ISSN 2005-8063
2017. 6. 30.
Vol.9 No.2
pp. 85-94

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

http://dx.doi.org/10.13064/KSSS.2017.9.2.085

PLDA 모델 적응과 데이터 증강을 이용한 짧은 발화 화자검증*

Short utterance speaker verification using PLDA model adaptation and data augmentation

윤 성 욱 · 권 오 욱**
Yoon, Sung-Wook · Kwon, Oh-Wook

Abstract

Conventional speaker verification systems using time delay neural network, identity vector and probabilistic linear discriminant analysis (TDNN-Ivector-PLDA) are known to be very effective for verifying long-duration speech utterances. However, when test utterances are of short duration, duration mismatch between enrollment and test utterances significantly degrades the performance of TDNN-Ivector-PLDA systems. To compensate for the I-vector mismatch between long and short utterances, this paper proposes to use probabilistic linear discriminant analysis (PLDA) model adaptation with augmented data. A PLDA model is trained on vast amount of speech data, most of which have long duration. Then, the PLDA model is adapted with the I-vectors obtained from short-utterance data which are augmented by using vocal tract length perturbation (VTLP). In computer experiments using the NIST SRE 2008 database, the proposed method is shown to achieve significantly better performance than the conventional TDNN-Ivector-PLDA systems when there exists duration mismatch between enrollment and test utterances.

Keywords: time delay neural network (TDNN), identity vector (I-vector), probabilistic linear discriminant analysis (PLDA), vocal tract length perturbation (VTLP)

1. 서론

현대의 화자검증 시스템은 Gaussian mixture model - universal background model (GMM-UBM)을 기반으로 제안되었다[1]. 그 후 화자의 특징을 저차원, 고정된 길이의 벡터로 표현하는 Iidentity vector (I-vector)기반 화자검증 시스템이 제안 되었고, 현대의 화자검증 시스템은 I-vector를 기반으로 구성된다[2]. I-vector는 가변적인 음성 발화의 프레임을 고정된 길이 벡터로 표현한다. 기존의 GMM 슈퍼 벡터에 비하여 훨씬 작은 차수를 가지고 있으며, 화자특징과, 채널 정보를 동시에 가지고 있기 때문에, 채널 보상이 필요하다. 이를 위해 I-vector의 채널 정보를 보상하며, 잘 스코어링 할 수 있는 확률적 선형판별분석 (probabilistic linear discriminant analysis; PLDA)이 사용된다[3]. 최근 I-vector 추출 시 필요한 충분통계량(sufficient statistics; SS)을 모으는데 사용되는 GMM을 deep neural network (DNN)으로 대체함으로써 화자검증 시스템의 성능향상을 이루었다[4],[5]. 최신의 DNN 기반 Ivector-PLDA 화자검증 시스템은 획기적인 성능향상을 이루어내었지만, 이는 등록화자와 테스트화자 사이에 발화길이 왜곡이 발생하지 않는 제약적인 상황에서의 연구 성과이다.

* 이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2015R1D1A3A01020817)
이 논문은 2015년도 충북대학교 학술연구지원사업의 교내연구비 지원에 의하여 연구되었음
** 충북대학교, owkwon@cbnu.ac.kr, 교신저자
Received 26 January 2017; Revised 2 May 2017; Accepted 8 June 2017

충분한 발화길이의 발화에서 추출된 I-vector는 두 가지 변이요인을 가진다. 첫 번째는 화자의 특성이고, 두 번째는 채널(channel) 환경이다. 그러나 최근의 연구에 따르면, 짧은 발화에서 추출된 I-vector는 위에서 기술한 두 가지 변이요인 외에 발화길이 왜곡에 따른 변이요인을 가진다. 같은 화자의 발화라도, 발화길이가 짧아질 때는 모든 음소를 포함하지 못하기 때문에, 포함되어 있는 음소의 종류에 따라 I-vector가 다르게 표현될 수 있다. 때문에 짧은 발화에서 추출된 I-vector는 학습 발화와 테스트 발화 사이에 발화길이 왜곡을 발생시켜 화자검증 시스템의 성능을 저하시킨다[6].

짧은 발화에서 발생하는 I-vector 발화길이 왜곡 문제는 최근 여러 그룹에서 연구되는 주제이다. 짧은 발화에서 생기는 I-vector의 변이요인을 보상하기 위해 짧은 발화를 개발 데이터베이스에 추가하여 시스템을 구축하는 방법과 발화 분산 정규화(short utterance variance normalization; SUVN)를 이용하는 방법을 통해 발화길이 왜곡 보상을 시도했다[7],[8]. 그 외에도 발화길이 왜곡을 보상하는 연구로, I-vector 발화길이에 따른 왜곡을 수량화하고 이를 PLDA 분류기에서 모델링하는 방식을 사용하였다[9]. 또한 I-vector 추출 과정에서 짧은 발화의 효과를 분석하고, PLDA를 이용해 I-vector의 발화길이에 따른 왜곡을 가산 잡음으로 모델링하였다[6].

본 연구에서는 화자검증 시스템에서 짧은 발화와 긴 발화의 차이로 인한 시스템 성능저하를 보상하기 위해, 도메인 적응 기법으로 사용되는 PLDA 적응 기법을 사용하여 기존에 긴 발화로 학습된 PLDA 모델을 짧은 발화에 적합한 PLDA모델로 적응한다. PLDA 적응 기술은 대량의 화자에 대한 정보가 있는 영역외 데이터로 학습된 신뢰성 있게 학습된 PLDA 모델이 있고, 화자 정보가 없는 영역내 PLDA 모델을 구축하고 싶을 때 영역외 PLDA 모델을 영역내 데이터(화자 정보가 없는)로 학습된 영역내 PLDA 모델로 적응시켜 사용하는 도메인 적응 기술로 활용되었다[10].

PLDA는 I-vector를 채널공간과 화자공간으로 분리하고 가장 잘 스코어링할 수 있는 최신의 기술이지만, 충분한 양의 화자정보가 있는 음성이 제공되지 못할 시에는 신뢰성 있는 PLDA 모델을 학습할 수 없다는 단점이 있다[11]. 하지만 화자인식 데이터베이스로 널리 사용되는 National Institute of Standards and Technology Speaker Recognition Evaluation (NIST SRE) 데이터베이스에서는 화자모델을 신뢰성 있게 모델링하기 위해 대부분의 음성파일이 긴 발화로 구성되어 있으며, 소량의 짧은 발화 음성파일만을 테스트용으로 제공한다. 때문에 신뢰성 있는 짧은 발화 PLDA 모델을 학습시키기 어렵다. 이를 극복하기 위하여 성도 길이 변화(vocal tract length perturbation; VTLP) 기법을 이용해 NIST SRE 데이터베이스에 부족한 짧은 발화를 증강시킨 다음 PLDA 적응 기법을 적용한다.

본 논문의 구성은 다음과 같다. 2절에서는 TDNN 기반 Ivector-PLDA 화자검증 시스템에 대해 소개하고, 3절에서 제안한 방법을 설명한다. 4절에서는 실험 및 결과를 분석하고, 5절에서 결론을 맺는다.

2. TDNN 기반 화자검증 시스템

본 연구에서는 화자검증 분야의 최신 시스템인 TDNN 기반 Ivector-PLDA 화자검증 시스템을 사용한다. 음성인식에서 성능이 좋은 DNN 모델이 DNN 기반 화자검증 시스템에서도 좋은 성능을 보임이 알려져 있다[4]. 최근 서브 샘플링을 적용한 time delay neural network (TDNN)과 recurrent neural networks (RNN)이 기존의 DNN에 비해 대어휘 음성인식에서 상대적으로 큰 성능 향상을 이루어냈다. 특히 서브 샘플링을 적용한 TDNN이 RNN에 비해 Switchboard (SWB) 태스크에서 상대적으로 11% 높은 성능을 보였다[12]. 기존에 제안된 TDNN 모델에 서브 샘플링 기법을 적용함으로써 연산량은 줄이면서, 대어휘 음성인식 테스크에서 성능 향상을 이루어 냈다[12],[13]. 기존 TDNN은 같은 층의 개수를 가진 DNN에 비해 10배의 연산 시간이 소요되며, 서브 샘플링을 적용한 TDNN은 DNN에 비해 2배의 연산 시간이 소요된다. 뿐만 아니라 음성인식기의 성능에 있어서도 서브샘플링을 적용한 TDNN이 기존의 TDNN보다 높은 성능을 보임이 보고되었다[12].

본 연구에서는 전반적인 화자검증 시스템의 정확도 향상을 위해 DNN을 TDNN으로 대체하여, TDNN 기반 Ivector-PLDA 화자검증 시스템을 베이스라인 시스템으로 사용한다. 전체 블록도는 <그림 1>에 나타나 있으며, 베이스라인 시스템의 학습과 화자 검증 방식을 보여준다. 전체 시스템은 훈련 과정, 평가 과정(등록 과정과 테스트 과정)으로 나눠 볼 수 있다.

훈련 과정에서는 개발 데이터베이스(대량의 화자인식 DB)와 전사정보가 있는 음성인식 DB와 음성인식기를 이용하여 TDNN을 학습시킨다. 훈련과정에서 개발 DB는 대량의 불특정 다수 화자들로부터 수집된 배경화자 음성데이터이다. 이는 일반적인 화자특징에 대한 모델링을 목적으로 하기 때문이다. 이후 TDNN과 특징벡터를 이용하여 충분통계량(sufficient statistics)을 수집하고 이를 바탕으로 전체 변이 행렬(total variability matrix; T-matrix)을 모델링한다.

평가 과정에서는 대량의 개발 데이터베이스로 훈련된 모델들에 대해 등록화자와 테스트화자에 대해 훈련 과정에서 학습된 모델을 등록화자와 테스트화자에 적합하게 적응한다. 등록 과정에서는 등록화자의 발화를 이용해 특징 벡터를 추출한다. 후에 훈련 과정에서 학습된 TDNN을 이용하여 프레임 사후확률과 등록DB에서 추출된 특징벡터를 이용해 충분통계량을 수집한 후 수집된 통계 정보와 훈련 과정에서 생성한 T-행렬을 이용하여 I-vector를 추출한다. 이를 등록화자 I-vector라고 한다.

테스트 과정에서는 앞서 설명한 등록과정과 같은 절차로 테스트화자 I-vector를 추출한다. 최종적으로 등록화자 I-vector 와 테스트화자 I-vector를 이용해 우도비를 계산하여 사전에 설정된 문턱치 값에 따라 등록 발화와 평가 발화의 일치 여부를 결정하게 된다. 우도비는 등록화자와 테스트화자가 동일한 화자라는 가정 하에 계산한 로그우도 값에서, 등록화자와 테스트 화자가 서로 다른 화자라는 가정 하에 계산된 로그우도 값을 뺀 값을 사용한다[24].

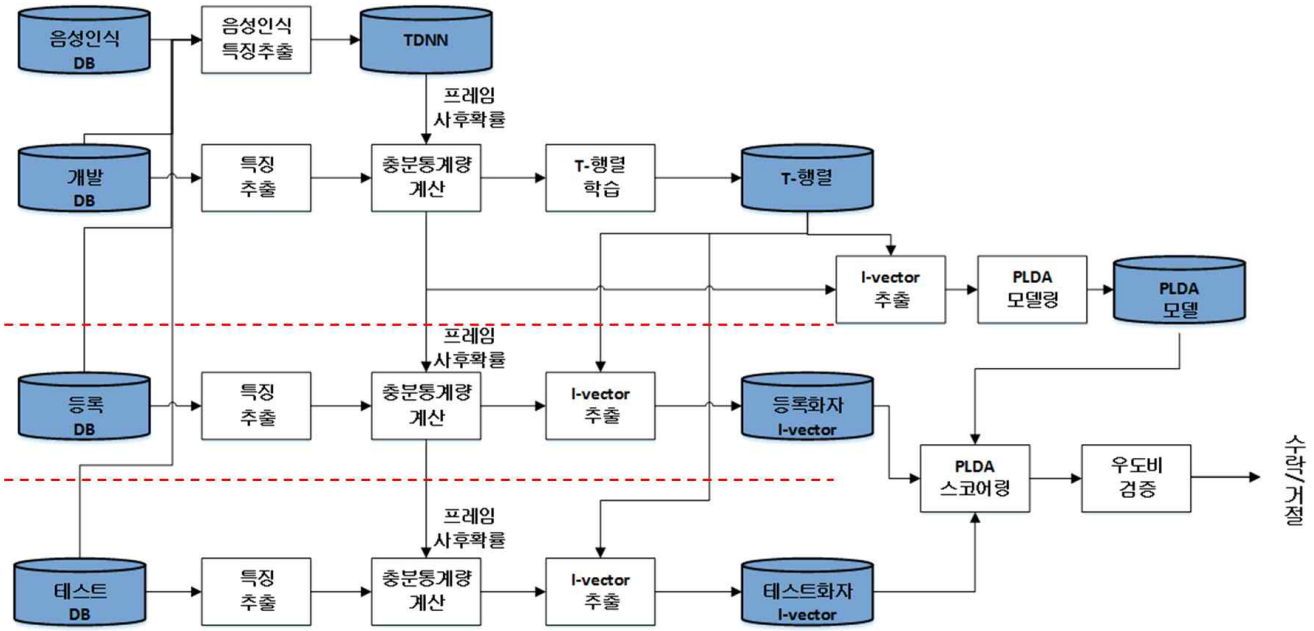

그림 1. 베이스 라인 시스템의 전체 블록도
Figure 1. Overall block diagram of the baseline system

2.1. I-vector model

I-vector는 Gaussian mixture model (GMM) 슈퍼벡터를 하나의 Total-variability 공간으로 표현한다. I-vector는 기존의 joint factor analysis (JFA) 기술에서 채널공간이 채널정보뿐만 아니라 화자 분별 정보를 포함하고 있다는 발견에서 시작되었다[14]. 화자, 채널 정보를 모두 포함하는 GMM 슈퍼벡터(가우시안 혼합 모델들의 평균을 연쇄시킨 벡터) μ는 다음과 같이 표현된다.

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

m은 universal background model (UBM) 슈퍼벡터이고 대량의 개발 데이터베이스로부터 학습된다. T는 전체 변이 행렬(total variability matrix; T-matrix) 이며, w는 전체 변이 요인(total-variability factor)으로서 I-vector를 나타낸다. 한 발화는 고정된 길이의 I-vector로 표현되며, 화자 특징을 표현하는 저차원의 강력한 벡터이다. t번째 음향벡터 $\mathbf{x}_t$에 대한 I-vector를 구하기 위해 계산되어야 하는 충분통계량은 다음과 같다.

$$N_c(u) = \sum_{t=1}^{L} P(c|\mathbf{x}_t, \Theta_{UBM}) = \sum_{t=1}^{L} \gamma_t(c) \tag{2}$$

$$F_c(u) = \sum_{t=1}^{L} P(c|\mathbf{x}_t, \Theta_{UBM})\mathbf{x}_t = \sum_{t=1}^{L} \gamma_t(c)\mathbf{x}_t \tag{3}$$

$$S_c(u) = \mathrm{diag}\left(\sum_{t=1}^{L} \gamma_t(c)\mathbf{x}_t\mathbf{x}_t^{T}\right) \tag{4}$$

N_c, F_c, S_c는 각각 0차(zeroth-order), 1차(first-order), 2차(second-order) 충분통계량(sufficient statistic)이며, u는 한 발화를 의미하며, t는 프레임 인덱스, L은 한 발화에 포함되는 프레임의 개수, c는 c번째 가우시안 요소의 인덱스를 의미한다. $\Theta_{UBM} = \{\mu_c, \Sigma_c, \pi_c\}$은 UBM의 평균, 공분산행렬, 가중치 파라미터를 의미한다. 충분통계량 계산에 필요한 사후확률은 다음 식과 같이 표현된다.

$$\gamma_t(c) = \frac{\pi_c P_c(\mathbf{x}_t|\mu_c, \Sigma_c)}{\sum_{i=1}^{C} \pi_i P(\mathbf{x}_t|\mu_i, \Sigma_i)} \tag{5}$$

여기서 $\gamma_t(c)$는 t번째 프레임의 c번째 가우시안 요소의 사후확률을 의미한다. 충분통계량들은 T-행렬의 학습과 I-vector 추출에 필요하다. 주어진 발화에 대한 I-vector는 다음 수식에 의해 계산된다[6].

$$w = (I + T^{T}\Sigma^{-1}N(u)T)^{-1}T^{T}\Sigma^{-1}F(u) \tag{6}$$

여기서 I는 단위행렬, $(\cdot)^{T}$는 행렬의 전치를 나타낸다. $N(u)$는 주대각선 성분이 $N_c(u)$들인 대각행렬(diagonal matrix)이며, $F(u)$는 $F_c(u)$들을 연쇄(concatenation)시켜 만든 행렬이다. 이렇게 추출된 I-vector는 화자 정보와 채널 정보를 동시에 포함하고 있기 때문에, PLDA를 통해 화자 정보와, 채널 정보로 분리되고, 스코어링된다.

2.2. TDNN

기존의 I-vector 시스템은 충분통계량들을 계산하기 위해 필요한 음향 프레임의 soft alignment를 제공하기 위해 GMM-UBM을 사용한다[15]. GMM의 각 mixture들은 화자들이 어떻게 다른 음향 특징을 가지는 지를 특징짓고 이를 클래스 별로 표현한다. 이상적으로 GMM의 각 클래스가 특정 음소에 대응된다면 같은 음소 내에서 화자간의 특징을 비교할 수 있다. 하지만 GMM의

한 클래스는 임의의 화자공간을 표현하기 때문에 이러한 효과를 기대할 수 없다. 이를 위해 tied triphone state (senone)로 학습된 음성인식 DNN모델로 기존의 GMM모델의 역할을 대체함으로써, 화자검증 시스템에 DNN을 결합시키는 연구가 이루어졌으며 UBM기반 화자검증 시스템에 비해 월등히 높은 성능을 보였다[16]. GMM의 각 mixture를 TDNN의 사후확률 클래스(senone)에 대응시켜 충분통계량을 구하여 I-vector를 추출한 supervised-GMM 기법 또한 TDNN의 성능보다는 떨어지지만 기존의 GMM -UBM 시스템에 비해 높은 성능을 보였다[26]. 이를 보면 TDNN 기반 시스템의 성능 우수성이 UBM모델 생성시 음소 정보를 반영한 방법에 있다고 볼 수 있다.

UBM기반 시스템과 TDNN의 기반 시스템의 차이점은 <그림 2>에서 볼 수 있다. UBM기반 시스템에서는 프레임 사후확률 계산을 위해 훈련과정에서 학습된 UBM을 활용해 충분통계량을 계산하고, I-vector를 계산하는데 사용한다. TDNN기반 화자검증 시스템에서는 전사정보가 있는 음성인식 DB를 이용하여, TDNN을 학습시키고, 이를 이용해 기존의 UBM시스템에서 프레임 사후확률과, 0차 통계량을 구하는 GMM의 역할을 대체한다. 본 연구에서는 음성인식 데이터베이스인 Wall Street Journal (WSJ) 데이터베이스를 이용하여 학습시킨 TDNN을 이용하였다[17]. 프레임 사후확률 값을 대체하여 기존 시스템 구조를 유지하면서 UBM의 역할만을 TDNN이 대체하였다. (b)에서 “음성인식 특징추출”의 의미는 프레임 사후확률 계산 시 TDNN의 입력으로 사용되는 벡터와, TDNN을 학습할 때 사용된 입력 벡터가 같은 방법으로 추출되게 됨을 의미한다.

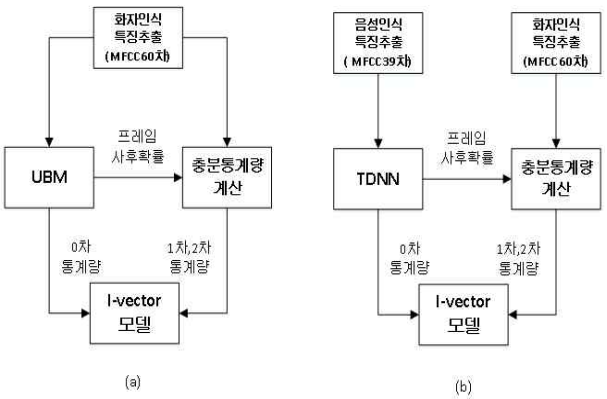

그림 2. I-vector 추출 블록도 (a)UBM 기반 (b)TDNN 기반
Figure 2. Block diagram of I-vector extraction
(a) UBM-based (b)TDNN-based

3. 짧은 발화 화자검증을 위한 제안 방법

본 연구에서는 최신의 TDNN-Ivector-PLDA 화자검증 시스템을 기반으로, 짧은 발화시 I-vector의 왜곡을 보상하기 위해 PLDA 적응 기법과 VTLP기법을 이용한다.

3.1. PLDA 적응

PLDA는 I-vector를 분류하고 스코어링할 수 있는 최신의 기술이지만, 방대한 양의 화자정보가 있는 음성으로 학습될 경우에 성능을 보장할 수 있다[11]. 하지만 현실적으로 화자정보가 포함된 음성정보를 제공 받을 수 있는 도메인은 흔치 않다. 예를 들어 유튜브나 페이스북과 같은 소셜 미디어에서 녹음된 화자의 음성에 대해 화자검증 시스템을 적용하는 경우 PLDA 모델을 신뢰성 있게 학습하기 많은 제약이 따른다. 왜냐하면 유튜브와 페이스북과 같은 소셜 미디어에는 무한에 가까운 음성정보가 존재하지만, 화자 정보를 포함한 음성은 극소수이다. 이런 경우 대량의 화자정보가 있는 음성으로 학습된 신뢰성 있는 영역외의 PLDA 모델을 이용해 목표 도메인에 PLDA 모델을 생성하는 용도로 PLDA 적응 기법을 사용한다[10]. 본 논문은 기존에 도메인 적응 분야에서 도메인에 따른 PLDA 모델 파라미터의 다른 경향성을 극복하기 위해 성공적으로 사용된 PLDA interpolation을 이용한 적응 기법을 긴 발화와 짧은 발화 사이의 PLDA모델 파라미터의 다른 경향성을 극복하기 위해 사용하였다. 이는 가장 단순하면서도 효과적으로 PLDA모델을 적응할 수 있는 방법이기 때문이다.

같은 화자가 발화한 음성이라도 긴 발화에서 추출된 I-vector와, 짧은 발화로 추출된 I-vector는 상당히 다른 경향을 가진다[6]. 때문에 긴 발화의 I-vector로 학습된 PLDA와 짧은 발화의 I-vector로 학습된 PLDA의 파라미터들 또한 상당히 다른 경향을 가지게 학습된다. 이점에 착안하여 본 논문에서는 기존에 도메인 영역 적응을 위해 사용되는 PLDA 적응 기법을 짧은 발화 화자검증 시스템에 적용하기 위해 사용한다. 화자인식에 널리 사용되는 NIST SRE 데이터베이스는 대량의 화자정보가 있는 음성파일로 구성 되어 있지만, 대부분이 긴 발화로 구성되어 있고, 극소량의 테스트를 위한 짧은 발화를 제공한다. 때문에 절대적인 데이터의 부족으로 짧은 발화 PLDA 모델을 신뢰성 있게 구축할 수 없다.

이 점을 극복하기 위해 본 논문에서는 기존에 긴 발화로 학습된 PLDA 모델을 짧은 발화 PLDA 모델로 적응하기 위해 PLDA 모델 적응 기법을 사용한다. PLDA모델 식은 다음과 같다.

$$w_{s,h} = w_0 + \Phi\beta_s + \epsilon_{s,h} \tag{7}$$

위의 수식에서 s는 화자, h는 세션을 표현한다. w_0는 개발 데이터베이스의 I-vector들의 평균값이고 $R\times 1$ 차원 벡터이다. R은 I-vector의 차수이다. Φ는 화자모델을 표현하는 저차원 행렬이고 각 열은 화자를 표현하는 기저벡터인 고유음성(eigenvoice) 벡터이며, $R\times N$ 차원을 가진다. $\beta_s \sim N(0,I)$는 $N\times 1$ 차원의 잠재 변수(latent variable)이고 화자 요인(speaker factor)이며 고유음성 벡터를 조정하여 화자특징을 결정한다. 그리고 $\epsilon_{s,h} \sim N(0,\Sigma)$는 랜덤 벡터이며 잔여 잡음 통계정보를 표현한다. I-vector는 다음과 같은 $N(w_0, \Phi\Phi^{\mathrm{T}} + \Sigma)$ 가우시안 분포를 따른다. $\Gamma = \Phi\Phi^T$는 화자클래스 간의 공분산 행렬로 정의되며, Σ는 화자클래스 내의 공분산 행렬을 표현한다. 적응 과정을 구

체적으로 보면 긴 발화 PLDA 파라미터($\Gamma_{long}, \Sigma_{long}$)와 짧은 발화 PLDA 파라미터($\Gamma_{short}, \Sigma_{short}$) 가 주어져 있을 때 PLDA 적응 수식은 다음과 같다[10].

$$\Gamma_{adapt} = \gamma\Gamma_{short} + (1-\gamma)\Gamma_{long} \quad (8)$$

$$\Sigma_{adapt} = \gamma\Sigma_{short} + (1-\gamma)\Sigma_{long} \quad (9)$$

($\Gamma_{adapt}, \Sigma_{adapt}$)는 적응된 PLDA 파라미터이다. 파라미터 γ는 긴 발화 PLDA 파라미터와, 짧은 발화 PLDA 파라미터의 영향을 결정하는 적응 가중치이다. 목표 도메인의 PLDA 파라미터는 짧은 발화 PLDA 파라미터에 해당하고, 영역외 PLDA 파라미터는 긴 발화 PLDA 파라미터에 해당한다. PLDA적응 기법을 발화 길이 왜곡 보상을 위해 사용할 때 NIST SRE 데이터베이스에 화자 정보가 있는 짧은 발화 데이터베이스의 부족으로 신뢰성 있는 짧은 발화 PLDA 파라미터를 얻을 수 없어, PLDA 적응 후에도, 짧은 발화 화자검증 시스템에 적합한 적응된 PLDA 파라미터를 생성할 수 없었다. 때문에 추가적으로 짧은 발화 음성을 VTLP를 이용해 증강한 후 짧은 발화 PLDA 파라미터를 생성하여, PLDA 적응 기법을 적용한다.

3.2. VTLN 과 VTLP

성도(vocal tract)는 인두와 구강으로 구성되는 발성기관의 일부이다. 성도 길이(vocal tract length)는 성문(glottis)에서 입술까지의 거리로 개인에 따라 차이가 나며 평균 17cm 정도 된다. 특히 남녀에 따라 큰 차이를 보인다. 일반적으로, 성도 길이 정규화(vocal tract length normalization; VTLN)는 위에서 설명한 성도 길이 차이 때문에 일어나는 음성의 변이를 제거하여 음성인식기의 성능을 향상하는 방법으로, 음성 신호 스펙트럼의 주파수 축을 워핑하여 성도길이를 정규화 하는 기술이다[18].

<그림 3>은 VTLN을 적용하여 MFCC를 구하는 과정이다. 기존의 MFCC를 구하는 과정에 선형(linear) 또는 이중선형(bilinear)과 같은 워핑 함수에 의한 성도길이 정규화 과정을 추가하여 성도길이 정규화를 한다. 이후에 필터뱅크와 로그함수가 취해지고 마지막 단계에서 이산 코사인 변환(discrete cosine transform; DCT)가 적용되어 MFCC가 만들어진다.

그림 3. VTLN을 적용한 MFCC 추출과정
Figure 3. MFCC extraction process with VTLN

워핑 함수의 형태는 여러 가지가 있다. 여러 가지 워핑 함수들 중 근사화된 선형(piece-wise linear) 함수가 워핑 함수로 가장 우수한 성능을 보임을 실험을 통해 증명하였다[19]. 때문에 본 논문에서는 근사화된 선형 함수를 워핑 함수로 사용하여 실험을 진행하였으며, 수식은 다음과 같다.

$$w(f) = \begin{cases} \alpha f & f \le f_{vtln} \\ \alpha f_{vtln} + \dfrac{f_0 - \alpha f_{vtln}}{f_0 - f_{vtln}} (f - f_{vtln}) & f \ge f_{vtln} \end{cases} \quad (10)$$

여기서 f_{vtln}은 VTLN 주파수, f_0는 제한주파수이다. f_{vtln}까지는 스펙트럼은 워핑 계수 α로 선형 워핑되고, f_{vtln} 이후부터 f_0 까지는 $w(f_0) = f_0$이 되도록 다른 워핑 계수가 적용되어, 생략되는 주파수 영역이 없도록 한다. 근사화된 선형 함수의 그래프는 <그림 4>와 같다.

앞서 설명한 VTLN은 성도의 길이 때문에 오는 음성인식 시스템의 변이요인을 적절한 워핑 계수 α를 구하여 성도길이 차이를 정규화 하는 기술이라면, VTLP 기법은 각 발화에 대해 랜덤 워핑 계수 α를 생성하고, (10)의 식과 같이 주파수 f를 새로운 주파수 f'으로 매핑하여 주파수 축을 왜곡 시키는 방법으로 원본 음성을 변환할 수 있다. 원본 음성을 적절한 워핑 계수(warping factor) α로 변환하면 데이터 증강효과를 볼 수 있다. VTLP는 음성인식에서 데이터 증강 방식으로 성공적으로 사용되었다[20].

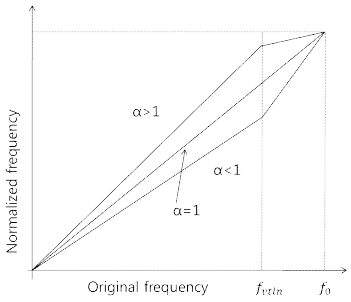

그림 4. 근사화된 선형 함수의 그래프
Figure 4. Graph of piece-wise linear function

3.3. VTLP 를 이용한 PLDA 적응

본 연구에서는 VTLP를 이용하여 워핑 계수의 개수에 비례해서 원본 음성을 증강한다. 하나의 원본음성파일에 대해 적절한 범위의 워핑 계수들을 정하고, 워핑 계수의 개수에 비례 하여 데이터를 증강한다. 증강된 짧은 발화로 짧은 발화 PLDA 파라미터를 생성하고, PLDA적응을 이용해 기존에 대부분이 긴 발화인 대량의 개발 데이터베이스로 신뢰성 있게 학습된 긴 발화 PLDA 파라미터를 짧은 발화 PLDA 파라미터를 이용하여 짧은 발화에 적합하게 적응한다.

PLDA는 지도 학습이기 때문에 VTLP를 이용하여 짧은 발화를 증강한 후 짧은 발화 PLDA 파라미터를 생성하려면, 증강된 짧은 발화에 대한 화자정보가 필요하다. VTLP는 성도길이를 변화시키는 효과를 주어 데이터를 증강하기 때문에, 원본 음성의 화자 특성을 변화시킨다. 원본 음성의 화자 정보는 VTLP로 증강된 음성의 화자정보로 쓸 수 없기 때문에 증강된 데이터에 대한 새로운 화자 정보가 필요하다. 이 때 증강된 짧은 발화에서 추출된 I-vector에 유사도 비교 함수로 간단한 코사인 유사도 스

코어링(cosine similarity score; CSS)을 활용하여 bottom-up 클러스터링 하면, 신뢰할 만한 화자 정보를 생성할 수 있다. 이렇게 하여 얻어진 새로운 화자정보로 짧은 발화 PLDA 파라미터를 학습한 후 PLDA 적응 기법을 적용할 수 있다[10].

PLDA 적응과 VTLP를 이용한 발화길이 왜곡 보상 기법은 <그림 5>에 나타나 있다. <그림 1>의 PLDA모델을 <그림 5>의 적응된 PLDA 모델로 대체하면, 본 논문에서 제안한 VTLP를 이용한 PLDA적응 짧은 발화 화자검증 시스템의 전체 블록도가 된다. 본 논문의 PLDA 적응 기법과 VTLP는 Kaldi 툴킷[21]을 활용하여 구현되었다.

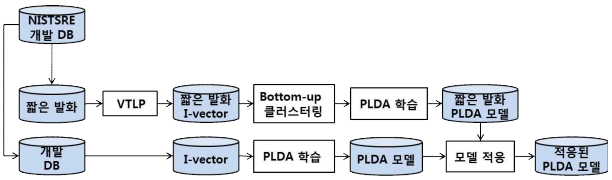

그림 5. VTLP를 활용한 PLDA 모델 적응
Figure 5. PLDA model adaptation using VTLP

4. 실험결과

4.1. GMM-UBM 성능

TDNN 기반 Ivector-PLDA 시스템은 GMM-UBM 기반 Ivector-PLDA 시스템에 비해 높은 성능을 보인다고 알려져 있지만, 이는 1,800시간 분량의 Fisher 데이터베이스를 사용하여 DNN을 학습시켰을 때이다[22]. 본 절에서는 비교적 훨씬 작은 양의 80시간 분량의 WSJ 데이터베이스로 TDNN을 학습시켜 TDNN 기반 화자검증 시스템을 구성했다. 데이터베이스 크기가 화자검증 시스템 성능에 미치는 영향을 조사해 본다. WSJ 데이터베이스는 TDNN을 학습시키는 용도로만 사용되기 때문에, 공정한 비교를 위하여 GMM-UBM 기반 화자검증 시스템에서 UBM 학습시에 WSJ을 추가적으로 사용할 필요는 없다[4],[5],[16],[26].

25ms 크기의 매 프레임에 대하여 20차 MFCC와 1차 미분 및 2차 미분을 추가하여 60차의 MFCC 벡터를 추출하였다. GMM-UBM은 2,048차 가우시안 혼합 모델이고, SWBⅡ- phaseⅡ와 NIST SRE 데이터베이스들(2004,2005,2006)로 학습시켰다. 400차원의 I-vector를 사용하였고, SWB와 NIST SRE 데이터베이스들을 모두 사용하여 성별 종속하게 학습하였다. 뒷단에서는 PLDA를 사용하여 스코어링하였으며 PLDA 역시 SWBⅡ-phaseⅡ와 NIST SRE 데이터베이스 모두를 사용하여 학습하였다. 성능 평가는 NIST SRE 2008을 사용하였다[25].

4.2. TDNN 기반 화자검증 시스템의 성능

Kaldi toolkit에서 제공하는 s5 recipe를 이용하여 학습된 음성인식기를 이용하여 프레임 사후확률을 계산하였다[21]. 본 연구에서 사용한 TDNN은 25ms 프레임 길이의 39차 MFCC를 입력으로 사용하고 WSJ로 학습되었다. TDNN은 4개의 은닉층을 가지며, 은닉층은 p-norm (p=2)를 활성함수로 사용한다[12]. TDNN의 샘플링 계수는 layer0[-2,1], layer1{-1,1}, layer2{-1,2}이다. layer0[-2,1]는 layer1의 한 노드를 중심으로 왼쪽 2개의 노드, 오른쪽 1개의 사이의 모든 노드의 값을 아핀 변환(affine transform)함을 의미하며, layer1{-1,1}는 layer2의 한 노드를 중심으로 왼쪽 1개, 오른쪽 1개의 노드 값만을 아핀 변환함을 의미한다. 소프트맥스(softmax) 출력층은 3048 senone에 대한 사후확률 값을 출력한다. TDNN -UBM은 3048차 senone에 대한 사후확률 값을 이용하여 I-vector를 추출한다. 화자검증 시스템은 600차원의 I-vector를 사용하였고, SWB와 NIST SRE 데이터베이스들을 모두 사용하여 성별 종속하게 학습하였다. 뒷단에서는 PLDA를 사용하여 스코어링 하였으며 PLDA 역시 SWB와 NIST SRE 데이터베이스 모두를 사용하여 성별 종속하게 학습하였다. 성능 평가는 NIST SRE 2008을 사용하였다. 본 논문에서는 WSJ DB를 이용해 음성인식기를 학습할 때 가장 적합한 senone의 수를 실험적으로 3048개로 정하여 사용하였으며, TDNN 기반의 화자검증 시스템의 또한 이에 맞게 3048차로 학습된다. GMM-UBM기반 시스템의 가우시안 차수는 일반적으로 가장 적합한 차수인 2048차로 학습하였다. GMM-UBM기반 시스템의 가우시안 차수는 2048차 이상을 사용하더라도 추가적인 성능 향상이 미미하다고 보고되어, 본 논문에서는 2048차 GMM을 사용하였다 [5],[26].

<표 1>의 첫 번째, 두 번째 행은 GMM-UBM 기반 화자검증 시스템과 TDNN 기반 화자검증 시스템의 SRE08 핵심 평가 데이터인 short2-short3의 공통 평가 조건(common evaluation condition)별 동일오류율(equal error rate; EER)을 보여준다. 평가 데이터는 등록 화자가 발성한 음성 DB-테스트 화자가 발성한 음성 DB로 구성된다[25]. short2, short3는 각각 5분 길이의 전화 대화체, 3분 길이의 인터뷰대화체 발췌 파일이다. EER이 낮을수록 신뢰성 있는 시스템을 의미한다. 표에서 알 수 있듯이 TDNN 기반 화자검증 시스템이 모든 공통 평가 조건에 걸쳐서 GMM -UBM 기반 시스템 화자검증 시스템에 비해 높은 성능을 보임을 알 수 있다. 공통 평가 조건 1~5는 실험에 사용되는 학습, 테스트 모든 음성이 인터뷰 혹은 전화 대화체 타입의 음성이고 공통 평가 조건 6~8은 학습, 테스트에 사용되는 모든 음성이 전화 대화 음성이다. 그 중에서 공통 평가 조건 7(모든 학습 테스트 음성들이 영어 전화 대화음성)을 살펴보면 상대적으로 42%의 성능 향상이 이루어졌음을 알 수 있다. 1,800시간의 Fisher 데이터베이스를 사용하지 않고, 비교적 작은 양인 80시간의 WSJ 데이터베이스로 학습한 TDNN 기반 화자검증 시스템이 기존의 GMM-UBM 기반 화자검증 시스템에 비해 월등히 좋은 성능을 보임을 확인할 수 있었다.

기존의 연구결과에 따르면, Fisher 데이터베이스로 학습한 TDNN기반 화자검증 시스템이 SRE10 평가 데이터를 대상으로, GMM기반 화자검증 시스템에 비해 상대적으로 50% 높은 EER을 보였다[26]. WSJ이 Fisher 데이터베이스에 비해 절대적으로 데이터베이스의 크기가 작음을 감안하면, 상대적으로 42% EER의 성능 향상은 신뢰할만한 수준의 결과로 보인다. 본 논문에서 다음 실험부터는 TDNN -UBM 기반 ivector-PLDA 화자검증 시

스템을 베이스라인 시스템으로 사용한다.

표 1. short2-short3(남성) ,short2-10sec(남성) 평가데이터의 EER(%)
Table 1. EER(%) of short2-short3(male), short2-10sec(male) evaluation data

공통 평가 조건	1	2	3	4	5	6	7	8
GMM-UBM short2-short3	9.51	1.21	9.61	8.43	5.94	6.52	4.33	2.63
TDNN short2-short3	7.77	0.81	7.91	7.29	4.22	5.03	2.73	1.75
TDNN short2-10sec	NA	NA	NA	NA	NA	8.27	6.92	6.58

4.3. 발화길이 왜곡의 영향

긴 발화와 짧은 발화로 구성된 평가 데이터에서의 화자검증 시스템의 성능 저하는 앞 절에서 기술한 것과 같이 자연스러운 현상이다. <표 1>의 3번째 행인 발화길이 왜곡이 나타나는 평가 데이터인 short2-10sec(남성) 평가 데이터에 대한 베이스라인 시스템의 EER을 보여준다. 10sec는 전화 대화체에서 10초간의 음성 발췌 파일이다. short2-10sec에서는 공통 평가 조건 1~5에 대한 평가는 제공하지 않는다. 예상한 바와 같이 짧은 발화(10sec) 데이터베이스에서 추출된 I-vector와, 긴 발화(short2) 데이터베이스에서 추출된 I-vector 사이의 내용 불일치가 발생함에 따라 발생하는 I-vector의 발화길이 왜곡이 화자검증 시스템의 EER 상승으로 이어짐을 확인할 수 있었다.

짧은 발화 상황에서 I-vector 발화길이 왜곡을 위한 선행연구를 보면, PLDA 모델을 학습할 때, 긴 발화와 짧은 발화를 동시에 학습하면 발화길이 왜곡이 있는 평가 셋에서 추가적인 성능 향상을 보였다[7]. 본 연구에서는 선행연구로 PLDA 모델을 학습할 때 사용한 NIST SRE 데이터베이스들에서 처음 10sec의 구간을 발취하여 PLDA 모델 학습 시에 긴 발화와 짧은 발화를 동시에 학습하였으나, short2-10sec 평가 셋의 공통 평가 조건 6~8에 대해 각각 8.86%, 7.31%, 7.58%로 베이스라인 시스템에 비해 EER이 소폭 상승함을 확인했다. 기존 Kanagsundaram(2012)의 연구에서 PLDA 모델 학습 시에 구체적으로 어떻게 짧은 발화를 추가하였는지 기술되어 있지 않아 같은 방법으로 확인해 볼 수 없었으나, 단순히 짧은 발화를 PLDA 모델 학습 시에 추가하는 방법으로는 추가적인 성능 향상을 확인할 수 없었다. 이는 기존에 긴 발화로 신뢰성 있게 학습된 화자모델에 단순히 소량의 10sec 발화를 발취해서 넣게 되면, 기존 화자모델에 noise로 작용할 수 있기 때문으로 보인다. 이러한 방법으로 추가적인 성능향상을 위해서는 긴 발화에서 통계적 신뢰성을 확보할 수 있을 만한 양의 짧은 발화를 발취하여 PLDA모델 학습 시 같이 사용해야할 것으로 예상된다.

본 연구의 베이스라인 화자검증 시스템이 신뢰성 있게 구축되어 있는지 확인하기 위해 타 연구의 EER 결과와 비교해본다[23]. 본 논문에서 비교 대상으로 설정한 타 연구를 보면 I-vector의 발화길이 왜곡 문제를 해결하기 위해 GMM-UBM 기반 Ivector-PLDA시스템을 기반으로, 추출된 I-vector에 LDA를 이용한 차원 축소 기법을 적용해 세션 변이(session variability)를 보상한 후 PLDA로 스코어링하였다[23]. 평가는 NIST SRE 2008 short2 -short3(남성)와 short2-10sec(남성)에 대해 이루어졌으며, 남성 화자에 공통 평가 조건 7에 대해 평가하였다. 나머지 공통 평가 환경에 대해서는 EER을 제공하지 않고 있다.

본 논문의 베이스라인 성능이 SRE08 short2-short3 공통 평가 조건 7에서 상대적으로 22%, short2-10sec 공통 평가 조건 7에서 상대적으로 40%의 EER 감소를 보임을 확인 하였다. 이는 TDNN 기반 시스템이 GMM-UBM 기반 시스템에 LDA 기법을 적용한 시스템 보다 세션 변이(session variability)보상 및 발화길이 왜곡 보상에 더욱 효과적임을 보이고 있다. 다음 실험부터는 발화길이 왜곡이 나타나는 평가데이터인 short2-10sec에 대해서만 실험한다.

4.4. 발화길이 왜곡 보상을 위한 PLDA 모델 적응의 성능

같은 화자가 여러 상황에서 발화할 경우, 발화 길이가 짧은 음성에서 추출된 I-vector들은 긴 발화에서 추출된 I-vector들에 비해 큰 분산을 가지게 되며, 그 크기 또한 발화길이가 짧아질수록 0에 가까워진다[6]. 이러한 동기에서 본 논문에서는 PLDA 적응 기법을 통해 기존에 대부분 긴 발화로 구성된 개발 데이터베이스로 학습된 PLDA 모델을, 짧은 발화로 학습된 PLDA 파라미터를 이용해 짧은 발화보상에 적합한 PLDA 파라미터로 적응시킨다. 짧은 발화 학습을 위하여 개발 데이터베이스 NIST SRE 2004, 2005, 2006에서 10sec, 30sec 만을 이용하였다.

표 2. PLDA 적응을 적용한 베이스라인 시스템의 short2-10sec(남성) 평가 데이터의 EER(%)
Table 2. EER(%) of Baseline system with PLDA adaptation using short2-10sec(male) evaluation data

공통 평가 조건 / 적응 가중치(γ)	6	7	8
0 (no adapt)	8.27	6.92	7.58
0.3	8.66	6.92	7.58
0.5	8.66	6.92	7.58
0.7	8.47	6.92	7.58
1.0	8.66	6.54	7.58

<표 2>는 베이스라인 시스템과 베이스라인 시스템에 PLDA 적응 기법을 적용한 시스템의 short2-10sec(남성) 평가 데이터에 대한 EER 결과이다. γ는 적응 가중치를 나타내며 γ가 0이면 적응이 이루어지지 않은 베이스라인 시스템의 PLDA 파라미터를 의미하며, 1에 가까울수록 짧은 발화 데이터베이스로 학습된 PLDA 파라미터에 크게 영향 받은 적응된 PLDA 파라미터를 의미한다. 결과를 보게 되면 PLDA 적응 후에 성능 향상이 이루어지지 않음을 알 수 있다. 개발 데이터베이스의 짧은 발화의 개수 자체가 부족하기 때문에 적절한 짧은 발화 PLDA 파라미터를 학습하지 못했기 때문으로 보인다. 이는 PLDA가 많은 양의 화자정보가 있는 데이터가 주어진 경우에 신뢰성 있는 모델을 학습할 수 있는 지도학습법이기 때문이다.

4.5. VTLP 를 이용한 데이터 증강 후 PLDA 모델 적응의 성능

본 절에서는 VTLP를 이용하여 짧은 발화 데이터를 증강한 후 PLDA 적응 기법을 적용한 화자검증 시스템의 성능 추이를 살펴본다. 본 논문에서 VTLP 증강시에 사용된 제한주파수와, VTLN 주파수는 각각 f_0=3,800Hz, f_{vtln}=3,300Hz이다. <표 3>은 VTLP 실험 파라미터로 1열은 데이터 증강비, 2열은 증강하는데 사용한 워핑 계수 α를 보여준다.

표 3. VTLP를 이용한 데이터 증강표

Table 3. Data augmentation table using VTLP

데이터 증강비	워핑 계수(α)
5배	0.8, 0.9, 1.0, 1.1, 1.2
9배	0.9, 0.95, 1.0, 1.05, 1.0, 1.15, 1.2, 1.25, 1.3
11배	0.8, 0.85, 0.9, 0.95, 1.0, 1.05, 1.1, 1.15, 1.2, 1.25, 1.3

워핑 계수 α의 범위가 1을 기준으로 비교적 큰 범위에서 작아지거나 커지면서 데이터가 증강되면, 화자의 고유 특징인 성도길이가 가변 되는 효과가 나타나, 증강된 음성 데이터의 화자가 원본 음성의 화자가 아닌 다른 화자로 labeling 될 것이고, 1을 기준으로 매우 작은 범위 안에서 α가 가변 되면, 동일 화자내의 변이로 볼 수 있는 음성 데이터 증강이 되어 원본 음성의 화자와 증강된 음성의 화자가 같은 화자로 labeling 되어 짧은 발화 PLDA 학습에 사용된다. <표 4>에서는 짧은 발화 증강에 사용된 2,615명의 화자의 발화에서 추출한 2,615 개의 I-vector와 워핑 계수 α로 가변된 발화에서 추출된 2,615개의 I-vector 사이의 CSS의 평균과 표준편차를 보여주고 있다. α가 1을 기준으로 크게 가변되면, 원본 발화와 가변된 발화 사이에서 추출된 I-vector 사이의 CSS의 평균은 작아지고, 표준편차는 커지는 경향성을 관찰할 수 있다.

표 4. 워핑계수에 따른 CSS의 평균과 표준편차

Table 4. The mean and standard deviation of CSS according to warping factor

워핑 계수(α)	평균	표준편차
0.8	90.5	62.9
0.9	328.8	46.1
0.95	513.5	20.1
1.0	600.0	0.2
1.05	522.5	18.5
1.1	367.5	41.6
1.2	150.9	63.4

2,615명의 짧은 발화에 대하여 가장 높은 CSS를 가지는 발화의 I-vector와 같은 클래스를 가지게 bottom-up clustering 했을 때, <표 5>에는 워핑 계수에 따른 화자분포가 나타나 있다. 특이한 점은 1과 거리 차이가 같은 워핑 계수라 할지라도 0.9와 0.95에서 1.1과 1.05보다 기존화자 클래스로 분류되는 짧은 발화의 수가 훨씬 많았다. 0.1이 차이나는 경우인 0.9와 1.1의 경우에서는 α의 값이 1보다 가까운 값이 있는 경우(α가 1.05와 0.95로 증강된 발화가 있는 경우)대부분 가장 가까운 α와 새로운 화자클래스를 생성한다. 그 외의 α값의 발화는 가장 가까운 α와 같은 화자클래스를 가지는 경향을 보인다.

표 5. 워핑계수에 따른 화자분포

Table 5. Speaker distribution according to warping coefficient

워핑 계수(α)	기존화자(명)	새로운 화자(명)
0.90	2030	585
0.95	2285	330
1.0	2615	0
1.05	1409	1206
1.1	328	2287

<표 6>에서는 개발 데이터베이스의 짧은 발화를 VTLP를 이용하여 워핑 계수의 개수에 비례하게 5, 9, 11배 증강한 후, PLDA 적응을 한 베이스라인 시스템의 EER을 보여주고 있다. VTLP를 이용하여 짧은 발화를 9배 증강을 한 후 PLDA 적응한 화자검증 시스템에서 가장 좋은 성능을 볼 수 있었다. 공통 평가 환경 컨디션 6,7,8에서 각각 2.5%, 12%, 11%의 상대적 성능 향상을 보임을 볼 수 있다. 가장 많은 데이터 증강이 이루어진 11배에서 오히려 EER이 상승하는 현상을 볼 수 있었다.

VTLP 기법이 기존의 원본 음성을 큰 왜곡 없이 증강한 경우인 5배,9배 증강에서는 γ가 1에 가까워지면 EER이 내려감을 관찰할 수 있고, γ가 0에 가까워지면 짧은 발화 PLDA 모델에 가중치가 낮아져 짧은 발화에 적합한 PLDA 모델이 생성되지 못하여, EER이 올라감을 관찰할 수 있다.

표 6. VTLP이용한 짧은 발화 증강 후 PLDA 적응의 성능

Table 6. Performance of PLDA adaptation after short utterance augmentation using VTLP

적응 가중치(γ) \ 공통 평가 조건		6	7	8
1배	**0 (no adapt)**	8.27	6.92	7.58
5배	**0.3**	8.66	6.92	7.58
	0.5	8.27	6.54	6.82
	0.7	8.27	6.15	6.82
	1.0	8.47	6,15	6.82
9배	**0.3**	8.66	6.92	7.58
	0.5	8.27	6.54	6.82
	0.7	**8.07**	**6.15**	**6.82**
	1.0	8.07	6.54	6.82
11배	**0.3**	9.06	6.92	6.82
	0.5	8.86	6.54	7.58
	0.7	8.86	6.92	7.58
	1.0	9.06	6.92	7.58

5. 결과

본 논문에서는 화자검증 시 발화길이에 따른 내용불일치 문제를 해결하기 위한 방법을 제시하였다. 일반적으로 긴 발화와 짧은 발화에 대한 PLDA 모델은 서로 다른 특징을 가진다는 점에 착안하여, 도메인 적응을 위해 사용되는 PLDA 모델 적응 기법을 짧은 발화에서 발생하는 내용 불일치 문제를 해결하기 위해 사용하였다. 또 화자인식 데이터베이스인 NIST SRE 데이터베이스에 짧은 발화 음성파일이 부족하여 신뢰성 있는 PLDA 적응 모델을 구축할 수 없는 문제를 해결하기 위해, 데이터 증강 기법인 VTLP를 사용하였다.

실험 결과는 TDNN을 적용한 화자검증 시스템이, 기존의 시스템에 비해 상대적으로 40% 낮은 EER을 보여줌을 확인하였다. 그 후 TDNN을 기반으로 한 화자검증 시스템에 PLDA 적응 기법을 적용한 후 성능 추이를 살펴본다. 실험 결과를 보면 NIST SRE 데이터베이스의 짧은 발화 데이터를 증강 없이 사용하였을 시에는 PLDA 적응 기법이 성능 향상을 보이지 못함을 확인하였다. VTLP 기법을 활용하여 짧은 발화 데이터들을 5배, 9배, 11배로 증강하여 PLDA 적응 기법을 적용함으로써 추가적인 성능 향상을 볼 수 있었고, 9배 증강한 데이터에서 가장 좋은 성능을 보여 주었다. 이는 VTLP 기법의 워핑 계수가 일정 범위 이상 커지면 원본 음성의 왜곡으로 이어지므로, 가장 많은 데이터 증강(11배 증강)이 이루어진 실험 환경에서 오히려 베이스라인 시스템보다 나쁜 성능을 보여준다.

향후 연구로는 VTLP에 적절한 워핑 계수를 자동으로 추정하는 방법이 필요할 것이다. 또한 음소 성분을 고려한 짧은 발화 보상 기법을 적용한다면 추가적인 성능향상을 가능하게 할 것으로 예상된다.

참고문헌

[1] Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1), 19-41.

[2] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4), 788-798.

[3] Prince, S. J., & Elder, J. H. (2007). Probabilistic linear discriminant analysis for inferences about identity. *Proceedings of the 11th IEEE International Conference on Computer Vision*. October, 2007.

[4] Garcia-Romero, D., Zhang, X., McCree, A., & Povey, D. (2014). Improving speaker recognition performance in the domain adaptation challenge using deep neural networks. *Proceedings of Spoken Language Technology Workshop*. December, 2014.

[5] Lei, Y., Scheffer, N., Ferrer, L., & McLaren, M. (2014). A novel scheme for speaker recognition using a phonetically-aware deep neural network. *Proceedings of International Conference on Acoustics, Speech and Signal Processing*. May, 2014.

[6] Hasan, T., Saeidi, R., Hansen, J. H., & van Leeuwen, D. A. (2013). Duration mismatch compensation for i-vector based speaker recognition systems. *Proceedings of International Conference on Acoustics, Speech and Signal Processing*. May, 2013.

[7] Kanagasundaram, A., Vogt, R. J., Dean, D. B., & Sridharan, S. (2012). PLDA based speaker recognition on short utterances. *Proceedings of Odyssey: The Speaker and Language Recognition Workshop.* June, 2012.

[8] Kanagasundaram, A., Dean, D., Sridharan, S., Gonzalez-Dominguez, J., Gonzalez-Rodriguez, J., & Ramos, D. (2014). Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques. *Speech Communication*, 59, 69-82.

[9] Kenny, P., Stafylakis, T., Ouellet, P., Alam, M. J., & Dumouchel, P. (2013). PLDA for speaker verification with utterances of arbitrary duration. *Proceedings of International Conference on Acoustics, Speech and Signal Processing*. May, 2013.

[10] Garcia-Romero, D., McCree, A., Shum, S., Brummer, N., & Vaquero, C. (2014). Unsupervised domain adaptation for i-vector speaker recognition. *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*. June, 2014.

[11] Dehak, N., Dehak, R., Kenny, P., Brümmer, N., Ouellet, P., & Dumouchel, P. (2009). Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. *Proceedings of INTERSPEECH*. September, 2009.

[12] Peddinti, V., Povey, D., & Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. *Proceedings of INTERSPEECH*. 2015.

[13] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3), 328-339.

[14] Kenny, P., Ouellet, P., Dehak, N., Gupta, V., & Dumouchel, P. (2008). A study of interspeaker variability in speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(5), 980-988.

[15] Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1), 19-41.

[16] Kenny, P., Gupta, V., Stafylakis, T., Ouellet, P., & Alam, J. (2014). Deep neural networks for extracting baum-welch statistics for speaker recognition. *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*. June, 2014.

[17] Paul, D. B., & Baker, J. M. (1992). The design for the wall street journal based csr corpus. *Proceedings of the workshop on Speech and Natural Language* (pp. 357-362).

[18] Pitz, M., & Ney, H. (2005). Vocal tract normalization equals linear transformation in cepstral space. *IEEE Transactions on Speech and Audio Processing*, 13(5), 930-944.

[19] Molau, S., Kanthak, S., & Ney, H. (2000). Efficient vocal tract normalization in automatic speech recognition. *Proceedings of the ESSV'00.* 2000.

[20] Jaitly, N., & Hinton, G. E. (2013). Vocal tract length perturbation (VTLP) improves speech recognition. *Proceedings of ICML Workshop on Deep Learning for Audio, Speech and Language.* June, 2013.

[21] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & vesely, K. (2011). The Kaldi speech recognition toolkit. *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*. 2011.

[22] Cieri, C., Miller, D., & Walker, K. (2004). The fisher corpus: resource for the next generations of speech-to-text. *Language Resources and Evaluation Conference*, 4, 69-71.

[23] Poddar, A., Sahidullah, M., & Saha, G. (2015). Performance comparison of speaker recognition systems in presence of duration variability. *Proceedings of IEEE India Conference(INDICON)*. December, 2015.

[24] Kenny, P., Boulianne, G., Ouellet, P.& Dumouchel, P. (2007). Speaker and session variability in GMM-based speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, 15(4), 1448-1460.

[25] National Institute of Standards and Technology. (2008). *The NIST year 2008 speaker recognition evaluation plan 2008*. Retrieved from http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf on December 11, 2016.

[26] Snyder, D., Garcia-Romero, D., & Povey, D. (2015). Time delay deep neural network-based universal background models for speaker recognition. *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding.* December, 2015.

• **윤성욱 (Yoon, Sung-Wook)**
충북대학교 제어로봇공학전공
충북 청주시 서원구 충대로1
Email: magi11@naver.com
관심분야: 화자인식, 음성인식
현재 제어로봇공학과 박사 재학 중

• **권오욱 (Kwon, Oh-Wook)** 교신저자
충북대학교 전자공학부
충북 청주시 서원구 충대로1
Tel: 043-261-3374
Email: owkwon@cbnu.ac.kr
관심분야: 음성인식, 화자인식, 감정인식, 음성신호처리
2003~ 현재 충북대학교 전자공학부 교수