pISSN 2005-8063
eISSN 2586-5854
2025. 9. 30.
Vol.17 No.3
pp. 15-22

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2025.17.3.015

Detection of AI-generated speech using acoustic parameters*

Yeonsoo Kim · Cheoljae Seong**

Department of Linguistics, Chungnam National University, Daejeon, Korea

Abstract

Voice cloning, also known as deep voice synthesis, is an artificial intelligence (AI) technology that extracts a speaker's vocal characteristics to replicate and generate speech. Until recently, cloned speech often sounded unnatural to human listeners. However, with the rapid advancement of AI technology, it is now possible to replicate voices using only a short audio sample, leading to real-world misuse, such as voice-phishing scams. This study aimed to identify significant acoustic-phonetic variables that help distinguish between natural and cloned speech generated using voice cloning technology. A total of 20 acoustic variables were selected across the spectral, cepstral, and prosodic domains, and analyzed using a two-way ANOVA and logistic regression. Perceptual evaluations and mathematical distance measurements (Euclidean and Mahalanobis distances) were conducted to compare the similarities between natural and cloned utterances. For male speech, the tilt and speech rate were significant variables for classification, whereas for female speech, the mean speech intensity was significant, with classification accuracies of 97.5% and 95%, respectively. In the similarity comparison, for female speech, the utterance perceived by humans as the most similar to natural speech was also the one closest in distance.

Keywords: acoustic parameters, voice cloning model, Deep Voice, auditory perceptual evaluation

1. 서론

딥보이스(Deep Voice)란 딥페이크(Deep Fake)의 일종으로, 생성형 AI 모델을 활용하여 만들어낸 음성이다. 최근 AI 기술이 빠르게 발전함에 따라 인간의 삶에 편리함을 가져다주지만, 동시에 악용 사례 또한 발생하고 있다. 특히 딥보이스가 보이스피싱에 활용되기 시작하면서 범죄 피해 사례가 급증하고 있다 (Kim & Lee, 2022). 생성한 음성만으로도 보이스피싱이 가능하지만, '보이스 클로닝', 즉 '음성 복제' 기술로 특정 인물의 목소리를 복제하여 보이스피싱에 활용하는 것이다. 이는 대다수가 아는 공인의 목소리, 혹은 지인의 목소리를 사칭할 수 있기 때문에, 큰 피해로 이어질 수 있다. 이전에는 TTS(Text-to-Speech) 기술을 사용하여 인간의 목소리를 비슷하게 만들어내긴 했지만, 기계 특유의 음색이 섞여있어 활용도가 떨어졌다. 하지만 최

* This paper is a revised and expanded version of a portion of the first author's master's thesis, supplemented with content that was previously presented at the 2025 Spring Conference of the Korean Society of Speech Sciences.

** cjseong49@gmail.com, Corresponding author

Received 30 June 2025; Revised 8 August 2025; Accepted 8 September 2025

근 개발된 보이스클로닝 모델로 만들어낸 음성은 실제 인간의 발화와 상당히 유사하다. 알지 못하는 인물의 목소리를 활용한 생성 발화보다는 짐작 가능한 지인의 목소리를 활용한 보이스 피싱이 훨씬 더 큰 범죄 사례로 이어질 수 있기 때문에, 본 연구에서는 발화의 유사성을 비교하는 데에 중점을 두었다.

기술로 인한 피해 사례가 급증함에 따라, 반대로 이를 탐지하기 위한 연구도 늘어나고 있다. 하지만 대부분 탐지 시스템을 개발하기 위한 제안이거나, 너무 적은 수의 음향변수를 분석했다는 한계가 있다. Kim et al.(2023b)에서는 뉴럴 보코더를 사용하여 딥보이스 음성을 만들고 자연 발화와 딥보이스 음성을 분석했으며, Kim et al.(2023a)에서는 pitch mean, intensity, first formant, second formant 변수를 선정하여 자연 발화, 복제 발화, 제 3자의 발화 수치를 비교하였다. 연구를 통해 음향 변수 값을 추출하고 비교하는 작업을 했지만, 각 발화의 수치를 단순 비교했다는 아쉬움이 있다. Han et al.(2023), Kim & Lee(2022), Lee & Yun(2024)은 딥보이스를 판별하기 위한 시스템 개발을 제안하였다. 공학 분야에서는 목소리의 특징을 벡터화하기 위해 주로 MFCC(Mel-Frequency Cepstral Coefficients)와 Mel-spectrogram을 사용하고(Han et al., 2023), 이를 CNN과 BiLSTM 같은 머신러닝 모델에 넣어 훈련시킨다(Han et al., 2023; Lee & Yun, 2024). 소개한 대개의 공학적 접근 방식은 복제 음성을 제작 및 탐지하는 모델 개발과 그 성능을 높이는 데에 목적을 두고 있다. 본 연구에서는 이러한 순수 공학적 접근과 달리 복제 음성 판별에 유의한 영향을 미치는 음향 변수를 음향음성학적 관점에서 찾아내고, 변수들의 통계적 가치를 머신러닝적 관점에서 밝힌 다음 보이스클로닝을 악용한 피해사례를 방지하는 데에 활용하는 방안을 찾는 것을 목표로 한다.

2. 이론적 배경

2.1. 보이스 클로닝

보이스클로닝, 즉 음성 복제 기술이란, 특정 화자의 음성 샘플을 학습하여 복제된 음성 발화를 만들어내는 기술이다. 이전의 자연어처리 방식이 대부분 거대한 양의 데이터셋을 활용하여 사전훈련(pre-training)을 시키는 방식이었다면, 최근에는 적은 양의 데이터로도 학습과 추론이 가능한 모델이 주목을 받고 있다. 음성의 경우 인간의 발화를 직접 녹음해야하기 때문에 전처리와 학습에 많은 어려움이 있었으나, 최근 즉각적인 음성 복제 기술(Instant Voice Cloning, IVC)이 등장하여 짧은 길이의 오디오 샘플만으로도 목소리를 복제할 수 있게 되었다. 1분 내외 정도의 특정 화자 목소리 샘플을 모델에 넣고 임의의 텍스트를 입력하면, 입력한 목소리가 텍스트를 읽는 output을 얻을 수 있는 것이다. IVC 기술은 자동 회귀 방식과 비자동 회귀 방식으로 나뉘는데, 자동회귀 방식은 이전 시점의 데이터 또는 이전 출력을 사용하여 다음 시점의 데이터를 예측하는 방식으로 속도가 느리고, 비자동 회귀 방식은 전체 문장을 한번에 처리하기 때문에 세부 조절이 어렵지만 속도가 훨씬 빠르다. 본 연구에서는 대표적인 비자동 회귀 방식 모델인 VITS(Kim et al., 2021)와 YourTTS(Casanova et al., 2022) 중 VITS 모델을 베이스로 삼은 음성 복제 모델을 선정하였다. 사용한 모델은 Myshell사에서 공개한 OpenVoice(Qin et al., 2023) 모델로, 사전학습 시킨 데이터 안에 특정 언어가 없더라도 추론이 가능한 zero-shot learning 기법(Xian et al., 2018)을 사용하기 때문에 속도가 빠르고 다국어 화자 간 복제가 가능하다. 또한, 모델 구조 안에 음성학적 개념인 IPA(International Phonetic Alphabet) 특성을 반영하였으며, 자유롭게 이용 가능한 공개 모델이라는 점에서 본 연구목적에 적합하다고 판단하였다.

2.2. MFCC(Mel-Frequency Cepstral Coefficients)

MFCC 변수는 음성인식, 화자 식별 등 공학적 음성처리에서 많이 쓰이는데, 주파수의 특성을 압축하여 벡터값으로 표현한다. 여러가지 특성을 압축해서 나타내기 때문에 음향음성학적 변수와는 차이가 있지만, 벡터값으로 추출되어 계산이 빠르고 모델에 입력하여 활용할 수 있다는 장점이 있다.

MFCC 값을 추출하는 과정은 다음과 같다. 우선 음성 신호를 프레임별로 나누어 푸리에 변환(Fourier Transform)을 적용하여 스펙트럼을 구한다. 스펙트럼에 Mel Filter Bank를 적용해 멜 스펙트럼을 구한 뒤, 멜 스펙트럼에 켑스트럴 분석(로그 계산, 역푸리에 변환)을 적용하면 MFCC 값이 추출된다.

2.3. 자연발화와 복제발화의 유사도 비교

유클리드 거리는 두 벡터값 사이의 거리를 측정하는 방식이다. 연구에서는 자연 발화와 복제 발화의 유사도 비교를 위해 유클리드 거리를 우선적으로 계산하였으나, 해당 거리식은 MFCC 값의 차원 특성을 반영하지 않는다는 아쉬움이 있었다. 이에 따라 차원 간 상관관계를 반영할 수 있는 마할라노비스 거리식을 사용하여 발화 간 유사도를 구해보았다(Lee, 2006). MFCC는 일반적으로 13개의 차원을 갖는 변수로, 음성의 특징을 벡터값으로 압축해 놓은 변수다. 따라서 마할라노비스 거리가 발화의 특성을 고려하여 좀 더 정확한 유사도 거리를 계산할 수 있다고 판단하였다. 유클리드 거리와 마할라노비스 거리 계산은 Python의 scipy 라이브러리를 활용하였으며, 아래는 두 거리의 공식이다. 수식 (1)은 유클리드 거리, 수식 (2)는 마할라노비스 거리를 나타낸다. 수식 (1)은 x, y 좌표로 표시되는 2차원에서 두 점 사이의 거리를 나타내며 수식 (2)에서 x, y는 n차원의 벡터를, ^{T}는 전치 행렬(transpose matrix)을, A^{-1}은 공분산 행렬 A의 역행렬을 의미한다.

$$d = \sqrt{(x_2 - x_1^2) + (y_2 - y_1^2)} \quad (1)$$

$$D(x,y) = \sqrt{(x-y)^T A^{-1}(x-y)} \quad (2)$$

2.4. 음향 변수

스펙트럼, 켑스트럼, 운율 변수를 포함하여 총 20개의 음향음성학적 관점에서의 음향 변수를 사용하였다. 스펙트럼 변수는

주파수에 따른 에너지 분포를 기술통계적 관점에서 보여주는 적률변수 4가지[COG(center of gravity), SD(standard deviation), skewness, kurtosis]와 기울기 변수(slope, tilt), 그리고 4,000 Hz를 경계로 하여 8,000 Hz 임계값 내에서의 두 영역 에너지 비율값인 LH(low to high) ratio로 구성하였다. 켑스트럼 변수는 cepstral peak, quefrency, CPP(Cepstral Peak Prominence), cepstral slope, cepstral intercept, RNR(Rhamonics to Noise Ratio) 6가지로, 주로 음성의 명료도와 주기성을 분석하는 데에 사용된다. 켑스트럼 변수의 경우, 복제 발화가 자연 발화를 기반으로 생성되긴 했지만, 명료도와 주기성 부분에서 자연발화와 다른 특성을 보일 것으로 추측하고 선정하였다. 운율 변수로는 기본주파수(F0)의 평균, 표준편차, 기본주파수 범위(range), 발화 강도(intensity)의 평균, 표준편차, 강도 범위, 그리고 발화 속도 변수 값을 추출하여 살펴보았다(표 1).

표 1. 20개 음향 변수의 의미와 단위
Table 1. Meanings and units of the 20 selected acoustic variables

	음향변수	의미	단위
Spectrum	COG	스펙트럼 에너지의 평균에 해당하는 주파수	Hz
	SD	에너지의 분산 정도	Hz
	skewness	에너지의 치우침 정도	dB ratio
	kurtosis	에너지의 뾰족한 정도	dB ratio
	slope	밴드대역 에너지차 기울기	dB
	tilt	추세 회귀 기울기	dB/Hz
	LH ratio	저주파영역에너지/고주파영역에너지	Pascal ratio
Cepstrum	cepstral peak	켑스트럼의 첫번째 피크 에너지값	dB
	quefrency	켑스트럼의 시간축	second
	CPP	켑스트럼 피크값과 선형 회귀선 상응값의 dB 차이	dB
	cepstral slope	켑스트럼 기울기	dB/s
	cepstral intercept	기울기선의 에너지 축 절편값	dB
	RNR	켑스트럼 내 조화성분과 노이즈 성분의 비율	dB
Prosody	mean F0	기본주파수(F0)의 평균값	Hz
	F0_*SD*	기본주파수(F0)의 표준편차	Hz
	F0_range	기본주파수(F0)의 범위	Hz
	mean_ intensity	강도(intensity)의 평균값	dB
	intensity_ *SD*	강도(intensity)의 표준편차	dB
	intensity_ range	강도(intensity)의 범위	dB
	speech rate	발화속도 (음절수/발화시간)	음절수/초

COG, center of gravity; CPP, Cepstral Peak Prominence; LH, low to high; RNR, Rhamonics to Noise Ratio.

3. 연구 방법

3.1. 연구 대상

본 연구는 보이스피싱과 관련되기 때문에 AI hub에 공개된 음성 데이터 중에서 '복지분야 콜센터 상담데이터'를 선정하여 사용하였다. 해당 데이터는 대학병원, 광역이동 지원센터, 정신건강 상담센터 3개의 복지시설에서 상담원과 고객이 통화한 내용을 바탕으로 연기자가 녹음한 자료다. 연구자는 발화가 명료하고 자연스러운 남자 40명, 여자 40명의 발화를 선정하였다. 청지각 평가는 AI 기술에 관심을 갖는 평균 29세의 남자 3, 여자 4명, 총 7명의 평가자를 모집하여 실시하였다.

3.2. 연구 절차

선정한 음성 데이터는 모델이 음성을 학습하기에 최적의 길이인 1분 내외로 다듬은 후에 복제 모델에 넣어 복제 시료로 제작하였다. 남녀의 자연 발화 80개와 복제 발화 80개를 대상으로 Praat 프로그램(ver. 6.3.20)을 사용하여 20개의 음향변수 값을 스크립트를 이용하여 추출하였으며(Seong, 2022) 추출한 음향변수 중 분산분석을 이용하여 유의변수 11개를 골라내었다. 이들 중 통계적 유의성 관점에서의 상위 변수 5개를 선택한 뒤 분류 통계 기법을 사용하여 복제 발화를 구분하는 유의 변수를 찾아내었다.

또한, 자연 발화와 복제 발화의 MFCC 값을 구해 유클리드 거리와 마할라노비스 거리를 구하였다. 청지각 평가는 같은 내용의 문장을 발화한 자연 발화와 복제 발화 음성을 연이어서 청취한 뒤, 우선 두 발화가 동일인이라고 생각하는지 예/아니오 응답을 선택하고(2AFC) 그 유사도 정도를 6점 척도로 평가하도록 하였다(goodness rating). 일반적 Likert scale이 5점과 7점으로 구성되는데 반해 6점을 선택한 이유는, 평가가 중간 점수에 과다하게 집중되어 올바른 인지 판단 결과를 도출하는데 방해가 될 수 있기 때문이다. 동일한 음성 데이터의 MFCC 값을 추출하여 수학적 거리 계산법(유클리드, 마할라노비스 거리)으로 수치적 유사도를 구한 뒤, 앞서 진행한 청지각 평가 결과와 비교하였다.

3.3. 통계

통계 분석은 SPSS(version 26, IBM) 프로그램을 사용하였다. 20개의 음향 변수를 종속 변수로, 성별과 복제 여부를 독립 변수로 두고 2원 분산 분석(2 way ANOVA)을 시행하여 복제 여부에 따른 발화의 특성을 알아보고자 하였다. 2원 분산 분석 결과 독립변수 복제 여부가 유의한 주효과(main effect)를 나타내는 음향 변수를 찾아낸 뒤, F 값이 크고 p 값이 작은 상위 변수를 골라내어 logistic regression을 진행하였다. 해당 통계는 복제 발화를 판별하는 유의한 음향 변수를 찾기 위해 진행하였다.

청지각 평가의 경우 감마와 카파 계수로 평가자 내 일치도를 측정하고, ICC(Intra class Correlation coefficient)로 평가자 간 신뢰도를 측정하였다.

4. 연구 결과

4.1. 유의한 판별 변수

전체 변수에 대한 기술통계 값은 다음과 같다(표 2).

표 2. 복제여부에 따른 각 변수의 기술통계
Table 2. Descriptive statistics of each variable by cloning status

변수	복제 여부	Mean	*SD*	N
COG	natural	2,408.9316	979.56411	80
	cloned	2,720.5487	736.14298	80
SD	natural	1,872.4725	486.07016	80
	cloned	2,161.9090	338.56224	80
skewness	natural	1.01410	1.042523	80
	cloned	.60575	.633926	80
kurtosis	natural	1.30879	3.597587	80
	cloned	−.54638	1.168271	80
slope	natural	−33.4101	6.42324	80
	cloned	−35.5087	4.36284	80
tilt	natural	−5.2729	1.52290	80
	cloned	−6.2681	.99343	80
LH ratio	natural	233,060.4	608,095.2	80
	cloned	134,900.5	155,220.6	80
cepstral peak	natural	62.6990	2.31663	80
	cloned	61.8211	1.72592	80
quefrency	natural	.00425	.001488	80
	cloned	.00471	.001950	80
CPP	natural	9.2416	1.19023	80
	cloned	8.7314	1.43427	80
cepstral slope	natural	−5.9666	.82646	80
	cloned	−6.0684	.66165	80
cepstral intercept	natural	22.3360	3.95619	80
	cloned	21.8319	2.86375	80
RNR	natural	.2303	.17375	80
	cloned	.3011	.13209	80
mean F0	natural	173.0993	53.70867	80
	cloned	173.1891	53.62889	80
F0 *SD*	natural	37.5691	15.55155	80
	cloned	34.5901	13.92515	80
F0 range	natural	221.7015	125.08452	80
	cloned	213.1456	133.68622	80
mean intensity	natural	69.2457	3.78478	80
	cloned	75.2713	.90006	80
intensity *SD*	natural	19.8545	15.76732	80
	cloned	21.3686	1.84847	80
intensity range	natural	67.1966	55.40319	80
	cloned	63.3636	4.03835	80
speech rate	natural	3.638	.6439	80
	cloned	4.166	.4119	80

COG, center of gravity; CPP, Cepstral Peak Prominence; LH, low to high; RNR, Rhamonics to Noise Ratio.

위 변수들로 2원 분산 분석을 진행한 결과, 복제 여부에서 유의한 주효과가 관찰된 음향 변수는 표 3과 같다.

표 3. 복제 여부에서 유의한 주효과가 관찰된 음향 변수
Table 3. Acoustic variables that showed significant main effects for cloning status

변수	*df*	F	Sig.
Cepstral peak	1	8.080**	.005
COG	1	5.335*	.022
CPP	1	6.195*	.014
Kurtosis	1	19.391***	<.0001
Mean intensity	1	194.747***	<.0001
RNR	1	8.487**	.004
SD	1	21.442***	<.0001
Skewness	1	9.093**	.003
Slope	1	5.783*	.017
Speech rate	1	39.098***	<.0001
Tilt	1	32.237***	<.0001

*p<.05, **p<.01, ***p<.001.
COG, center of gravity; CPP, Cepstral Peak Prominence; RNR, Rhamonics to Noise Ratio.

COG, SD, RNR, mean intensity, speech rate 변수의 값은 자연 발화보다 복제 발화가 높았으며, 그외 변수인 skewness, kurtosis, slope, tilt, cepstral peak, CPP 변수의 값은 자연발화에서 더 높게 나타났다. 표 3의 결과 중, F값이 크고 p 값이 작은 상위 5개의 변수를 골라내고 Enter 방식으로 logistic regression을 진행하였다. 남성의 경우 SD, kurtosis, tilt, mean intensity, speech rate 변수를 입력하였으며, 표 4와 같이 높은 분류 정확도를 보이는 변수 모형을 구할 수 있었다. 표의 결과를 살펴보면 자연발화의 경우 100%의 분류 정확도를 보이며, 복제 발화보다 자연 발화의 분류 정확도가 조금 더 높은 것을 확인할 수 있다.

표 4. 남성 발화의 복제 여부에 대한 분류 정확도
Table 4. Classification accuray for cloning status in male speech

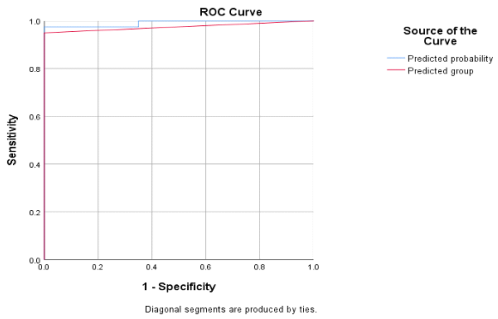
		Predicted		Percentage correct
		Natural	Cloned	
Observed	Natural	40	0	100.0
	Cloned	2	38	95.0
Overall Percentage				97.5

SD, tilt, speech rate 변수 모두 복제 여부를 구분하는 유의 변수로 판명되었지만(p<.05), 오즈비[Exp(B)]와 신뢰구간(CI)도 고려하여 판단해야 한다. 오즈비는 해당 독립변수의 값이 1단위 증가하면 복제 발화일 확률이 자연 발화일 확률보다 Exp(B)만큼 증가함을 의미하며, 신뢰구간(CI)에는 1이 포함되지 않아야 한다. 또한, Wald 통계값까지 유의한 변수를 감안하면 최종적으로 tilt와 speech rate 변수가 남성의 복제 발화를 구분하는 유의 변수로 가려진다. tilt의 경우 오즈비가 0.03으로, 1단위 증가함에 따라 집단 2, 즉 복제 발화로 분류되는 확률이 3% 증가(97% 감소)한다고 해석할 수 있다. 마찬가지로 speech rate의 경우 오즈비가 27.23으로, 1단위 증가에 따라 복제 발화로 분류되는 확률이 27.23배 정도 높아진다고 해석할 수 있다(표 5).

표 5. 남성 발화의 로지스틱 회귀분석 결과
Table 5. Logistic regression results for male speech

	B	Sig.	Exp(B)	95% C.I. for EXP(B)	
				Lower	Upper
SD	0.012	0.042	1.012	1.000	1.024
Kurtosis	−0.026	0.972	0.974	0.222	4.273
Tilt	−3.441	0.041	0.032	0.001	0.869
Mean intensity	1.290	0.063	3.632	0.933	14.142
Speech rate	3.304	0.023	27.233	1.592	465.753
Constant	−151.081	0.012	0.000		

그림 2는 위에서 구한 분류 모델의 ROC(Receiver Operating Characteristic) 곡선이다. ROC 곡선은 왼쪽 상단 모서리에 가까울수록 높은 TPR(true positive ratio, 양성을 양성으로 예측)과 낮은 FPR(false positive ratio, 양성으로 잘못 예측)을 의미하며, 이는 모델의 성능이 우수함을 나타낸다. 또한 AUC(Area Under the Curve)는 ROC 곡선 아래의 면적을 의미하는데, 모델의 전반적인 분류 능력을 숫자로 요약한 지표다. 0에서 1의 값을 가지며, 1에 가까울수록 우수한 성능임을 나타낸다(Fawcett, 2006). 따라서, 그림 1의 ROC Curve와 표 6의 AUC 결과를 보면 남성 발화에 대한 회귀 모델은 상당히 높은 성능의 분류 모델임을 확인할 수 있다.



ROC, Receiver Operating Characteristic.

그림 1. 남성 발화에 대한 ROC curve
Figure 1. ROC curve for male speech

표 6. 남성 발화에 대한 회귀 모델의 분류 능력 지표(AUC)
Table 6. Classification performance metric of the regression model

Test result variable(s)	Area
Predicted probability	0.991
Predicted group	0.975

AUC, Area Under the Curve.

여성의 경우 tilt, mean intensity, speech rate 변수를 넣어 입력하였으며, 표 7과 같이 높은 분류 정확도를 보이는 변수모형을 구할 수 있었다. 표의 결과를 살펴보면 복제 발화의 경우 97.5%의 분류 정확도를 보이며, 자연 발화보다 복제 발화의 분류 정확도가 더 높은 것을 확인할 수 있다.

표 7. 여성 발화의 복제 여부에 대한 분류 정확도
Table 7. Classification accuray for cloning status in female speech

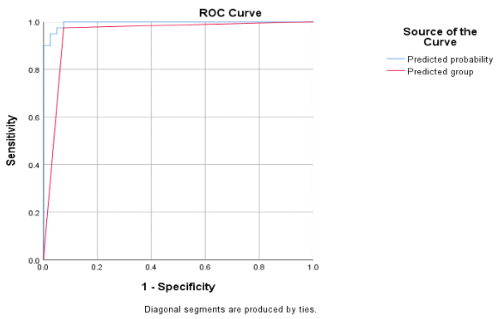
		Predicted		Percentage Correct
		Natural	Cloned	
Observed	Natural	37	3	92.5
	Cloned	1	39	97.5
Overall Percentage				95.0

Tilt, mean intensity, speech rate 변수 모두 복제 여부를 구분하는 유의 변수로 나타났지만(p<.05), 남성의 경우와 마찬가지로 오즈비[Exp(B)]와 신뢰구간(CI)도 고려하여 판단해야 한다. 역시 Wald 통계값까지 유의한 변수를 감안하면 최종적으로 mean intensity 변수가 여성의 복제 발화를 구분하는 유의 변수로 결정되었다. mean intensity 변수의 경우 오즈비가 8.95로, 1단위 증가함에 따라 집단 2, 즉 복제 발화로 분류되는 확률이 8.95배 정도 증가한다고 해석할 수 있다(표 8).

표 8. 여성 발화의 로지스틱 회귀분석 결과
Table 8. Logistic regression results for female speech

	B	Wald	Sig.	Exp(B)	95% C.I. for EXP(B)	
					Lower	Upper
Tilt	−1.511	2.280	0.131	0.221	0.031	1.569
Mean iten-sity	2.191	7.086	0.008	8.945	1.782	44.901
Speech rate	−1.024	0.252	0.616	0.359	0.007	19.609
Constant	−165.859	7.876	0.005	0.000		

다음은 여성 발화 분류 모델의 ROC 곡선이다. 그림 2의 ROC Curve와 표 9의 AUC 결과를 보면 여성 발화에 대한 회귀 모델 역시 상당히 높은 성능의 분류 모델임을 확인할 수 있다.



ROC, Receiver Operating Characteristic.

그림 2. 여성 발화에 대한 ROC Curve
Figure 2. ROC Curve for female speech

표 9. 여성 발화에 대한 회귀 모델의 분류 능력 지표(AUC)
Table 9. Classification performance metric of the regression model

Test result variable(s)	Area
Predicted probability	0.996
Predicted group	0.950

AUC, Area Under the Curve.

4.2. 유사도 계산

자연 발화와 복제 발화에 있어 인간의 지각, 인지적 능력과 수학적 계산법의 비교를 위해 청지각 평가와 MFCC 값 추출 및 거리 계산을 진행하였다. 유클리드 거리와 마할라노비스 거리는 숫자가 작을수록 두 발화가 유사한 것이고, 청지각 점수는 숫자가 작을수록 두 발화가 유사하지 않음을 나타낸다. 가장 유사하게 나타난 문항을 확인해보면, 남성 발화의 경우 유클리드 거리상 가장 가까운 문항은 8번, 마할라노비스 거리상 가장 가까운 문항은 3번, 지각평가 상 가장 유사한 문항은 5번이었다(표 10). 여성의 경우 유클리드와 마할라노비스 거리 모두 가장 가까운 문항은 2번이었으며, 유사도 점수상 가장 유사한 문항은 2번과 8번이었다. 동점인 문항이 있긴 했지만, 여성 발화의 경우에는 벡터 값을 활용하여 구한 계산값과 사람이 인지한 발화의 유사도 정도가 상당히 일치함을 확인할 수 있다.

평가자 내 신뢰도 검정 결과 감마와 카파 계수 모두 유의하지 않았는데, 유사도에 대한 평가자의 주관에 일관성이 없었다고 해석할 수 있다. 평가자 간 신뢰도 검정 결과(ICC), Cronbach의 알파 값 중 평균측도가 0.572(p<.01)로 유의하게 나타났지만 낮은 값이었다. 유사도 평가가 상대적으로 어려운 작업임을 알 수 있다.

표 10. MFCC 값의 거리 값과 청지각 유사도 점수 비교
Table 10. Comparison between MFCC-based distance values and perceptual similarity scores

성별	문항	Euclidean	Mahalanobis	유사도 점수
남성	1	5,369.99	0.8030	3
	2	4,076.66	0.8133	2.71
	3	5,436.58	0.7805	2.57
	4	5,320.75	0.8091	2.71
	5	3,692.08	0.7973	3.43
	6	4,283.62	0.8148	1.43
	7	5,168.52	0.8085	2
	8	3,587.12	0.8144	2.29
	9	4,306.47	0.8108	2.71
	10	3,844.83	0.8078	1.71
여성	1	5,747.18	0.8396	3
	2	2,655.81	0.7968	3.57
	3	3,783.72	0.8031	3.43
	4	4,685.94	0.8137	1.29
	5	4,628.98	0.8116	1.43
	6	4,094.72	0.8172	2.14
	7	4,964.40	0.8001	1.57
	8	3,689.93	0.8152	3.57
	9	3,831.09	0.8109	2
	10	2,831.85	0.8139	2.86

MFCC, Mel-Frequency Cepstral Coefficients.

5. 논의 및 결론

음향 변수를 활용하여 딥페이크 음성을 판별하기 위해 복제 발화 시료를 제작하여 20개의 음향음성학적 변수를 선정하고 통계적 분석과 청지각적 평가를 진행하였다. 자연 발화의 경우 AI Hub의 콜센터 상담전화 데이터를 선정하여 남녀 각 40개의 음성 발화를 수집하였으며, 이와 대조하기 위한 복제 발화 시료의 경우 2024년에 공개된 OpenVoice 보이스클로닝 모델을 사용하여 제작하였다. 연구에서는 복제 발화를 판별하는 유의한 음향변수와 발화 간 유사도(지각적, 수학적) 두 가지를 중점적으로 살펴보았다. 각 결과에 대한 해석은 다음과 같다.

첫째, 2원 분산 분석을 통해 복제 여부에 유의한 영향을 주는 11개의 음향 변수를 골라내고, 그중 상위 5개의 변수(*SD*, kurtosis, tilt, mean intensity, speech rate)를 넣어 logistic regression을 진행하였다. 2원 분산 분석 결과 복제 발화의 COG와 SD 변수의 값이 자연 발화보다 높다는 것은 복제 발화가 고주파에 더욱 집중되어 있음을 나타낸다. skewness, slope, kurtosis, tilt 값은 자연 발화가 더 큰데, 이는 자연 발화가 대체로 일상 언어의 주요 사용 영역인 4,000 Hz 아래쪽 저주파 대역에 더욱 집중되어 있고(skewness, slope) 특정 에너지 대역에 집중되어 있음(kurtosis)을 의미한다. 켑스트럼 변수인 RNR 값은 복제 발화가 더 크고, cepstral peak와 CPP 값은 자연 발화가 더 큰 것으로 보아 복제 발화의 하모닉 성분이 노이즈에 비해 더 많지만, 음질 관점에서는 자연 발화가 더 명료함을 알 수 있다.

종합해보면, 복제 발화는 자연 발화에 비해 에너지가 고주파에 집중되어 있고 하모닉 성분이 상대적으로 노이즈보다 많으며, 발화속도가 더 빠르고 더 강한 에너지로 발화되었다고 정리할 수 있다. 발화의 강도를 나타내는 mean intensity 변수는 복제 발화가 자연 발화보다 높은 값으로 나타났는데, 이는 복제 발화의 intensity 값이 자연 발화보다 낮다는 Kim et al.(2023b)의 연구와 상반되는 결과로, 사용한 모델과 모델 내 파라미터 설정 차이를 원인으로 볼 수 있다. logistic regression 결과 남성 발화자의 경우 스펙트럼의 에너지 기울기 특성을 나타내는 tilt와 발화 속도인 speech rate, 여성 발화자의 경우 발화의 강도를 나타내는 mean intensity가 자연 발화와 복제 발화를 구분하는 유의한 음향변수로 판별되었다. 이는 2원 분산 분석 결과 복제 여부에 따라 유의한 영향을 받는 11개의 변수 중 3개의 변수 모두 중복되는 결과로, 이 3개의 변수는 자연 발화와 복제 발화를 구분하고자 할 때 유의미한 변수로 활용될 수 있을 것이다. 본 연구에서 활용한 발화 속도나 음성의 강도와 같은 운율 변수의 경우 모델에 따라 parameter setting을 조절해 볼 수 있겠지만, 스펙트럼이나 켑스트럼 변수의 경우 단순히 모델의 setting 값을 변경함으로써 변수 값의 차이를 좁히긴 어려울 것으로 보인다.

두번째로, 청지각 실험 결과 여성 발화의 경우 MFCC 값을 활용하여 발화 간 유사도를 구한 결과와 사람이 유사하다고 평가한 문항이 일치함을 확인할 수 있었다. 남성 발화의 경우에는 유클리드 거리상 두 번째로 유사한 문항이 평가자들이 가장 유사하다고 느낀 문항이었다. 이는 여성 발화에 대한 지각 평가가

더 정확했다고 해석할 수 있으며, 모델에 입력된 raw data의 품질 차이도 있을 것으로 보인다. 실제로 다수의 청지각 평가자들이 여성 발화를 더욱 자연스럽게 느꼈다는 피드백을 주었는데, 후속 연구에서는 발화의 품질을 더욱 엄격하게 걸러낸 후 모델에 넣는 과정이 필요할 수 있겠다. 연구 결과 MFCC 값이 발화 특성을 상당히 잘 추출한다고 할 수 있지만 MFCC는 본 연구에서 진행한 다수의 음향 음성학적 변수와 같은 분석적 변수가 아닌 통합적 변수라는 한계가 있기 때문에, 향후에는 MFCC의 장단점과 음향음성학적 변수의 특성을 상호 보완 및 개발할 필요가 있다.

본 연구는 크게 세 가지 한계점을 갖는다. 우선, 청지각 평가에 사용한 복제 발화의 속도가 부자연스럽다는 피드백이 있었다. 유사도 평가를 위해 자연 발화와 복제 발화의 발화속도를 거의 비슷하게 운율 변조 기법을 이용하여 조정하였지만 사람이 인지하기에는 부자연스러웠을 것으로 추측한다. 예를 들어, 한 문장을 발화할 때 특정 부분은 느릴 수도 혹은 빠를 수도 있는데, 운율 변조에 의해 전체를 일률적 비율로 조정한 것이므로 자연성의 관점에서 어색할 수 있기 때문이다. 또한, 사용한 데이터에 대한 아쉬움이 있다. 선정한 AI Hub의 데이터 중에서 최대한 잡음이 적고 자연스러운 음성을 골라내긴 했지만, 실험실의 방음 부스환경에서 녹음을 진행한 것이 아니기 때문에 한계가 있었다. 주제가 보이스피싱인 만큼 어느 정도의 잡음은 실제 전화 음성과 유사하게 들리는 요소가 될 수 있지만, 모델에 넣어 학습 데이터로 사용하는 데에는 단점으로 작용하였다. 마지막으로, 해당 연구에서 제작한 복제 발화 시료는 특정 보이스클로닝 모델을 사용한 것으로, 본 연구의 결과는 선정 모델에 한정된 결과임을 고려해야 한다(Kim, 2025).

음성 복제 기술의 발전 속도가 빠른 만큼 이를 활용한 범죄도 급증하고 있으며, 검찰청이나 국립과학수사연구원과 같은 국가 기관에서도 딥보이스를 탐지하기 위한 연구를 지속하고 있다. 복제 발화를 판별하는 데에 본 연구 결과가 활용될 수 있을 것으로 기대한다(Kim & Seong, 2025).

References

Casanova, E., Weber, J., Shulby, C. D., Junior, A. C., Gölge, E., & Ponti, M. A. (2022). Yourtts: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. *Proceedings of Machine Learning Research*, *162*, 2709-2720.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861-874.

Han, S., Han, S., You, S., Song, D., & Seo, C. (2023, July). Deep voice detection system based on voice feature extraction and deep learning. *Proceedings of the Summer Conference of the Korean Institute of Electrical Engineers* (pp. 2487-2488), Yongpyong, Korea.

Kim, J., Kong, J., & Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *Proceedings of Machine Learning Research, 139,* 5530-5540.

Kim, J., Ryu, S., Han, C., Lee, C., Lee, J., Lee, M., ...Park, S. (2023a, November). Voice frequency analysis for response to voice phishing using deep voice. *Proceedings of the 2023 Fall Conference of the Institute of Electronics and Information Engineers* (pp. 573-576), Siheung, Korea.

Kim, S., & Lee, S. (2022). Development of voice phishing damage prevention service misusing deep voice. *The Journal of Korean Institute of Communications and Information Sciences, 47*(10), 1677-1685.

Kim, S., Park, N., Park, E., Park, J., Song, D., Sim, H., Jo, I., ... Jo, D. (2023b, June). Identification of difference between deep voice and real voice for preventing voice phishing crime by impersonating an acquaintance. *Proceedings of the 2023 Summer Conference of the Korean Institute of Communications and Information Sciences,* (pp. 773-774), Jeju, Korea.

Kim, Y. (2025). *The role of acoustic parameters in deep voice detection* (Master's Thesis). Chungnam National University, Daejeon, Korea.

Kim, Y., & Seong, C. (2025, May). Detection of AI generated speech using acoustic parameters. *Proceedings of the 2025 Conference of the Korean Society of Speech Sciences* (p. 55).

Lee, C. Y. (2006). A study on the optimal mahalanobis distance for speech recognition. *Speech Sciences, 13*(4), 177-186.

Lee, C. H., & Yun, C. L. (2024, October). A study on machine learning-based deep voice analysis for crime prevention using generative AI. *Proceedings of KIIT Conference* (pp. 25-29), Jeju, Korea

Qin, Z., Zhao, W., Yu, X., & Sun, X. (2023). Openvoice: Versatile instant voice cloning. arXiv. https://doi.org/10.48550/arXiv.2312.01479.

Seong, C. J. (2022). Guidance to the Praat, a software for speech and acoustic analysis. *Journal of the Korean Society of Laryngology, Phoniatrics and Logopedics, 33*(2), 64-76.

Xian, Y., Lampert, C. H., Schiele, B., & Akata, Z. (2018). Zero-shot learning: A comprehensive evaluation of the good, the bad and the ugly. arXiv. https://doi.org/10.48550/arXiv.1707.00600

• **김연수 (Yeonsoo Kim)**
충남대학교 언어학과 석사
대전 유성구 대학로 99
Tel: 042-821-6391
Email: emilya60@naver.com
관심분야: 화자식별, 합성 및 복제음성 분석

• **성철재 (Cheoljae Seong)** 교신저자
충남대학교 언어학과 교수
대전 유성구 대학로 99
Tel: 042-821-6395
Email: cjseong49@gmail.com
관심분야: 분절음 및 운율 분석

음향 변수를 활용한 딥페이크 음성 판별*

김연수 · 성철재
충남대학교 언어학과

국문초록

보이스클로닝은 특정 화자의 목소리 특징을 추출하여 발화를 복제 및 생성, 즉 딥보이스(Deep Voice)를 생성하는 AI 기술이다. 이전에는 사람이 듣기에 어색함이 있었지만, 최근 AI 기술이 빠르게 발전함에 따라 짧은 시간의 오디오 샘플만으로도 음성 복제가 가능해지면서 이는 보이스피싱 피해 사례로 이어지고 있다. 이에 본 연구에서는 보이스클로닝을 활용해 생성된 복제 발화를 판별하는 데에 유의한 음향음성학적 변수를 찾아내고자 하였다. 스펙트럼, 켑스트럼, 운율 총 20개의 음향 변수를 선정하여 2원 분산 분석과 로지스틱 회귀분석을 진행하였으며, 자연 발화와 복제 발화에 대한 유사도를 비교하기 위해 청지각 평가와 발화 간 수학적 거리 계산(유클리드, 마할라노비스)을 진행하였다. 남성 발화의 경우 스펙트럼의 추세 회귀 기울기(tilt)와 발화 속도(speech rate), 여성 발화의 경우 평균 발화 세기(mean intensity)가 발화 구분에 유의한 변수였으며, 분류 모델은 각각 97.5%, 95%의 정확도를 보였다. 발화 간 유사도 비교 결과, 여성 발화의 경우 벡터값의 거리상 가장 가까운 발화와 인간이 유사하다고 느낀 발화가 일치하였다.

핵심어: 음향 변수, 보이스클로닝, 딥보이스, 청지각 평가

참고문헌

김소운, 이성택(2022). 딥보이스를 악용한 보이스 피싱 피해방지 서비스개발. *한국통신학회논문지, 47,* 1677-1685.

김수민, 박나희, 박은빈, 박지수, 송도연, 심혜지, 조일영, 김경배, 정연호,이우용, 정연만, 조동욱(2023b). 지인 사칭 보이스 피싱 범죄 예방을 위한 딥보이스와 실제 목소리와의 차이 규명. *한국통신학회학술대회논문집,* 773-774.

김연수(2025). Deep voice 판별에서 음향 변수의 역할: 인지실험과 관련하여. *충남대학교 석사 학위논문.*

김연수, 성철재(2025). 음향 변수를 활용한 딥페이크 음성 판별. *한국음성학회 2025 년 봄 학술대회 발표논문집,* 55.

김지후, 류세민, 한채림, 이창의, 이종현, 이민재, 오민규, 박세진(2023a). 딥보이스를 이용한 보이스피싱의 대응방안을 위한 음성 주파수 분석. *대한전자공학회 학술대회,* 573-576.

이치훈, 윤철희(2024). 생성형 AI 를 활용한 범죄 예방을 위한 머신러닝 기반 딥보이스 판별에 관한 연구. *Proceedings of KIIT Conference,* 25-29.

성철재(2022). 음성 및 음향분석 프로그램 Praat 의 임상적 활용법. *대한후두음성언어의학회지*, *33*(2), 64-76.

한승우, 한성훈, 유성민, 송동호, 서창진(2023). 음성 특징 추출 및 딥러닝 기반 딥보이스 탐지 시스템. *대한전기학회 학술대회 논문집,* 2487-2488.

Bright_Dev(2019). 이미지(간략한 MFCC 추출과정). Retrieved from https://brightwon.tistory.com/11

* 이 논문은 제 1 저자의 2025 석사학위 내용과 2025년 한국음성학회 봄 학술대회 발표논문의 일부를 수정, 보완한 것입니다.