말소리와 음성과학 제6권 제2호 (2014.06.30)
ISSN 2005-8063, pp. 21~28
www.speechsciences.or.kr
http://dx.doi.org/10.13064/KSSS.2014.6.2.021

법음성학에서의 오디오 신호의 위변조 구간 자동 검출 방법 연구

An Automatic Method of Detecting Audio Signal Tampering in Forensic Phonetics

양 일 호[1] · 김 경 화[2] · 김 명 재[3] · 백 록 선[4] · 허 희 수[5] · 유 하 진[6]
Yang, IL-Ho · Kim, Kyung-Wha · Kim, Myung-Jae · Baek, Rock-Seon · Heo, Hee-Soo · Yu, Ha-Jin

ABSTRACT

We propose a novel scheme for digital audio authentication of given audio files which are edited by inserting small audio segments from different environmental sources. The purpose of this research is to detect inserted sections from given audio files. We expect that the proposed method will assist human investigators by notifying suspected audio section which considered to be recorded or transmitted on different environments. GMM-UBM and GSV-SVM are applied for modeling the dominant environment of a given audio file. Four kinds of likelihood ratio based scores and SVM score are used to measure the likelihood for a dominant environment model. We also use an ensemble score which is a combination of the aforementioned five kinds of scores. In the experimental results, the proposed method shows the lowest average equal error rate when we use the ensemble score. Even when dominant environments were unknown, the proposed method gives a similar accuracy.

Keywords: forgery detection, digital audio authentication, GMM, SVM, classifier ensemble

1. 서론

디지털 기술이 발달함에 따라 최근 범죄와 관련된 증거를 촬영하거나 녹음하여 수사기관에 제출하는 경우가 증가하고 있다. 이와 함께 사건의 피의자가 증거 녹음자료가 편집되었다고 주장하는 사례도 꾸준히 늘고 있다. 최근 언론에 보도된 사건에서도 증거 녹음파일의 편집여부가 쟁점이 되면서, 관련 학계를 비롯하여 법조계에서도 위변조 식별 방법에 대한 관심이 높아지고 있다. 녹음 파일이 결정적 증거인 사례가 늘면서, 음성감정관이 법정에서 녹음 파일 분석 결과에 대하여 증언을 하고, 재판부가 이를 토대로 증거 채택 여부를 결정하는 경우가 종종 있다. 이에 따라 녹음 파일에 대한 정밀 분석 결과를 도출하고, 분석 결과와 녹음 파일이 법정에서 증거로 인정받기 위해서 오디오 파일의 위변조 식별 기술에 대한 지속적인 연구가 필요하다.

범죄 수사와 관련된 음성분석 분야에서는 증거로 제출된 녹음 파일에 인위적으로 편집이 가해졌는지를 판단한다. 편집 여부 분석을 위해서는 청취분석, 음향분석, 언어학적 분석, 위변조 식별 장비를 이용한 분석 등 다양한 방법을 사용한다. 또한 분석을 수행하는 음성감정관은 오디오 포렌식에 대한 전문 지식과 기술을 습득하고 실무 경험을 축적하는 것이 필수적이다.

최근 디지털 녹음기와 스마트폰의 기능이 발달하면서 10시간 이상의 장시간 녹음이 가능해졌고 실제로 증거 녹음파일이 2~3시간 이상인 경우가 크게 늘고 있다. 음성감정관이 오디오 파일을 분석하는 경우 파일 전체를 여러 번 듣고 오디오 신호를 분석하기 때문에, 분석 시간이 녹음 시간에 비례하고 오랜 시간과 노력이 필요하다.

그러므로 음성공학적 기술을 바탕으로 위변조 구간을 자동으로 검출하는 기술을 개발하여 기존의 분석 방법과 함께 사용함으로써, 오디오 파일 분석에 소요되는 시간을 단축하고,

1) 서울시립대학교, heisco@hanmail.net
2) 대검찰청 음성분석실, savoix@spo.go.kr
3) 서울시립대학교, mj@uos.ac.kr
4) 서울시립대학교, whites86@naver.com
5) 서울시립대학교, zhasgone@naver.com
6) 서울시립대학교, hjyu@uos.ac.kr, 교신저자

이 논문은 2012~2013년 대검찰청 연구용역의 지원으로 수행되었습니다. (과제명: 디지털 음성신호 위변조 식별기법 연구 I, II)

접수일자: 2014년 5월 19일
수정일자: 2014년 6월 10일
게재결정: 2014년 6월 16일

감정 결과의 정확도를 제고할 수 있을 것이다.

자동 위변조 검출에 대한 많은 연구가 수행되었는데, 최근 주목 받고 있는 방법 중 하나는 음성 데이터 내의 전기 네트워크 주파수(ENF, electric network frequency)를 분석하는 것이다[1]-[4]. ENF는 발전 및 송전 과정에서 발생하는 60Hz 안팎(일부 외국에서는 50Hz)의 교류 신호인데, 이를 지역 · 시간대별로 데이터베이스화한 후 음성 데이터에서 추출한 ENF와 비교하여 증거 제출자가 주장하는 데이터의 녹음 장소 및 시간이 일치하는지, 혹은 음성 데이터 내에 삭제되거나 삽입된 위변조 구간은 없는지 확인할 수 있다[1]-[3]. 하지만 이를 위해서는 ENF 데이터베이스가 필요하므로, 이러한 ENF 데이터베이스를 수집하는 시스템이 구축되어 있지 않은 지역에서는 이 방법을 사용할 수 없다. 또 다른 접근 방법은 음성 데이터에서 추출한 ENF 위상의 불연속점이 있는지 확인하는 것이다[1], [4]. 만약 음성 데이터의 특정 구간이 삭제되거나 삽입되었다면 이러한 ENF 불연속점이 발생할 수 있다. 하지만 교류 전자기장에서 멀리 떨어져 녹음하는 경우나 녹음 기기의 종류에 따라, 혹은 사용한 코덱 등의 영향을 받아 ENF 추출이 어려울 수 있다[1], [3], [5]. ENF를 사용하지 않고 불연속점을 찾는 다른 방법으로 참고문헌[4]에서는 LPC(linear prediction coefficient)와 같은 스펙트럼 기반 특징의 거리를 추출하여 인접한 구간의 스펙트럼 거리가 커지는 지점을 불연속점으로 검출하는 방법을 제안하였다.

본 연구에서는 참고문헌[4]와 유사하게 스펙트럼 특징을 사용하되, 불연속점을 찾는 것이 아니라 삽입 위변조시 불연속점과 불연속점 사이에 인위적으로 삽입된 구간을 검출하고자 하였다. 서로 다른 두 종류의 음성 데이터를 이어 붙인 경우 두 음성 데이터를 녹음한 환경이나 전송 채널, 압축 코덱 등에 의해 구간별 스펙트럼에 차이가 발생할 것이므로, 절단면만을 찾는 것보다 삽입된 구간을 찾는 것이 더 유리할 것으로 기대하였다. 예를 들어, 전체 음성 데이터 중 연속된 1구간이 다른 채널을 거친 음성을 삽입한 위변조라고 할 때, 불연속점은 최대 2지점(맨 앞 · 맨 뒤에 삽입한 경우에는 1지점)이 발생하지만 삽입된 구간 내의 스펙트럼 특징을 활용하면 연속된 삽입 구간의 길이가 길면 길수록 보다 많은 정보를 활용할 수 있으므로, 보다 높은 정확도를 보이거나, 불연속점을 검출하는 방법과 복합적으로 사용함으로써 상호 보완할 수 있을 것으로 기대하였다.

이와 관련된 연구는 참고문헌[6]-[9] 등이 있다. 참고문헌[6]에서는 시간 영역 및 멜 켑스트럼 영역 기반 특징을 추출하여 베이지안 분류기로 녹음에 사용한 마이크 및 녹음 장소를 식별하였다. 참고문헌[7]에서는 녹음에 사용한 유선전화 혹은 마이크가 다양한 오디오 신호로부터 MFCC(mel-frequency cepstral coefficient)s 및 LFCC(linear-frequency cepstral coefficient)s 특징을 추출하여 GMM-UBM(Gaussian mixture model - universal background model)[10] 방식으로 개별 오디오 파일의 GSV(Gaussian supervector)를 추출한 후 SVM (support vector machine)[11]으로 식별하는 연구를 수행하였다. 참고문헌[8]에서는 MP3 혹은 WMA로 압축하였다가 WAV로 복원한 파일로부터 이전에 거친 압축 방식 및 압축시 비트율을 식별할 수 있음을 보였고, 참고문헌[9]에서는 G.711, G.726, G.728, G.729, iLBC, AMR, Silk 등의 코덱을 거친 오디오 파일을 식별하였다. 그러나 이 연구들은 개별 오디오 파일의 녹음 환경 및 코덱 등을 식별한 것으로서 동일한 오디오 파일 내에서 위변조된 구간을 검출하는 방법은 아니다.

본 연구에서는 참고문헌[7]과 유사하게 GMM-UBM으로 녹음 환경을 모델링하되, 한 오디오 파일 내에서 삽입 위변조된 구간을 찾고자 하였다. 이 때 녹음 환경은 원본 오디오 파일을 다양한 샘플링레이트와 코덱을 거쳐 시뮬레이션 하여 실험하였다. 또한, 위변조 검출을 하기 위해서 사전에 모든 녹음 환경에 대한 모델을 학습해야 하는 부담을 최소화 하고, 미지의 녹음 환경을 거친 위변조 구간도 검출할 수 있도록 UBM(universal background model) 및 점수 정규화를 위한 레퍼런스 모델들을 제외하고는 오프라인 단계에서 인식 대상 모델을 별도로 학습하지 않는 방법을 고안하였다.

본 연구에서의 가정은 다음과 같다.

음성 데이터를 음성감정관이 수작업으로 분석하려면 많은 시간이 필요하다. 전체 음성 데이터 중 일부는 다른 환경에서 녹음 및 전송된 음성을 삽입한 위변조 구간이다. 동일한 음성 데이터 내에서 일부를 복사하여 붙여 넣었거나 삭제한 경우는 고려하지 않는다(기존의 불연속점 검출 방법을 활용한다고 가정). 제안한 위변조 검출 시스템은 위변조되었을 가능성이 높은 순으로 의심 구간을 음성감정관에게 제시함으로써, 감정관이 위변조 여부의 최종 결정을 내리는데 도움을 줄 수 있다.

이를 통해, 음성감정관이 음성 데이터 전체를 순차적으로 정밀 분석하는 대신 위변조 가능성이 높은 구간을 우선적으로 집중 분석함으로써 분석에 소요되는 시간을 단축할 수 있을 것이다.

본 논문의 구성은 다음과 같다. 2장에서 제안한 위변조 검출 시스템을 소개하고, 3장에서 성능을 평가한 뒤, 4장에서 결론을 맺는다.

2. 제안한 위변조 검출 시스템

2.1 전체 시스템 구조

제안한 시스템의 개략적인 구조는 GMM-UBM[10] 방식으로 통계적 모델을 구성한 후 일정한 길이의 프레임별로 log LR등의 점수를 계산하여 전체 오디오 파일과 특징 분포가 상이한 구간을 검출하는 것이다. <그림 1>은 전체 오디오 파일을 일정한 길이와 간격의 여러 프레임으로 나누어 각 프레임

별 유사도 점수를 계산하는 테스트 과정을 개략적으로 나타낸 것이다.

이 때, 유사도 점수는 오디오 파일의 주된 녹음(혹은 전송 및 저장) 환경 모델과 각 프레임의 특징 분포가 얼마나 유사한지를 계산하는 척도로, 이 값이 작을수록 위변조된 구간일 가능성이 높을 것이라고 가정한다. 유사도 점수를 계산할 때에는 서로 다른 5 종류의 척도를 이용하여 독립적으로 위변조 의심 구간을 검출한 후 계산된 점수를 앙상블 결합한다.

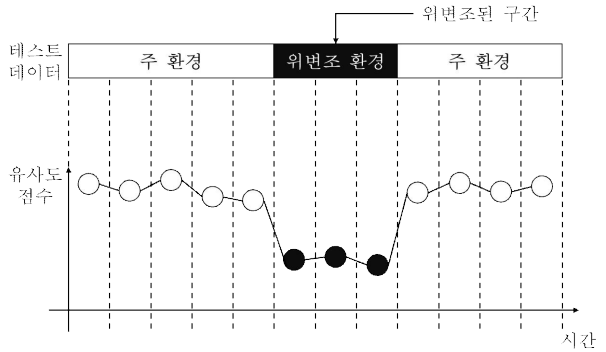

그림 1. 제안한 위변조 검출 시스템의 원리
Figure 1. Concept of proposed forgery detection system

2.2 테스트 데이터의 환경 모델 학습

사전에 다양한 환경의 오디오 데이터로부터 추출한 특징들로 하나의 UBM을 학습하고, 입력된 테스트 데이터(위변조 검출 대상)를 이용하여 테스트 데이터의 환경 모델을 얻는다. 유사도 점수 계산 척도에 따라 LR 기반 척도인 경우에는 전체 특징을 이용하여 UBM으로부터 1회 MAP 적응한 GMM(Gaussian mixture model)을 얻고(<그림 2>), SVM 점수 기반 척도인 경우에는 프레임별 특징으로 MAP(maximum a posteriori) 적응한 GMM 평균으로 GSV를 추출하여 1-class SVM[12]을 학습한다(<그림 3>). 이렇게 학습한 테스트 데이터의 환경 모델은 테스트 데이터의 주된 환경 특성을 잘 나타낼 것이라고 가정한다.

2.3 유사도 점수 계산 척도

2.3.1 GMM log LR

화자 확인 분야에서 널리 사용하는 log LR[10]을 유사도 점수로 사용하였다. 이 때, 테스트 데이터 환경 모델과 더 유사하다면 주 환경으로, UBM과 더 유사하다면 위변조 환경으로 간주하였다. i번째 프레임의 테스트 데이터 환경 모델에 대한 log likelihood가 $LL_{test}(i)$,UBM에 대한 log likelihood가 $LL_{ubm}(i)$ 라 할 때, GMM log LR 점수는 다음과 같다.

$$score_{LLR}(i) = LL_{test}(i) - LL_{ubm}(i) \quad (1)$$

2.3.2 T-normalized log LR

다양한 환경의 레퍼런스 모델을 사전에 학습하고 이를 이용하여 각 프레임별로 계산한 log LR을 정규화(T-norm)하였다. 본 연구에서는 레퍼런스 환경 모델의 종류를 UBM 학습 시 사용한 환경과 동일하게 하였다.

2.3.3 GMM log LR의 편차

만약 UBM 학습시 환경의 종류가 충분하지 못했을 경우, 미지의 환경을 지닌 테스트 데이터에 대해 위변조 구간 프레임의 유사도 점수가 정상 구간 프레임의 유사도 점수보다 높게 나올 가능성도 있다. 만약 테스트 데이터 환경 모델이 잘 학습되었고 프레임의 길이가 충분히 길다면 모든 정상 구간 프레임의 log LR이 비슷한 값으로 추출될 것이라 기대하였다. 따라서, 평균 log LR과의 차이가 큰 프레임은 위변조 되었을 가능성이 높다고 판단할 수 있을 것이다.

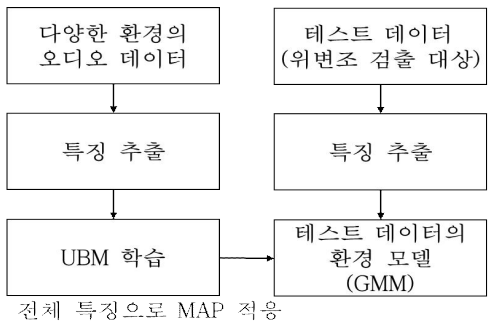

그림 2. 테스트 데이터의 환경 모델 학습 과정 (LR 기반 척도)
Figure 2. Training process of test environment model (for LR based scores)

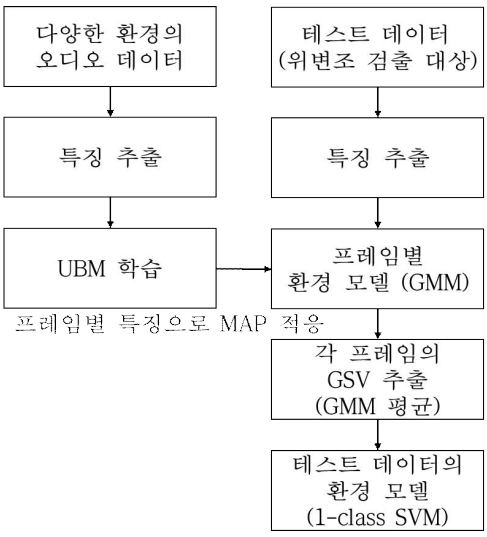

그림 3. 테스트 데이터의 환경 모델 학습 과정 (SVM 기반 척도)
Figure 3. Training process of test environment model (for SVM based score)

전체 오디오 파일이 총 K개 프레임으로 나누어질 때, GMM log LR의 편차 점수는 편차에 -절대값을 취하여 다음과 같이 계산하였다.

$$score_{LLR-DIFF}(i) = -\left| score_{LLR}(i) - \left(\sum_{j=1}^{K} score_{LLR}(j) \right) / K \right| \quad (2)$$

T-normalized log LR의 편차도 이와 동일한 방법으로 정규화한 log LR의 편차를 계산하였다.

2.3.4 GSV - 1 class SVM 점수

이는 참고문헌[7]에서와 유사하게 각 프레임별로 GSV를 추출하여 SVM으로 테스트 데이터의 환경 모델을 학습하는 방법이다. 이 때, 비교사 방식으로 적용하기 위해 참고문헌[7]과는 달리 1-class SVM[12]으로 학습하고, 각 프레임별 GSV의 유사도를 척도로 사용하였다. 1-class SVM을 사용한 이유는 오프라인 단계에서 별도의 모델 학습을 수행하지 않고 위변조 의심 구간을 검출하기 위해서이다. 즉, 각 프레임별 GSV로 학습한 1-class SVM이 테스트 데이터의 환경 모델을 잘 나타내며, 이 모델에 대한 SVM 점수가 낮은 프레임은 다른 환경에서 녹음된 구간일 가능성이 높을 것으로 보았다. 이 때, GSV는 각 프레임별 특징으로 학습한 GMM으로부터 평균을 취하여 구성하였다.

2.3.5 분류기 앙상블 점수

각 프레임별로 앞서 정의한 5가지 유사도 점수(GMM log LR, T-normalized log LR, GMM log LR의 편차, T-normalized log LR의 편차, GSV - 1 class SVM 점수)를 계산한 후 합산하여 최종적인 유사도 점수로 사용하였다. 단, 합산 전에 각 점수를 정규화하였다.

$$score_{ENSEMBLE}(i) = \sum_{s \in \{S\}} \left(score_s(i) - \mu_s \right) / \sigma_s \quad (3)$$

이 때, $\{S\}$ 는 앞서 열거한 5가지 유사도 점수 집합을 나타내며, $score_s$, μ_s , σ_s 는 각각 점수 s ,모든 프레임에 대한 점수 s 의 평균 및 표준 편차를 의미한다.

3. 성능 평가

3.1 데이터베이스

3.1.1 원본 데이터베이스

ETRI 중가마이크 화자인식용 데이터베이스를 <표 1>과 같이 구분하여 실험에 사용하였다. 해당 데이터베이스는 주차, 월차, 3개월차의 3종류 간격으로 녹음된 화자 총 250명의 숫자 및 문장 발성으로 구성되어 있으며, 4번의 시차를 두고 각 시차별로 5회씩 발성한 것이다. 각 발성의 저장 형식은 16kHz, mono, headerless PCM이며, 실험에 사용한 문장 발성 1개의 길이는 약 2초 가량이고 시작과 끝 부분에 각각 0.15초 내외의 무음 구간을 포함한다. 화자 인식 분야의 연구에 사용[13]되는 데이터베이스이나 본 연구에서는 이 데이터를 가공하여 가상의 위변조 파일을 생성하였다.

표 1. 원본 데이터베이스 구분
Table 1. Partitioning of original database

항목	내용	총 발성 수
UBM 학습	월차 화자 100명의 1시차 1회차 문장 발성 10개	1000개
T-norm을 위한 레퍼런스 모델 학습	3개월차 화자 50명의 1시차 1회차 문장 발성 10개	500개
테스트	주차 화자 100명의 1~4시차 1~5회차 문장 발성 10개	20000개

3.1.2 다양한 환경의 데이터로 시뮬레이션

다양한 환경의 데이터가 필요하므로 원본 데이터를 시뮬레이션하였다. 원본 데이터베이스에 다운샘플링(8kHz) 및 MP3, AMR, OPUS 등의 코덱을 거쳐 시뮬레이션하였다. 다운샘플링, 업샘플링, AMR 코덱 시뮬레이션은 FFmpeg[14]을 이용하여 수행하였고, MP3 코덱은 lame[15](3.98.2 버전), OPUS는 opus audio tools[16](0.1.8 버전)를 사용하였다. 전체 시뮬레이션 과정은 <그림 3>과 같다.

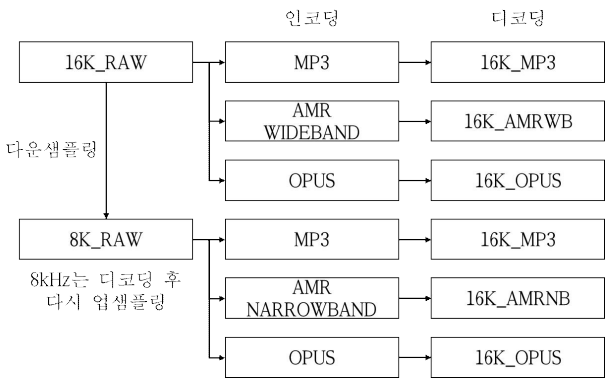

그림 3. 환경 시뮬레이션 과정
Figure 3. Simulating process for various environments

실험의 각 단계에서 사용한 환경은 <표 2>와 같다.

표 2. 시뮬레이션한 환경 구분
Table 2. Partitioning of simulated environments

항목	사용한 환경	비고
UBM 학습	16K_RAW, 8K_RAW 16K_MP3, 16K_MP3 16K_OPUS, 16K_OPUS	1개의 GMM
T-norm을 위한 레퍼런스 모델 학습	16K_RAW, 8K_RAW 16K_MP3, 16K_MP3 16K_OPUS, 16K_OPUS	환경별 모델 학습, 총 6개의 GMM
테스트	16K_RAW, 8K_RAW 16K_MP3, 16K_MP3 16K_OPUS, 16K_OPUS 16K_AMRWB, 8K_AMRNB	시뮬레이션 후 인공데이터 생성

각 환경은 [샘플링레이트]_[코덱]으로 구분하였다. 즉, 원본 데이터베이스는 16kHz(16K), headerless PCM(RAW)이므로 “16K_RAW”와 같이 표기하였다. 테스트 환경 중 “16K_AMRWB”와 “8K_AMRNB”는 UBM 학습 및 레퍼런스 모델 학습시 제외된 미지의 환경에 대해 위변조 검출 정확도를 평가하기 위해 추가한 것이다.

3.1.3 인공적인 위변조 데이터 생성(테스트용)

테스트의 경우는 각 환경별로 화자별 200개의 발성을 순차적으로 연결하되, 이 중 일정 비율의 발성을 선택하여 다른 환경으로 시뮬레이션한 데이터로 대체하였다. 주 환경과 위변조 환경의 쌍으로 구분하여 모든 환경 조합에 대해 총 56종류의 실험 SET을 인공적으로 생성하였다. 1 SET은 100개 위변조 파일로 구성하였으며 한 파일의 길이는 약 8분 내외이다. 한 파일 내 위변조 구간의 비율은 전체 길이 대비 5, 10, 20%로 하였다(이 때, 위변조 구간의 길이는 각각 약 24, 48, 96초).

3.2 실험 설계

3.2.1 특징 추출 및 위변조 검출시 프레임 길이

16차 켑스트럼(20ms window, 10ms shift)을 특징으로 사용하였다. 테스트 파일에 대한 위변조 검출시 한 프레임의 길이 및 간격은 10초(1000개 특징 벡터 포함)로 하였다. 검출 프레임의 길이가 다소 긴 이유는, 너무 짧게 설정할 경우 채널 등의 환경 정보에 비해 상대적으로 빠르게 변화하는 음성 등의 다른 요인에 민감하게 영향을 받아 위변조 검출에 악영향을 미치는 것을 방지하기 위해서이다.

3.2.2 UBM 및 레퍼런스 모델 학습

GMM혼합 수 1, 4, 16, 64개로 학습한 UBM과 이로부터 1회 MAP 적응한 레퍼런스 모델을 각각 사용하였다. GMM 학습 및 log likelihood 계산은 HTK(3.4.1 버전)[17]를 이용하였다.

3.2.3 GSV-SVM

입력된 오디오 파일의 각 구간의 특징으로 UBM으로부터 1회 MAP 적응한 GMM 평균의 GSV를 추출하고 1 class SVM을 학습한 후 각 프레임별 점수를 계산하였다. SVM의 커널은 가우시안 RBF 커널을 사용하였다. 1-class SVM의 모델링 및 점수 계산에는 libsvm[18](3.17 버전)을 이용하였다.

3.3 실험 결과 및 분석

<표 3>은 56종류 실험 SET에 대해 각 오디오 파일별로 EER(equal error rate)을 계산한 평균치이다. GMM log LR은 “LLR”, GMM log LR의 편차는 “LLR-DIFF”, T-normalized log LR은 “TNORM”, T-normalized log LR의 편차는 “TNORM-DIFF”, GSV - 1 class SVM 점수는 “GSV-1CSVM”, 분류기 앙상블 점수는 “ENSEMBLE”로 표기하였다.

실험 결과 분류기 앙상블 점수(“ENSEMBLE”)를 제외한 모든 척도 사용시 위변조 구간 비율이 5%일 때 가장 오류가 적었다. 반면 위변조 구간 비율이 20%일 때에는 전체적으로 오류가 증가하는 경향을 보였다. 이는 테스트 데이터 환경 모델을 학습할 때, 위변조 구간이 포함되기 때문에 위변조 구간의 길이가 길 경우 학습된 모델이 주 환경을 잘 표현하지 못하기 때문으로 판단된다.

표 3. 실험 결과 (평균 EER)
Table 3. Experimental results (average EER)

위변조 구간 비율 (%)	혼합 수	LLR (A)	TNORM (B)	LLR -DIFF (C)	TNORM -DIFF (D)	GSV-1C SVM (E)	ENSEMBLE (A+B+C+D+E)
5	1	16.56	15.74	18.70	18.04	26.94	13.90
	4	12.71	**12.80**	15.03	13.78	18.62	11.67
	16	**10.77**	13.27	**12.70**	**13.60**	**12.73**	9.73
	64	11.47	16.81	13.49	16.90	12.82	10.34
10	1	16.78	16.00	19.57	17.78	27.24	13.94
	4	12.78	13.26	15.69	15.02	19.73	11.33
	16	11.13	13.50	13.58	15.38	14.14	**9.63**
	64	12.44	18.11	14.44	17.22	14.27	10.84
20	1	18.26	18.72	20.53	19.86	32.67	16.25
	4	14.26	16.29	18.38	18.69	26.02	13.85
	16	13.11	17.02	17.95	18.11	21.14	12.41
	64	14.93	21.02	19.48	19.02	20.88	13.78

T-normalized log LR(“TNORM”)을 제외한 모든 척도 사용 시 혼합 수 16개일 때 가장 오류가 적었다.

앙상블 결합하지 않은 각 척도별 최저 오류율은 “LLR”, “LLR-DIFF”, “TNORM”, “TNORM-DIFF”, “GSV-1CSVM” 순으로 각각 10.77, 12.80, 12.70, 13.60, 12.73%로 GMM log LR(“LLR”)이 가장 낮았으며, 앙상블 결합시 최저 오류율은 9.63%로 개별 척도만 사용하였을 때 보다 더 낮았다. 앙상블 결합시 상대 오류 감소율은 개별 척도(“LLR”, “LLR-DIFF”, “TNORM”, “TNORM-DIFF”, “GSV-1CSVM” 순) 대비 각각 10.58%, 24.77%, 24.17%, 29.19%, 24.35%였다.

앙상블 결합시 최저 오류율을 보인 경우(위변조 구간 비율 10%, 혼합 수 16개일 때) 실험 SET의 환경 조건에 따라 평균 EER을 계산한 결과는 <표 4>와 같다. 환경 조건은 총 56개 실험 SET에 대해 주 환경과 위변조 환경의 샘플링레이트가 동일하거나 다른 경우(동일한 경우: O, 다른 경우: X), 전체 오디오 파일의 주 환경이 UBM 및 레퍼런스 모델에 포함된 경우(포함된 경우: O, 포함되지 않은 경우: X), 위변조 구간의 환경이 UBM 및 레퍼런스 모델에 포함된 경우(포함된 경우: O, 포함되지 않은 경우: X)로 구분하였다.

표 4. 환경 조건에 따른 실험 결과 (평균 EER)
Table 4. Experimental results according to environmental conditions (average EER)

주 환경과 위변조 환경의 샘플링레이트가 동일	주 환경이 UBM 및 레퍼런스 모델에 포함	위변조 환경이 UBM 및 레퍼런스 모델에 포함	실험 SET 수 (개)	ENSEMBLE (A+B+C+D+E)
X	X	X	18	0.04
X	X	O	6	0.00
X	O	X	6	0.00
X	O	O	2	0.00
O	X	O	12	20.39
O	O	X	6	29.14
O	O	O	6	20.73

실험 결과, 주 환경과 위변조 환경의 샘플링레이트가 다른 경우 매우 낮은 오류율을 보였으며, 주 환경과 위변조 환경의 샘플링레이트가 동일한 경우에는 20%대의 오류율을 보였다. 주 환경과 위변조 환경의 샘플링레이트가 동일하고 테스트 데이터의 모든 환경을 알고 있을 때(주 환경 및 위변조 환경이 UBM 및 레퍼런스 모델에 포함된 경우) 20.73%의 오류율을 보였다. 마찬가지로 샘플링레이트가 동일할 때 미지의 주 환경(주 환경이 UBM 및 레퍼런스 모델에 포함되지 않은 경우)에 대해서는 20.39%로 별 차이가 없었으나, 미지의 위변조 환경(위변조 환경이 UBM 및 레퍼런스 모델에 포함되지 않은 경우)에 대해서는 29.14%로 오류율이 증가하였다. 미지의 주 환경에 대해 모든 환경을 알고 있을 때와 비슷한 오류율을 보이는 이유는, 테스트 데이터 모델을 학습하는 과정에서 주 환경으로 간주할 수 있는 모델을 생성하게 되기 때문이라고 판단된다. 반면, 미지의 위변조 환경에 대해서는 오류율이 다소 증가하는 것을 확인하였다.

다음은 실험에 사용한 데이터 중 한 파일에 대한 분류기 앙상블 점수의 그래프이다(주 환경: 16kHz, AMRWB / 위변조 환경: 16kHz, MP3 / 위변조 구간 비율: 5% / 실험에 사용한 화자 id: bsa00). 실제 위변조 구간을 점선으로 표시하였으며 분류기 앙상블 점수를 통해 위변조 구간을 잘 검출하는 것을 확인할 수 있다.

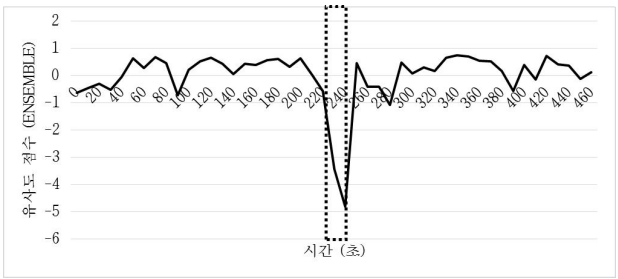

그림 4. 한 실험 데이터 파일에 대한 분류기 앙상블 점수 그래프
Figure 4. A graph of classifier ensemble score of the one experimental data file

4. 결론 및 향후 연구

본 연구에서는 음성 데이터의 일부 구간이 다른 환경에서 녹음된 데이터로 삽입 위변조된 경우, 이를 자동으로 검출하기 위한 방법을 제안하였다. 제안한 시스템은 최종 결정을 음성감정관이 내린다는 가정 하에 위변조 가능성이 높은 프레임을 우선적으로 제시하는 기능을 수행한다. 관련된 연구들에서 사용하는 4종류의 LR 기반 척도들과 GSV-SVM 점수를 각 프레임별로 계산한 후 앙상블 결합하는 형태로 고안하였으며, 이는 사전에 UBM 및 점수 정규화(T-norm)를 위한 레퍼런스 모델 학습을 제외하고는 오프라인 단계에서 인식 대상 모델을 별도로 학습하지 않으므로 미지의 환경에서 녹음된 음성 데이터에 대해서도 위변조 검출이 용이할 것으로 기대하였다.

실험 결과 제안한 앙상블 시스템을 이용하여 위변조 검출 시 개별 척도를 사용하였을 때(10.77~13.60%) 보다 더 낮은 오류(9.63%)를 보이는 것을 확인하였다. 환경 조건에 따른 실험 결과 음성 데이터의 주 환경과 위변조 환경의 샘플링레이트가 동일한 경우 20%대의 오류율을 보였으나, 미지의 주 채널에 대해서도 비슷한 수준으로 위변조 검출이 가능함을 보였다. 하지만 미지의 위변조 채널에 대해서는 오류율이 다소 증

가하였고, 위변조 구간의 비율이 커질 때에도 취약함을 보였는데 이러한 약점은 향후 성능 개선을 통해 극복해야 할 부분이라고 판단된다.

향후 연구로서 앞서 언급한 상황에서의 성능 개선, 서론에서 소개한 불연속점 검출 방법과의 결합, 보다 다양한 상황의 데이터에 대한 성능 평가 등을 수행할 것이다.

참고문헌

[1] Brixen, E. B. (2007). Techniques for the authentication of digital audio recording, *Audio Engineering Society Convention 122.*

[2] Ojowu, O. (2012). ENF extraction from digital recordings using adaptive techniques and frequency tracking, *Information Forensics and Security.* Vol. 7, 1330-1338.

[3] Grigoras, C. (2009). Applications of ENF analysis in forensic authentication of digital audio and video recording, *Journal of the Audio Engineering Society.* Vol. 57, Issue 9, 643-661.

[4] Nicolade, D. P. & Apolinario, J. A. (2009). Evaluating digital audio authenticity with spectral distances and ENF phase change, *in proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,* 1417 – 1420.

[5] Malik, H. (2013). Acoustic environment identification and its applications to audio forensics, *IEEE Transactions on Information Forensics and Security.* Vol. 8, No. 11, 1827-1837.

[6] Kraetzer, C., Oermann, A., Dittmann, J. & Lang, A. (2007). Digital audio forensics: a first practical evaluation on microphone and environment classification, *in proceedings of the 9th workshop on Multimedia & security,* 63-74.

[7] Garcia-Romero, D. & Espy-Wilson, C. Y. (2010). Automatic acquisition device identification from speech recordings, *in proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing.* 1806-1809.

[8] Luo, D., Luo, W., Yang, R. & Huang, J. (2012). Compression history identification for digital audio signal, i*n proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.* 1733-1736.

[9] Jenner, F. & Kwasinski, A. (2012). Highly accurate non-intrusive speech forensics for codec identifications from observed decoded signals, *in proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.* 1737-1740.

[10] Reynolds, D. A., Quatieri, T. F. & Dunn, R. B. (2000). Speaker Verification Using Adapted Gaussian Mixture Models, *Digital Signal Processing.* Vol. 10, 19-41.

[11] Campbell, W. M., Sturim, D. E. & Reynolds, D. A. (2006). Support Vector Machines Using GMM Supervectors for Speaker Verification, *IEEE Signal Processing Letters,* Vol. 13, No. 5, 308-311.

[12] Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J. & Platt, J. C. (1999). Support Vector Method for Novelty Detection, *Advances in Neural Information Processing Systems.* Vol. 12, 582-588.

[13] Kim, M. J., Yang, I. H., So, B. M., Kim, M. S. & Yu, H. J. (2012). Histogram Equalization Using Background Speakers' Utterances for Speaker Identification, *Journal of the Korean society of speech sciences.* Vol. 4, No. 2, 79-86.
(김명재, 양일호, 소병민, 김민석, 유하진. (2012). 화자 식별에서의 배경화자데이터를 이용한 히스토그램 등화 기법. 말소리와 음성과학, 4권 2호, 79-86.)

[14] FFmpeg, http://www.ffmpeg.org/index.html.

[15] LAME MP3 Encoder, http://lame.sourceforge.net/.

[16] Opus Codec, http://www.opus-codec.org/downloads/.

[17] Young, S., Evermann, G., Gales, M.H.T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. & Woodland, P. (2009). The HTK Book (for HTK Version 3.4), *Cambridge University Engineering Department.*

[18] Chang, C. & Lin, C. (2014). LIBSVM -- A Library for Support Vector Machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

● **양일호 (Yang, IL-Ho)**
서울시립대학교 컴퓨터과학부
서울시 동대문구 전농동 90번지
Tel: 02-6490-5697 Fax: 02-6490-2444
Email: heisco@hanmail.net
관심분야: 화자 인식, 음성 인식, 법음성학
현재 컴퓨터과학과 대학원 박사과정 재학 중

● **김경화 (Kim, Kyung-Wha)**
대검찰청 과학수사담당관실
서울시 서초구 반포대로 157
Tel: 02-3480-2150 Fax: 02-3480-2707
Email: savoix@spo.go.kr
관심분야: 법음성학, 화자 식별
현재 대검찰청 과학수사담당관실 음성분석실장

● **김명재 (Kim, Myung-Jae)**
서울시립대학교 컴퓨터과학부
서울시 동대문구 전농동 90번지
Tel: 02-6490-5697 Fax: 02-6490-2444
Email: mj@uos.ac.kr
관심분야: 화자 인식, 음성 인식
현재 컴퓨터과학과 대학원 박사과정 재학 중

• **백록선 (Baek, Rock-Seon)**
서울시립대학교 컴퓨터과학부
서울시 동대문구 전농동 90번지
Tel: 02-6490-5697 Fax: 02-6490-2444
Email: whites86@naver.com
관심분야: 화자 인식, 음성 인식
현재 컴퓨터과학과 대학원 석사과정 재학 중

• **허희수 (Heo, Hee-Soo)**
서울시립대학교 컴퓨터과학부
서울시 동대문구 전농동 90번지
Tel: 02-6490-5697 Fax: 02-6490-2444
Email: zhasgone@naver.com
관심분야: 화자 인식, 음성 인식, 법음성학
현재 컴퓨터과학과 대학원 석사과정 재학 중

• **유하진 (Yu, Ha-Jin), 교신저자**
서울시립대학교 컴퓨터과학부
서울시 동대문구 전농동 90번지
Tel: 02-6490-2448 Fax: 02-6490-2444
Email: hjyu@uos.ac.kr
관심분야: 화자 인식, 음성 인식
현재 컴퓨터과학부 교수