문제를 더욱 가중시킨다.

이 문제에 대한 해결책으로 화자점수 평준화 방법에서 적용했던 화자군집 모델과 유사한 방법을 MLP에 도입한다. 즉, 충분한 화자로 구성된 전체 배경화자를 서로 유사한 화자들로 이루어진 군집으로 재구성하여 화자 등록과 증명 처리를 해당 군집에서만 수행하게 한다는 것이다. 이렇게 하면 인식률의 큰 하락없이 등록시간을 단축할 수 있다. 본 연구에서는 이 방법을 기본지역모델이라고 정의한다. 이 모델에서 학습에 사용될 군집을 선택하는 방법과 의뢰화자의 검증시도를 그림 3에서 예시한다.

기본지역모델의 유효성은 사전에 각 군집을 얼마나 적절히 형성하는가에 달려있다. 형성된 군집들이 전체 화자 영역을 완전히 포함하고, 각 군집에 속하는 배경화자의 수가 해당 군집 내에서 등록화자와의 차이를 양호하게 나타낼 수 있을 만큼 충분해야 한다. 기본지역모델의 화자군집은 다음과 같은 과정을 통해 형성할 수 있다.

(1) 충분히 큰 화자집단을 학습용과 시험용으로 나눈다.

(2) 학습용을 다시 무작위로 이등분하여 A집단과 B집단으로 나눈다.

(3) A집단 내의 화자마다 MLP의 출력 노드를 할당하여 학습시킨 뒤 인식률이 일정 수준을 넘는 것만 남기고 나머지 화자를 B집단에 넣는다.

(4) B집단의 각 화자를 학습된 MLP에 입력하여 각 출력 노드의 값을 확인한다. 이 때 출력노드 값이 일정 수준에 못 미치는 것을 MLP의 새로운 출력 노드에 할당하여 재학습시킨다.

(5) 새로 추가된 노드의 인식률이 단계 (3)의 수준에 미치지 못하면 이 화자를 다시 B집단에 넣는다.

(6) 더 이상 MLP의 출력노드로 추가할 수 있는 화자가 없을 때까지 단계 (4)~(5)를 반복한다.

이와 같이 화자집단을 형성하는 데 사용되는 MLP를 MLP-1이라고 한다. MLP-1에서 각 출력노드는 개별 화자군집과 대응한다(그림 4). MLP-1으로 분류된 군집 내에서의 화자증명을 처리하는 MLP는 MLP-2라고 정의한다. MLP-2는 배경화자가 해당 군집 내의 화자로 한정된다는 점을 제외하면 광역

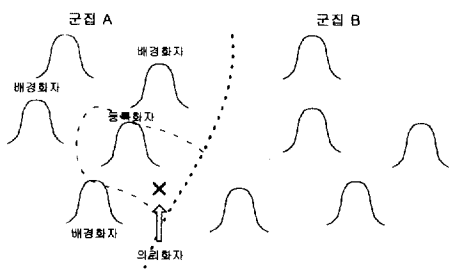


그림 3. 기본지역모델 MLP 화자증명

Fig. 3. Speaker verifying scheme for narrow local speaker MLP modeling.

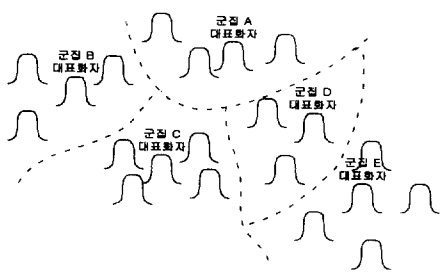


그림 4. 기본지역모델 MLP-1을 통해 형성된 화자군집

Fig. 4. Speaker cohorts arranged by MLP-1 for narrow local areas.

모델의 MLP와 학습과 증명 방법에서 동일하다. 즉,오프라인 학습을 통해 문턱값을 결정하고 온라인 학습으로 화자를 등록한 뒤 의뢰화자의 데이터에 문턱값을 적용하여 수락 여부를 결정한다.

기본지역모델의 동작 과정을 서술하면 다음과 같다. 화자등록 단계에서는 MLP-1을 이용하여 등록화자의 군집을 선택하고, 선택된 군집 내의 배경화자와 등록화자의 데이터를 이용하여 MLP-2를 학습시킨다. 화자증명 단계에서는 역시 MLP-1을 이용하여 의뢰화자의 군집을 판별하고, 판별된 군집이 등록화자의 군집과 같은 경우 MLP-2에 다시 의뢰화자의 데이터를 입력하여 MLP-2의 출력결과가 문턱값을 넘는지의 여부에 따라 수락을 결정한다.

4-4 확장지역모델 MLP 화자증명

군집모델의 유효성은 군집 내 배경화자의 상호관계가 충분히 반영될 수 있도록 의뢰화자의 위치가 특정 화자군집의 중앙부근에 있을 것이란 가정 하에 유지된다. 그러나 실제로는 군집의 가장자리에 등록화자가 위치할 수 있으므로 등록화자의 분포가 여러 인접 화자군집에 걸쳐 있을 수 있다. 지역모델에서는 해당 화자군집 내에서만 처리가 이루어지므로 이 경우 오인 거부율(false acceptance)이 급격히 증가하게 된다.

이 문제의 가장 직관적인 해결 방법은 등록화자의 분포를 일정수준 이상 포함하는 화자군집들을 하나의 새로운 화자군집으로 재설정하여 이에 대해 등록화자를 학습시키는 것이다. 하지만 등록학습시간은 군집 내 배경화자의 수에 비례하므로 여러 군집을 통합할 때 지나치게 많은 배경화자가 포함될 수 있다. 따라서 화자군집을 형성할 때 군집의 수는 최대한 많게 하는 대신 군집 내 배경화자의 수는 최대한 적게 만들어야 한다. 이것은 화자군집을 형성하는 과정에서 단계 (3)과 (4)의 수준을 적절히 조정함으로써 해결한다. 그림 5에서는 군집 D와 E의 경계에 등록화자가 위치할 때 이 두 군집을 하나의 군집으로 재설정하는 모습을 보인다.

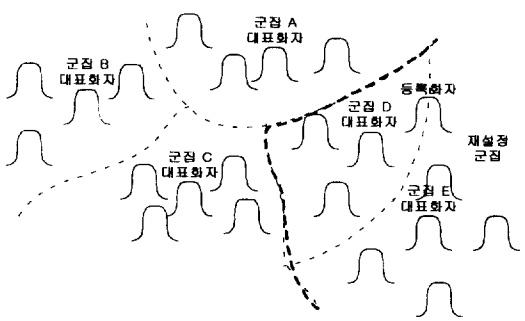


그림 5. 확장지역모델에서 군소 화자군집 통합

Fig 5. Integrating narrow local areas at extended local speaker MLP modeling.

V. 실 험

본 연구에서 제안한 방법을 시험하기 위해 한정된 범위에서 실험을 실시하였다. 이 실험에서는 지속음 가운데 15명의 한국어 /a/ 모음을 사용해 9명의 남성화자를 검증한다.

5-1 데이터베이스

실험에 사용한 데이터베이스는 지속음인식을 위한 것과 화자인식을 위한 것으로 구별된다.

지속음인식용은 ETRI에서 제작한 611DB이며 남성화자 3명의 611개 PBW(phone balanced word) 발성을 한국어 음소기준에 맞춰 레이블링해 놓은 것이다. 이 실험에서는 이들 음소 중에서 /a/ 음을 추출하도록 지속음 추출기의 MLP를 학습시켰다.

화자인식용은 ETRI 611DB와 함께 ETRI 445 PBW DB를 사용하였다. ETRI 445 PBW DB는 음소 레이블링이 되어 있지 않기 때문에 미리 학습시킨 지속음인식 MLP로 /a/ 모음을 추출하여 사용했다. 실험에 사용된 화자는 611DB의 3명에 445DB의 21명을 더한 24명이고, 이중 배경화자에 15명을 할당하고 남은 9명을 화자증명 시험에 사용하였다. 지역모델에서 화자군집은 배경화자 5명씩 3개를 만들었다. /a/ 모음은 배경화자 당 97개와 증명화자 당 130개를 사용했으며, 학습에 사용한 /a/ 모음은 화자 당 10개이다.

음성특징은 16Bit, 16kHz로 샘플링되어 있는 각

음성에 32ms의 Hamming 필터를 10ms마다 씌워 프레임을 만든 후 각 프레임에서 16차 Mel-scaled 필터뱅크 특징을 계산하였다[16].

5-2 시스템 구조

본 논문에서는 화자증명 문제만 다루고 있으므로 각 화자의 음성에서 음소를 인식하여 추출하는 부분의 설명을 생략한다. 음소추출 부분을 포함하여 시스템은 4-2에서 설명했듯이 각 음소마다 화자군집을 결정하는 MLP(MLP-1)와 재구성된 군집에 대해 학습되는 MLP(MLP-2) 부분으로 구성된다. MLP-1은 사전에 미리 학습시키며 재구성 군집에 대한 MLP-2는 화자 등록시 학습시킨다.

이 실험에서 사용된 MLP는 기본적인 MLP의 구조와 오류 역전파 알고리즘과 달리 입력 계층에 특징 프레임 3개 만큼의 시간지연을 두고, 한 패턴이 프레임마다 차례로 모두 입력되었을 때 MLP 내의 전체 가중치 연결이 갱신되는 방식을 취한다. 이들은 입력 계층 지연이 3이고 첫 번째 은닉 계층 지연이 1인 TDNN(time-delay neural network)[17]의 방식을 도입한 것이다. 입력 시간지연의 효과로 인식률이 향상되는 것을 [5]에서 확인하였으며, 위의 가중치 갱신 방식은 각 모음에 대한 시간 차이로 인해 가중치 갱신 기회에 차이가 나는 것을 막기 위해서이다.

5-3 실험결과

실험은 세 방법으로 나누어 실시하였다.

(1) 기본지역모델

(2) 확장지역모델

(3) 광역모델

광역모델의 실험결과는 나머지 두 방법의 기준이 된다. 지역모델의 MLP-1은 20개의 은닉노드와 3개의 출력노드를 가지고 3개의 화자군집을 모델링한다. MLP-2는 40개의 은닉노드와 1개의 출력노드를 갖는다. 이에 비해 광역모델의 MLP는 40개의 은닉노드와 1개의 출력노드를 갖는다. 각 은닉노드의 수는 반복실험을 통해 결정했다.

시험화자 9명에 대한 오류율은 오인 거부율과 오인 수락률로 나누고 /a/ 모음개수를 1에서 5개까지 늘려가며 측정하여 표 1에 정리해 놓았다.(이 표에서 FR은 오인 거부율을, FA는 오인 수락률을 나타낸다.)

결과에서 볼 수 있듯이 인식률 면에서 광역모델에 비해 기본지역모델의 오인 거부율이 크게 상승한 반면 확장지역모델에서 상당 부분 복원되는 것을 알 수 있다. 전반적인 오인 수락률이 높은 것은 사용한 데이터베이스의 경쟁학습 화자수가 부족하기 때문인 것으로 보인다. 계산량 면에서는 광역모델을 기준으로 등록에 사용된 배경화자의 수가 기본지역모델의 경우 1/3, 확장지역모델의 경우 2/3으로만 한정되므로 이들의 학습에 필요한 계산량의 감소를 확인하였다.

Ⅵ. 결 론

본 연구에서는 차량 운행시 안전과 보안에 사용할 어구독립 화자증명 시스템의 개발을 위해 지속음 단위의 화자인식모델과 인식률과 제약 독립성 면에서 우수한 MLP를 사용하였다.

인간의 음성은 언어정보를 기준으로 그 범위 내

표 1. 실험 결과

Table 1. Experiment results.

/a/ 수	기본지역모델		확장지역모델		광역모델	
	FR	FA	FR	FA	FR	FA
1	40.6%	11.4%	14.3%	14.1%	10.8%	12.8%
2	35.6%	11.9%	9.8%	14.0%	7.3%	11.2%
3	32.2%	11.7%	7.1%	13.9%	5.1%	10.6%
4	28.0%	11.3%	6.1%	13.9%	5.0%	10.3%
5	29.4%	10.4%	4.4%	13.8%	4.8%	10.3%

에서 화자간 차이를 보인다. 이에 따라 먼저 언어정보에 따른 분류를 수행한 뒤 다시 그 범위 내에서 화자증명을 처리하는 방법이 유리하다. 화자간 차이를 드러내는 성능을 여러 음소범주에 대해 조사한 기존 연구결과에 의거하여 본 연구에서는 지속적인 음성부분이 화자간 인식에 많은 기여를 한다고 설정하고 각 지속음별로 화자증명을 수행하도록 하였다. 이 방법은 구현 시스템을 어구독립 방식으로 만들지만, 차후 음성인식 기술을 추가함으로써 어구지시 방식으로의 개선이 가능하다.

화자증명을 수행하는 인식기술로는 최근 음성인식 분야에서 연구되고 있는 MLP를 사용하였다. MLP는 기존의 확률적 방법에 비하여 경쟁학습을 통한 높은 인식률과 확률분포에 대한 사전지식이 필요 없다는 점 때문에 관심을 모으고 있다. 하지만 MLP의 학습에는 많은 계산이 필요하다. 이 점은 오프라인에서 인식대상을 학습하는 음성인식에서는 그다지 큰 문제가 되지 않지만, 화자가 반드시 온라인에서 등록해야 하는 화자인식에서는 그렇지 못하다. 이에 따라 본 연구에서는 합리적인 수준에서 화자등록을 처리할 수 있도록 기존의 화자점수 평준화 방법에서 고안된 화자군집 모델을 도입하였다. 이 모델은 등록화자와 유사한 배경화자들로 군집을 형성하여 이들에 대해서만 평준화(MLP의 경우 학습)를 수행하게 한다. 이로써 모든 배경화자를 사용하는 방법에 비해 인식률의 차이는 적고 계산량을 크게 감소시킬 수 있었고 실험을 통해 이 사실을 확인하였다.

차후 연구과제로 차량주행 환경에서의 환경변화에 강인한 시스템 특성을 개발하는 것과 다른 방법을 통해 화자등록시간을 더욱 단축하는 방법을 계획하고 있다.

참 고 문 헌

[1] L. Gwennap, "Linley on Linux: Linux on Wheels: A New Opportunity", Linuxjournal, SSC Publication, Aug., 2000.

[2] http://www.onstar.com

[3] L. Boves and E. den Os, "Speaker Recognition in Telecom Applications", *IEEE 4th Workshop on Interactive Voice Technology for Telecommunications Applications*, pp. 203-208, 1998.

[4] J. P. Eatock and J. S. Mason, "A Quantitative Assessment of the Relative Speaker Discriminating Properties of Phonemes", *ICASSP*, vol. 1, pp. 133-136, 1994.

[5] D. P. Delacretaz and J. Hennebert, Text-Prompted Speaker Verification Experiments with Phoneme Specific MLPs", *ICASSP*, vol. 2, pp. 777-780, 1998.

[6] S. Furui, "An Overview of Speaker Recognition Technology", Automatic Speech and Speaker Recognition, Kluwer Academic Publishers, pp. 31-56, 1996.

[7] R. P. Lippmann, "An Introduction to Computing with Neural Nets", *IEEE ASSP Magazine*, Apr., 1987.

[8] A. L. Higgins, et al., "Speaker Verification Using Randomized Phrase Prompting", *Digital Signal Processing*, vol. 1, pp. 89~106, 1991.

[9] T. Zeppenfeld, et al., Improving the MS-TDNN for Word Spotting, *ICASSP*, vsol. 2, pp. 475-478, 1993.

[10] M. Franzini, et al., Connectionist Viterbi Training: A New Hybrid Method for Continuous Speech Recognition, *ICASSP*, vol. 1, pp. 425-428, 1990.

[11] N. Fakotakis and J. Sirigos, A High Performance Text Independent Speaker Recognition System Based on Vowel Spotting and Neural Nets, *ICASSP*, vol. 2, pp. 661- 664, 1996.

[12] H. Liou and R. J. Mammone, Speaker Verification Using Phoneme-Based Neural Tree Networks and Phonetic Weighting Scoring Method, *Proceedings of the 1995 IEEE Worksohp Neural Networks for Signal Processing* V, pp. 213-222, 1995.

[13] J. M. Naik and D. M. Lubensky, A Hybrid HMM-MLP Speaker Verification Algorithm for Telephone Speech, *ICASSP*, vol. 1, pp. 153-156, 1994.

[14] J. Oglesby and J. S. Mason, Optimization of Neural Models for Speaker Identification, *ICASSP*, pp. 216-264, 1990.

[15] R. P. Lippmann, An Introduction to Computing with Neural Nets, *IEEE ASSP Magazine*, Apr., 1987.

[16] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[17] A. Waibel, et al., Phoneme Recognition Using Time-Delay Neural Networks, *ASSP*, vol. 2, pp. 475-478, 1993.

이 태 승(李泰承)

1974년 6월 11일생

1997년 2월 : 한국항공대학교 항공전자공학과(공학석사)

2000년 3월~현재 : 한국항공대학교 대학원 항공전자공학과 박사과정 재학 중

관심분야 : 음성인식, 화자인식, 패턴인식, 자연언어처리, 인공 지능

최 호 진(崔虎珍)

1959년 3월 16일생

1982년 2월 : 서울대학교 전자계산기공학과(공학사)

1985년 12월 : 영국 Univ. of Newcastle, Computing Laboratory (MSc)

1995년 12월 : 영국 Imperial College, Dept. of Computing(Ph.D)

1982년 6월~1989년 8월 : 한국 데이터통신(주) 정보통신연구소 선임연구원

1995년 9월~1996년 11월 : Imperial College, Planning Applications Research Centre 연구원

1997년 3월~현재 : 한국항공대학교 항공전자공학과 조교수

관심분야 : 인공지능, 논리프로그래밍, 구속조건 만족문제, 소프트웨어 공학