말소리와 음성과학 제7권 제4호 (2015.12.31)
ISSN 2005-8063, pp. 9~16
ksss.jams.or.kr
http://dx.doi.org/10.13064/KSSS.2015.7.4.009

깊은 신경망 특징 기반 화자 검증 시스템의 성능 비교

Performance Comparison of Deep Feature Based Speaker Verification Systems

김 대 현[1] · 성 우 경[2] · 김 홍 국[3]
Kim, Dae Hyun · Seong, Woo Kyeong · Kim, Hong Kook

ABSTRACT

In this paper, several experiments are performed according to deep neural network (DNN) based features for the performance comparison of speaker verification (SV) systems. To this end, input features for a DNN, such as mel-frequency cepstral coefficient (MFCC), linear-frequency cepstral coefficient (LFCC), and perceptual linear prediction (PLP), are first compared in a view of the SV performance. After that, the effect of a DNN training method and a structure of hidden layers of DNNs on the SV performance is investigated depending on the type of features. The performance of an SV system is then evaluated on the basis of I-vector or probabilistic linear discriminant analysis (PLDA) scoring method. It is shown from SV experiments that a tandem feature of DNN bottleneck feature and MFCC feature gives the best performance when DNNs are configured using a rectangular type of hidden layers and trained with a supervised training method.

Keywords: speaker verification, deep neural network, tandem feature

1. 서론

최근 들어 다양한 이점과 응용 분야를 갖는 화자 검증 시스템의 중요성이 대두되고 있다[1], [2]. 화자 검증은 접근성이 높아 신체가 불편한 사람들도 쉽게 사용 가능하며, 목소리 샘플 매칭 등을 이용한 과학 수사에 사용될 수 있다. 또한, 토론회와 같은 여러 화자의 음성이 존재하는 경우에 각 화자에 대한 음성을 인지하는 오디오 인덱싱 기술에 응용이 가능하다.

그 동안 화자 검증 시스템을 위한 여러 연구가 수행되었다. 초기에는 가우시안 혼합 모델-공통 배경 모델(GMM-UBM: Gaussian mixture model-universal background model) 기반의 화자 검증 시스템이 주를 이루었다. 이 후, 공동 요인 분석(JFA: joint factor analysis) 방법을 이용한 화자 검증이 도입되었다[3], [4]. 근래에는 Dehak에 의한 I-벡터 방법이 제안되었다[5]. 이러한 I-벡터 기반 화자 검증 시스템에서는 일반적으로 mel-frequency cepstral coefficient (MFCC)를 특징 벡터로 하여 I-벡터 추출에 사용한다. 최근에는 I-벡터 기반 화자 검증 시스템에서 딥러닝(deep learning) 기법을 적용한 깊은 신경망(deep neural network) 기반 특징 추출에 대한 연구가 수행되었다[6]. 하지만, 깊은 신경망에 대해 화자 검증에 최적화된 훈련 방법, 은닉층 수, 은닉층 별 유닛수와 같은 깊은 신경망의 구조 및 깊은 신경망 기반 특징 벡터와 타 특징 벡터와의 결합 여부에 따른 변화 등에 대한 연구가 다소 부족한 편이다.

각각을 살펴보면 은닉층 수의 경우, 감독 훈련 시 은닉층 수가 증가할수록 성능은 저하된다. 이는 은닉층 수가 증가할수록 각각의 은닉층은 더 적은 정보를 담고 있기 때문이다[7]. 은닉층 별 유닛수의 경우, 은닉층 별 유닛수가 클수록 높은 성능을 보인다. 이는 은닉층 별 유닛수가 클수록 깊은 신경망의 비선형적 분별력이 강화되기 때문이다[8]. 은닉층 구조의 경우, 크게 병목 구조와 직사각형 구조 두 가지로 나뉜다. 직사각형 구조가 병목 구조 대비 높은 성능을 보인다. 왜냐하면 병목 구조의 경우 낮은 차수로 인해 깊은 신경망 훈련 시 비선형적 분별력이 떨어지는 현상이 발생하기 때문이다[8]. 특징 벡터 결합 여부의 경우, 단일 특징 벡터보다는 탠덤 특징 벡터를 이용하

1) 광주과학기술원, wmelonw88@gmail.com
2) 광주과학기술원, wkseong@gist.ac.kr
3) 광주과학기술원, hongkook@gist.ac.kr, 교신저자

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2015년도 문화기술연구개발지원사업의 연구결과로 수행되었음.

접수일자: 2015년 8월 29일
수정일자: 2015년 11월 15일
게재결정: 2015년 12월 9일

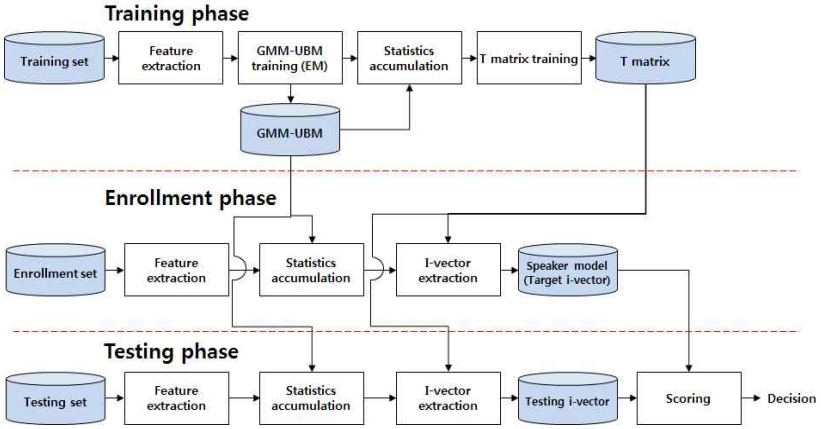

그림 1. I-벡터 기반의 화자 검증 시스템
Figure 1. I-vector based speaker verification system

였을 경우 높은 성능을 보인다. 왜냐하면 MFCC에서 발생하는 에러 패턴과 깊은 신경망으로부터 추출된 특징 벡터에서 발생하는 에러 패턴이 다르게 나타나는데, 이 때 두 가지 특징 벡터를 조합을 통해 에러 패턴이 상호 보완되기 때문이다[9].

따라서 본 논문에서는 I-벡터를 기반으로 하는 화자 검증 시스템을 구축하고, 깊은 신경망 기반 특징 추출에 따른 화자 검증 성능의 비교 연구를 수행한다. 또한 이를 위해 I-벡터와 확률적 선형 판별 분석 (PLDA: probabilistic linear discriminant analysis) 평가 방법[6]을 기반으로 화자 검증 시스템을 구축한다. 다음으로는 최적화된 화자 검증 시스템을 이용하여 최적의 깊은 신경망 특징 추출을 위한 깊은 신경망 훈련 방법, 깊은 신경망 구조 및 깊은 신경망 기반 특징 벡터와 타 특징 벡터와의 결합 여부에 따른 비교 실험을 수행한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 기존 I-벡터 기반의 화자 검증 기법 및 화자 검증을 위한 깊은 신경망 특징 벡터에 대해 서술한다. 3장에서는 제2장에서 기술된 방법들에 대한 비교 연구를 수행한다. 4장에서는 실험 환경을 기술하고 비교 실험을 수행한 후, 그 결과에 대해 논의한다. 마지막으로 5장에서는 본 논문의 결론을 맺는다.

2. I-벡터 기반 화자 검증 기법 및 깊은 신경망 특징 벡터

2.1 I-벡터 기반 화자 검증 기법

2.1.1 I-벡터 추정

<그림 1>은 I-벡터 기반의 화자 검증 시스템의 학습과 이를 이용한 화자 검증 방식을 보여 준다. 우선 훈련 과정에서는 훈련 데이터베이스(DB: database)를 이용하여 특징 벡터를 추출하고 이를 이용해 GMM-UBM을 모델링한다. 이 후 GMM-UBM과 특징 벡터를 이용하여 통계 정보를 수집하고 이를 토대로 전체 가변 행렬(T matrix: total variability matrix)을 모델링한다. 등록 과정에서는 등록된 발화를 이용해 특징 벡터를 추출한다. 이 후 훈련 과정에서 생성한 GMM-UBM과 특징 벡터를 이용하여 통계 정보를 수집한 후 수집된 통계 정보와 훈련 과정에서 생성한 전체 가변 행렬을 이용하여 I-벡터를 추출한다. 이를 화자 모델 또는 타겟 I-벡터라고 한다. 평가 과정은 등록 과정과 마찬가지의 방법으로 평가 발화에 대해 I-벡터를 추출하고, 등록 I-벡터와 평가 I-벡터를 이용해 스코어링을 하여 사전 설정된 문턱치값(threshold)에 따라 등록 발화와 평가 발화의 일치 여부를 결정하게 된다.

I-벡터는 전체 가변 행렬을 이용해 GMM 슈퍼 벡터를 낮은 차수로 표현한 것이다[10]. 화자 및 채널 종속 GMM 슈퍼 벡터 $\mathbf{M}(s)$은 수식 (1)과 같이 표현된다.

$$\mathbf{M}(s) = \mathbf{m} + \mathbf{T}\mathbf{w}(s) \tag{1}$$

여기서, s는 각각의 화자를 나타내며, $\mathbf{m}$은 화자 및 채널 독립 배경 UBM 슈퍼 벡터이다. 또한 $\mathbf{T}$는 전체 가변 행렬로서 낮은 랭크만으로도 수집된 개발 데이터 간의 분별력을 나타낸다. 이때 $\mathbf{T}$는 $CD \times R$의 차원으로 표시되며, 여기서 C는 전체 가우시안 혼합 개수, D는 특징 벡터의 차수, R은 전체 가변 행렬의 랭크로서 $R \leq S$(등록 화자 수)가 된다. 하지만, 등록 화자 수와 같은 크기의 랭크는 모든 정보를 담기에 충분하기 때문에 대개 $R = S$로 설정한다[11]. 수식 (1)에서 $\mathbf{w}(s)$는 정규분포 $N(\mathbf{0},\mathbf{I})$를 갖는 I-벡터를 의미한다. 한 명의 화자의 하나의 발화에 대해 수식 (2)와 같이 한 개의 고유한 I-벡터를 추출할 수 있다.

$$\mathbf{w}(s) = \mathbf{I} + (\mathbf{T}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{N}\mathbf{T})^{-1}\mathbf{T}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{F} \tag{2}$$

여기서, $\mathbf{I}$는 $CD \times CD$ 차원의 단위행렬이며, $\mathbf{N}$은 0차 통계정

보를 갖는 $D \times D$ 크기의 대각 행렬을 나타낸다. $\mathbf{F}$는 정규화된 1차 통계정보를 직렬로 이어 생성한 슈퍼 벡터이다. 공분산 행렬 $\boldsymbol{\Sigma}$는 전체 가변 행렬에서 포착되지 않은 잔여값들을 나타낸다. 전체 가변 행렬의 추정 방법은 참고문헌[5]에 상세히 기술되어 있다.

2.1.2 PLDA를 이용한 I-벡터 화자 검증 평가

타겟 I-벡터 $\mathbf{w}_{target}$와 평가 I-벡터 $\mathbf{w}_{test}$가 주어졌을 때, 화자 검증 점수는 수식 (3)과 같이 두 개의 가설$\{H_0, H_1\}$의 로그-우도비로 계산할 수 있다[12]-[14].

$$score = \log p(\mathbf{w}_{target}, \mathbf{w}_{test}|H_0) - \log p(\mathbf{w}_{target}, \mathbf{w}_{test}|H_1)$$
$$= \log p(\mathbf{w}_{target}, \mathbf{w}_{test}|H_0) - \log \{p(\mathbf{w}_{target}|H_1)p(\mathbf{w}_{test}|H_1)\} \quad (3)$$

여기서, H_0와 H_1은 $\{\mathbf{w}_{target}, \mathbf{w}_{test}\}$이 같은 화자에 속해 있을 경우와 서로 다른 화자에 속해 있을 경우에 대한 가설을 각각 의미한다. 수식 (3)의 로그-우도비에 대해 PLDA 기법을 적용하면 아래와 같이 표현된다.

$$score = \frac{1}{2}\left[\sum_{j=1}^{2}\mathbf{w}_j'\right](2\mathbf{K}+\mathbf{I})^{-1}\left[\sum_{j=1}^{2}\mathbf{w}_j'\right] - \frac{1}{2}\sum_{j=1}^{2}(\mathbf{w}_j')^{\mathrm{T}}(2\mathbf{K}+\mathbf{I})^{-1}(\mathbf{w}_j') \quad (4)$$

여기서, $\mathbf{K}$와 $\mathbf{w}_j'$는 아래의 식과 같이 정의된다.

$$\mathbf{K} = \mathbf{B}^{\mathrm{T}}(\mathbf{G}\mathbf{G}^{\mathrm{T}} + \boldsymbol{\epsilon})^{-1}\mathbf{B} \quad (5)$$
$$\mathbf{w}_j' = \mathbf{B}(\mathbf{G}\mathbf{G} + \boldsymbol{\epsilon})^{-1}(\mathbf{w}_j - \boldsymbol{\mu}) \quad (6)$$

수식 (5)와 (6)에서 B와 G는 각각 화자 모델, 채널 모델을 나타내는 행렬이며 추정법은 참고문헌[15]에 자세히 기술되어 있다. 또한 $\mathbf{w}_j$는 타겟 또는 평가 I-벡터를 의미하며, $\boldsymbol{\epsilon}$과 $\boldsymbol{\mu}$는 잔여 통계정보와 I-벡터의 평균 벡터를 각각 의미한다.

2.2 깊은 신경망 특징 벡터

깊은 신경망은 <그림 2>에서 보는 바와 같이 한 개 이상의 은닉층을 갖는 앞먹임(feed-forward) 인공 신경망(artificial neural network)을 의미한다. 이는 패턴 분류에 주로 사용되며, 깊은 신경망의 출력층은 분류하고자 하는 객체로 구성된다. 이를 감독 훈련 기법(supervised training approach)이라 일컫는다. 감독 훈련의 경우, <그림 3>에서 보는 바와 같이 분류하고자 하는 대상에 따라 화자 또는 음소 등을 깊은 신경망의 출력으로 나타낼 수 있다[16]. 이러한 감독 훈련을 위한 깊은 신경망의 파라메타는 다음의 criterion을 통해 최적화된다.

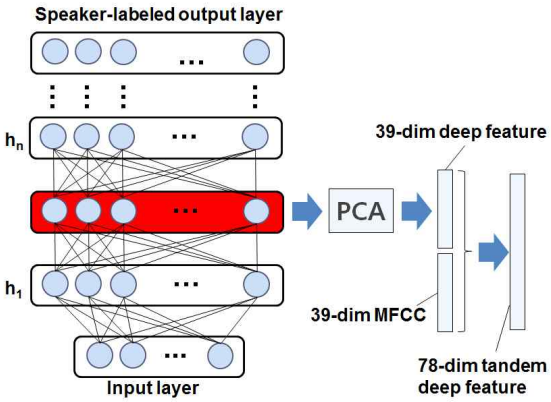

그림 2. 깊은 신경망 기반 특징 벡터 추출
Figure 2. DNN-based feature vector extraction

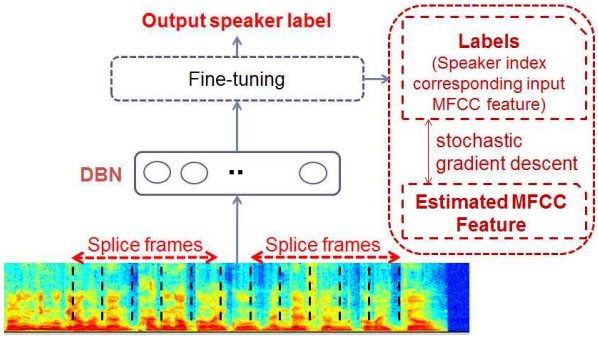

그림 3. 깊은 신경망 미세 조정 (감독 훈련)
Figure 3. DNN fine-tuning (supervised training)

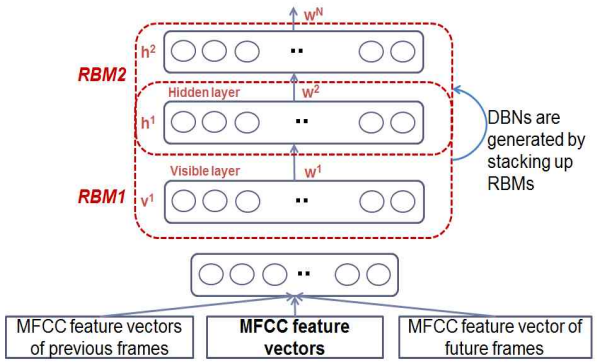

그림 4. 심층 신뢰망 선훈련 (무감독 훈련)
Figure 4. DBN pre-training (unsupervised training)

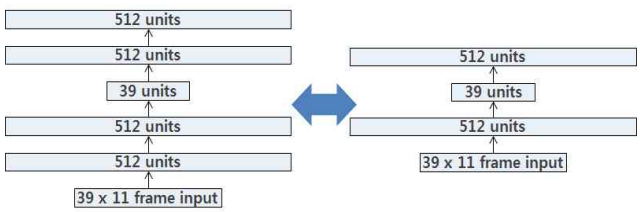

그림 5. 깊은 신경망 은닉층 수 비교
Figure 5. Comparison of the number of DNN hidden layers

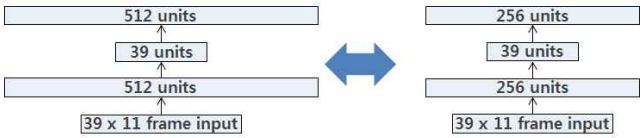

그림 6. 깊은 신경망 은닉층 별 유닛수 비교
Figure 6. Comparison of the number of DNN hidden units

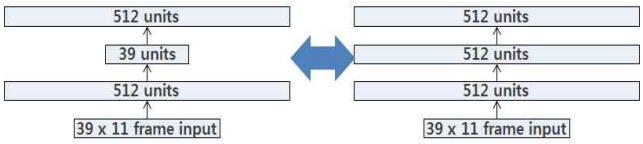

그림 7. 깊은 신경망 은닉층 구조 비교
Figure 7. Comparison of the structure of hidden layers

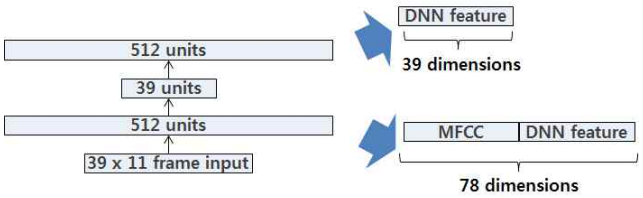

그림 8. 단일 또는 탠덤 특징 벡터 비교
Figure 8. Comparison of single or tandem feature vector

$$L(\theta) = -\sum_{s} d_s \log P(s|\mathbf{o},\theta) \tag{7}$$

여기서, θ는 깊은 신경망 파라메터 집합을 나타내고, $P(s|\mathbf{o})$는 출력 클래스 s에 대한 사후 확률을 나타낸다. 또한 d_s는 타겟 클래스에 대해서는 1을, 비타겟 클래스에 대해서는 0을 가진다.

깊은 신경망은 제한된 볼츠만 기계(RBM: restricted Boltzmann machine)으로 잘 알려진 무감독 훈련 기법(unsupervised training approach)을 적용할 수 있다. <그림 4>에서 보는 바와 같이 RBM은 감독 훈련 기법과 달리 분류하고자 하는 대상의 정보를 훈련 시 필요로 하지 않으며, 은닉층(hidden layer)와 가시층(visible layer)로 구성된다[7]. 두 층 사이의 확률 분포는 아래와 같이 나타낸다.

$$p(\mathbf{v},\mathbf{h}|\theta) = \frac{e^{-E(\mathbf{v},\mathbf{h}|\theta)}}{Z(\theta)} \tag{8}$$

여기서, $E(\mathbf{v},\mathbf{h}|\theta)$는 에너지 함수, $Z(\theta) = \Sigma_v \Sigma_h e^{-E(\mathbf{v},\mathbf{h}|\theta)}$는 정규화 변수를 나타낸다. 수식 (8)을 통해 구해진 $p(\mathbf{v},\mathbf{h}|\theta)$에 대해 contrastive divergence 기법[6]을 적용하여 깊은 신경망의 파라메터를 추정한다. 여기서 RBM을 여러 층으로 쌓아 생성된 망을 심층 신뢰망(DBN: deep belief network)이라고 한다.

표 1. 화자 검증을 위한 데이터베이스 구성
Table 1. Database specification for speaker verification

항목	내용
음성 데이터베이스	Sitec CleanSent01 speech DB
데이터베이스 상세내용	
총 재생시간	32시간 (남: 16시간, 여: 16시간)
화자 수	200명 (남: 100명, 여: 100명)
발화 수	21,000발화 (1명 당 105발화)
파라메터 설정	
가우시안 혼합 개수	64
I-벡터 차수	40
PLDA Eigenvoice 차수	20
PLDA Eigenchannel 차수	10
평가 방법	
화자 모델 수	60
시도 횟수	개발 DB 이용 시: 2,100(타겟)+123,900(비타겟) = 126,000 평가 DB 이용 시: 2,100(타겟)+123,900(비타겟) +42,000(사칭) = 168,000

다음으로 무감독 또는 감독 훈련 기법에 의해 훈련된 깊은 신경망을 통해 화자 검증을 위한 특징 벡터 추출에 대해 서술한다. 일반적으로, 입력층에서 N개의 프레임에 대한 MFCC 값을 입력으로 받는다. 입력된 MFCC 값에 대해 앞먹임 기법을 적용하여 상위 층에 대한 출력 값들을 획득한다. 중간 은닉층의 출력 값에 대해 주성분 분석(PCA: principal component analysis)을 통해 39차원의 최종 특징 벡터를 추출한다. 이 뿐만 아니라, 깊은 신경망을 통해 추출된 특징 벡터와 39차원의 perceptual linear prediction (PLP) 또는 MFCC 특징 벡터를 결합한 탠덤(tandem) 특징 벡터를 화자 검증에 적용한다.

3. DNN 특징 벡터와 화자 검증 성능 분석 방법

본 절에서는 2장에서 설명된 깊은 신경망 기반 특징 벡터에 따른 화자 검증 시스템의 성능에 대한 영향을 분석하기 위한 방법을 기술한다. 즉, 깊은 신경망의 훈련 방법, 깊은 신경망의 구조, 특징 벡터 종류 및 타 특징 벡터와의 결합 여부 등 총 3가지의 관점에서 분석이 수행된다. 우선, 깊은 신경망의 훈련 방법의 관점에서는 2장에서 언급된 무감독 또는 감독 훈련 방법에 따른 영향을 분석한다. 다음으로, 깊은 신경망의 구조에 대한 분석을 위해 <그림 5>에서 <그림 7>에서 보는 바와 같이 깊은 신경망의 은닉층 수, 은닉층 별 유닛수 및 은닉층 구조를 변화시킨다. 특히, 은닉층 수를 3개 또는 5개로 변경하며, 이 중 2, 3, 4번째 은닉층에서의 출력값을 화자 검증 시스템의 입

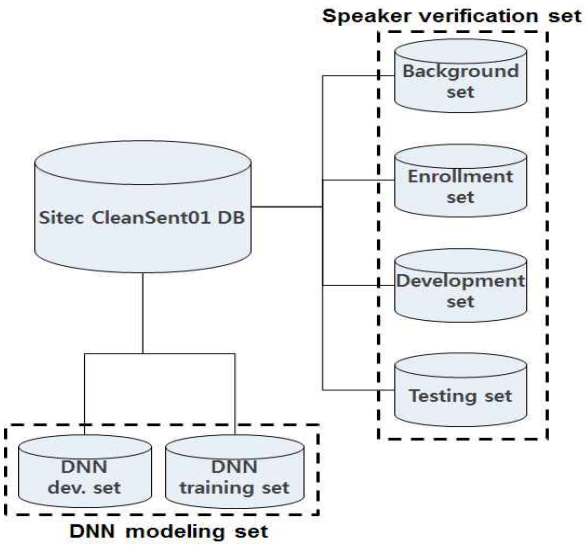

그림 9. 화자 검증 시스템의 평가를 위한 데이터베이스 구성
Figure 9. Database for a speaker verification system

력으로 이용한다. 은닉층 별 유닛수는 128, 256, 512개로 변경하며, 은닉층 구조는 병목(bottleneck) 구조 또는 직사각형 구조를 비교한다. 마지막으로, 특징 벡터 종류 및 타 특징 벡터와의 결합에 따른 영향을 분석하기 위해 <그림 8>에서와 같이 탠덤 특징 벡터를 고려한다. 즉, 39차원의 MFCC 또는 PLP 특징 벡터와 39차원의 신경 구조망 특징 벡터가 결합된 78차원의 탠덤 특징 벡터와 39차원의 신경 구조망 단일 특징 벡터에 따른 영향을 분석한다.

4. 실험 및 논의

4.1 실험 환경

화자 검증 시스템의 성능 평가를 위해, ALIZE 화자 인식/검증 toolkit을 사용하였다[17], [18]. ALIZE는 가우시안 혼합 모델링뿐만 아니라 공동 요인 분석, support vector machine (SVM), I-벡터 기법을 사용할 수 있으며, 스코어링에서는 최신 기법인 PLDA까지 제공한다. 깊은 신경망 특징 벡터의 경우 Kaldi toolkit을 사용하여 추출하였다[19]. <표 1>은 데이터베이스의 상세 내용과 시도 방법에 대한 내용을 나타낸다[20], [21]. <그림 9>는 데이터베이스의 분할에 대한 그림을 나타낸다. <그림 9>에서 보는 바와 같이 크게 화자 검증, 깊은 신경망 모델링의 두 부분으로 나뉜다. 화자 검증 DB의 경우, 배경 DB는 남, 여 각각 50명×70발화로 구성되며 이는 총 7,000발화, 10시간 분량에 해당한다. 등록 DB의 경우, 남, 여 각각 30명×1발화로 구성된다. 여기서 화자는 배경 DB의 화자와 중복되지 않는다. 개발 DB의 경우, 남, 여 각각 30명×35발화로 구성되며, 총 2,100발화이다. 평가 DB의 경우, 남, 여 각각 40명×35발화로 구성되며, 총 2,800발화이다. 이 중 60명 화자(남,여 각각 30명)는 등록 DB와 중복되며, 나머지는 중복되지 않는다. 깊은 신경망 모델링 DB의 경우, 깊은 신경망 훈련 DB와 개발 DB 두 부분으로 나뉜다. 깊은 신경망 훈련 DB의 경우, 남, 여 각각 50명×100발화로 구성되며, 이는 총 10,000발화, 15시간 분량에 해당한다. 깊은 신경망 개발 DB와 경우, 남, 여 각각 50명×5발화로 구성되어, 총 500발화, 50분 분량에 해당한다. 깊은 신경망 훈련 DB의 화자와 중복되지 않는다. 또한 깊은 신경망 DB의 경우, 모든 화자가 화자 검증 DB와 중복되지 않는다. <표 2>와 <표 3>은 각각 무감독 및 감독 깊은 신경망 훈련 환경을 나타낸다. 성능 측정법으로는 동일오류율(EER: equal error rate)을 사용하였다. 동일오류율이란 오인식률(FAR: false acceptance rate)과 오거부률(FRR: false rejection rate)이 같아지는 비율을 말한다.

표 2. 심층 신뢰망 훈련 환경
Table 2. DBN training specification

항목	내용
특징 벡터	MFCC
특징 벡터 차수	39차×11프레임
은닉층 수	3개 또는 5개
은닉층 별 유닛수	128, 256, 512
학습률	0.0008
Splice 개수	5

표 3. 깊은 신경망 훈련 환경
Table 3. DNN training specification

항목	내용
학습률	0.0008
훈련 반복 횟수	30~50 (포화시점에 따라 다름)
레이블	타겟 특징 벡터의 화자 인덱스
오류 함수	역전파 알고리즘을 이용한 최대 엔트로피
디코딩 알고리즘	앞먹임 전파 (Feed-forward propagation)

4.2 실험 결과

4.2.1 특징 벡터 종류 별 성능 비교

깊은 신경망의 입력 특징 벡터 종류를 결정하기 위해 화자 검증에서 일반적으로 사용되는 MFCC, linear-frequency cepstral coefficient (LFCC), PLP 특징 벡터에 대한 화자 검증 실험을 수행하였다. 이 때 사용한 배경 DB 발화 개수는 7,000개로 고정하였다. 가우시안 혼합 개수는 16, 32, 64, 128개를 이용하였다. <표 4>에서 보는 바와 같이, MFCC와 PLP 특징 벡터는 비슷한 성능을 보였으나, LFCC의 경우 MFCC와 PLP 특징 벡터에 비해 성능이 저하되었다. 또한 가우시안 혼합 개수의 경우 64개에서 높은 성능을 보였다. 따라서, 깊은 신경망 특징 벡터 기반 실험에서는 64개의 가우시안 혼합 개수의 MFCC와 PLP를 사용하였다.

표 4. I-벡터 기반 화자 검증 시스템의 가우시안 혼합수의 변화에 따른 특징 벡터 종류별 성능 비교 (동일오류율, %)
Table 4. Performance comparison of a I-vector-based speaker verification system using different feature vectors according to the number of Gaussian mixtures (EER, %)

가우시안 혼합 개수 / 특징벡터 종류	16	32	64	128
MFCC	7.43	8.48	7.33	7.86
LFCC	9.98	8.77	8.51	8.78
PLP	7.76	7.81	7.33	7.67

4.2.2 깊은 신경망의 훈련 방법에 따른 성능 비교

깊은 신경망 훈련 방법에 따른 성능을 비교하기 위해 MFCC를 입력으로 병목 구조의 서로 다른 은닉층 수, 타겟 은닉층, 은닉층 별 유닛수에 따라 깊은 신경망을 무감독 또는 감독 훈련하였고, 깊은 신경망으로부터 추출된 특징 벡터를 화자 검증 시스템에 이용하였다. 실험 결과, <표 5>에서 보는 바와 같이 무감독 훈련과 감독 훈련은 비슷한 성능을 보였다. 하지만 무감독 훈련 시에는 은닉층의 수가 3개이고 은닉층 별 유닛수가 256개일 때 동일오류율이 17.00%을 보이며, 은닉층의 수가 3개이고 은닉층 별 유닛수가 512개일 때는 동일오류율이 12.10%로 현저하게 성능이 저하되는 문제가 발생하였다. 은닉층 수 관련해서는 감독 훈련 시 은닉층의 수가 5개보다는 3개일 때 높은 성능을 보였다. 반면 은닉층 별 유닛수의 경우, 은닉층 유닛수가 많을수록 높은 성능을 보였다. 이는 화자 검증의 경우에도 은닉층 별 유닛수가 많을수록 깊은 신경망의 변별력이 강화되는 특성을 보이는 것을 의미한다. 따라서 대체적으로 감독 훈련이 무감독 훈련에 비해 나은 성능을 보였기 때문에, 나머지 실험에서는 감독 훈련만을 고려하였다.

표 5. 다른 구조에 따라 MFCC에 적용한 무감독 또는 감독 훈련된 깊은 신경망으로 추출된 특징 벡터를 사용한 화자 검증 시스템의 성능 비교 (동일오류율, %)
Table 5. Performance comparison of a speaker verification system using feature vectors extracted from unsupervised or supervised DNN applied to MFCC depending on different structure of DNNs (EER, %)

은닉층 수 (타겟 은닉층) / 은닉층 별 유닛수		128	256	512
깊은 신경망 무감독 훈련	3 (2)	7.14	17.00	12.10
	5 (2)	9.24	7.52	7.38
	5 (3)	9.29	8.24	7.81
	5 (4)	10.00	10.29	9.48
깊은 신경망 감독 훈련	3 (2)	8.38	8.00	8.00
	5 (2)	7.91	8.19	8.00
	5 (3)	8.71	8.62	8.43
	5 (4)	9.14	8.97	8.05

4.2.3 깊은 신경망의 구조에 따른 성능 비교

깊은 신경망의 구조에 따른 성능을 비교하기 위해 MFCC를 입력으로 병목 구조 또는 직사각형 구조에서 서로 다른 은닉층 수, 타겟 은닉층, 은닉층 별 유닛수에 따라 깊은 신경망을 감독 훈련하였고, 깊은 신경망으로부터 추출된 특징 벡터를 화자 검증 시스템에 이용하였다. 실험 결과, <표 6>에서 보는 바와 같이 직사각형 구조가 병목 구조보다 높은 성능을 보였다. 이는 병목 구조의 경우 은닉층의 유닛수가 적어서 깊은 신경망 훈련 시 변별력이 떨어지는 현상이 발생하기 때문으로 판단된다. 또한 직사각형 구조에서 은닉층의 수가 적을수록 높은 성능을 보였다. 하지만, 은닉층의 수가 1개와 같이 너무 적은 경우에는 정보가 함축되어 오히려 변별력이 저하된다. 한편, 은닉층 수가 과도하게 많을 경우에도 각각의 은닉층은 더 적은 정보를 담고 있기 때문에 성능이 저하된다. 이는 데이터베이스 크기에 맞는 가능한 적은 수의 은닉층을 쓰되 필요 이상으로 증가하면 성능 저하가 발생할 수도 있다는 것을 의미한다. 따라서 다음의 실험에서는 감독 훈련된 직사각형 구조의 은닉층 수가 3개인 깊은 신경망 기반 특징 벡터만을 고려하였다.

표 6. 다른 구조에 따라 MFCC에 적용한 감독 훈련된 깊은 신경망으로 추출된 특징 벡터를 사용한 화자 검증 시스템의 성능 비교 (동일오류율, %)
Table 6. Performance comparison of a speaker verification system using feature vectors extracted from supervised DNN applied to MFCC depending on different structure of DNNs (EER, %)

은닉층 수 (타겟 은닉층) / 은닉층 별 유닛수		128	256	512
깊은 신경망 병목 구조	3 (2)	8.38	8.00	8.00
	5 (2)	7.91	8.19	8.00
	5 (3)	8.71	8.62	8.43
	5 (4)	9.14	8.97	8.05
깊은 신경망 직사각형 구조	3 (2)	6.48	5.48	6.10
	5 (2)	7.00	6.33	5.86
	5 (3)	6.71	6.52	6.00
	5 (4)	9.43	7.48	6.57

4.2.4 특징 벡터 종류 및 타 특징 벡터와의 결합 여부에 따른 성능 비교

특징 벡터 종류 및 타 특징 벡터와의 결합 여부에 따른 성능을 비교하기 위해 MFCC 또는 PLP를 입력 특징 벡터로 하여 은닉층 수가 3개, 은닉층 별 유닛수가 128개, 직사각형 구조에서 깊은 신경망을 감독 훈련하였고, 깊은 신경망의 2번째 타겟 은닉층으로부터 추출된 단일 또는 탠덤 특징 벡터를 화자 검증 시스템에 이용하였다. 실험 결과, <표 7>에서 보인 바와 같이 PLP보다는 MFCC를 이용하였을 경우 높은 성능을 보였다. 또한 단일 특징 벡터보다는 탠덤 특징 벡터를 이용하였

표 7. MFCC 혹은 PLP에 적용한 감독 훈련된 직사각형 구조의 깊은 신경망으로 추출된 단일 및 탠덤 특징 벡터를 사용한 화자 검증 시스템의 성능 비교 (동일오류율, %)
Table 7. Performance comparison of a speaker verification system using single or tandem feature vectors extracted from supervised DNN applied to MFCC or PLP using a rectangular type of hidden layers (EER, %)

타 특징 벡터와 결합 여부 / 특징 벡터 종류	단일	탠덤
MFCC	6.48	4.29
PLP	7.33	5.76

을 경우 높은 성능을 보였다. 이는 탠덤 특징 벡터의 경우, MFCC 특징 벡터와 깊은 신경망으로부터 추출된 특징 벡터의 값들을 살펴보면 에러 패턴이 다르게 나타나는데, 이 때 두 가지 특징 벡터의 조합을 통해 상호 보완되기 때문이다. 종합적으로 볼 때, MFCC를 깊은 신경망의 입력 특징 벡터로 하여 깊은 신경망을 통해 추출된 특징 벡터와 MFCC를 결합하였을 경우 가장 높은 성능을 보였다.

5. 결론

본 논문에서는 깊은 신경망에서 추출한 특징 벡터의 변화가 화자 검증에 미치는 영향을 분석하기 위한 비교 연구를 수행하였다. GMM-UBM 기반의 화자 검증 시스템에 있어서는 가우시안 혼합 개수 및 배경 DB 크기를 증가시킬수록 화자 검증 시 높은 성능을 보였다. 특징 벡터 종류 관련해서는 MFCC와 PLP가 LFCC 대비 우수한 성능을 보였다. 깊은 신경망 기반 특징 벡터 추출의 경우, 깊은 신경망의 훈련 방법, 구조, 특징 벡터 종류 및 타 특징 벡터와의 결합 여부라는 3가지 측면으로 나누어 화자 검증 시스템에 미치는 영향을 분석하였다. 깊은 신경망 훈련 방법에서는 무감독 훈련보다는 화자 인덱스 정보를 이용한 감독 훈련 시 높은 성능을 보였다. 깊은 신경망 구조 측면에서는 직사각형 구조, 은닉층 수는 3개일 때 높은 성능을 보였다. 특징 벡터 종류 및 타 특징 벡터와의 결합 여부 관점에서는 PLP보다는 MFCC를 사용하였을 때가 높은 성능을 보였으며, MFCC만을 개별적으로 사용하는 방법에 비해 탠덤 특징 벡터를 사용하였을 경우 더 높은 성능을 보였다. 종합적으로 볼 때, MFCC를 입력으로 한 감독 훈련된 직사각형 모양 은닉층을 갖는 깊은 신경망 기반 탠덤 특징 벡터 이용 시 화자 검증에서 가장 좋은 성능을 보였다.

참고문헌

[1] Kinnunen, T. & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, Vol. 52, No. 1, 12-40.

[2] Reynolds, D. A., Quatieri, T. F. & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, Vol. 10, No. 1, 19-41.

[3] Kenny, P., Boulianne, G., Ouellet, P. & Dumouchel, P. (2007). Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing,* Vol. 15, No. 4, 1435-1447.

[4] Matrouf, D., Scheffer, N., Fauve, B. G. & Bonastre, J. F. (2007). A straightforward and efficient implementation of the factor analysis model for speaker verification. In *Proceedings of Interspeech*, Antwerp, Belgium, 1242-1245.

[5] Dehak, N., Dehak, R., Glass, J. R., Reynolds, D. A. & Kenny, P. (2010). Cosine similarity scoring without score normalization techniques. In *Proceedings of Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, 71-75.

[6] Fu, T., Qian, Y., Liu, Y. & Yu, K. (2014). Tandem deep features for text-dependent speaker verification. In *Proceedings of Interspeech,* Singapore, Singapore, 1327-1331.

[7] Yu, D. & Seltzer, M. L. (2011). Improved bottleneck features using pretrained deep neural networks. In *Proceedings of Interspeech*, Florence, Italy, 237-240.

[8] Zhang, Y., Chuangsuwanich, E., & Glass, J. (2014). Extracting deep neural network bottleneck features using low-rank matrix factorization. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 185-189.

[9] Liu, Y., Fu, T., Fan, Y., Qian, Y., & Yu, K. (2014). Speaker verification with deep features. In *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, Beijing, China, 747-753.

[10] Kanagasundaram, A. (2014). Speaker verification using I-vector features. Ph.D. Dissertation, Queensland University of Technology.

[11] Kenny, P., Boulianne, G. & Dumouchel, P. (2005). Eigenvoice modeling with sparse training data. *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 3, 345-354.

[12] Bishop, C. M. (2007). *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer.

[13] Prince, S. J. & Elder, J. H. (2007). Probabilistic linear discriminant analysis for inferences about identity. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, Rio de Janeiro, Brazil, 1-8.

[14] Lee, K. A., Larcher, A., You, C. H., Ma, B. & Li, H. (2013). Multi-session PLDA scoring of i-vector for partially open-set

speaker detection. In *Proceedings of Interspeech*, Lyon, France, 3651-3655.

[15] Kenny, P. (2010). Bayesian speaker verification with heavy tailed priors. In *Proceedings of Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, paper no 014.

[16] Sainath, T. N., Kingsbury, B. & Ramabhadran, B. (2012). Auto-encoder bottleneck features using deep belief networks. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, 4153-4156.

[17] Larcher, A., Bonastre, J. F., Fauve, B. G., Lee, K. A., Lévy, C., Li, H. & Parfait, J. Y. (2013). ALIZE 3.0-open source toolkit for state-of-the-art speaker recognition. In *Proceedings of Interspeech*, Lyon, France, 2768-2772.

[18] Bonastre, J. F., Wils, F. & Meignier, S. (2005). ALIZE, a free toolkit for speaker recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, PA, 737-740.

[19] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N. & Vesel, K. (2011). The Kaldi speech recognition toolkit. In *Proceedings of IEEE ASRU*, Honolulu, HI, 1-4.

[20] Brümmer, N. & De Villiers, E. (2010). The speaker partitioning problem. In *Proceedings of Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, 194-201.

[21] Greenberg, C. S., Stanford, V. M., Martin, A. F., Yadagiri, M., Doddington, G. R., Godfrey, J. J. & Hernandez-Cordero, J. (2013). The 2012 NIST speaker recognition evaluation. In *Proceedings of Interspeech,* Lyon, France, 1971-1975.

• **김대현 (Kim, Dae Hyun)**
광주과학기술원 정보통신공학부
광주광역시 북구 첨단과기로 123 (오룡동)
Tel: 062-715-3121
Email: wmelonw88@gmail.com
관심분야: 음성인식, 화자검증
광주과학기술원 정보통신공학부 석사

• **성우경 (Seong, Woo Kyeong)**
광주과학기술원 정보통신공학부
광주광역시 북구 첨단과기로 123 (오룡동)
Tel: 062-715-3121
Email: wkseong@gist.ac.kr
관심분야: 음성인식
광주과학기술원 정보통신공학부 석박사 통합과정

• **김홍국 (Kim, Hong Kook), 교신저자**
광주과학기술원 정보통신공학부
광주광역시 북구 첨단과기로 123 (오룡동)
Tel: 062-715-2228 Fax: 062-715-2204
Email: hongkook@gist.ac.kr
관심분야: 음성/오디오 코덱, 3D-오디오, 음성인식
현재 광주과학기술원 정보통신공학부 교수