JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 27, No. 7, pp. 1883-1895, Jul. 2026

Check for updates

AI 기반 한국어 가창 음성 변환에서 가창자 음향 특성의 차원별 보존 및 전이 분석

정 다 훈[1] · 이 철 희[2*]
[1]경희대학교 포스트모던음악학과 석사과정 [2]경희대학교 포스트모던음악학과 조교수

Dimension-Level Analysis of Singer Acoustic Feature Preservation and Transfer in AI-Based Korean Singing Voice Conversion

Da-Hun Jeong[1] · Chul-Hee Lee[2*]

[1]Master's Course, Department of Postmpdern Music, Kyung Hee University, Yongin 17104, Korea
[2]Assistant Professor, Department of Postmpdern Music, Kyung Hee University, Yongin 17104, Korea

[요 약]

생성형 AI로 실존 인물의 음성을 무단 활용한 가창 생성이 확산되는 가운데, 본 연구는 한국어 가창 음원의 가창자 음향 특성을 7축 58차원으로 정량화하는 측정 체계를 구축하고, 85명, 총 2,803곡을 대상으로 가창자 간 변별력과 가창자 내 안정성을 확인하였다. 이 가운데 비브라토 축을 제외한 6개 축 56차원을 AI 음성 변환 결과물에 동일 적용한 결과, 음색 엔벨로프와 성도 공명은 전이되고 한국어 음운 조음은 보존되며, 피치 동역학은 원본 F0 궤적을 반영하는 변환 설정에서 보존되는 영역별 분화가 관찰되었다. 음운 조음 4차원에서는 받침 파열음만 혼합으로, 마찰음, 파찰음, 받침 유성음은 보존으로 분류되었다. 학습 양식이 대비되는 두 모델의 최다 판정 일치율은 83.9%로, 이 패턴이 모델 설계와 무관한 음향 차원의 속성임을 뒷받침한다. 가창자 간 변별력은 전이 방향성과 양의 상관을 보이나 단독 결정 요인은 아니며, 음소 경계 안정성은 변환 품질의 독립 지표로 집계된다. 본 측정 체계는 한국어 가창 음원의 가창자 변별 음향 특성 분석과 변환 품질 평가의 측정 토대로 기능할 수 있다.

[Abstract]

Advances in generative AI have made it possible to reproduce real vocal identities in synthesized singing without authorisation. This study constructed a seven-axis 58-dimension framework for quantifying the acoustic features of Korean singing, and validated it on 85 singers and 2,803 songs. When six of the axes and 56 of the dimensions, excluding vibrato, were applied to voice conversion, the timbral envelope and vocal tract resonance were transferred, whereas Korean phoneme articulation was preserved and pitch dynamics remained close to the source under an F0-preserving conversion setting. Of the 4 phoneme articulation dimensions, fricatives, affricates, and voiced codas were preserved in both models, while plosive codas gave mixed results. Agreement of 83.9% between two models built on opposing paradigms indicates that this pattern is a model-independent acoustic property. Singer discriminability correlates with the direction of transfer but does not determine it on its own; and phoneme boundary stability serves as an independent indicator of conversion quality. The framework offers a measurement foundation for singer-identifying analysis and the evaluation of conversion quality.

색인어 : AI 음성 변환, 가창자 식별, 한국어 가창 음운, 가창 변환 평가, 딥페이크 탐지

Keyword : AI Voice Conversion, Singer Identification, Korean Vocal Phoneme, Singing Voice Conversion Evaluation, Deepfake Detection

http://dx.doi.org/10.9728/dcs.2026.27.7.1883

Received 27 May 2026; **Revised** 16 June 2026
Accepted 13 July 2026

***Corresponding Author; Chul-Hee Lee**

Tel: +82-31-201-2863
E-mail: ch@khu.ac.kr

Ⅰ. 서 론

AI 음성 변환(VC; voice conversion) 모델이 가창 영역으로 빠르게 확산되면서, 특정 가창자의 목소리를 모방한 2차 창작물이 원작자의 동의 없이 유통되는 사례가 늘어나고 있다[1]. 2026년 4월, 팝가수 Taylor Swift는 미국 특허상표청(USPTO; United States Patent and Trademark Office)에 자신의 가창 음성을 사운드 마크로 출원했다[2]. 이는 현재의 저작권법으로는 규율하기 어려운 실존 인물의 음성을 활용한 무분별한 AI 음성 생성 문제에서 본인의 음성을 상표권으로 보호하려는 시도로 볼 수 있다. 미국 판례에서는 일찍이 가창 음성의 모방을 인격권 침해로 인정한 Midler v. Ford 사건이 선례로 자리잡았으며[3], 한국에서도 음성을 인격권의 일부로 본 판결이 있다[4]. 또한, 한국저작권보호원은 AI 커버곡의 권리 귀속 문제를 정책 의제로 다루고 있다[1]. 학술 영역에서도 작곡 AI를 포함한 AI 음악의 저작권 쟁점이 논의되고 있다[5].

이러한 분쟁의 기술적 근거는 '어느 기준까지가 가창자의 고유한 특성인가'에 대한 정량적 기준이다. 그러나 한국어 가창 보컬에 한정된 객관적 고유성 지표 체계는 충분히 정립되어 있지 않은 실정이다. 가창 발성은 발화에 없는 가창 고유 성분을 포함하며, 가창의 고유성은 다차원이므로 단일 지표(MOS; mean opinion score)[6]로 환원하면 차원별 보존 및 전이를 분해할 수 없다. 또한, 기존의 객관 평가 지표는 스펙트럼 거리와 청취 점수에 집중되어 있으며[7],[8], 기존 평가 체계가 변환 품질의 핵심 축 중 하나인 음소 경계 안정성을 별도 측정 차원으로 포함하지 않는다[9].

본 연구는 이러한 문제를 해결하기 위해 한국어 가창 보컬의 다차원 고유성을 정량화하는 7축 58차원 측정 체계를 제안한다. 또한, 한국어 가창에 특화된 음운 조음 측정 차원을 새로 정의하고, 음소 경계 안정성을 독립 측정 차원으로 포함하여 기존 평가 체계의 한계점을 보완한다. 나아가, 이를 음성 변환 결과물에 동일 적용하여 차원별 보존 및 전이 패턴을 분석하고, 가창자 음향 특성의 차원별 보존 및 전이를 측정 가능한 형태로 제시하고자 한다.

Ⅱ. 이론적 배경 및 관련 연구

2-1 발화 음성의 음향 측정

보컬 개인성을 음향 측면에서 측정하는 접근은 전통적으로 개별 음향 지표의 분석에 기반해 왔다. 성도의 공명 특성을 반영하는 포먼트(F1, F2)[10], 후두의 진동 특성을 반영하는 기본 주파수(F0)와 변동성(jitter, shimmer)[11], 발성의 주기성을 반영하는 HNR(harmonics-to-noise ratio)[12] 등이 대표적이다. 이러한 개별 지표는 발성의 물리적 기제와 직접 대응되어 명확한 물리적 해석이 가능하다.

음색의 전반적 형태는 스펙트럼 통계로 측정된다. MFCC (Mel-Frequency Cepstral Coefficients)는 인간 청각의 주파수 지각 특성에 맞춘 멜 척도 (mel scale) 위에서 스펙트럼 엔벨로프를 압축적으로 표현하는 계수로[13], 음성 및 음악 영역에서 표준 음색 측정 도구로 사용된다. 스펙트럼 중심(spectral centroid), 롤오프(rolloff), 대역폭(bandwidth)은 스펙트럼 에너지 분포의 형태를 요약하는 통계량이다[14].

2-2 가창 발성의 음향 측정

가창 발성은 발화에 없는 특화 음향 성분을 포함한다. 비브라토(vibrato)는 음높이의 주기적 변조로, 변조의 빠르기를 나타내는 비브라토율(vibrato rate)과 변조의 크기를 나타내는 비브라토 폭(vibrato extent)으로 측정된다[15]. 피치 동역학 (pitch dynamics)은 F0의 평균, 표준편차, 범위로 가창의 음역과 음정 변화 폭을 표현한다[16]. 성악 공명은 가창 발성에서 2~4 kHz 대역의 에너지 집중으로 나타나는 음향 특성으로[17], SPR(Singer's Power Ratio)[18]로 정량화된다.

가창 변환 영역의 평가 표준으로는 SVCC(Singing Voice Conversion Challenge)가 사용되어 왔다[7],[8]. SVCC 2023[7]은 보컬 변환 모델의 객관 및 주관 평가 체계를 제시했으며, SVCC 2025[8]는 이를 다국어 및 다장르로 확장했다. SVCC 평가 체계는 발화 영역의 음성 식별 패러다임을 가창에 적용한 형태로, 변환 결과의 종합 평가에 중점을 둔다. 단, SVCC의 종합 평가는 변환 품질을 단일 지표로 요약하기 때문에, 어느 음향 차원이 원본 가창자의 특성으로 보존되고 어느 차원이 대상 가창자로 전이되는지를 차원별로 분해하지는 않는다. 본 연구는 이러한 차원별 보존 및 전이 분석을 목적으로, SVCC 평가 체계 대신 자체 7축 58차원 측정 체계를 적용한다.

2-3 한국어 가창 음운의 음향 측정

한국어의 음운 체계는 발성의 시간 구조를 결정하는 요소다. 자음은 조음 방식에 따라 장애음(obstruent)과 공명음(sonorant)으로 분류되며, 장애음은 다시 파열음, 마찰음, 파찰음으로 나뉜다. 종성 받침은 파열음(ㄱ, ㄷ, ㅂ), 비음(ㄴ, ㅁ, ㅇ), 유음(ㄹ)으로 분류되며, 음절 말 위치에서 음향적으로 뚜렷한 시간 패턴을 형성한다[19]. 이러한 자음과 받침의 분류 체계는 한국어 발성의 음향 분석에서 기본 단위로 작동한다.

한국어 가창의 자음, 모음, 받침 음성적 특징을 분석한 선행 연구[19]는 가창에서의 한국어 음운 조음이 발화와 다른 양상을 보인다고 보고하였다. 가창은 음악적 시간 구조에 음운을 정렬하는 과정이므로, 발화에서 명확히 구현되는 종성 받침의 파열 및 폐쇄가 가창에서는 약화되거나 생략되기도

한다. 또한 유성음에 해당하는 비음과 유음은 성대 진동의 영향으로 가창의 창법 및 음높이에 따라 음성적으로 변형될 수 있다.

한국어 음운 조음 측정은 음소 단위의 강제 정렬(forced alignment)을 전제로 한다. 다만, 선행 연구[19]는 가창의 한국어 음운 조음 양상을 정량화하는 음향 측정 변수를 제공하지 않으므로, 본 연구는 MFA(Montreal Forced Aligner)[20]의 한국어 사전학습 모델로 강제 정렬을 수행한 다음, 그 출력을 입력으로 4차원의 측정 변수를 자체 정의한다.

2-4 AI 음성 변환 모델

1) 음성 변환(VC) 개요

음성 변환은 원본(source) 화자의 음성을 대상(target) 화자의 음향 특성으로 변환하면서 언어 내용을 유지하는 작업이다[21]. 가창을 대상으로 하는 가창 음성 변환(SVC; singing voice conversion)은 여기에 음높이, 멜로디, 비브라토 등 가창 고유 성분의 처리가 더해진다는 점에서 일반 음성 변환과 구분된다[7]. 음성 변환은 학습 양식에 따라 화자 종속 학습(speaker-dependent training)으로 대상 화자의 음향 특성을 모델 파라미터에 학습시키는 방식과, 대상 학습 없이 짧은 참조 음성만으로 대상 특성을 추정하는 방식(zero-shot)으로 나뉜다. 본 연구는 이러한 두 방식의 대표 모델인 RVC v2와 SEED-VC를 비교 분석 대상으로 선정한다. 두 모델은 가창 전용으로 설계된 SVC 모델이 아니라 범용 음성 변환 모델이지만, 이를 가창에 적용한 결과물이 무단 복제와 저작권 분쟁 등 사회적 문제로 이어지고 있어[1] 분석 대상으로서 의의를 가진다. 다만, 두 모델이 가창 특화 모델이 아니라는 점은 본 결과 해석의 조건으로 함께 고려하며, 이에 따라 본 연구의 위치는 SVC 모델의 성능 평가가 아니라 범용 음성 변환 모델을 가창에 적용한 변환 결과물의 음향 특성 분석으로 한정된다.

2) RVC v2

RVC v2(Retrieval-based Voice Conversion v2)는 화자 종속 학습 기반의 검색 합성 모델로, 세 단계로 이루어진다. 첫째, 임베딩 단계에서는 자기지도학습(SSL; self-supervised learning)으로 사전 훈련된 ContentVec[22] 임베더(embedder)가 원본 음성에서 화자 정보를 분리한 언어 내용 표현을 추출한다. 둘째, 검색 단계에서는 대상 화자의 학습 음성에서 구축한 특징 색인으로부터 원본 임베딩에 가장 가까운 대상 특징을 검색한다. 셋째, 합성 단계에서는 VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech)[23] 기반 보코더(vocoder)가 검색된 특징을 시간 영역의 음성 파형으로 합성한다. RVC Project[24]는 이 세 단계를 통합한 오픈소스 음성 변환 프레임워크로, 화자 종속 학습 모델 계열의 대표적 구현으로 자리잡았다.

3) SEED-VC

SEED-VC[25]는 확산(diffusion) 기반의 zero-shot 음성 변환 모델이다. 확산 모델은 잡음을 점진적으로 제거하는 역방향 과정을 학습하여 데이터를 생성하는 생성 모델 계열로[26], 음성 합성 분야에서 자연스러운 음향 품질을 달성하며 표준 생성 모델로 자리잡았다. SEED-VC는 30초 길이의 대상 참조 음성으로부터 대상 화자의 음향 특성을 추정하고, 확산 과정을 통해 원본 음성을 대상 특성으로 변환한다. 모델 설계는 원본 피치 궤적의 보존을 지원하며, 가창 변환에서 멜로디 정보를 유지한다.

Ⅲ. 가창자 고유성 측정 체계와 분포 분석

3-1 가창자 고유성의 정의와 측정 차원 구성

본 연구에서 가창자 고유성은 가창자 간에는 변별되고 같은 가창자의 여러 곡에서 안정적으로 유지되는 음향 특성군으로 정의한다. 이러한 고유성은 발화 발성, 가창 발성, 한국어 음운의 세 음향 측면을 포괄하는 다층 구조로 구성되며, 본 연구는 이를 차원별 가창자 간 변별력과 가창자 내 안정성으로 정량화한다. 각 차원이 가창자 고유성 지표로 해석되기 위한 기준은 두 가지다. 같은 가창자의 여러 곡에서 곡 단위 변동 계수(CV; coefficient of variation)가 0.5 미만, 즉 곡 간 표준편차가 평균의 절반 미만으로 유지되어야 하고, 가창자 쌍 간 표준화 효과 크기(Cohen's d)의 중앙값이 효과 크기 분류에서 large에 해당하는 0.8을 초과해야 한다. 두 임계값은 본 연구가 표준 통계 해석에 근거하여 설정한 자체 기준이며, 차원별 충족 현황은 뒤의 차원별 분포 분석에서 전수 보고한다. 두 기준을 모두 충족하는 차원에 한하여 가창자 고유성 지표로 해석한다. 축별 차원 구성은 아래 표 1에 제시한다.

표 1. 7축 58차원 한국어 가창자 고유성 측정 체계
Table 1. The 7-axis 58-dimensional Korean singer identity measurement framework

Axis	Dimension	Count
1. Anatomical vocal tract	F1, F2	2
2. Vibrato	vibrato_rate, vibrato_extent	2
3. Voice quality	HNR, jitter, shimmer	3
4. Singer's formant	SPR	1
5. Pitch dynamics	f0_mean, f0_std, f0_range	3
6. Timbre envelope	MFCC1–20 mean, MFCC1–20 std, spectral centroid, rolloff, bandwidth	43
7. Korean phoneme articulation	fric, aff, plo, voi	4
Total		58

본 체계의 7개 축과 58개 세부 차원은 앞서 언급한 세 음향 측면의 음성학 및 음향학 표준 측정 변수에서 이론적으로 도출한 것이다. 가창에도 공통으로 적용되는 발화 발성의 해부학적 성도 특성, 발성질, 음색 엔벨로프를 축 1, 3, 6으로 구성하며, 가창 특화 성분인 비브라토, 성악 공명, 피치 동역학으로 축 2, 4, 5를 구성한다.

한국어 가창 음운 측면에서는 자음 및 받침의 음성적 특징을 분석한 선행 연구[19]의 분류 체계를 적용한다. 다만, 해당 연구에서 보고한 약화 및 시간 변형 양상을 음향 측정 변수로 정량화한 선례가 없으므로, 본 연구는 마찰음, 파찰음, 받침 파열음, 받침 유성음에 각각 대응하는 fric(fricative onset-offset timing), aff(affricate onset-offset timing), plo(batchim plosive drop), voi(batchim voiced loss) 4차원을 정의하여 축 7로 사용한다. 이는 선행 연구의 한국어 음운 분류를 음향 측정 변수로 조작화(operationalization)한 것이다. 또한, 본 체계는 측정된 변별력에 맞추어 차원을 선별하지 않았다.

3-2 다음색 가이드보컬 데이터 및 측정 대상

분석 대상은 과학기술정보통신부와 한국지능정보사회진흥원이 운영하는 AI Hub에서 제공하는 “다음색 가이드보컬 데이터”[27]이다. 다음색 데이터셋은 92명 가창자의 3,609곡을 44.1 kHz mono PCM 16 bit 형식으로 수록하며, 가창자당 37~45곡(평균 39.2곡), 곡당 평균 약 150초의 구성을 가진다. 본 데이터셋은 한국지능정보사회진흥원의 AI 학습용 데이터 활용 약관에 따라 제공되며, 본 연구는 해당 약관의 학술 연구 영역에 한정 사용한다.

데이터셋은 5단 계층 메타데이터 라벨을 제공한다. 가창자 단위 라벨은 성별, 연령, 장르, 음색, 가창자 ID로 구성되며, 각 가창자는 단일 조합에 고정된다. 가창자 단위 메타데이터 분포는 표 2에 제시한다. 곡 단위로는 음표(note) 단위 라벨이 제공되어, 곡당 중앙값 약 405개의 음표에 대해 시작 및 종료 시간, 길이, MIDI 음고, 한국어 음절 단위 가사, 그리고 비브라토(is_vibrt), 벤딩(is_bending), 호흡(is_breath) 플래그가 부여된다. 축 2의 비브라토 측정은 ‘is_vibrt’ 라벨을 직접 활용한다.

다음색 데이터셋은 가창자 단위 메타데이터가 단일 조합으로 정합되어, 성별, 연령, 음색에 따른 층화 분석을 가능하게 한다. 다만, 스튜디오와 마이크 같은 녹음 환경 메타데이터와 가창자별 환경 차이 정보는 공개되지 않아, 환경 변인의 통제는 본 연구 범위 밖이다.

본 연구의 측정 대상은 다음색 92명에서 두 단계 제외 절차를 거쳐 확정된다. UNDER10 그룹 6명은 모두 CHILDREN 장르로 분류되어 발성 발달 영역과 장르 영역이 분리되지 않으므로 측정 이전에 제외하며, 정렬 정확도 검증에 미달한 1명은 측정 이후 추가 제외한다. 따라서, 최종 분석 대상은 85명이다.

표 2. 다음색 데이터셋 가창자 단위 메타데이터 분포
Table 2. Singer-level metadata distribution of the Daeumsaek dataset

Demographic	Category	Count	Ratio (%)
Gender	FEMALE	46	50.00
	MALE	46	50.00
Age	UNDER10	6	6.52
	10TO29	42	45.65
	30TO49	36	39.13
	OVER49	8	8.70
Genre	BALLAD	48	52.17
	DANCE	38	41.30
	CHILDREN	6	6.52
Timbre	NORMAL	41	44.57
	CLEAR	26	28.26
	HUSKY	25	27.17

3-3 측정 도구 및 방법

앞서 정의한 7축 측정 체계에 따라 58차원을 측정하기 위해, librosa 0.11.0[28], parselmouth(Praat backend)[29], torchcrepe[30],[31], MFA 3.3.9[20]를 사용한다. 축별 측정 도구와 라이브러리 및 버전은 표 3에 제시한다.

축 3 발성질의 pitch ceiling(600 Hz)은 가창 F0 측정 영역 (fmax = 1000 Hz)과 분리하여 발성 안정성 측정에 집중하기 위한 설정이며[11], 주기 범위(0.1 ms~20 ms)는 인간 발성 F0(50 Hz~10 kHz)[16] 검출 한계를 포괄한다. 축 4 SPR은 단시간 푸리에 변환(STFT; Short-Time Fourier Transform) 크기 제곱의 2~4 kHz 대역 합을 전 대역 합으로 나눈 비율로 산출한다[18]. 축 5 F0 측정은 torchcrepe[30]로 수행하며, 주기성(periodicity) 0.5 이상의 프레임만 유효 F0로 채택하여 무성 구간을 배제한다. fmin = 50 Hz로 설정하여 인간 발성 F0 하한[16]에 해당하는 50 Hz 미만 영역을 배제한다. 축 6 MFCC[13]는 표준 발화 음성 13차원에서 20차원으로 확장하며, 이는 음악 정보 검색 영역의 음색 정밀 표현 관행을 따른 것이다[14]. librosa STFT 공통 설정(‘n_fft = 2048’, ‘hop_length = 512’)은 75% 중첩 음향 분석 표준이고[28], sr(sample rate)은 librosa 기반 측정에 22,050 Hz를 적용한다. torchcrepe 기반 F0 측정은 원본 sr로 로드하여 내부 16 kHz로 변환하며, 이는 각 도구의 표준 영역을 그대로 적용한 것이다[28],[30].

축 7 한국어 음운 조음 4차원은 강제 정렬 출력의 음소 카테고리를 입력으로 본 연구가 정의한 측정식으로 산출한다. fric과 aff는 음절 초성이 마찰음 또는 파찰음인 경우의 자음 시작 시점(onset)과 첫 모음 시작 시점의 시간 차이(ms)이며, plo는 받침 파열음 구간 parselmouth로 산출한 강도

(intensity)의 시작값과 최저값의 차(dB), voi는 받침 유성음 구간 강도의 시작값과 끝값의 차(dB)이다. 축 1, 3, 4, 5, 6은 곡 단위 평균을 가창자 단위 곡 평균으로 집계하고, 축 7은 음소 출현 단위 값을 가창자 단위로 전수 합산하여 평균한다.

표 3. 7축 측정 도구 및 라이브러리
Table 3. Measurement tools and libraries for the 7-axis framework

Axis	Tool	Library / Version
1. Anatomical vocal tract	Praat formant (burg)	parselmouth
2. Vibrato	Daeumsaek `is_vibrt` label	–
3. Voice quality	Praat (HNR, jitter, shimmer)	parselmouth
4. Singer's formant	librosa STFT	librosa 0.11.0
5. Pitch dynamics	torchcrepe (viterbi)	torchcrepe
6. Timbre envelope	librosa MFCC + spectral	librosa 0.11.0
7. Korean phoneme articulation	MFA forced alignment + custom measurement	MFA 3.3.9 + parselmouth

3-4 한국어 음운 측정을 위한 전처리

축 7 한국어 음운 조음 차원의 측정은 강제 정렬 결과를 전제한다. 전처리는 .lab 파일 생성, MFA 정렬, 정렬 정확도 검증의 세 단계로 수행한다.

첫째, MFA는 오디오에 대응하는 텍스트 파일(.lab)을 입력으로 요구하므로, 다음색 라벨의 'note.lyric' 필드에서 한글 음절을 순서대로 추출하고 공백 분리하여 곡 단위 .lab 파일로 저장한다.

둘째, 가창자별 단일화자(single_speaker) 모드로 'korean_mfa' 음향 모델과 발음 사전을 적용하여 정렬을 수행한다[20]. 단일화자 모드는 한 가창자의 모든 곡을 동일 화자로 처리하여 가창자 내 음향 일관성을 활용하는 절차이다. 정렬 결과는 단어(words) 계층과 음소(phones) 계층을 포함한 Praat 표준 주석 형식인 TextGrid로 산출되며, 단어 계층의 음절 시작 시간은 정렬 정확도 검증의 입력으로, 음소 계층의 구간 경계는 축 7 차원 측정의 입력으로 각각 사용된다. MFA가 정렬 과정에서 자동 삽입하는 묵음 토큰인 sil(silence), sp(short pause), spn(spoken noise)는 단어, 음소 계층에서 제외한 다음 측정 입력으로 사용한다.

셋째, 발화 음성 기반으로 학습된 MFA 음향 모델[20]이 가창 음성에서 발생시키는 정렬 시간 오차[32]에 대해 정렬 정확도 검증을 수행한다. MFA 단어 계층 음절 시작 시간과 다음색 음표 시작 시간의 쌍별 부호 오차(signed error)를 산출하고, 곡 단위 중앙값의 절댓값이 200 ms 이하인 곡을 통과로 판정한다. 200 ms 임계는 음운 단위 시간 안정성에 관한 문헌 기준[32]에 따라 사전에 설정한 값으로, 표본 확보를 위해 조정한 값이 아니다. 이 문헌 기준을 적용한 결과로 가창자당 통과 곡 수가 N_min = 10 이상 확보되었으며, N_min = 10은 차원별 통계 추정에 필요한 최소 곡 표본 수를 보장하는 보수적 하한으로, 통과 곡 수 분포(중앙값 33곡)에서 대부분의 가창자를 유지한다. 본 기준에 미달한 SINGER_16(통과 9곡)을 추가 제외하여 측정 대상 확정이 마무리된다.

3-5 차원별 분포 분석

1) 가창자 내 곡 간 안정성

가창자 내 안정성은 곡 단위 CV로 측정한다. CV는 한 가창자의 차원 측정값이 곡별로 얼마나 안정한가를 표준편차와 평균의 비율로 표현하며, CV < 0.5를 안정 임계값으로 적용한다. CV < 0.5는 곡 간 표준편차가 평균의 절반 미만임을 뜻하며, 표준편차가 평균과 같아지는 CV = 1보다 엄격한 기준이다. 이 기준을 충족하는 차원은 측정값이 곡에 따라 평균에서 크게 벗어나지 않기 때문에, 특정 곡에 의존하기보다 가창자의 안정적 특성을 반영한다. 차원별 통과율(pass_rate)은 85명 중 CV < 0.5를 충족하는 가창자의 비율이다.

58차원 중 34차원이 85명 전원 통과(pass rate = 1.0)를 보였다. 동률 그룹 내에서 평균 CV가 가장 낮은 5차원은 F2_mean(0.048), spectral_bandwidth_mean(0.049), mfcc1_mean(0.056), HNR(0.060), F1_mean(0.076)으로, 발성의 물리적 기제와 직접 대응되는 차원이 가창자 내 안정성에서 우위를 보인다. 반대로 통과율 최저 차원은 vibrato_rate(0.141)이며, vibrato_extent(0.329)와 함께 비브라토 2차원이 가창자 내 변동성이 큰 차원군에 속한다. 축별 패턴을 보면, 음색 엔벨로프 축의 MFCC 평균 차원 11개가 0.5 미만이지만 MFCC 표준편차 차원은 mfcc20_std(0.988)를 제외한 19/20차원이 1.0을 기록하여, 같은 음색 축이라도 표준편차가 평균보다 안정적이다. 축 4 SPR은 0.541로 단일 차원 축 중 통과율이 가장 낮으며, 축 7 한국어 음운 조음 4차원은 fric 및 aff가 1.0, plo 및 voi가 각각 0.976, 0.988로 통과율이 높은 차원군에 속한다. 이로써 본 측정 체계의 다수 차원이 가창자 고유 특성을 곡 단위로 안정적으로 포착함을 확인한다. 차원별 안정성 분포는 그림 1에 제시한다.

2) 가창자 간 변별력

가창자 간 변별력은 가창자 쌍별 Cohen's d로 측정한다. Cohen's d는 두 가창자의 측정값 분포를 표준화하여 단위에 무관한 비교를 가능하게 한다. 85명의 가창자 쌍 3,570쌍에 대해 Cohen's d를 산출하고, 차원별 |d| 분포의 중앙값(abs_d_median)을 주요 지표로 사용한다. 중앙값은 일부 가창자 쌍의 극단값에 덜 민감하기 때문에 분포 대표성이 높다. 변별력 임계값은 효과 크기 분류에서 large에 해당하는 d > 0.8을 적용한다[33].

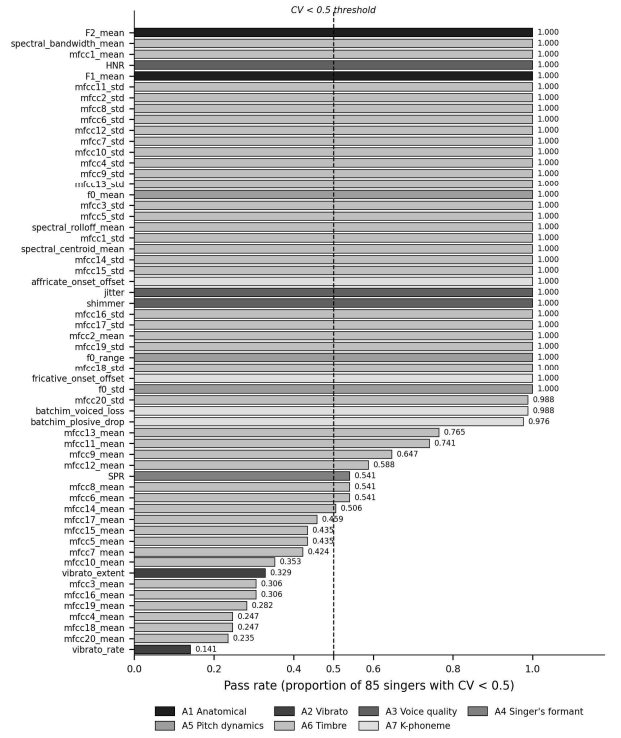

그림 1. 58차원 가창자 내 안정성 분포

Fig. 1. Distribution of within-singer stability across 58 dimensions

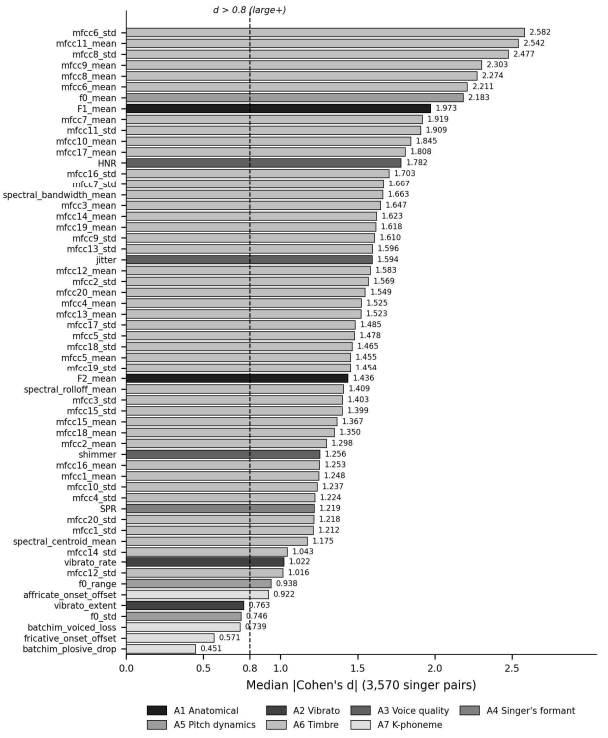

그림 2. 58차원 가창자 간 변별력 분포

Fig. 2. Distribution of between-singer discriminability across 58 dimensions

58차원 중 53차원이 d > 0.8 임계를 통과했다. 변별력이 가장 높은 5차원은 mfcc6_std, mfcc11_mean, mfcc8_std, mfcc9_mean, mfcc8_mean으로 모두 음색 엔벨로프 축에 속한다. 변별력이 가장 낮은 5차원은 plo, fric, voi, f0_std, vibrato_extent이며, 한국어 음운 조음 차원이 하위에 집중적으로 분포한다. 음색 엔벨로프 축은 가창자 변별의 핵심 차원군이고, 한국어 음운 조음 축은 변별력이 상대적으로 약하다. 이로써 본 측정 체계의 다수 차원이 가창자 변별에 높은 유효성을 보임을 확인한다. 앞서 정의한 두 기준을 모두 충족하여 가창자 고유성 지표로 해석되는 차원은 58차원 중 32차원이며, 축 7 한국어 음운 조음은 aff를 제외한 3차원이 변별력 기준에 미달하여 여기에서 제외된다. 가창자 간 차원별 변별력 분포는 그림 2에 제시한다.

3) 인구통계 변인 독립성

가창자 간 변별력은 가창자 개인 수준 효과와 인구통계 군집 효과를 분리하지 않는다. 한 차원이 성별, 연령, 음색 그룹 간 평균 차이로 변별력을 얻는다면, 그 차원은 가창자 고유성이 아닌 인구통계 의존 지표일 가능성이 있다. 본 항은 차원별 인구통계 독립성을 일원 분산 분석(one-way ANOVA)[34]으로 검정한다. 본 검정은 다음색 가창자 단위 메타데이터 4변인을 독립 변인으로 사용한다. 변인별 카테고리는 성별 MALE 43, FEMALE 42, 연령 10TO29, 30TO49, OVER49, 음색 NORMAL, CLEAR, HUSKY, 장르 BALLAD 47, DANCE 38이다. 58차원과 4 변인에 대해 총 232회의 검정을 수행하는 다중 비교(multiple comparisons) 상황에서는, 개별 검정의 유의수준을 그대로 적용하면 우연에 의한 거짓 양성이 누적된다. 이를 통제하기 위해 본페로니 교정(Bonferroni correction)[35]으로 유의수준을 검정 수로 나눈다($\alpha = 0.05 / 232 \approx 2.16 \times 10^{-4}$).

본페로니 교정 후 유의 차원은 성별 38개, 장르 11개, 연령 및 음색은 모두 0개로 수렴한다. 단순 임계($p < 0.05$) 기준에서는 연령 및 음색에도 일부 유의 차원이 분포하나, 본페로니 교정 임계를 통과하지 못한다. 인구통계 의존성은 성별과 장르 효과가 주요하며, 성별 효과가 가장 강한 차원은 f0_mean($F = 456.52$, $p = 1.74 \times 10^{-35}$)로 ANOVA 검정 중 최대 효과를 보인다. 장르 효과가 가장 강한 차원은 jitter($F = 110.30$, $p = 6.65 \times 10^{-17}$)이며, 발성질, 비브라토, 한국어 음운 영역에서 주로 관찰된다. 축 7 한국어 음운 조음 4차원은 성별, 연령, 음색에서 모두 0/4 유의이나, 장르에서는 3/4 유의로 BALLAD와 DANCE 영역에서 음운 조음 차원이 차이를 보인다. 연령과 음색 의존성이 본페로니 교정 후 소멸하는 점은, 본 측정 체계의 차원들이 연령 및 음색 군집과 독립적이며 해당 변인에 대해 가창자 개인 수준의 고유성을 반영함을 시사한다. 축별 분포 분석 요약은 표 4에 제시한다.

표 4. 차원별 분포 분석 축별 요약(G: 성별, A: 나이, T: 음색, Gn: 장르)
Table 4. Per-axis summary of distribution analysis (G: gender, A: age, T: timbre, Gn: genre)

Axis	Count	Stability	Discriminability	Sig. dims (G / A / T / Gn)
1. Anatomical vocal tract	2	2 / 2	2 / 2	2 / 0 / 0 / 0
2. Vibrato	2	0 / 2	1 / 2	0 / 0 / 0 / 2
3. Voice quality	3	3 / 3	3 / 3	0 / 0 / 0 / 3
4. Singer's formant	1	0 / 1	1 / 1	1 / 0 / 0 / 0
5. Pitch dynamics	3	3 / 3	2 / 3	3 / 0 / 0 / 2
6. Timbre envelope	43	24 / 43	43 / 43	32 / 0 / 0 / 1
7. Korean phoneme articulation	4	2 / 4	1 / 4	0 / 0 / 0 / 3
Total	58	34 / 58	53 / 58	38 / 0 / 0 / 11

Ⅳ. 가창 음성 변환 실험과 차원별 분석

4-1 실험 설계

1) 변환 실험 대상

가창자 간 교차 변환 실험은 성별, 연령, 음색의 인구통계 조합으로 가창자를 층화한다. 음색이 NORMAL 단일로 구성되어 층화가 불가능한 OVER49 8명은 층화 격자에서 제외하여, 연령은 10TO29와 30TO49 2수준으로 구성된다. 대상 24명은 성별(2), 연령(2), 음색(3)의 12 조합에서 각 조합을 대표하는 조합 내 최대 변별 쌍 2명씩을 결정론적 규칙으로 선택하여 인구통계 분포의 대표성을 확보한다. 원본 12명은 대상 24명을 제외한 후보군에서 성별(2), 연령(2)의 4 조합에 조합당 3명씩, 재현 가능성 확보를 위해 시드값 (seed value)을 42로 고정하여 무작위 추출한다. 원본 음성의 음색은 변환 과정에서 대상 음색으로 치환되므로 음색 층화는 적용하지 않는다. 실험 대상 구성은 표 5에 제시한다.

변환 입력은 각 가창자의 60초 원본 클립으로, 이 길이는 음운 단위 측정에 필요한 음소 출현을 충분히 포함하며 정렬 정확도 검증의 200 ms 임계를 충족한다. 대상 24명의 학습 및 검증 곡 단위 분할은 정렬 정확도 검증을 통과하고 정렬에 성공한 가용 곡에서 7:3의 비율로 무작위 수행하여[36], 24명 합계 학습 563곡, 검증 257곡을 산출한다. 곡 분할 시 가창자마다 시드값을 42로 고정하여 재현 가능성을 확보하며, 학습 곡이 검증에 포함되지 않도록 한다. 학습 곡은 RVC v2 학습에 사용하며, 검증 곡은 56차원 측정과 변환 원본 클립 추출에 사용한다.

표 5. 변환 실험 설계
Table 5. Voice conversion experiment design

Item	Value
Target singers	24 (12 cells × 2)
Source singers	12 (4 cells × 3)
Conversion pairs	288 (12 × 24)
VC models	RVC v2 / SEED-VC
Total conversions	576 (288 × 2)
Source clip	60.0 s
Same-gender pairs	144
Cross-gender pairs	144

2) 모델 학습 및 추론 설정

RVC v2는 화자 종속 학습으로 대상 가창자당 1개 모델을 학습하고, SEED-VC는 30초 대상 참조 음성으로 zero-shot 변환을 수행한다. 두 모델의 학습 양식 대조는 변환 결과의 보존 및 전이 패턴이 학습 양식에 따라 차별화되는지 또는 일관 재현되는지를 확인하기 위함이다. 두 모델 공통으로 추론 환경을 통일하여 하드웨어 의존성을 제거한다. 두 모델은 기본 F0 처리가 다르므로, 동일 기준 비교를 위해 원본 F0 궤적이 변환 출력에 직접 반영되도록 설정한다. RVC v2 학습은 Applio 기본 설정을 적용하며, 'overtraining_detector'를 활성화하여 검증 손실이 50회 연속 개선되지 않으면 조기 종료한다. 모델 학습 및 추론 설정은 아래 표 6에 제시한다.

표 6. RVC v2 와 SEED-VC 의 모델 학습 및 추론 설정
Table 6. Training and inference settings of RVC v2 and SEED-VC

Item	RVC v2	SEED-VC
Training	Per-singer pre-training	Zero-shot
Training framework	Applio[37] (c70e3b9f)	–
Training epochs	300	–
Batch size	16	–
Embedder	ContentVec	built-in
F0 extractor	crepe	built-in
Inference environment	vast.ai RTX 4090	vast.ai RTX 4090
Model version	Applio c70e3b9f	seed-vc 51383ef
Pitch preservation	pitch = 0	f0_condition = True; auto_f0_adjust = False
Conversion time	11.1 min	189.2 min

3) 보존 및 전이 판정 기준

576개의 변환 wav 음원에 앞서 정의한 7축 측정 도구와 파라미터를 동일 적용하여 변환 결과의 차원별 값을 산출한다. 축 2 비브라토 2차원은 변환 wav 음원에 'is_vibrt'이 없어 측정이 불가하므로 분석에서 제외한다. 58차원 중 56차원에 대해 분석을 수행한다.

차원별 보존 및 전이 판정은 분류 비율의 통계적 안정성을 정량화하기 위해 부트스트랩(bootstrap) 방법을 적용한다[38]. 재표집은 차원 유형에 따라 분기하여, 음소 단위 출현값을 갖는 축 7 4차원은 변환 결과의 음소 출현값을 복원추출해 반복마다 평균을 산출하고, 가창자당 단일 값을 갖는 축 1, 3, 4, 5, 6은 변환 값을 고정한 채 원본 및 대상 기준 분포를 재표집한다. 두 방식 모두 시드값 42로 1,000회 반복하며, 각 반복에서 변환 결과가 원본 기준에 더 가까운지, 대상 기준에 더 가까운지를 비교한다. 비교 통계량은 수식 (1)의 z 거리이며, $Z_{src} < Z_{tgt}$ 이면 보존, $Z_{src} > Z_{tgt}$ 이면 전이로 집계한다.

$$z_{\mathrm{src}} = \frac{|o_{\mathrm{mean}} - \mathrm{src_b}|}{\mathrm{src_s}}, \quad z_{\mathrm{tgt}} = \frac{|o_{\mathrm{mean}} - \mathrm{tgt_b}|}{\mathrm{tgt_s}} \tag{1}$$

여기서 O_{mean}은 변환 출력의 차원 값이고, 원본 기준과 대상 기준은 각각 원본 가창자와 대상 가창자가 앞서 산출한 차원별 측정 분포다. src_b와 tgt_b는 재표집에서 반복마다 측정 평균을 중심으로 하는 정규분포에서 추출한 기준 평균이고, src_s와 tgt_s는 측정 표준편차를 표본 수의 제곱근으로 나눈 값과 측정 표준편차로 각각 고정한다. 거리를 표준편차로 나누어 차원마다 다른 측정 척도를 정규화함으로써, 보존 및 전이 비교를 차원 간 동일 기준으로 수행한다.

신뢰구간은 참 비율이 위치할 법한 범위이며, 윌슨 스코어 신뢰구간(Wilson score interval)은 비율이 0이나 1에 가깝거나 표본이 작을 때에도 안정적인 구간을 제공하여 비율 추정에 적합하다[39],[40]. 1,000회 반복 결과에서 보존 또는 전이 방향의 윌슨 스코어 95% 신뢰구간 하한이 0.5를 초과하는 쪽을 차원별 판정으로 확정하며, 양쪽 모두 임계를 통과하지 못하면 혼합으로 분류한다. 0.5는 보존 및 전이 이항 분류의 과반 기준으로, 신뢰구간 하한이 이를 초과함은 해당 방향의 과반이 부트스트랩 표집 변동에 견고함을 뜻한다. 축 7 음소 출현 횟수가 3 미만이거나 변환 결과의 측정값이 산출되지 않는 사례는 미판정으로 분류한다. 본 장의 보존 및 전이 판정에는 차원 간 다중 비교 보정(FWER; family-wise error rate)을 적용하지 않으며, 통계적 표현은 각 차원의 분류 신뢰구간에 한정한다. 이는 집단 차이의 통계적 유의성을 판정하는 검정에서는 다중 비교 보정이 요구되는 것과 달리, 본 장의 보존 및 전이 판정은 각 차원을 독립적 기술 분류로 보고할 뿐 56차원을 묶어 통계적 결론을 내리지 않기 때문이다.

4-2 차원별 보존 및 전이 판정

1) 56차원 종합 판정

두 모델의 56차원 종합 판정 결과, RVC v2는 보존 19개, 전이 25개, 혼합 12개, SEED-VC는 보존 19개, 전이 24개, 혼합 13개로 산출된다. 두 모델 모두 전이 판정이 가장 많은 비중을 차지하며, 보존과 혼합의 분포는 유사하다. 차원별 판정 결과는 다음 그림 3에 제시한다.

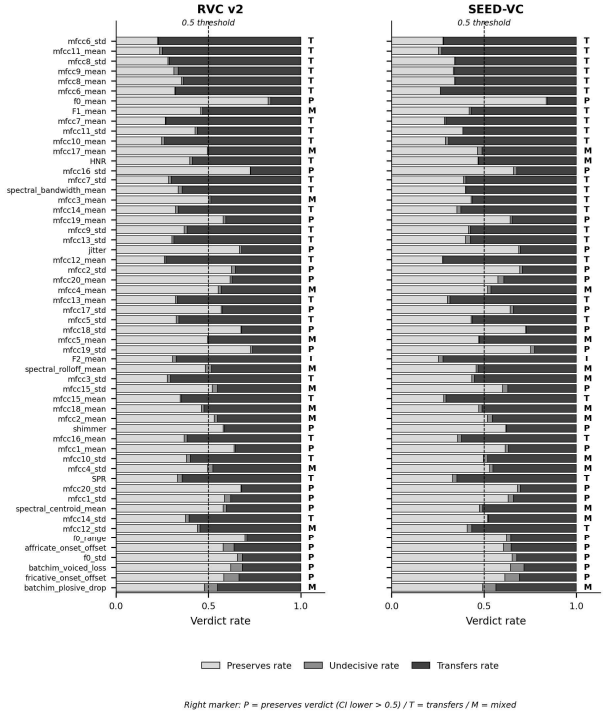

그림 3. 56차원 × 2 모델 보존, 전이, 혼합 판정
Fig. 3. Per-dimension verdict by VC model

2) 발화 음성의 음향 차원

발화 음성의 음향 영역에 해당하는 축 1, 축 3, 축 6의 48차원은 다수 차원이 전이로 분류된다. 축 1 해부학적 성도 2차원은 SEED-VC에서 양 차원 모두 전이로, RVC v2에서 1차원 전이, 1차원 혼합으로 분류된다. 축 3 발성질 3차원에서 jitter와 shimmer는 두 모델 모두 보존, HNR은 RVC 전이, SEED-VC 혼합으로 분류된다. 축 6 음색 엔벨로프 43차원은 두 모델 모두 전이가 가장 우세한 분류이나, 영역에 따라 차이가 있다. MFCC 평균 20차원 중 mfcc6~16 평균은 두 모델 모두 강한 전이를 보이고, mfcc1, 19, 20 평균은 두 모델 모두 보존이다. MFCC 표준편차 20차원은 mfcc16~20 표준편차에서 두 모델 모두 보존으로 분류되어, MFCC 표준편차가 평균보다 보존 비율이 높다. 스펙트럼 통계 3차원에서는 spectral_bandwidth_mean이 두 모델 모두 전이, spectral_centroid_mean이 RVC 보존, SEED-VC 혼합, spectral_rolloff_mean이 두 모델 혼합이다.

3) 가창 발성의 음향 차원

가창 발성의 음향 영역에 해당하는 축 4와 축 5의 4차원은 전이와 보존이 축별로 대비된다. 축 4 성악 공명 SPR은 두 모델 모두 전이로 분류된다. 축 5 피치 동역학 3차원은 두 모델 모두 보존으로 분류되며, f0_mean의 보존 비율은 RVC v2 82.3%, SEED-VC 83.7%로 56차원 중 가장 높다. f0_std의 보존 비율은 RVC v2 65.6%, SEED-VC 65.3%, f0_range는 RVC v2 69.8%, SEED-VC 62.2%이다. 다만, 축 5의 3차원 보존 판정은 앞서 두 모델 모두 원본 F0 궤적이 변환 출력에 직접 반영되도록 한 설정의 영향을 받으므로, 해당 설정에 따른 결과로 해석한다.

4) 한국어 가창 음운의 음향 차원

정렬 정확도 검증은 변환 결과 wav 음원의 음운 타이밍 안정성을 변환 품질의 독립 지표로 평가하며, 축 7 한국어 음운 조음 차원 검정의 입력 자격을 결정한다. 576개 변환 wav 음원에 동일하게 적용한 결과, RVC v2 269/288(93.40%), SEED-VC 272/288 (94.44%)가 통과하며, 통과 분포는 아래 그림 4에 제시한다.

한국어 가창 음운의 음향 영역에 해당하는 축 7 음운 조음 4차원에서 fric, aff, voi가 두 모델 모두 보존으로 분류되며, 보존 비율은 fric RVC 58.4%, SEED-VC 61.4%, aff RVC 58.0%, SEED-VC 60.7%, voi RVC 62.1%, SEED-VC 64.3%이다. plo는 보존 및 전이 비율이 RVC 125/118, SEED-VC 130/115로 1:1에 가까워 양 모델 모두 혼합으로 분류된다.

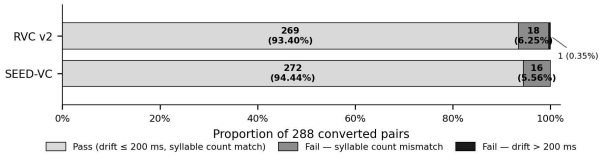

그림 4. 변환 wav 정렬 정확도 검증 통과 및 미통과 분포

Fig. 4. Alignment time error check pass/fail distribution for converted wav

4-3 차원별 보존 및 전이 해석

1) 한국어 음운 조음의 보존 비대칭

축 7 한국어 음운 조음 4차원의 판정은 비대칭을 보인다. 앞서 fric, aff, voi는 두 모델 모두 보존으로 분류된 반면 받침 파열음에 대응하는 plo만 혼합으로 분류되었다. 이는 받침 파열음이 일으키는 소리의 단절[19]이 짧은 구간에서 일어나고, 그 정도가 가창마다 일정하지 않기 때문에, 변환 모델이 이를 일관되게 재현하지 못해 보존과 전이 어느 한쪽으로도 수렴하지 않은 것으로 해석된다.

이러한 비대칭은, 가창에서 받침 유성음이 소리의 단절 없이 음을 지속하고 마찰음과 파찰음은 지속 시간이 길어 비교적 안정적으로 실현된다는 선행 연구[19]의 보고와 정합한다. 한편 앞의 표 4 가창자 간 변별력 분포에서 plo, fric, voi는 하위 5차원에 분포하며, 이는 축 7 차원이 가창자 변별보다 한국어 가창 음운 시스템의 시간 구조 안정성을 추적하는 측정 영역으로 기능함을 시사한다.

2) 두 모델 판정 결과의 일관성

차원별 최다 판정 일치율은 56차원 중 47개로 83.9%에 이르며, 9개 불일치 차원은 모두 한 모델에서는 보존 또는 전이로, 다른 모델에서는 혼합으로 분류된 경계 사례이며, 두 모델 간 보존과 전이가 뒤바뀐 차원은 없다. 축별로는 축 5 피치 동역학 3차원과 축 7 한국어 음운 조음 4차원에서 두 모델의 최다 판정이 전수 일치하며, 축 6 음색 엔벨로프에서도 43차원 중 36개가 일치한다. 축별 보존 및 전이 판정 분포는 표 7에 제시한다. 이러한 높은 일치율은 본 7축 측정 체계가 포착하는 차원별 보존 및 전이 패턴이 특정 학습 양식이 아니라 음향 차원 자체의 속성임을 뒷받침한다.

이 불일치 9개 차원 중 7개가 음색 엔벨로프 축에 속해, 불일치는 전이가 우세한 음색 영역의 혼합 경계에 집중된다. 나아가 변환 쌍 단위로 두 모델의 차원별 판정 일치율을 산출하여 가창자 성별, 연령, 음색, 장르별로 비교한 결과, 장르와 음색에서는 유의한 차이가 없었고 성별과 연령에서 검출된 차이도 일치율 중앙값 기준 0.05 미만에 그쳤다. 따라서 두 모델의 판정 불일치는 특정 가창자 집단이나 장르에 집중되기보다 음색 엔벨로프 축의 경계 차원에 구조적으로 분포하며, 측정 체계의 정밀화는 이 영역에 집중되어야 함을 시사한다.

표 7. RVC v2 및 SEED-VC의 축별 보존 및 전이 판정 분포(P: 보존, T: 전이, M: 혼합)

Table 7. Per-axis verdict distribution of RVC v2 and SEED-VC (P: preserves, T: transfers, M: mixed)

Axis	count	RVC P / T/ M	SEED-VC P / T / M	Agreement
1. Anatomical vocal tract	2	0 / 1 / 1	0 / 2/ 0	1 / 2
2. Vibrato	–	–	–	–
3. Voice quality	3	2 / 1 / 0	2 / 0 / 1	2 / 3
4. Singer's formant	1	0 / 1 / 0	0 / 1 / 0	1 / 1
5. Pitch dynamics	3	3 / 0 / 0	3 / 0 / 0	3 / 3
6. Timbre envelope	43	11 / 22/ 10	11 / 21 / 11	36 / 43
7. Korean phoneme articulation	4	3 / 0 / 1	3 / 0 / 1	4 /4
Total	56	19 / 25 / 12	19 / 24 / 13	47 / 56 (83.9%)

3) 가창자 간 변별력과 변환 후 방향성의 상관

가창자 간 변별력이 차원별 보존 및 전이 방향성의 결정 요인인지 검토하기 위해 점이연 상관계수(point-biserial

correlation)[41]를 산출한다. 보존과 전이 건수가 동률인 차원은 이항 변수화 불가로 산출에서 제외하며, RVC v2의 n이 SEED-VC보다 한 차원 적은 것은 축 6의 spectral_rolloff_mean이 보존과 전이 결괏값이 동률로 제외되었기 때문이다. 두 모델 모두 통계적으로 유의한 양의 상관을 보이며, 효과 크기는 medium 수준[42]에 해당한다. 95% 신뢰구간 하한이 모두 0을 상회하여 양의 상관의 강건함을 시사하며, F0 직접 보존 통제 변수로 고정한 축 5 3차원을 제외한 민감도 분석에서도 $\Delta r \approx +0.013 \sim +0.017$에 그쳐 본 상관 결과가 통제 변수에 의존하지 않음을 나타낸다. 산출 결과는 표 8에, 변별력과 방향성의 산포도 및 회귀선은 그림 5에 제시한다.

표 8. 가창자 간 변별력과 변환 후 방향성의 점이연 상관계수 산출 결과

Table 8. Point-biserial correlation between discriminability and verdict direction

Metric		RVC v2	SEED-VC
Baseline (56 dims)	Correlation r	+0.4553	+0.4910
	p-value	4.79×10^{-4}	1.22×10^{-4}
	Sample size n	55	56
	95% CI lower	+0.22	+0.26
	Effect size	Medium	Medium
	R^2	0.21 (21%)	0.24 (24%)
Sensitivity (Axis 5 excluded)	Correlation r	+0.4686	+0.5083
	Sample size n	52	53
	Δr	+0.0133	+0.0173

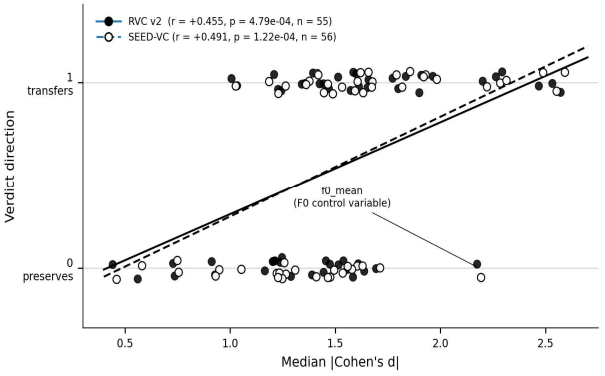

그림 5. 차원별 가창자 간 변별력과 보존 및 전이 판정 방향성의 점이연 상관계수

Fig. 5. Point-biserial correlation between median |Cohen's d| and verdict direction

Ⅴ. 결 론

5-1 연구 요약

본 연구는 한국어 가창 보컬 고유성의 7축 58차원 측정 체계를 구축하고, 두 가지 학습 양식이 대비되는 음성 변환 모델에서 비브라토 축을 제외한 6개 축 56차원의 보존 및 전이 패턴을 분석하였다.

첫째, 보존과 전이는 음향 영역에 따라 다르게 나타난다. 음색 엔벨로프와 성도 공명은 대상 가창자의 특성으로 전이되는 반면, 피치 동역학은 원본 F0 궤적이 변환 출력에 반영되도록 한 변환 설정에서 원본 가창자의 특성을 유지한다.

둘째, 선행 연구[19]를 기반으로 축 7로 정의한 한국어 음운 조음 4차원의 변환 후 보존 및 전이 양상을 분석하였다. 마찰음, 파찰음, 유성 받침은 두 모델 모두 보존으로, 받침 파열음만 혼합으로 분류되는 비대칭이 관찰되었다. 또한, 변환 결과물의 음소 경계 안정성을 음운 단위 이항 측정으로 집계하는 독립 지표를 제안하며, 두 모델 모두 90% 이상의 통과율을 보인다.

셋째, 가창자 간 변별력은 보존 및 전이 방향성과 유의한 양의 상관을 보이며, 효과 크기는 양 모델 모두 medium 수준에 해당한다. 변별력이 높을수록 전이 방향으로 수렴하는 경향이 있으나, 변별력이 방향성을 단독으로 결정하지는 않는다. 나아가 차원별 보존율 분포가 두 모델에서 일관되게 나타나는 점은, 본 측정 체계가 실제 음성 변환 적용 시 차원별 보존 및 전이 경향을 다루는 측정 토대로 기능할 수 있음을 시사한다.

5-2 한계점 및 향후 연구

1) 한계점

본 연구의 한계는 다음과 같다. 첫째, 가창자 간 변별력 산출에 성별 층화 분석을 적용하지 않았다. 일원 분산 분석에서 성별 유의 차원이 38개로 확인되었음에도 성별을 별도 집단으로 분리하여 변별력을 재산출하는 층화 분석은 수행하지 않았으므로, 일부 차원의 변별력에 성별 효과가 혼입되었을 가능성을 완전히 배제하지 못한다. 둘째, 가창 특화 성분인 축 2 비브라토는 변환 결과 분석에서 축 전체가 제외되었다. 비브라토 2차원 (vibrato_rate, vibrato_extent)은 원본 가창자 측정에서 다음색 데이터의 `is_vibrt` 라벨로 산출하였으나, 변환 wav 음원에는 해당 라벨이 없어 원본과 동일 기준의 라벨 기반 측정이 불가하였다. 이로 인해 본 변환 실험의 보존 및 전이 결론은 비브라토를 제외한 6개 축 56차원에 한정되며, 가창 고유의 음높이 변조 성분이 변환 과정에서 보존되는지 또는 전이되는지는 본 연구에서 검증되지 않았다. 셋째, 본 연구의 변환 실험은 데이터셋 구조에 따라 BALLAD와 DANCE 장르에 한정되어 수행하였으므로, 다른 장르에서 동일한 보존 및 전이 패턴이 나타나는지는 확인되지 않았다. 장르별 음운 조음 및 발성 특성이 상이할 수 있어 본 결과의 장르 간 일반화는 제한적이다. 넷째, 피치 동역학의 보존은 변환 모델이 가창자 고유성을 독립적으로 보존한 결과로 단정하기 어렵고, 변환 설정과 음악적 멜로디 정보의 영향을 함께 반영하므로, 이러한 보존 양상에 대한 해석은 본 실험 조건에 국한된다.

2) 향후 연구

향후 연구는 네 영역을 제안한다. 첫째, 본 측정 체계의 평가 확장이 후속 과제로 남는다. 음역대 라벨을 갖춘 데이터셋에서 음역대별 보존 및 전이 패턴을 층화 분석하고, 특정 차원을 추가하거나 제거하여 분류 기반 식별 정확도의 변화를 평가하는 것이다. 둘째, 본 측정 체계의 축 1부터 축 6은 언어 비의존적 음향 특성에 기반하는 반면, 축 7 한국어 음운 조음은 한국어에 특화되므로, 타 언어권 가창으로 확장하기 위해서는 해당 언어의 음운 분류에 맞춘 축 7 재정의가 필요하다. 셋째, 본 측정 체계를 생성형 AI 가창 모델에 적용하는 것이다. 사용자 보컬 데이터를 학습하여 가창을 생성하는 TTM(text-to-music) 모델은 변환 모델과 근본적으로 다른 패러다임이다. 원본과 대상을 비교한 본 연구의 판정 구조를 적용하기 어려우며, TTM 모델의 패러다임에 맞는 별도의 측정 및 판정 절차 설계가 필요하다. 넷째, 음성 변환 기술의 확산으로 가창 보컬의 무단 복제 및 도용 가능성이 높아지는 가운데, 본 연구의 차원별 보존 및 전이 패턴이 AI를 통해 생성되거나 변환된 가창의 식별 특징으로 기능하는지 검증하고, 이를 한국어 가창 딥페이크 탐지[43] 모델 개발로 확장하는 것이 후속 과제이다.

5-3 연구 기여 및 시사점

본 연구는 한국어 가창 보컬의 음향 특성을 7축 58차원으로 정량화하는 측정 체계를 구축하고, AI 음성 변환이 이 가운데 어느 차원을 보존하고 어느 차원을 대체하는지를 모델 비의존적으로 특성화하였다.

특히, 한국어 음운 조음 차원을 음향 측정 변수로 직접 정의하여 변환 후 보존 및 전이 양상을 분석한 것은, 언어 특화 발성 특성이 변환 평가에서 독립적으로 다루어질 수 있음을 보인 시도이다. 변환 결과물의 음소 경계 안정성을 음운 단위 이항 측정 지표로 제안한 것도 같은 맥락으로, 프레임 단위 평균 지표가 포착하지 못하는 음운 타이밍 정보를 명시적으로 집계한다.

보존 및 전이 패턴이 학습 양식이 대비되는 두 모델에서 일관되게 재현된다는 점은, 본 측정 체계가 포착하는 차원별 패턴이 특정 모델 설계가 아닌 음향 차원 자체의 속성임을 뒷받침한다. 본 체계는 후속 변환 모델 평가와 가창 음원 저작권 및 딥페이크 논의에서 가창자 변별 음향 특성을 다루는 측정 토대로 기능할 수 있다.

참고문헌

[1] Korea Copyright Protection Agency (KCOPA), AI Cover Songs and Copyright Issues, Seoul, Technical Report, pp. 7-9, 2024.

[2] The Brussels Times with Belga. Taylor Swift Moves to Trademark Her Voice Amid Concerns over AI [Internet]. Available: https://www.brusselstimes.com/art-culture/2101076/taylor-swift-moves-to-trademark-her-voice-amid-concerns-over-ai.

[3] Justia U.S. Law. Midler v. Ford Motor Co., 849 F.2d 460 (9th Cir. 1988) [Internet]. Available: https://law.justia.com/cases/federal/appellate-courts/F2/849/460/37485.

[4] S. T. Shim, “A Study on the Right of Voice of Public Figures: Focusing on the Analysis of the ‘Newstapa Decision’,” *Justice*, No. 181, pp. 143-168, December 2020. https://doi.org/10.29305/tj.2020.12.181.143

[5] D. Y. Lee, “Present and Key Issues of AI Composer and the Direction of Music Copyright Discussions in Korea,” *Korean Journal of Popular Music*, No. 33, pp. 241-274, 2024. https://doi.org/10.36775/kjpm.2024.33.241

[6] International Telecommunication Union, Methods for Subjective Determination of Transmission Quality, Geneva, Switzerland, ITU-T Recommendation P.800, August 1996.

[7] W.-C. Huang, L. P. Violeta, S. Liu, J. Shi, and T. Toda, “The Singing Voice Conversion Challenge 2023,” in *Proceedings of the 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Taipei, Taiwan, pp. 1-8, December 2023. https://doi.org/10.1109/ASRU57964.2023.10389671

[8] L. P. Violeta, X. Zhang, J. Shi, Y. Yasuda, W.-C. Huang, Z. Wu, and T. Toda, “The Singing Voice Conversion Challenge 2025: From Singer Identity Conversion to Singing Style Conversion,” in *Proceedings of the 2026 IEEE International Conference on Acoustics, Speech and Signal Proceessing (ICASSP),* Barcelona, Spain, pp. 17707-17711, 2026.

[9] T. Walczyna and Z. Piotrowski, “Overview of Voice Conversion Methods Based on Deep Learning,” *Applied Sciences*, Vol. 13, No. 5, 3100, February 2023. https://doi.org/10.3390/app13053100

[10] G. Fant, *Acoustic Theory of Speech Production*, The Hague, The Netherlands: Mouton, 1960.

[11] J. P. Teixeira, C. Oliveira, and C. Lopes, “Vocal Acoustic Analysis - Jitter, Shimmer and HNR Parameters,” *Procedia Technology*, Vol. 9, pp. 1112-1122, 2013. https://doi.org/10.1016/j.protcy.2013.12.124

[12] E. Yumoto, W. J. Gould, and T. Baer, “Harmonics-to-Noise Ratio as an Index of the Degree of Hoarseness,” *Journal of the Acoustical Society of America*, Vol. 71, No. 6, pp. 1544-1550, June 1982. https://doi.org/10.1121/1.387808

[13] S. Davis and P. Mermelstein, “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 28, No. 4,

pp. 357-366, August 1980. https://doi.org/10.1109/TASSP.1980.1163420

[14] G. Tzanetakis and P. Cook, “Musical Genre Classification of Audio Signals,” *IEEE Transactions on Speech and Audio Processing*, Vol. 10, No. 5, pp. 293-302, July 2002. https://doi.org/10.1109/TSA.2002.800560

[15] J. A. Diaz and H. B. Rothman, “Acoustical Comparison Between Samples of Good and Poor Vibrato in Singers,” *Journal of Voice*, Vol. 17, No. 2, pp. 179-184, June 2003. https://doi.org/10.1016/S0892-1997(03)00002-X

[16] J. Sundberg, *The Science of the Singing Voice*, DeKalb, IL: Northern Illinois University Press, 1987.

[17] J. Sundberg, “Level and Center Frequency of the Singer’s Formant,” *Journal of Voice*, Vol. 15, No. 2, pp. 176-186, June 2001. https://doi.org/10.1016/S0892-1997(01)00019-4

[18] K. Omori, A. Kacker, L. M. Carroll, W. D. Riley, and S. M. Blaugrund, “Singing Power Ratio: Quantitative Evaluation of Singing Voice Quality,” *Journal of Voice*, Vol. 10, No. 3, pp. 228-235, September 1996. https://doi.org/10.1016/S0892-1997(96)80003-8

[19] C.-H. Lee, Phonetic Characteristics of Korean Phonemes in Singing and Its Relationship to Lyrics and Pronunciation in Korean Popular Music, Ph.D. Dissertation, Kyung Hee University, Gyeonggi, 2020.

[20] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi,” in *Proceedings of the Interspeech 2017*, Stockholm, Sweden, pp. 498-502, August 2017. https://doi.org/10.21437/Interspeech.2017-1386

[21] B. Sisman, J. Yamagishi, S. King, and H. Li, “An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 29, pp. 132-157, November 2020. https://doi.org/10.1109/TASLP.2020.3038524

[22] K. Qian, Y. Zhang, H. Gao, J. Ni, C.-I. Lai, D. Cox, ... and S. Chang, “ContentVec: An Improved Self-Supervised Speech Representation by Disentangling Speakers,” in *Proceedings of the 39th International Conference on Machine Learning (ICML)*, Baltimore: MD, PMLR 162, pp. 18003-18017, July 2022.

[23] J. Kim, J. Kong, and J. Son, “Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech,” in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, PMLR 139, pp. 5530-5540, July 2021.

[24] GitHub. RVC Project, Retrieval-Based-Voice-Conversion-WebUI [Internet]. Available: https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI.

[25] S. Liu, “Zero-Shot Voice Conversion with Diffusion Transformers,” arXiv:2411.09943, November 2024. https://doi.org/10.48550/arXiv.2411.09943

[26] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, pp. 6840-6851, December 2020.

[27] AI Hub. Daeumsaek Guide-Vocal Dataset (Dataset ID 473) [Internet]. Available: https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=473.

[28] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, “Librosa: Audio and Music Signal Analysis in Python,” in *Proceedings of the 14th Python in Science Conference*, Austin: TX, pp. 18-24, July 2015. https://doi.org/10.25080/Majora-7b98e3ed-003

[29] Y. Jadoul, B. Thompson, and B. de Boer, “Introducing Parselmouth: A Python Interface to Praat,” *Journal of Phonetics*, Vol. 71, pp. 1-15, November 2018. https://doi.org/10.1016/j.wocn.2018.07.001

[30] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, “CrePE: A Convolutional Representation for Pitch Estimation,” in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, pp. 161-165, April 2018. https://doi.org/10.1109/ICASSP.2018.8461329

[31] M. Morrison. Torchcrepe [Internet]. Available: https://github.com/maxrmorrison/torchcrepe.

[32] G. Doras, Y. Teytaut, and A. Roebel, “A Linear Memory CTC-Based Algorithm for Text-to-Voice Alignment of Very Long Audio Recordings,” *Applied Sciences*, Vol. 13, No. 3, 1854, January 2023. https://doi.org/10.3390/app13031854

[33] S. S. Sawilowsky, “New Effect Size Rules of Thumb,” *Journal of Modern Applied Statistical Methods*, Vol. 8, No. 2, pp. 597-599, November 2009. https://doi.org/10.22237/jmasm/1257035100

[34] R. A. Fisher, *Statistical Methods for Research Workers*, Edinburgh, Scotland: Oliver and Boyd, 1925.

[35] O. J. Dunn, “Multiple Comparisons Among Means,” *Journal of the American Statistical Association*, Vol. 56, No. 293, pp. 52-64, March 1961. https://doi.org/10.1080/01621459.1961.10482090

[36] V. R. Joseph, “Optimal Ratio for Data Splitting,” *Statistical Analysis and Data Mining: The ASA Data Science Journal*, Vol. 15, No. 4, pp. 531-538, August 2022. https://doi.org/1

0.1002/sam.11583

[37] IAHispano. Applio: A Simple, High-Quality Voice Conversion Tool [Internet]. Available: https://github.com/IAHispano/Applio.

[38] B. Efron, “Bootstrap Methods: Another Look at the Jackknife,” *The Annals of Statistics*, Vol. 7, No. 1, pp. 1-26, January 1979. https://doi.org/10.1214/aos/1176344552

[39] E. B. Wilson, “Probable Inference, the Law of Succession, and Statistical Inference,” *Journal of the American Statistical Association*, Vol. 22, No. 158, pp. 209-212, June 1927. https://doi.org/10.1080/01621459.1927.10502953

[40] L. D. Brown, T. T. Cai, and A. DasGupta, “Interval Estimation for a Binomial Proportion,” *Statistical Science*, Vol. 16, No. 2, pp. 101-133, May 2001. https://doi.org/10.1214/ss/1009213286

[41] R. F. Tate, “Correlation between a Discrete and a Continuous Variable. Point-Biserial Correlation,” *The Annals of Mathematical Statistics*, Vol. 25, No. 3, pp. 603-607, September 1954. https://doi.org/10.1214/aoms/1177728730

[42] J. Cohen, “A Power Primer,” *Psychological Bulletin*, Vol. 112, No. 1, pp. 155-159, July 1992. https://doi.org/10.1037/0033-2909.112.1.155

[43] Y. Zang, Y. Zhang, M. Heydari, and Z. Duan, “SingFake: Singing Voice Deepfake Detection,” in *Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, pp. 12156-12160, April 2024. https://doi.org/10.1109/ICASSP48485.2024.10448184

정다훈(Da-Hun Jeong)

2014년: 명지대학교 예술체육대학 음악학부 작곡전공(음악학사)
2026년 : 경희대학교 대학원 (포스트모던학 석사)
2024년～현 재: 경희대학교 포스트모던 음악학과 석사과정
※관심분야 : AI 음성 변환(AI Voice Conversion), 오디오 소프트웨어(Audio Software), 컴퓨터음악(MIDI) 등

이철희(Chul-Hee Lee)

2009년 : 경희대학교 Post Modern 음악전공 (음악학사)
2013년 : 경희대학교 아트퓨전디자인 대학원 (음악학석사)
2020년 : 경희대학교 대학원 (응용예술학박사)
2007년～2021년: 마인드오프너 음악감독
2021년～현 재: 경희대학교 포스트모던음악학과 조교수
※관심분야 : 컴퓨터음악(MIDI), 작곡(Composition), 사운드디자인(Sound Design) 등