Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.29, NO.6, Dec. 2019 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2019.29.6.1393

STFT와 RNN을 활용한 화자 인증 모델

김 민 서,† 문 종 섭‡
고려대학교 정보보호대학원

Speaker Verification Model Using Short-Time Fourier Transform and Recurrent Neural Network

Min-seo Kim,† Jong-sub Moon‡
Graduate School of Information Security, Korea University

요 약

최근 시스템에 음성 인증 기능이 탑재됨에 따라 화자(Speaker)를 정확하게 인증하는 중요성이 높아지고 있다. 이에 따라 다양한 방법으로 화자를 인증하는 모델이 제시되어 왔다. 본 논문에서는 Short-time Fourier transform(STFT)를 적용한 새로운 화자 인증 모델을 제안한다. 이 모델은 기존의 Mel-Frequency Cepstrum Coefficients(MFCC) 추출 방법과 달리 윈도우 함수를 약 66.1% 오버랩하여 화자 인증 시 정확도를 높일 수 있다. 새로운 화자 인증 모델을 제안한다. 이 때, LSTM 셀을 적용한 Recurrent Neural Network(RNN)라는 딥러닝 모델을 사용하여 시변적 특징을 가지는 화자의 음성 특징을 학습하고, 정확도가 92.8%로 기존의 화자 인증 모델보다 5.5% 정확도가 높게 측정되었다.

ABSTRACT

Recently as voice authentication function is installed in the system, it is becoming more important to accurately authenticate speakers. Accordingly, a model for verifying speakers in various ways has been suggested. In this paper, we propose a new method for verifying speaker verification using a Short-time Fourier Transform(STFT). Unlike the existing Mel-Frequency Cepstrum Coefficients(MFCC) extraction method, we used window function with overlap parameter of around 66.1%. In this case, the speech characteristics of the speaker with the temporal characteristics are studied using a deep running model called RNN (Recurrent Neural Network) with LSTM cell. The accuracy of proposed model is around 92.8% and approximately 5.5% higher than that of the existing speaker certification model.

Keywords: Speaker verification, STFT, Deep Learning, Recurrent Neural Network(RNN)

I. 서 론

최근 IoT, 임베디드 시스템 등에 사용자 인증 기능이 탑재됨에 따라 사용자를 정확하게 인증하는 중요성이 높아지고 있다. 화자 인증(Speaker Verification)은 화자의 음성을 활용하여 화자가 시스템에 등록된 정당한 사용자인지 여부를 인증하는 것이다[1].

화자 인증은 화자 인식(Speaker Recognition)의 범주에 속하는데, 이는 텍스트 독립적인 유형(text-independent)과 텍스트 종속적인 유형(text-dependent)으로 구분된다[2]. 텍스트 종속적인 화자 인증 유형은 화자 등록 과정과 테스트 과정의 음성 내용이 동일하기에 상대적으로 높은 정확도를 나타낸다. 그러나 텍스트 독립적인 화자 인증 유형은 화자 인증 시, 화자가 등록한 음성 내용과 무

Received(2019. 10. 23), Modified(1st: 11. 14. 2019, 2nd: 11. 20. 2019), Accepted(12. 03. 2019)
† 주저자, piggyminseo@gmail.com
‡ 교신저자, jongsubmoon@korea.ac.kr(Corresponding author)

관하게 검증하고, 수많은 변형을 포함할 수 있기 때문에 상대적으로 높은 정확도를 나타내기 어렵다.

일반적으로 화자 인증 모델은 화자의 목소리를 등록(enrollment)하는 과정과 화자의 음성을 인증(verification)하는 과정이 필요하다[2]. 화자의 목소리 등록 시, 화자의 음성에서 특징을 추출하기 위해 Linear Predictive Coding(LPC)[3], Mel-Frequency Cepstrum Coefficients(MFCC) [4], Perceptual Linear Predictive Analysis(PLP)[5] 등의 방법을 사용한다. 화자의 음성 특징 추출 시 많이 사용되는 방법은 MFCC로써, 이 방법은 입력된 신호에서 실제 유효한 소리의 특징을 추출한다. MFCC는 입력된 소리 전체를 대상으로 하는 것이 아니라, 일정 구간(Short-time)으로 나누어 해당 구간에 대한 스펙트럼을 분석하여 특징을 추출하는 방법이다.

본 논문에서는 화자 인증 시, 기존의 MFCC 방법과 달리 시간적인 데이터 단위인 윈도우를 특정 비율로 오버랩하여 화자의 음성을 Short-time Fourier transform(STFT)에 적용하고, 딥 러닝 모델인 Recurrent Neural Network(RNN) 모델을 사용하여 음성 데이터를 학습하여 기존의 텍스트 독립적인 화자 인증 모델보다 높은 정확도를 나타낼 수 있는 화자 인증 모델을 제안한다. 정확도 측면에서, 기존의 화자 인증 모델은 일반적인 MFCC 방법을 사용한다[4]. 그러나 본 논문은 화자의 음성 특징 추출 시, 윈도우 함수를 일부분 겹치게 하여 화자의 음성에서 손실되는 정보가 최소화되도록 한 후, STFT를 적용한다.

본 논문의 구성은 다음과 같다. 2장에서는 화자 인증 모델 관련 연구로써 화자의 음성을 전처리하기 위한 방법, 딥 러닝 모델을 적용한 화자 인증 모델에 대해 설명한다. 3장에서는 변경한 MFCC 추출 방법을 사용하여 화자의 음성 특징을 추출한 후, RNN 모델을 사용한 화자 인증 모델을 제안한다. 4장에서는 3장에서 제안한 방법을 통한 실험 결과를 보이고 마지막 5장에서는 결론과 향후 연구 방향을 제시한다.

II. 관련 연구

2.1 Short-time Fourier transform 연구

STFT란 시간이 지남에 따라 변화하는 신호의 사인파 주파수와 위상 성분을 결정하는 데 사용되는 푸리에 관련 변환이다[6]. STFT는 시간에 따라 변화하는 긴 신호를 짧은 시간 단위로 분할한 다음에 푸리에 변환을 적용하기에 결과적으로 각 시간 구간마다 어떤 주파수들이 존재하는지 알 수 있다.

2.2 화자 인증(Speaker Verification) 연구

일반적으로 화자 인증 연구에서는 화자 음성 모델을 등록하고, 화자를 인증하기 위해 음성 데이터에서 화자의 음성 특징을 추출하는 과정을 수행한다.

이 때, 가장 널리 쓰이는 방법은 MFCC이다. 이 방법은 사람의 청각이 저주파수 대역에서 민감한 반면 고주파수 대역에서 상대적으로 둔감한 특성을 표현한 멜 스케일(mel scale)에 기반한 음성 특징이다[4]. MFCC 추출 과정은 Fig. 1.과 같다. 우선 음성 신호로부터 매 프레임 단위로 윈도우 함수를 씌운 다음 Discrete Fourier Transform(DFT) 과정을 통해 시간 영역에서 주파수 영역으로 변환시킨다. 그 다음으로 멜 스케일을 가지도록 주파수 축을 변환한 후, 이 스케일에서 동일한 대역폭을 가지는 삼각 필터뱅크를 통해 필터뱅크 별 에너지를 구한다. 여기에 로그 함수를 취한 다음 Discrete Cosine Transform(DCT)를 통해 최종적인 MFCC 값을 구하게 된다.

2.3 딥 러닝 모델을 적용한 화자 인증

음성 데이터는 발화 기관의 상태에 따라, 그리고 각 단어의 문장 내 위치에 따라 큰 변동성을 보이는 특징을 갖는다. 이 특징은 음성 인식 시스템의 실용화를 어렵게 한다. 이러한 문제를 해결하기 위해 음성 데이터 전처리 단계와 화자 모델 생성 단계에 딥 러닝 모델을 적용하면 주변 환경의 잡음을 제거하고,

Fig. 1. Mel-Frequency Cepstrum Coefficients Process

화자 음성 이외의 노이즈 부분을 제거할 수 있다. 대표적으로 Deep Neural Network(DNN)를 활용한 d-vector[7]의 경우 화자의 음성 특징을 추출하기 위해 음성 데이터를 사용해 심층 신경망의 특정 은닉 층의 출력 값을 취함으로써 d-vector를 추출하였다. 신경망 기반의 학습 기법을 사용하면 주변 환경의 잡음이 제거되고, 화자의 음성 특징이 반영된 d-vector를 생성해 낼 수 있다.

2.4 순환 신경망 모델

RNN은 인공 신경망의 한 종류로, 유닛 간의 연결이 순환적 구조를 갖는 특징을 갖고 있다. 이러한 구조는 시변적 동적 특징을 모델링 할 수 있도록 신경망 내부에 상태를 저장할 수 있게 해준다. 전방전달 신경망(Feedforward neural network)과 달리 RNN은 내부의 메모리를 이용해 시퀀스 형태의 입력을 처리할 수 있기에 음성 인식이나 필기체 인식과 같이 시변적 특징을 가지는 데이터를 처리하는데 적합하다.

Long-Short Term Memory(LSTM)[8]는 RNN에서 사용되는 은닉 층 활성화 유닛으로, ReLu 활성화 함수를 사용하여 RNN 모델 학습 시 시간 간격이 커질수록 오류에 의한 기울기 값이 잘 전달되지 않는 vanishing gradient 현상을 해결하기 위해 이용한다.

일반적인 RNN의 구조는 Fig. 2.와 같다. t는 시간을 의미하며, x_t, y_t, h_t는 각각 t 시간에서의 입력 값, 출력 층의 활성화 값, 은닉 층의 활성화 값을 의미한다. V, U, W는 각각 입력 층과 은닉 층을 연결하는 가중치, 은닉 층과 출력 층을 연결하는 가중치, 그리고 은닉 층을 서로 연결하는 가중치를 의미한다. c_t는 t 시간 단계에서의 상태 값으로, LSTM 셀에서 은닉 층의 활성화 값의 연산에 사용하는 내부적인 값이다.

현재 시간에서 상태 값과 활성화 값은 여러 게이트들에 의하여 결정된다. 게이트에는 입력 게이트, 출력 게이트, 그리고 망각 게이트가 있다. 입력 게이트는 상태 값에 입력을 얼마나 반영할지, 출력 게이트는 상태 값을 얼마나 활성화 값으로 보낼지, 망각 게이트는 과거의 상태 값을 얼마나 남길 것인지 결정한다. 각각의 게이트 값은 현재 시간의 입력 값과 이전 시간의 은닉 층 활성화 값에 의하여 결정된다.

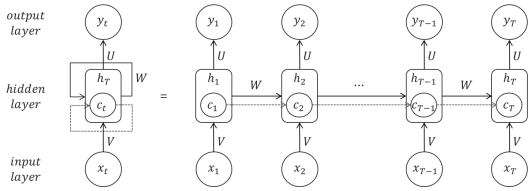

Fig. 2. Structure of RNN with LSTM Cell

III. 제안 방법

3.1 화자 인증 모델 개요

본 논문에서는 RNN을 활용한 새로운 화자 인증 모델을 제안한다.

화자의 목소리를 등록하는 과정에서는 화자의 음성 데이터에서 화자의 특징을 추출하고, 추출한 특징을 기반으로 화자의 음성 모델을 학습한다. 화자의 음성을 인증하는 과정에서는 인증을 위해 입력으로 들어오는 음성 데이터에서 화자의 음성 특징을 추출한 후, 추출한 특징과 기존에 학습이 완료된 화자의 음성 특징 간의 유사도를 비교하여 화자가 시스템에 등록되어 있는 화자인지 인증한다.

3.2 음성 특징 추출

본 논문에서는 화자의 음성 특징을 추출하기 위해 기존의 MFCC와 달리 윈도우 함수를 특정 비율로 오버랩하고, STFT를 사용하여 화자의 음성 특징을 추출한다. 변경된 부분은 Fig. 3.과 같고, 변경된 단계는 아래와 같다.

3.2.1 윈도우 함수 적용

본 논문에서는 오디오나 음성 처리에서 가장 많이 사용하는 윈도우 함수인 해밍 윈도우 함수

Fig. 3. Mel-Frequency Cepstrum Coefficients Process

(Hamming Window Function)를 사용한다. 해밍 윈도우 함수는 식(1)과 같고, Fig.4.와 같은 모양을 나타낸다.

$$w(n) = 0.54 - 0.46 \times \cos\left(\frac{2n\pi}{W}\right) \quad (1)$$

여기에서 $w(n)$은 해밍 윈도우 함수의 크기, W는 윈도우의 길이, n은 윈도우의 개수를 의미한다.

일반적으로 화자의 음성 특징 추출 시, 윈도우 함수의 크기에 따라 음성 데이터를 분리하여 특정 구간에서 음성 특징을 추출하기 때문에 음성 데이터에서 화자의 음성 특징이 손실되거나 왜곡되는 현상이 발생한다[8]. 본 논문에서는 이러한 현상을 최소화하기 위해 Fig.5.와 같이 해밍 윈도우 함수를 특정 비율로 오버랩하여 음성 신호와 컨볼루션(convolution)한다.

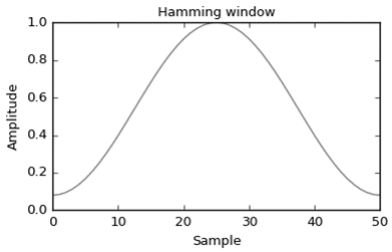

Fig. 4. Hamming Window Function

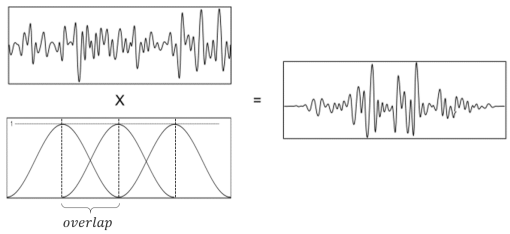

Fig. 5. Convolution

3.2.2 Short-time Fourier Transform

일반적으로 음성 데이터를 기반으로 하는 시스템에서는 위상 스펙트럼에 대한 정보를 고려하지 않기에 MFCC 추출 시, DFT 과정으로 Fast Fourier Transform(FFT)을 사용한다[9].

하지만, 위상 스펙트럼에 관한 Liu[10]과 Paliwal [11] 그리고 Alsteris [12], [13]의 연구를 살펴보면 짧은 시간에서도 위상 스펙트럼이 크기 스펙트럼과 비슷한 명료성을 가지는 것을 보여주기에 본 논문에서는 MFCC의 DFT 과정으로 STFT를 사용한다. 시간, 주파수와 음량 간의 관계를 손실하지 않고 화자의 음성 특징을 추출하기 위해서 시간 구간별로 STFT 알고리즘을 적용하였다.

$$STFT(t, w) = \int_{-\infty}^{\infty} f(t)w(t-t')e^{-jwt}dt \quad (2)$$

여기에서 시간에 대한 신호 함수 $f(t)$와 윈도우 $w(t)$에 대한 STFT 알고리즘은 식(2)와 같다.

STFT 적용 이후의 과정은 기존의 MFCC 추출 과정과 동일하게 멜 스케일을 가지도록 주파수 축을 변환한 후, 이 스케일에서 동일한 대역폭을 가지는 삼각 필터뱅크를 통해 필터뱅크 별 에너지를 구하여 로그 함수를 취한 다음 DCT를 통해 최종적인 MFCC 값을 구한다[4].

3.3 RNN을 활용한 화자 인증 모델

본 논문에서는 화자의 음성 특징을 학습하기 위해 LSTM 셀을 사용한 RNN 모델을 활용하며, 이 모델은 일반적인 기계학습이나 딥 러닝 모델과 마찬가지로 학습 과정을 거친 후, 화자 검증에 이용된다. 전체적인 RNN 구조는 Fig. 6.과 같고, 각 단계는 다음과 같다.

3.3.1 학습 단계

본 논문에서 제시하는 RNN 모델은 화자의 음성 특징을 담고 있는 프로파일을 생성하고, 임의의 화자의 음성에서 추출한 MFCC 값을 모델이 임베딩한 값과 화자 자신의 음성 프로파일과 비교했을 때, 가장 높은 유사도를 갖도록 학습된다. 이 절에서는 이와 같은 학습을 정형적으로 기술한다.

$X \in R^{N \times M \times T \times D}$은 모델을 학습하기 위한 음성 데이터 셋이며, 여기에서 N은 데이터 셋에 포함된 화자의 수를 의미하며, 각 화자는 M개의 음성을 갖는다. 또한 T는 각 음성 별 타임 스텝(time step) 길이, D는 각 타임 스텝에서 LSTM 셀의 입

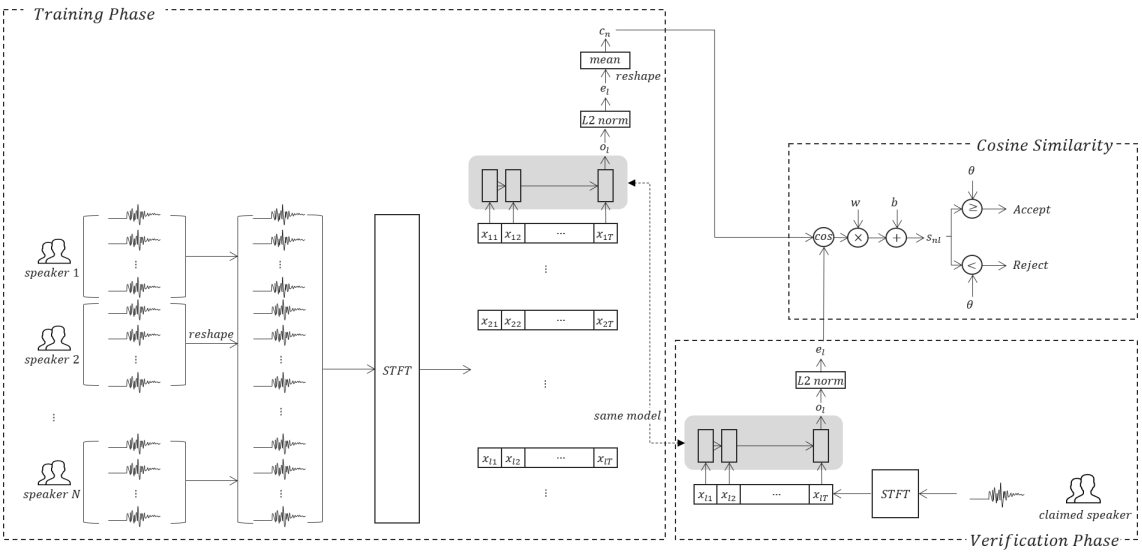

Fig. 6. RNN-Based Speaker Verification Model

력으로 주어질 MFCC 벡터의 차원을 의미한다. 데이터 처리의 편의상 데이터 셋을 $X' \in R^{L \times T \times D}$ 형태로 재구성(reshape)했으며, $L = N \times M$이다. LSTM 셀의 입력 데이터인 $\boldsymbol{x_{lt}} \in X'$ 는 D차원을 갖는 l번째 음성 중 타임스텝 t번째 MFCC 벡터를 의미하며, l과 t는 각각 1과 L사이, 1과 T사이의 정수형 인덱스이다. 마지막으로 LSTM 셀의 연산을 r_θ로 기술하며, 이때 θ는 LSTM 셀의 은닉 계층을 구성하는 가중치이다. 따라서 각 타임 스텝 별 연산을 식(3)과 같이 표기한다.

$$\boldsymbol{v_t^l}, \boldsymbol{z_t^l} = r_\theta(\boldsymbol{x_{lt}}, \boldsymbol{v_{t-1}^l}, \boldsymbol{z_{t-1}^l}) \tag{3}$$

여기에서 $\boldsymbol{v_t^l} \in R^H$와 $\boldsymbol{z_t^l} \in R^H$는 각각 l번째 음성에 대한 타임스텝 t에서의 셀 상태(cell state)와 은닉 상태(hidden state)를 의미하며, H는 셀의 은닉 계층 크기이다. 이때 $\boldsymbol{z_t^l}$는 t 타임스텝까지의 음성 특징이 누적된 임베딩 값이 된다.

본 논문에서는 화자의 모든 음성 특징이 누적된 마지막 타임스텝의 임베딩 값인 $\boldsymbol{z_T^l}$만 사용하며, 이후 기술의 편의상 $\boldsymbol{o_l}$로 표기한다. 이후 식(4)와 같이 L2 노옴(norm)을 이용한 정규화(normalization)를 수행하여 l번째 화자 음성에 대한 정규화된 임베딩 $\boldsymbol{e_l} \in [0,1]^H$을 계산한다.

$$\boldsymbol{e_l} = \frac{\boldsymbol{o_l}}{||\boldsymbol{o_l}||_2} \tag{4}$$

위와 같은 연산을 전체 데이터 셋에 적용하여 정규화된 임베딩 데이터 셋은 $E \in [0,1]^{L \times H}$ 획득한다. 그 다음 다시 화자별 음성을 구분하기 위해 E를 재구성하여 $E' \in [0,1]^{N \times M \times H}$를 획득한다. 마지막으로 식(5)와 같이 각 화자 별 M개의 정규화된 임베딩 값의 평균을 계산하여 화자 별 인증을 위한 프로파일로 사용한다.

화자의 음성 인증은 생성된 프로파일을 이용과 정규화된 임베딩 값과의 유사도 비교를 통해 진행된다.

$$c_n = \frac{1}{M} \sum_{m=1}^{M} e_{nm} \tag{5}$$

본 논문에서 유사도 $s_{nl} \in R$는 식(6)과 같이 n번째 화자의 프로파일 c_n과 인증을 위해 입력으로 들어오는 음성 e_l 사이의 코사인 유사도를 사용한다.

여기에서 $w \in R$와 $b \in R$는 각각 코사인 유사도 측정에 사용되는 가중치와 편형치이다.

$$s_{nl} = w \bullet \cos(\boldsymbol{c_n}, \boldsymbol{e_l}) + b \quad (6)$$

따라서 데이터 셋에 N명의 화자가 포함된 경우, 임의의 화자의 음성은 각 화자의 프로파일과 비교되어 총 N개의 유사도가 계산된다. 입력된 음성은 유사도가 가장 높은 사용자로 분류되며, 이상적으로, 음성을 생성한 화자 자신의 프로파일과 비교했을 때, 유사도가 가장 높게 학습이 되어야 한다.

본 논문에서는 위와 같이 계산된 유사도를 이용해 모델을 학습하기 위해 log softmax 손실 함수[14]를 사용한다. log softmax는 기본적인 softmax에 비해 모델의 예측 결과에 민감하므로 학습 소요시간이 감소하여 효율적이다. N명의 화자의 음성으로 구성된 크기가 $L(= N \times M)$인 데이터 셋에 대해 이 손실 함수는 식(7)과 같이 정의된다.

$$Loss = \sum_{l=1}^{L}(s_{jl} - \ln \sum_{i=1}^{N} \exp(s_{nl})) \quad (7)$$

여기에서 s_{jl}은 l번째 음성에 대한 정규화된 임베딩 값 e_l을 N개의 프로파일과 유사도 비교하여 유사도가 가장 높았던 j번째 유사도 값을 의미한다.

이 모델은 위 손실 함수를 이용하여 Stochastic Gradient Descent(SGD)[15]나 AdaGrad[16], Adam[17]과 같은 경사도 기반 최적화 알고리즘으로 학습될 수 있다.

3.3.2 검증 단계

화자 인증 모델은 3.3.1. 절에서 기술한 것과 같이 생성된 음성 프로파일 c_n과 인증을 위해 새롭게 입력되는 음성 e_l의 유사도를 식(6)과 같이 비교하여 입력된 음성이 등록되어 있는 화자의 프로파일과 일치하는 화자의 음성인지 확인한다.

본 논문에서는 입력으로 들어오는 음성이 등록되어있는 프로파일에 대해 측정한 유사도 s_{nl}이 일정 임계치(threshold) ψ보다 클 경우($s_{nl} \geq \psi$)이라면, 인증에 성공하고, 그렇지 않으면($s_{nl} < \psi$) 인증에 실패한다.

IV. 실험 및 평가

이 장에서는 3장에서 제안한 화자 인증 모델의 성능을 측정하기 위해 화자 인증 정확도를 측정하고, 그 결과를 제시한다.

4.1 실험 환경 및 실험 데이터

본 논문은 음성 데이터 전처리와 RNN 학습을 위해 Ubuntu 16.04에서 실험을 진행하였다. 그리고 이 때 사용한 딥 러닝 라이브러리 TensorFlow[18]의 버전은 1.14.0 이다.

실험에 사용한 데이터 셋은 음성 인식 연구에서 많이 사용되는 Voxceleb1[19] 데이터 셋을 사용하였고, 형태는 Table 1.과 같다. Voxceleb1 데이터 셋은 화자의 성별이 균형을 이루고, 다양한 민족과 억양의 음성을 포함하고 있다. 그리고 유튜브에서 업로드 된 비디오에서 추출되었으며 채널 소음, 녹음 등의 다른 노이즈 특성이 특정 비율로 혼합되어 있어 본 논문이 제안하는 화자 인증 모델에 적합하다.

Table 1. Dataset for speaker verification

Set	#speaker	#utterances
Training	1,211	148,642
Test	40	4,874
Total	1,251	153,516

4.2 실험 구성

실험은 Voxceleb1 데이터 셋에 대한 화자 인증을 수행하고, 이에 대한 정확도를 측정하는 방식으로 진행하였다. 정확도는 인증을 위해 입력으로 들어오는 음성이 등록되어 있는 화자 음성 모델이 맞을 경우 시스템 접근을 허가하고, 등록되어 있는 화자 음성 모델이 아닐 경우 시스템 접근을 거부하도록 측정하였다. 이 때 err는 화자를 정확히 인증하지 못한 경우를 의미하며, 시스템에 등록된 사용자 외 다른 사람을 등록자로 오인하고 인증을 수행하는 오류인 False Acceptance Ratio(FAR)와 시스템에 등록된 사용자가 사용 시 본인임을 확인하지 못하고 인증을 거부하는 오류인 False Rejection Ratio(FRR) 값의 교차점을 사용하였다.

인증을 위한 음성 데이터 셋은 Voxceleb1 데이

터 셋의 4,874개 테스트 데이터를 사용하였다. 본 논문에서는 음성 데이터 전처리 시, 윈도우 함수의 길이를 $25ms$로 설정하고, 화자의 음성 특징이 손실되는 것을 최소화하기 위해 윈도우 함수를 약 $66.1\%(25ms \times 0.661\% = 16.525ms)$오버랩하여 사용하고, STFT를 적용하였다. 음성 데이터 전처리 완료 후, 화자 음성 모델 학습을 위한 학습은 총 600,000번을 반복하고, 이 때 N은 4, M은 5, 러닝 레이트(learning rate)는 10^{-2}으로 설정하여 학습을 수행하였다.

4.3 실험 결과

Table 2에서 볼 수 있듯이, 제안된 화자 인증 모델은 정확도 측면에서 다른 모든 모델을 상당히 능가한다.

본 논문에서 제안한 화자 인증 모델은 화자의 음성 데이터에서 손실되는 화자의 음성 특징을 최소화하여 특징을 추출한 후, 화자의 음성 특징을 학습하였기에 높은 정확도를 나타낸다.

Table 2. The comparison of different methods

model	Accuracy
LCN[20]	82.6%
CNN[21]	83.1%
LSTM[22]	86.0%
3D-CNN[23]	87.3%
our method	**92.8%**

V. 결 론

본 논문에서는 음성 데이터 전처리 시, 윈도우 함수를 약 66.1% 오버랩하여 STFT를 적용할 경우, 화자의 음성 특징이 손실되는 것을 최소화할 수 있다는 점에 착안하여 RNN 모델을 사용한 화자 인증 모델을 제안하였다. RNN 모델에 LSTM 셀을 적용하여 기존 연구에 비해 화자 인증 시 높은 정확도를 나타내는 것을 확인하였다.

하지만 화자 인증 시 화자의 음성과 변조된 화자의 음성을 구분하여 화자를 인증하는 것이 힘들다는 한계점과 기존 화자 인증 시스템에 사용자가 추가되거나 삭제될 경우, 변경 사항을 시스템에 적용할 수 없다는 한계점이 있었다. 이러한 한계점은 향후 연구에서 발전시킬 수 있을 것이다.

References

[1] Z. Zhang & A. Subramaya, "Text-dependent speaker verification," U.S. Patent no. 8, 2012

[2] F. Bimbot, J. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier & D. Reynolds, "A tutorial on text-independent speaker verification," EURASIP Journal on Advances in Signal Processing, Apr. 2004

[3] J. Hai & E. M. Joo, "Improved linear predictive coding method for speech recognition," Fourth International Conference on Information, vol. 3, pp. 1614-1618, Dec. 2003

[4] L. Muda, M. Begam & I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," Journal of Computing, vol. 2, Issue 3, Mar. 2010

[5] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," Journal of the Acoustical Society of America, vol. 87, pp. 1738-1752, 1990

[6] T. Baba, "Time-Frequency Analysis Using Short Time Fourier Transform," Open Acoustics Journal, 2012

[7] E. Variani, X. Lei, E. XDermott, I. L. Moreno & J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP) IEEE, pp. 4052-4056, May. 2014

[8] H. Sak, A. Senior & F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," Fifteenth annual conference of the international speech

communication association, 2014
[9] C. Van Loan, "Computational frameworks for the fast Fourier transform," vol. 10, 1992
[10] L. Liu, J. He & G. Palm, "Effects of phase on the perception of intervocatic stop consonants," Speech Communication, vol. 22, pp. 403-417, Sep. 1997
[11] K. Paliwal & L. Alsteris, "Usefulness of phase spectrum in human speech perception," EUROSPEECH, pp. 21187-2120, Sep. 2003
[12] L. D. Alsteris & K. Paliwal, "Importance of window shape for phase-only reconstruction of speech," IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP) IEEE, pp. 573-576, May. 2004
[13] L. D. Alsteris & K. Paliwal, "Further intelligibility results from human listening tests using the short-time phase spectrum," Speech Communication, vol. 48, pp. 727-736, Jun. 2006
[14] S. Kanai, Y. Fujiwara, Y. Yamanaka & S. Adachi, "Sigsoftmax: Reanalysis of the softmax bottleneck," Advances in Neural Information Processing Systems, pp. 286-296, 2018
[15] I. Loshchilov & F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," arXiv preprint arXiv:1608.03983, 2016
[16] J. Duchi, E. Hazan & Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," Journal of Machine Learning Research, pp. 2121-2159, Jul. 2011
[17] D. P. Kingma & J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014
[18] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean & M. Kudlur, "Tensorflow: A system for large-scale machine learning," 12th USENIX Symposium on Operating Systems Design and Implementation, pp. 265-283, 2016
[19] A. Nagrani, J. S. Chung & A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," arXiv preprint arXiv:1706.08612, 2017
[20] Y. H. Chen, I. Lopez-Moreno, T. N. Sainath, M. Visontai, R. Alvarez & C. Rarada, "Locally-connected and convolutional neural networks for small footprint speaker recognition," Sixteenth Annual Conference of the International Speech Communication Association, 2015
[21] H. Salehghaffari, "Speaker Verification using Convolutional Neural Networks," arXiv, August. 2018
[22] H. Georg, M. Ignacio, B. Samy & S. Noam, "End-to-end text-dependent speaker verification." IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP) IEEE, pp. 5115 - 5119, 2016
[23] T. Amirsina, D. Jeremy & M. Nasser, "Text-Independent Speaker Verification Using 3D CNN," arXiv, June. 2018

〈저자소개〉

김 민 서 (Min-seo Kim) 학생회원
2018년 2월: 상명대학교 정보통신공학과 학사
2018년 3월~현재: 고려대학교 정보보호대학원 석사과정
〈관심분야〉 정보보호, 시스템 보안, 딥 러닝

문 종 섭 (Jong-sub Moon) 종신회원
1981년 2월: 서울대학교 계산통계학과 학사
1983년 2월: 서울대학교 계산통계학과 석사
1991년 2월: Illinois Institute of Technology 전산학과 박사
1993년 3월~현재: 고려대학교 전자 및 정보공학부 교수
2001년 2월~현재: 고려대학교 정보보호대학원 겸임교수
〈관심분야〉 정보보호, 운영체제, 침입탐지