잡음을 활용한 효과적인 화자 인식 기술

고수완[O], 강민지[*], 방세희[*], 정원태[**], 이경률[*]
[O]목포대학교 정보보호학과,
[*]목포대학교 정보보호학과,
[**]대구가톨릭대학교 컴퓨터소프트웨어학과
e-mail: {gpffh123, alswl3016, 011016h}@mokpo.ac.kr[O*], dnjsxo4354@cu.ac.kr[**], carpedm@mnu.ac.kr[*]

Effective Speaker Recognition Technology Using Noise

Suwan Ko[O], Minji Kang[*], Sehee Bang[*], Wontae Jung[**], Kyungroul Lee[*]
[O]Dept. of Information Security, Mokpo National University,
[*]Dept. of Information Security, Mokpo National University,
[**]Dept. of Computer Software, Daegu Catholic University

● 요 약 ●

정보화 시대 스마트폰이 대중화되고 실시간 인터넷 사용이 가능해짐에 따라, 본인을 식별하기 위한 사용자 인증이 필수적으로 요구된다. 대표적인 사용자 인증 기술로는 아이디와 비밀번호를 이용한 비밀번호 인증이 있지만, 키보드로부터 입력받는 이러한 인증 정보는 시각 장애인이나 손 사용이 불편한 사람, 고령층과 같은 사람들이 많은 서비스로부터 요구되는 아이디와 비밀번호를 기억하고 입력하기에는 불편함이 따를 뿐만 아니라, 키로거와 같은 공격에 노출되는 문제점이 존재한다. 이러한 문제점을 해결하기 위하여, 자신의 신체의 특징을 활용하는 생체 인증이 대두되고 있으며, 그중 목소리로 사용자를 인증한다면, 효과적으로 비밀번호 인증의 한계점을 극복할 수 있다. 이러한 화자 인식 기술은 KT의 기가 지니와 같은 음성 인식 기술에서 활용되고 있지만, 목소리는 위조 및 변조가 비교적 쉽기에 지문이나 홍채 등을 활용하는 인증 방식보다 정확도가 낮고 음성 인식 오류 또한 높다는 한계점이 존재한다.

상기 목소리를 활용한 사용자 인증 기술인 화자 인식 기술을 활용하기 위하여, 사용자 목소리를 학습시켰으며, 목소리의 주파수를 추출하는 MFCC 알고리즘을 이용해 테스트 목소리와 정확도를 측정하였다. 그리고 악의적인 공격자가 사용자 목소리를 흉내 내는 경우나 사용자 목소리를 마이크로 녹음하는 등의 방법으로 획득하였을 경우에는 높은 확률로 인증의 우회가 가능한 것을 검증하였다. 이에 따라, 더욱 효과적으로 화자 인식의 정확도를 향상시키기 위하여, 본 논문에서는 목소리에 잡음을 섞는 방법으로 화자를 인식하는 방안을 제안한다. 제안하는 방안은 잡음이 정확도에 매우 민감하게 반영되기 때문에, 기존의 인증 우회 방법을 무력화하고, 더욱 효과적으로 목소리를 활용한 화자 인식 기술을 제공할 것으로 사료된다.

키워드: 화자 인식(Speaker Recognition), 머신러닝(Machine Learning), MFCC(Mel Frequency Cepstral Coefficients), 사용자 인증 (User Authentication)

I. 서론

정보화 시대에 스마트폰이 대중화되고 실시간 인터넷 사용이 가능해짐에 따라, 본인을 식별하기 위한 사용자 인증이 필수적으로 요구된다. 사용자 인증 기술로는 아이디와 비밀번호를 이용한 비밀번호 인증 기술과 얼굴이나 지문 등을 활용한 생체 인증 기술 등이 있다[2, 5]. 하지만, 얼굴 인식이나 지문 등의 생체 인증은 손상이나 복제의 위험성이 높은 편이며, 비밀번호 인증의 경우, 시각 장애인이나 고령층을 포함하여 키보드를 입력하기에 한계가 있는 사람들이 많은 서비스로부터 요구되는 아이디와 비밀번호를 기억하고 입력하기에는 불편함이 따른다[3]. 그뿐만 아니라, 금융 상담 전화나 통신사 상담 전화 등 기존 상담원과 통화하며 주민등록번호와 같은 인증을 위한 정보를 입력하는 절차의 번거로움이 있다. 따라서 사용자 목소리의 음성 정보를 등록하고, 등록된 음성 정보와 통화 중인 사용자 목소리의

음성을 비교하여 화자를 인식한다면, 시각 장애인이나 고령층에게 더욱 편리한 인증 방안을 제공할 것으로 판단된다[1].

이를 위하여, 카카오의 카카오 미니와 KT의 기가 지니와 같은 음성 인식 기술을 활용한 화자 인식 기술을 제공한다[1]. 하지만, 목소리는 위조 및 변조가 비교적 쉽기에 지문이나 홍채 등을 활용하는 인증 방식보다 정확도가 낮고 음성 인식 오류 또한 높다는 한계점이 존재한다.

본 논문에서는 우선 화자 인식 기술의 정확도를 측정하기 위하여, 사용자 목소리를 학습시켰으며, 목소리의 주파수를 추출하는 MFCC(Mel Frequency Cepstral Coefficient) 알고리즘을 이용하여 테스트 목소리와의 정확도를 측정하였다[6]. 하지만, 사용자 목소리를 학습시키더라도, 딥 보이스 프로그램과 같이 사용자 목소리의 주파수가 비슷하도록 테스트 목소리를 조작한다면, 악의적인 공격자가 높은 확률로 인증을 우회하는 것이 가능하다. 이러한 문제점을 해결하기 위하여, 본 논문에서는 잡음을 인증을 위한 정보로 활용함으로써, 목소리에 잡음을 섞는 방법으로 화자를 인식하는 방안을 제안한다.

II. 관련 연구

1. 화자 인식 기술

화자 인식 기술은 주어진 음성으로부터 그 화자에 대한 정보를 찾아내는 기술로, 일반적으로 화자 검증과 화자 식별로 나누어진다. 화자 식별 기술은 그림 1과 같이 특정 사람이 시스템에 음성을 입력하면, 시스템에 저장된 화자 음성과 유사도를 비교 후 가장 일치하는 화자를 찾는다[6].

2. MFCC 알고리즘

MFCC(Mel Frequency Cepstral Coefficients) 알고리즘은 음성 인식에서 많이 사용되는 알고리즘으로, 소리의 특징을 추출한다. 이 알고리즘은 입력된 소리의 특징을 전부 추출하는 것이 아닌, 일정 구간, 일반적으로 20ms~40ms 정도의 작은 프레임으로 나누는 과정을 거치고, 나누어진 프레임들의 스펙트럼을 분석함으로써, 소리의 특징을 추출한다. MFCC를 이용한 특징 추출은 음정이 변해도 추출된 특징은 일정하다는 장점이 있어, 음성 인식에 효과적이다[2, 6].

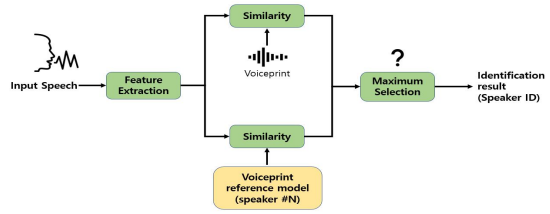

Fig. 1. 화자 인식 기술

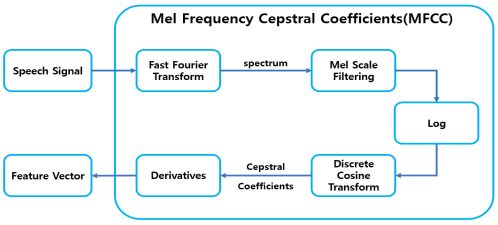

Fig. 2. MFCC 알고리즘

그림 2는 MFCC의 전체 알고리즘을 나타내었으며, 사람마다 목소리의 고유 주파수나 진폭 등이 다르다는 특징을 활용함으로써 화자를 인식한다[6]. 화자 인식에 대한 기술적인 접근 방법으로는 GMM(Gaussian mixture model)과 SVM(Support vector machine) 기법이 대표적이며, 본 논문에서는 SVM 기법을 이용하여 학습하였다[4].

3. 화자 인식 응용 서비스

IBK 기업은행은 20년 6월부터 국내 최초로 음성 본인확인 서비스인 Voice ID를 도입하였다. 음성 본인확인 기술은 개개인이 가진 100가지 이상의 목소리 특징을 수집한 정보로 고객을 식별함으로써, 상담과 금융거래에 활용한다. 음성 정보를 처음으로 등록하기 위해서는 영업점을 방문하여야 하며, 키오스크(무인 창구)나 직원에게 본인을 확인하고 음성 정보를 등록한다. 이후 은행은 수집된 고객 목소리의 특징을 인공지능으로 분석하고 음성 인증 서버에 보관한다.

음성 인증 과정은 고객이 상담원과 통화를 하며 음성 정보를 제공하면, 은행은 서버에 보관된 고객의 음성 정보를 검색한 후 해당 음성과 통화 중인 고객의 음성을 비교함으로써 본인을 확인한다.

기업은행은 뉘앙스 커뮤니케이션즈라는 미국의 음성 인식 전문 세계적 기업의 음성 소프트웨어를 사용하였고, 위와 같은 서비스 도입하기 전, 성능 검증 및 시스템 보안 테스트를 진행하였다. 그 결과, 다양한 생활 환경과 사칭 테스트에서 문제없이 방어에 성공하였다.

하지만, 어떠한 방식으로 사칭 테스트를 방어하였는지에 대한 내용은 알려지지 않았으며, 악의적인 공격자가 나날이 발전하는 인공지능 기술을 이용함으로써, 사용자 목소리를 흉내 내거나 마이크와 같은 녹음 장치로 목소리를 획득하였을 경우에는 충분히 본인 인증이 가능할 것으로 사료된다. 이에 따라, 본 논문에서는 악의적인 공격자가 정당한 사용자 목소리를 흉내 내었을 경우와 실제 사용자 목소리를 녹음하고 녹음된 목소리로 사용자를 인증하는 경우의 사용자 인증 성능을 평가함으로써, 기존 기술의 문제점을 도출하였다.

III. 조건에 따른 화자 인식 성능 평가

1. 실제 사용자 목소리에서의 화자 인식 성능 평가

첫 번째로, 실제 사용자 목소리로 화자 인식의 성능을 평가하기 위하여, 사용자 목소리 데이터를 약 20개 정도 수집하였고, 정확도를 측정하기 위하여, 머신러닝을 활용하여 수집된 목소리 데이터를 학습하였다. 학습된 모델을 통하여, 테스트 목소리의 정확도가 95% 이상이

면 사용자를 인증하는 것으로 나타났다.

정확도를 비교하기 위하여, MFCC 알고리즘을 이용하여 사용자 목소리의 주파수를 추출하였고, 그 결과, 그림 3과 같이 주파수의 시각화 및 데이터화가 가능하며, 이를 활용하여 정확도를 비교하였다 [6].

2. 사용자 목소리 흉내 및 녹음에서의 화자 인식 성능 평가

두 번째로, 악의적인 공격자가 정당한 사용자 목소리를 흉내 내었을 경우와 녹음기 등을 통하여 사용자 목소리를 획득하였을 경우를 가정한다. 이러한 상황에서 흉내 및 녹음에서의 화자 인식의 성능을 평가하였다. 먼저 스피커에서 출력되는 사용자 목소리를 스마트폰의 음성 녹음 기능을 이용하여 획득하였을 경우, 정확도가 76%로 나타났으며, 스피커 출력 및 마이크 입력에서의 잡음을 제거하는 기술을 활용한다면, 이러한 결과보다 더욱 향상된 정확도가 도출될 것으로 판단된다.

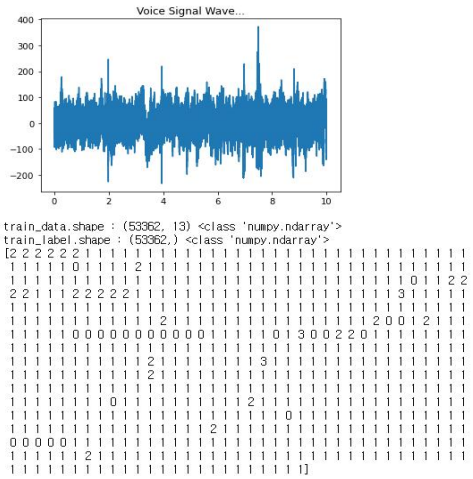

Fig. 3. MFCC를 활용한 추출된 사용자 목소리의 시각화 및 데이터화

다음은 사용자 목소리를 흉내 내었을 경우 정확도를 도출하였다. 흉내 내는 목소리를 활용하기 위하여, "이선균" 씨의 목소리를 선정하였으며, "이선균" 씨의 목소리 데이터를 약 20개 정도 수집하여 학습하였다. 이와 마찬가지로, 목소리를 흉내 내는 데이터를 수집하기 위하여, "이선규" 씨 성대모사 데이터도 수집하였으며, 성대모사 데이터의 정확도는 30% 정도로 나타났다. 성대모사의 경우에는 정확도가 상당히 높을 것으로 예상하였지만, 생각보다 낮은 정확도를 보였다. 하지만 이는 악의적인 공격자가 "네이버 클로바의 내 목소리로 동화 읽어주기"와 같이 인공지능을 활용하여 사용자 목소리를 학습시키고, 잡음을 제거하고 목소리를 가공하는 기술을 활용한다면, 이러한 결과보다 더욱 향상된 정확도가 도출될 것으로 판단된다.

3. 잡음을 섞은 목소리에서의 화자 인식 성능 평가

대부분의 화자 인식에서의 잡음은 제거 대상으로 연구가 진행되었다. 하지만, 이러한 잡음을 인증 정보 중 하나로 활용한다면, 화자 인식의 성능을 향상시킬 것으로 판단된다. 이 방안은 사용자와 서버만 알고 있는 특별한 소리인 "잡음"을 함께 녹음하여 학습함으로써, 목소리를 흉내 내거나 녹음을 통하여 목소리로 사용자를 인증하는 것을 우회하는 공격에 대응이 가능할 것으로 판단된다.

이를 토대로, "이선균" 씨 목소리에 사이렌 소리를 잡음으로 추가한 데이터를 22개 수집하여 학습하였으며, 잡음이 녹음된 "이선균" 씨 목소리의 정확도는 83%로 나타났다.

하지만 사이렌 소리를 잡음으로 학습시킨 "이선균" 씨 목소리를 토대로, 잡음이 없는 "이선균" 씨 목소리의 정확도를 측정한 결과, 4%로 나타났으며, 성대모사를 한 "이선균" 씨 목소리는 1%의 정확도, 그리고 스피커에서 출력되는 "이선균" 씨 목소리를 스마트폰의 음성 녹음 기능을 이용한 경우는 1%의 정확도로 나타났다. 전체 성능 평가 결과를 표 1에 나타내었다.

Table 1. 조건에 따른 화자 인식 성능 평가 결과

조건에 따른 화자 인식		정확도
실제 목소리 인식		95% 이상
(공격) 목소리 녹음		76% (향상 가능)
(공격) 목소리 흉내		30% (향상 가능)
잡음 활용 목소리 인식		83%
(방어) 잡음 활용	(공격) 잡음 없는 실제 목소리	4%
	(공격) 목소리 녹음	1%
	(공격) 목소리 흉태	1%

이와 같은 결과를 토대로, 잡음을 포함하여 목소리를 학습시킨 경우에는 악의적인 사용자가 목소리를 흉내 내거나 녹음기로 목소리를 획득하더라도 정확도가 명백하게 차이가 나타나기 때문에, 목소리로 사용자를 인증하는 것을 우회하는 공격에 효과적으로 대응한다. 이러한 장점에도 불구하고, 해당 방안은 사용자의 목소리를 등록할 때, 잡음을 추가하여 목소리를 등록한 것으로 가정한다면, 이후, 인증 과정에서 항상 잡음이 요구되므로, 보안성은 향상되지만, 사용성 측면에서는 한계점이 있을 것으로 판단된다.

IV. 결론

본 논문에서는 기존의 화자 인식 기술의 한계점을 극복하기 위하여 잡음을 섞음으로써 정확도를 향상시키는 방안을 제안하였다. 조건에 따른 화자 인식 성능 평가의 결과를 살펴보면, 실제 사용자 목소리로 화자 인식의 성능을 평가한 결과, 정확도가 95% 이상으로, 목소리로 사용자를 인증할 수 있을 것으로 판단된다.

하지만, 목소리를 흉내 내거나 목소리를 녹음하는 경우에는 취약한 부분이 존재하며, 목소리를 흉내 내는 경우의 성능을 평가한 결과, 정확도가 30% 정도로 나타났다. 또한, 목소리를 녹음하는 경우의 성능을 평가한 결과, 정확도가 76%로 나타났으며, 두 경우 모두 잡음을 제거하거나 목소리를 가공한다면 더욱 향상된 정확도가 도출될 것으로 판단된다.

이러한 문제점을 해결하고자, 잡음을 섞은 목소리의 성능을 평가한 결과, 정확도가 83%로 나타났으며, 이는 잡음이 없는 실제 목소리의 정확도인 4%, 목소리를 녹음하는 경우의 정확도인 1%, 목소리를

흉내 내는 경우의 정확도인 1%인 점을 감안할 때, 정확도의 차이가 명백하므로, 효과적으로 화자 인식이 가능할 것으로 판단된다. 그럼에도 불구하고, 사용자 목소리는 주변 환경 및 사람의 상태, 예를 들면 심한 감기나 성대결절에 걸린 경우에는 매우 심하게 변하는 특징이 있으므로, 잡음을 활용하더라도 화자를 인식하기에는 한계가 있을 것으로 판단된다. 향후, 이러한 한계점을 극복하기 위한 연구를 진행할 예정이다.

ACKNOWLEDGEMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2021R1F1A1050542).

REFERENCES

[1] Y. Kim and H. Yun, "A Study on the Improvement Plan of Voice Recognition Security Vulnerability," Proceedings of The Korea Information Processing Society Conference, pp. 746-748, Nov. 2017.

[2] N. Kim and J. Choi, "A study on user authentication method using speaker authentication mechanism in login process," Journal of The Smart Media Journal, Vol. 8, No. 3, pp. 23-30, Sep. 2019.

[3] B. Cho, S. Cheon, K. Kim, and H. Yuk, "A policy study for the voice recognition technology based on elderly health care," Journal of Digital Convergence, Vol. 16, No. 2, pp. 9-17, Feb. 2018.

[4] K. Lee, "A Study on SVM-Based Speaker Classification Using GMM-supervector," Journal of The Institute of Korean Electrical and Electronics Engineers, Vol. 24, No. 4, pp. 1022-1027, Dec. 2020.

[5] P. Jeong and Y. Cho, "User Authentication Mechanism using Smartphone," Journal of the Korea Institute of Information and Communication Engineering, Vol. 21, No. 2, pp. 301-308, Feb. 2017.

[6] Dydtjr1128, "Speaker-Recognition-using-NN(2019)," Retrieved Jun. 23, 2022, from *https://github.com/dydtjr1128/Speaker-Recognition-using-NN.*