논문 16-41-08-18 The Journal of Korean Institute of Communications and Information Sciences '16-08 Vol.41 No.08
http://dx.doi.org/10.7840/kics.2016.41.8.958

I-벡터 기반 오픈세트 언어 인식을 위한 다중 판별 DNN

강 우 현◆, 조 원 익*, 강 태 균*, 김 남 수°

Multiple Discriminative DNNs for I-Vector Based Open-Set Language Recognition

Woo Hyun Kang◆, Won Ik Cho*, Tae Gyoon Kang*, Nam Soo Kim°

요 약

본 논문에서는 여러 개의 이원 support vector machine (binary SVM)을 사용하여 세 개 이상의 클래스를 분류하는 multi-class SVM과 유사하게 다중의 판별 deep neural network (DNN) 모델을 사용하는 i-벡터 기반의 언어 인식 시스템을 제안한다. 제안하는 시스템은 NIST 2015 i-vector Machine Learning Challenge 데이터베이스에 포함된 i-벡터들을 이용하여 학습 및 테스트 되었으며, 오픈 세트에서 기존의 cosine distance, multi-class SVM 및 단일 neural network (NN) 기반의 언어 인식 시스템에 비하여 높은 성능을 보임이 확인되었다.

Key Words : I-vector, language recognition, deep learning, machine learning, multi-class classification

ABSTRACT

In this paper, we propose an i-vector based language recognition system to identify the spoken language of the speaker, which uses multiple discriminative deep neural network (DNN) models analogous to the multi-class support vector machine (SVM) classification system. The proposed model was trained and tested using the i-vectors included in the NIST 2015 i-vector Machine Learning Challenge database, and shown to outperform the conventional language recognition methods such as cosine distance, SVM and softmax NN classifier in open-set experiments.

Ⅰ. 서 론

최근 몇 년간 음성 인식 및 언어 인식 분야와 같은 음성 처리 분야에서 딥 러닝의 활용에 대한 연구가 활발하게 진행되고 있다. 음성 인식 분야에서 deep neural network (DNN)를 기존에 사용되어 온 음향 모델인 Gaussian mixture model (GMM) 대신 사용하거나 특징 보상에 활용함으로써 높은 성능을 보였다.[1,2] 이러한 DNN의 음성 인식에서의 성공은 음성이 가지고 있는 복잡한 특성을 딥 러닝을 통하여 모델링할 수 있다는 가능성을 열어주었다.

더 나아가 DNN은 음성 인식에서의 음향 모델링과 유사하게 화자 인식을 위한 특징 추출에 활용돼 높은 성능을 보였다[3]. 이 뿐만 아니라, 입력 시퀀스와 출력 라벨 사이의 비선형적 관계를 표현할 수 있는 DNN의 능력을 활용하기 위하여 DNN을 화자 인식에서 인식

※ 이 연구는 방위사업청 및 국방과학연구소의 재원에 의해 설립된 신호정보 특화연구센터 사업의 지원을 받아 수행되었음.
◆ First Author : Seoul National University Department of Electrical and Computer Engineering and Institute of New Media and Communications, whkang@hi.snu.ac.kr, 학생회원
° Corresponding Author : Seoul National University Department of Electrical and Computer Engineering and Institute of New Media and Communications, nkim@snu.ac.kr, 종신회원
* Seoul National University Department of Electrical and Computer Engineering and Institute of New Media and Communications, wicho@hi.snu.ac.kr, tgkang@hi.snu.ac.kr, 학생회원
논문번호 : KICS2016-04-066, Received April 25, 2016; Revised June 27, 2016; Accepted July 15, 2016

기로 직접 사용하는 방법에 대한 연구도 진행되어왔다.[4,5] 화자 인식에서 DNN을 이용한 분류 기법은 기존에 사용되어온 support vector machine (SVM) 및 cosine distance 기반의 분류 방식에 비하여 높은 성능을 보였다.

화자 인식에서와 마찬가지로 DNN은 연령 인식[6,7]과 언어 인식에서도 높은 성능을 보였다. 화자 인식에서와 유사하게 softmax 출력층을 가지고, i-벡터 입력을 받는 DNN 분류기는 언어 인식에서 기존의 cosine distance 방식에 비하여 우수한 성능을 보였다[8].

본 논문에서는 학습 데이터에 포함되있지 않은 언어인 out-of-set (OOS) 언어를 판별하기 위하여 기존의 멀티 클래스 SVM 분류기[9]와 유사한 방식을 따르는 DNN 분류 모델을 제안한다. 제안하는 시스템은 one-vs-all SVM 기반의 언어 분류 시스템[10]과 같이, 입력 i-벡터가 특정 언어를 발화했을 때의 음성으로부터 추출되었을 확률을 출력하는 판별 모델을 분류하고자 하는 모든 언어에 대하여 각각 학습한다. 판별 모델로는 단일 시그모이드 출력 노드를 가진 DNN이 사용되었으며, 모든 DNN은 화자 판별 분야에서 연구된 판별 DNN 학습 기법인 universal deep belief network (UDBN) 방식[5]을 이용하여 학습되었다. NIST 2015 i-vector Machine Learning Challenge 데이터 셋에 포함된 학습 데이터 및 테스트 데이터[11]를 이용하여 제안된 알고리즘을 학습하였으며, 기존의 cosine distance, 단일 neural network (NN) 분류기, SVM 기반의 분류 알고리즘에 비하여 OOS 언어를 고려하는 오픈 세트 실험에서 우수한 성능을 보였다.

Ⅱ. I-벡터

2.1 I-벡터 추출

I-벡터는 현재 화자 인식 및 언어 인식 분야에서 가장 널리 사용되는 특징 중 하나로, 음성이 가지고 있는 다양한 변이성을 낮은 차원의 고정된 크기의 벡터로 표현할 수 있다는 장점을 가지고 있다[12]. Principle component analysis (PCA)나 eigenvoice 분리 기법과 마찬가지로, i-벡터 추출은 행렬 분리 기법으로 볼 수 있으며, 이상적인 GMM 슈퍼벡터와 i-벡터의 관계는 다음과 같은 식으로 정의된다.

$$M = m + Tw \tag{1}$$

여기에서 M은 특정 화자 혹은 음성에 종속적인 이상적인 GMM 슈퍼벡터를 나타내며, m, T, w는 각각 universal background model (UBM), 전체 변이성 행렬 (total variability matrix), 그리고 i-벡터를 의미한다. T 행렬은 낮은 행렬 계수의 행렬이며, GMM 슈퍼벡터가 갖는 변이성을 저차원 벡터로 투영시키는 역할을 한다. T 행렬은 eigenvoice 기법과 유사하게 EM 알고리즘을 이용하여 학습한다.

2.2 I-벡터 정규화

I-벡터는 음성이 가지고 있는 여러 변이성을 표현하기 때문에 화자나 언어에 대한 정보뿐만 아니라 잡음과 같이 언어 인식에서 불필요한 정보도 포함한다. 그렇기에 판별 시스템의 입력으로 사용하기 위해서는 i-벡터가 갖는 불필요한 정보를 제거해주는 정규화 과정을 거쳐야 한다.

Within-class covariance normalization (WCCN)은 널리 사용되는 특징 보상 기법으로, 같은 군내의 특징들에 대한 공분산을 감소시킴으로써 특징을 판별에 최적화 시켜준다[13]. WCCN의 투영 행렬 A는 아래와 같은 수식으로 구할 수 있다.

$$A = B^{-1} = S_w^{-1/2} \tag{2}$$

위 수식에서 B는 군내 공분산 S_w에 Cholesky decomposition을 적용함으로써 구해지는 행렬로, 군내 공분산은 아래의 식으로 구할 수 있다.

$$S_w = \sum_c \sum_{j \in c} (w_j - \mu_c)(w_j - \mu_c)' \tag{3}$$

위 수식에서 w_j는 학습 데이터에 포함된 j번째 i-벡터를 의미하며, μ_c는 군 c에 포함된 모든 i-벡터들의 평균을 의미한다.

WCCN를 위해서는 계산된 투영 행렬 A를 아래와 같이 i-벡터에 곱해주면 된다.

$$\phi_{WCCN} = Aw \tag{4}$$

여기에서 ϕ_{WCCN}와 w는 각각 정규화된 i-벡터와 원본 i-벡터를 의미한다.

Ⅲ. UDBN 기반 판별 DNN 학습

UDBN 학습은 화자 판별을 위하여 제안된 판별 DNN 학습 방법이다[5]. UDBN이란 이름이 붙은 이유

는 기존의 UBM-GMM[14]과 같이, 여러 군이 포함된 대량의 데이터를 사용하여 군 독립적인 모델 (class independent model)을 학습하기 때문이다. 또한, UDBN 방식은 SVM과 같은 이원 분류기 (binary classifier)의 학습 과정과 같이 과적응 (over-fitting)을 방지하기 위해 타겟 데이터 셋과 imposter 데이터 셋의 균형을 맞추는 과정을 거친다.

3.1 UDBN 학습

많은 경우, DNN의 파라미터는 무작위로 초기화하는 경우보다 pre-training을 통하여 초기화했을 때 더 좋은 최적 값으로 수렴한다. 가장 널리 사용되는 pre-training 기법으로는 restricted Boltzmann machine (RBM)을 여러 개 쌓은 모델인 deep belief network (DBN)를 학습하여 이를 DNN의 초기 모델로 사용하는 방식이다[15]. UDBN은 군에 독립적인 DBN이며, 다양한 군이 포함된 많은 데이터를 사용하여 학습한다. 일반적인 DBN과 마찬가지로 UDBN은 contrastive divergence라는 층별 탐욕 알고리즘 (greedy layer-wise algorithm)을 사용하여 비교사 학습 (unsupervised learning)한다.

3.2 UDBN 적응화

UDBN이 학습된 후, 판별하고자 하는 군 종속적 학습 데이터 (class dependent training set)를 이용하여 contrastive divergence 알고리즘으로 UDBN을 다시 한 번 학습 한다. 이 과정은 군 독립적인 모델을 특정 군에 종속적으로 변환시키므로 적응화(adaptation) 과정이라고 한다. 군 종속적 학습 데이터는 관심 대상인 군의 음성으로부터 추출된 i-벡터와 관심 대상에 포함되지 않은 군들로부터 추출된 i-벡터가 같은 비율로 포함되어있는 데이터 세트를 의미한다. 적응화가 끝난 군 종속적인 DBN은 이를 초기 파라미터로 하여 다시 군 종속적 학습 데이터를 이용해 오류 역전파 알고리즘 (error back-propagation algorithm)을 통해 DNN으로 학습된다. 이 과정을 미세조정 (fine-tuning)이라 하며, 이 때 판별 DNN의 학습을 위하여 출력 층으로는 단일 시그모이드 값을 이용하고, 관심 군으로부터 생성된 i-벡터의 라벨은 1, 그 외의 라벨은 0으로 구성한다.

이렇게 학습된 판별 DNN은 관심 군으로부터 생성된 i-벡터에 유사한 입력일수록 1에 가까운 출력을 가지며, 이는 입력이 해당 모델이 타겟으로 하는 군에 포함될 확률을 나타낸다고 볼 수 있다.

Ⅳ. 다중 판별 DNN을 이용한 언어 분류

본 논문에서 제안하는 시스템의 학습 알고리즘은 그림 1과 같이 수행된다. 학습을 마친 이후에는 그림 2에서와 같이 N개의 판별 DNN을 사용해 N개의 언어를 분류한다. 분류 모델을 학습하기 이전에, 우선 각 언어에 대하여 학습 데이터를 생성하여 총 N개의 데이터 셋을 만든다. 각 데이터 셋은 해당 셋이 타겟

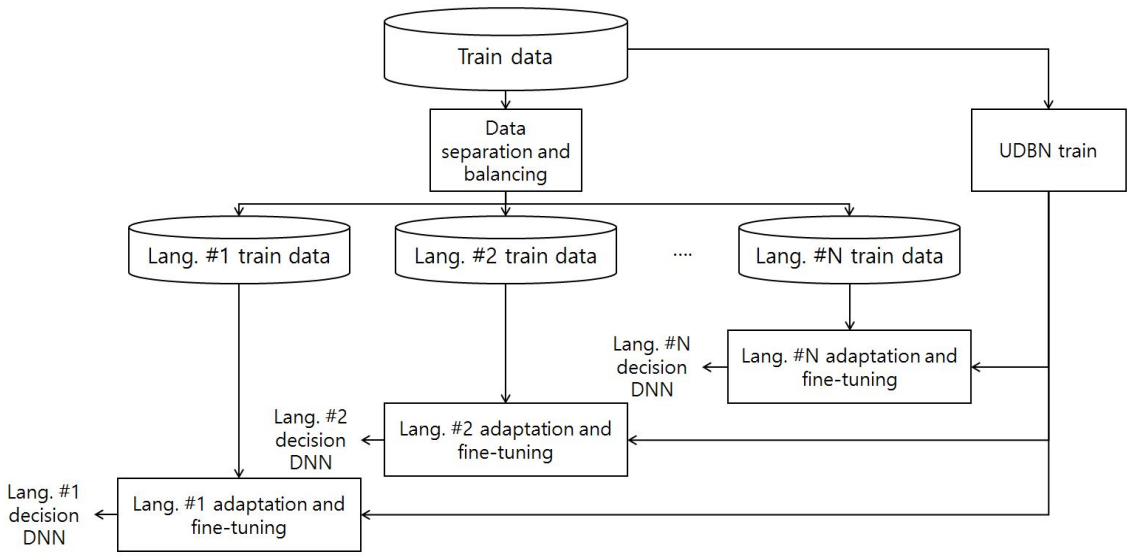

그림 1. 제안하는 학습 알고리즘의 블록도
Fig. 1. Block diagram for the proposed learning algorithm

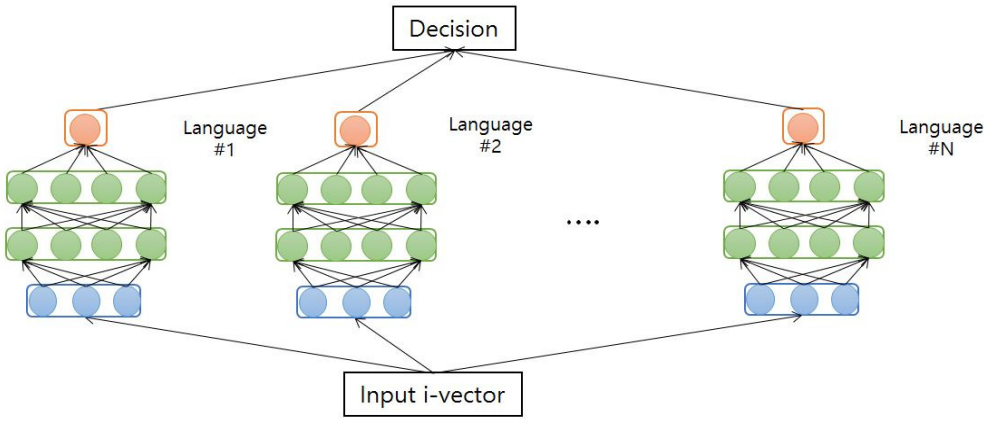

그림 2. 제안하는 다중 DNN 언어 인식 시스템의 구조
Fig. 2. Basic scheme of the proposed multiple DNN based language recognition system

으로 하는 언어의 음성들로부터 추출된 i-벡터들과 타겟 언어가 아닌 imposter 데이터들로부터 추출된 i-벡터들로 구성되어있다. 타겟 데이터와 imposter 데이터의 비율을 같게 맞추기 위하여 타겟 i-벡터들은 중복적으로 포함되었다.

학습 과정에서는 첫 번째 단계로 라벨이 있는 i-벡터 데이터뿐만 아니라 라벨이 없는 데이터도 함께 사용하여 UDBN을 학습한다. UDBN이 학습된 이후에는 각 언어에 대한 군 종속적 학습 데이터를 이용하여 적응화 과정을 거쳐, 각 언어에 종속적인 총 N개의 DBN을 생성한다. 모든 언어에 대한 DBN들이 학습되면, 각 DBN에 단일 시그모이드 노드를 가진 출력층을 추가하고, 다시 군 종속적 학습 데이터를 이용하여 미세 조정과정을 거쳐 각 언어에 대한 판별 DNN을 학습한다. 미세 조정과정에서 출력 라벨은 타겟 언어인 경우 1을, 아닌 경우에는 0을 준다. 이러한 언어별 DNN을 학습한 결과물로는 N개의 판별 DNN이 생성된다.

테스트 단계에서는 입력 i-벡터를 N개의 판별 DNN의 입력으로 준다. 각 판별 DNN의 출력은 0과 1사이의 시그모이드 값이며, 이는 입력된 i-벡터가 해당 판별 DNN이 종속된 언어를 발화한 음성으로부터 추출 되었을 확률을 의미한다. 이렇게 N개의 판별 DNN에서 최종 출력은 N개의 0과 1사이의 스칼라 값으로, N 차원의 벡터로 생각할 수 있다. 이는 N개의 시그모이드 노드를 출력층으로 사용하는 단일 DNN에서의 출력과 유사하다. 하지만 제안된 기법에서는 모든 DNN을 독립적으로 적응화 및 미세조정 하였다. 이는 각 DNN의 파라미터들이 단일 언어를 판별하기 위해서만 학습이 되었음을 의미한다. 이로 인한 출력은 모든 언어를 분류하기 위하여 학습된 단일 DNN에서의 출력보다 특정 언어에 대한 확률을 더욱 강인하게 표현할 수 있다.

테스트 데이터의 언어가 학습 데이터에 포함된 언어 중 하나일 보장이 있는 클로즈 세트 (closed-set) 상황의 경우, 소프트맥스 출력과 마찬가지로 최대 출력을 선택하여 이에 해당하는 언어를 정답으로 판별할 수 있다. 하지만 오픈 세트인 경우, 학습된 N개의 판별 DNN이 종속되지 않은 언어가 테스트 언어에 포함될 수 있으므로, 단순히 최대 출력을 확인하는 것으로는 OOS 언어를 판별할 수 없다. 이 경우에는 최대 출력 값의 OOS 여부를 판단하기 위하여 문턱 값 (threshold)을 설정해야 한다. 최종 출력 결과물인 N 차원 벡터 Y와 0과 1사이의 문턱 값 τ에 대하여, OOS 여부는 다음과 같이 판단 할 수 있다.

$$l = \begin{cases} in-set & , \text{ if } \max(Y) > \tau \\ OOS & , otherwise \end{cases} \quad (5)$$

여기에서 문턱 값 τ은 각 언어에 대한 확률이 언어의 판별에 유효하기 위한 최소값을 의미하며, 주어진 데이터 셋 및 분류 대상에 따라 실험적으로 정해줄 수 있다. 즉, 입력된 i-벡터가 학습된 언어일 확률이 모두 일정 값보다 작거나 같으면, 학습된 언어가 아니라고 판단하는 것이다. 최대 출력 값이 문턱 값보다 큰 경우, 클로즈 세트 상황과 마찬가지로 최대 출력 값에 해당하는 언어를 정답으로 판단한다.

V. 실험 및 결과

5.1 베이스라인 및 데이터 셋

5.1.1 데이터 셋

본 연구에서는 NIST 2015 i-vector Machine Learning Challenge에 포함된 training, development, test 데이터 셋을 사용하여 학습 및 테스트를 진행하였다. Training 데이터에는 50가지 언어 라벨이 있는 15000개의 i-벡터가 포함되어있으며, development 데이터에는 라벨이 없는 6431개의 i-벡터가 포함되어있다. 이 두 데이터 셋은 베이스라인 및 제안된 기법의 학습에 사용하였다. Test 데이터 셋은 OOS 언어가 포함된 6500개의 i-벡터로 구성되어있으며, OOS 언어에 대한 라벨은 언어의 종류에 무관하게 'out-of-set'이라 라벨링 되었다. 데이터 셋에 포함된 모든 i-벡터는 400차원이다. 본 논문에서의 실험들에서는 training set에 포함된 i-벡터들로부터 구한 군내 공분산을 이용하여 WCCN을 적용한 i-벡터들을 사용하였다.

5.1.2 베이스라인 시스템

본 논문에서는 비교를 위하여 베이스라인으로서 추가적으로 네 가지 언어 분류 시스템을 실험하였다. 첫 번째는 cosine distance를 이용한 분류 시스템이며, 모든 i-벡터를 단위 구좌표에 투영시킨 후, 각 언어의 평균 i-벡터와 테스트 i-벡터 사이의 cosine distance를 계산하여 분류하였다[11].

두 번째는 multi-class SVM 기반의 언어 분류이며, 트레이드 오프 (trade-off) 비용 상수 C를 1000으로 설정하고, radial basis function (RBF) 커널을 사용하여 50개의 언어 분류를 위하여 하나의 언어 군과 나머지 언어 군에 대하여 분류하는 SVM 50개를 학습하였다.

세 번째 베이스라인은 시그모이드 출력층을 가진 neural network (NN)이며, 200차원의 단일 시그모이드 은닉층으로 구성되었다. 본 논문에서 제안한 기법의 OOS 여부 판단 방식을 이용하여 오픈 세트를 고려하였다. 은닉층의 개수 및 크기는 실험을 통하여 가장 좋은 결과를 보이는 구성으로 설정되었으며, 학습 과정에서 50% fraction의 dropout 기법[16]이 적용되었다.

본 논문에서 실험한 베이스라인 cosine distance, SVM, 그리고 시그모이드 NN은 각각 CD, SVM, NN이라 지칭하였다.

표 1. 제안하는 알고리즘과 베이스라인 기법들의 EER 성능 비교
Table 1. Equal error rate performance of the proposed method and baseline systems

Classifier	EER (%)
CD	1.41
SVM	1.41
NN	1.39
UDBN-1	**1.38**
UDBN-2	**1.31**
UDBN-3	**1.35**

5.2 제안하는 기법 설정

제안한 기법은 50개 언어에 대하여 50개의 판별 DNN을 학습하였으며, 각 판별 DNN은 단일 시그모이드 출력으로 구성되었다. UDBN의 학습 과정에는 모든 development 데이터 셋과 training 데이터 셋이 사용되었으며, epoch은 10으로 설정하였다. 각 DBN 및 DNN의 학습에 사용된 군 종속적 데이터 셋에는 300개의 타겟 언어 i-벡터가 49회 중복되어 총 14700개로 구성되었으며, 타겟 언어가 아닌 i-벡터는 한 언어당 300개씩 49가지 언어가 포함되어, 총 14700개로 구성되었다.

각 군 종속적 데이터 셋은 해당 데이터 셋의 타겟 언어에 해당하는 DBN 및 판별 DNN을 학습하기 위해 사용되었다. DNN 구조에 따른 성능 비교를 위하여 은닉층으로 512개 노드를 한 개 사용하는 구성, 100개 노드를 두 개 사용하는 구성, 50개 노드를 3개 사용하는 구성, 총 3가지의 은닉층 구성으로 DNN을 학습하였고, 본 실험에서는 이들을 각각 UDBN-1, UDBN-2, UDBN-3로 지칭한다. 제안된 시스템들의 모든 DNN을 학습할 때는 50% fraction의 dropout 기법이 적용되었다. 또한, 적응화 과정에서는 10 epoch 동안 학습하였으며, DNN 미세 조정 학습 과정에서는 학습률 0.05로 50 epoch만큼 학습하였다. 문턱 값은 0.5로 설정하였다.

5.3 결과

본 실험에서는 베이스라인 시스템들과 제안하는 알고리즘을 50개의 언어와 OOS를 포함하는 총 51개 군을 분류하는 오픈 세트 환경에서 equal error rate (EER)을 이용하여 비교하였다. 분류 시스템의 문턱 값을 가변하게 되면 결과의 오인식률(false alarm rate, FAR)과 오거부율(false reject rate, FRR)이 변화하게 되는데, 이 두 수치가 같아질 때의 오류율을 EER이라 정의한다. EER은 분류 시스템이 동작하는 최적의 문턱 값에서의 성능을 나타내며, 화자 인식 분야에서 가장 통상적으로 사용되는 비교 수치 중 하나다[17].

표 1에서와 같이 제안하는 시스템은 오픈 셋 환경에서 기존의 베이스라인 기법들에 비해 우수한 성능을 보이는 것을 확인하였다. 우선 베이스라인 기법들끼리의 결과를 비교하여보면 단순한 유사도 기반의 분류 기법인 CD나 선형적 분류 기법인 SVM에 비하여 비선형적인 특성을 갖는 NN이 우수한 성능을 보임을 볼 수 있다. 이는 음성 특징이 언어 군에 따라 단순히 선형적으로 분포하지 않고, 서로 다른 언어 군들끼리도 유사한 특성을 공유하며 복잡하게 분포하기 때문이다. 제안하는 기법은 여기서 더 나아가서 각 언어 군에 종속적으로 파라미터를 학습한 DNN들을 사용하였기 때문에 모든 언어를 분류하기 위하여 통합적으로 학습된 베이스라인 NN보다 각 언어에 대한 확률을 확실하게 표현할 수 있으며, 보다 높은 성능을 보이는 것을 확인할 수 있다.

실험한 세 가지 구조의 제안된 기법 중에서 10개 노드를 갖는 은닉층 두 개를 사용하는 UDBN-2가 가장 높은 성능을 보였다. UDBN-1은 앞서 기술한 제안된 기법의 장점으로 인하여 기존의 베이스라인 기법들에 비하여 높은 성능을 보였지만, 단일 은닉층을 사용하여 비선형성이 크지 않기 때문에 UDBN-2에 비하여 큰 성능 개선을 보이지는 않았다. 반면 UDBN-3은 UDBN-2에 비하여 많은 수의 은닉층을 사용하여 보다 비선형적인 분류를 할 수 있지만, 너무 많은 파라미터를 사용하기에 과적응으로 인해 성능 저하를 보이는 것이라 볼 수 있다.

Ⅵ. 결 론

본 논문에서는 WCCN을 적용한 i-벡터 특징을 입력으로 하는 다중 DNN 기반의 화자 언어 분류 기법을 실험하였다. NIST 2015 i-vector Machine Learning Challenge 데이터 셋을 이용하여 실험해본 결과, 기존의 SVM 및 단일 NN을 사용한 언어 분류 기법에 비하여 오픈 세트 테스트에서 월등한 성능을 보이는 것을 확인할 수 있었다.

References

[1] G. Hinton, L. Deng, D. Yu, A. Mohamed, et. al., “Deep neural networks for acoustic modeling in speech recognition,” *IEEE Sig. Process. Mag.*, vol. 29, no. 6, pp. 82-97, Nov. 2012.

[2] K. H. Lee, S. J. Kang, W. H. Kang, N. S. Kim, and S. J. Yang, “DNN-based feature compensation using environmental parameter,” in *Proc. KICS ICC 2015*, pp. 72-73, Gangwon, Korea, Jan. 2016.

[3] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, “A novel scheme for speaker recognition using a phonetically-aware deep neural network,” in *Proc. ICASSP 2014*, pp. 1714-1718, Florence, Italy, May 2014.

[4] J. Wang, D. Wang, T. F. Zheng, and F. Bie, *DNN-based discriminative scoring for speaker recognition based on i-vector*, CSLT, Tech. Rep. 20150002, Jan. 2015.

[5] O. Ghahabi and J. Hernando, “Deep belief networks for i-vector based speaker recognition,” in *Proc. ICASSP 2014*, pp. 1700-1704, Florence, Italy, May 2014.

[6] W. H. Kang, K. H. Lee, T. G. Kang, S. J. Kang, N. S. Kim, and K. J. Shin, “Speaker age regression using i-vectors trained with MFCC and pitch,” in *Proc. KICS ICC 2015*, pp. 967-968, Jeju, Korea, Jun. 2015.

[7] W. H. Kang, K. H. Lee, T. G. Kang, and N. S. Kim, “NN based speaker age classification using i-vectors,” in *Proc. KICS ICC 2015*, pp. 589-590, Seoul, Korea, Nov. 2015.

[8] I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, et. al., “Automatic language identification using deep neural networks,” in *Proc. ICASSP 2014*, pp. 5374-5378, Florence, Italy, May 2014.

[9] C. Chang and C. Lin, “LIBSVM: a library for support vector machines,” *ACM TIST*, vol. 2, no. 3, pp. 1-39, Apr. 2011.

[10] W. M. Campbell, E. Singer, P. Torres-Carrasquillo, and D. A. Reynolds, “Language recognition with support vector machines,” in *Proc. Odyssey 2004*, pp. 41-44, Toledo, Spain, May-Jun. 2004.

[11] *The 2015 Language Recognition i-Vector Machine Learning Challenge*(2015), Retrieved Dec. 29, 2015, from http://www.nist.gov/itl/iad/mig/upload/lre_ivectorchallenge_rel_v2.pdf

[12] N. Dehak, P. Kenny, R. Dehak, P. Dumouchei, and P. Ouellet, “Front-end factor

analysis for speaker verification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 788-798, May 2011.

[13] A. O. Hatch, S. S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. Interspeech*, pp. 2-5, 2006.

[14] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Sign. Process.*, vol. 10, pp. 19-41, Jan. 2000.

[15] R. Salakhutdinov, "Learning deep generative models," Ph. D. Dissertation, University of Toronto, 2009.

[16] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, et. al., "Dropout: a simple way to prevent neural networks from overfitting," *JMLR,* vol. 15, no. 1, pp. 1929-1958, Jun. 2014.

[17] S. Furui, *Speaker recognition*(2008), Retrieved Jul., 12, 2016, from http://www.scholarpedia.org/article/Speaker_recognition

강 우 현 (Woo Hyun Kang)

2014년 2월 : 국민대학교 전자공학과 학사 졸업
2014년 3월~현재 : 서울대학교 전기정보공학과 석박통합과정 박사과정
<관심분야> 음성 신호처리, 음성 인식, 화자 인식

조 원 익 (Won Ik Cho)

2014년 8월 : 서울대학교 전기정보공학과 학사 졸업
2014년 9월~현재 : 서울대학교 전기정보공학과 석박통합과정 박사과정
<관심분야> 음원 분리, 자동 채보, 음향 환경 인지

강 태 균 (Tae Gyoon Kang)

2010년 2월 : 서울대학교 전기정보공학과 학사 졸업
2012년 2월 : 서울대학교 전기컴퓨터공학부 석사 졸업
2012년 3월~현재 : 서울대학교 전기컴퓨터공학과 박사과정
<관심분야> 음성 신호처리

김 남 수 (Nam Soo Kim)

1988년 2월 : 서울대학교 전자공학과 학사 졸업
1990년 2월 : 한국과학기술원 전기 및 전자공학과 석사
1998년 3월~현재 : 서울대학교 교수
<관심분야> 음성 인식, 음성 향상, 음성 합성, 머신러닝, 인공지능, 실감 음향, 음원 분리