JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 11, pp. 37-62, November 2025
https://doi.org/10.9708/jksci.2025.30.11.037

Multi-Domain ESQ Metrics for Quality Assessment of Augmented Emotional Speech

Do Kyung Shin*, Young Dae Kim*

*Chief Research Engineer, Maritime Lab., LIG NEX1 Co., Ltd., Gyeonggi-do, Korea

[Abstract]

Recent advances in Automatic Speech Recognition (ASR) technology have driven active research in Speech Emotion Recognition (SER) applications. While SER performance heavily depends on data quality and quantity, data scarcity remains a persistent challenge, making data augmentation techniques essential. Existing voice quality evaluation metrics such as PESQ and STOI are single-dimensional evaluation methods, and have the disadvantage of not being able to comprehensively evaluate the quality of audio data with high-dimensional and multi-dimensional characteristics, such as emotional speech. This study proposes ESQ (Emotion-Specific Quality Assessment) metrics for evaluating the quality of augmented emotional speech data. To validate the ESQ metrics, we utilized the EMO dataset augmented across quality levels using MetricGAN. Experimental results demonstrate consistent score improvements across all seven groups as quality levels increase, achieving an overall average improvement rate of 90.75%.

▸ **Key words**: SER, Audio Quality Assessment, Data Augmentation, PESQ, STOI, MetricGAN

[요 약]

최근 자동음성인식(ASR) 기술 발전과 함께 음성 감정 인식(SER) 기반 응용 연구가 활발히 진행되고 있다. 음성 감정 인식 성능은 데이터의 품질과 양에 크게 의존하며, 여전히 데이터 부족 문제가 존재하여 데이터 증강 기법 연구가 필수적이다. PESQ, STOI 등의 기존 음성 품질 평가 메트릭은 단일 차원의 평가 방식으로써, 감정 발화 음성과 같은 고차원적, 다차원적 특성을 갖는 오디오 데이터의 품질을 종합적으로 평가하지 못하는 단점이 존재한다. 본 연구는 증강된 감정 음성 데이터의 품질 평가를 위한 ESQ(Emotion-Specific Quality Assessment) 메트릭을 제안한다. ESQ 메트릭의 유효성 검증을 위해 MetricGAN으로 품질 레벨별 증강을 수행한 EMO 데이터 셋을 활용하였다. 실험 결과, 7개 그룹에서 품질 레벨 증가에 따른 일관된 평가 점수 향상을 확인하였으며, 전체 평균 향상률은 90.75%를 달성하였다.

▸ **주제어**: 음성 감정 인식, 오디오 품질 평가, 데이터 증강, PESQ, STOI, MetricGAN

• First Author: Do Kyung Shin, Corresponding Author: Do Kyung Shin
*Do Kyung Shin (dokyung.shin@lignex1.com), Maritime Lab., LIG NEX1 Co., Ltd.
*Young Dae Kim (youngdae.kim@lignex1.com), Maritime Lab., LIG NEX1 Co., Ltd.
• Received: 2025. 09. 17, Revised: 2025. 10. 16, Accepted: 2025. 11. 17.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

최근 인공지능(AI)의 발전과 함께 음성 신호를 활용하여 화자의 감정을 자동으로 인식하는 음성 감정 인식(SER, Speech Emotion Recognition) 기술은 인간-컴퓨터 상호작용(HCI), 대화형 에이전트, 헬스케어, 감정 기반 맞춤형 서비스 등 다양한 응용 분야에서 중요한 연구 주제로 부상하고 있다[1-3]. SER의 성능은 학습에 활용되는 데이터의 품질과 양에 크게 좌우되며, 특히 감정 발화 음성은 수집 과정에서 화자의 주관적 표현, 상황적 맥락, 언어적 제약 등의 요인으로 인해 대규모로 확보하기 어렵다. 이러한 이유로 SER 분야에서는 데이터 부족 문제를 해결하기 위한 다양한 데이터 증강(data augmentation) 기법이 활발히 연구되고 있다. 대표적인 음성 데이터 증강 기법에는 신호 변형 기반(예: 피치 변환, 시간 늘이기/줄이기, 잡음 추가), 특징 변환 기반(예: 스펙트럼 왜곡, 보코더 변환), 그리고 생성 모델 기반(예: GAN, VAE, Diffusion 모델 등) 방법이 있다. 이와 같은 방법들은 부족한 감정 발화 데이터를 보완하고 학습의 일반화 성능을 향상에 기여하고 있다. 그러나 동일한 증강 방법이라 하더라도 학습 데이터의 특성과 응용 목적에 따라 결과에 미치는 영향은 크게 달라질 수 있다. 증강 데이터를 적용하여 학습 데이터의 양을 증가시켜도 증강 유형, 강도, 데이터 조건에 따라 성능에 미치는 영향이 크게 달라진다. 따라서 적용 분야 및 과제 특성에 맞는 증강 선정이 성능을 좌우하기 때문에 증강된 데이터의 품질 검증 시스템이 필요하다[4]. 이와 같이, 증강된 데이터의 유효성을 사전에 검증하는 과정이 필수적임에도 불구하고, 실제로 증강된 데이터의 품질을 체계적으로 측정·분석하는 연구는 상대적으로 부족한 실정이다. 기존에 사용되는 대표적인 음성 품질 측정 지표로는 PESQ (Perceptual Evaluation of Speech Quality)[5], STOI (Short-Time Objective Intelligibility)[6-7], SNR (Signal-to-Noise Ratio)이 있다. PESQ는 ITU-T 표준으로 제안된 음질 평가 지표로, 원본 음성과 처리된 음성을 비교하여 사람의 청각적 지각 특성을 반영한 점수가 산출된다. 하지만 통신 음성의 전송 품질을 평가하기 위한 목적으로 설계되었기 때문에 감정 표현의 핵심인 운율 변화, 억양 패턴, 발화 리듬 등의 음향 특성은 구분하지 못함으로 음성의 세부 특성의 품질을 평가할 수 없는 단점이 존재한다. STOI는 주로 음성의 명료도(Intelligibility)를 평가하는 데 사용된 다차원적 특징 정보를 종합적으로 분석할 수 없는 단점이 존재한다. 따라서 본 연구에서는 기존의 PESQ, STOI, SNR 품질 지표를 이용하여 잡음 환경에서의 음성 이해도를 정량화한다. 본 지표는 음소(phoneme) 수준의 이해 및 음성 내용의 인식 가능성을 평가하지만, 감정의 표현 요소인 음색, 음량 변화 등의 품질 평가가 불가능한 단점이 존재한다. SNR은 물리적 신호대 잡음비를 나타내는 단순한 지표로, 음질의 지각적 특성을 직접적으로 반영하지는 못한다. 기존 음질 평가 지표는 통신 음성이나 잡음 제거 알고리즘의 성능 평가에는 유용하지만, 감정 발화 음성과 같이 운율적(prosodic), 음향적(acoustic), 시간적(temporal), 지각적(perceptual), 감정적(emotional)의 다차원적 특성들이 상호 독립적이지 않고 복잡하게 얽혀 있기 때문에 단일 차원 평가 지표로는 종합적인 감정 음성 품질 평가가 불가능한 단점이 존재한다. 따라서 감정 발화 음성의 주요 세부 특성에 대한 사전 품질 검증 및 선별이 필수적이다[8-11].

최근 감정 발화 오디오의 품질을 검증하기 위한 다양한 접근이 시도 되어 왔다. 첫째, 가장 널리 활용되는 방법은 전통적 음성 품질 지표(PESQ, STOI 등)를 적용하는 방식이다. Wu[12] 등은 증강된 감정 발화 음성의 품질을 비교하거나 데이터 증강 기법의 효과를 검증하기 위해 PESQ와 STOI를 그대로 활용하였다. 그러나 이러한 지표들은 본래 통신 음성 품질 평가를 목적으로 설계되었기 때문에, 감정 발화에서 중요한 감정 단서나 프로소디(prosody) 변이를 충분히 반영하지 못하는 단점이 존재한다. 둘째, 일부 연구에서는 주관적 평가(MOS, Mean Opinion Score)를 자동화하기 위해 딥러닝 기반 품질 예측 모델을 도입하였다. 대표적으로 Lo 등[13]은 MOSNet을 통해 대규모 주관 평가 데이터를 학습하여 사람이 평가한 MOS 점수를 예측하는 방법을 제안하였으며, Mittag 등[14]은 비침습적 음성 품질 평가 모델(NISQA)을 제안하여 잡음, 왜곡, 명료도 등 다차원적인 품질 요소를 동시에 예측할 수 있음을 보였다. 이러한 학습 기반 MOS 예측기는 감정 발화에 확장 적용될 가능성이 있으나, 여전히 일반 음성의 음질 품질 예측에 최적화되어 있으며, 감정 표현의 자연스러움이나 일관성과 같은 고차원적 속성을 직접 반영할 수 없는 단점이 존재한다. 셋째, 감정 발화에서 중요한 특성으로 알려진 프로소디 요소(F0, 에너지, 지속 시간)를 활용한 품질 분석 접근 방법이 있다. Burkhardt와 Sendlmeier[15]는 감정 발화에서 F0와 에너지 패턴이 감정 구별에 핵심적임을 제시하였으며, 이후 연구들에서도 증강된 발화의 품질을 검증하기 위해 프로소디 변화를 정량적으로 분석하는 방법이 사용되었다[16]. 그러나 이 접근은 특정 감정 단서를 개별적으로 측정할 수는 있으나, 전체적인 청취 품질이나 감정과 같은 전통적 음성 품질 지표의 단점을 보완

하고, 감정 발화 오디오 데이터의 특성을 반영할 수 있는 새로운 품질 측정 메트릭을 제안한다. 제안하는 메트릭은 크게 시간 도메인(Time Domain), 주파수 도메인(Frequency Domain), 운율(Prosodic), 지각적 품질(Perceptual Quality), 감정 일관성(Emotional Consistency), 분포 유사성(Distribution Similarity), 클러스터링 품질(Clustering Quality)의 총 7개 그룹 지표 항목을 기반으로 구성되며, 그룹별 중요도에 따른 차등 가중치를 적용하여 0~1 범위의 종합 품질 점수를 산출한다. 최종 점수는 Excellent(S), Very Good(A), Good(B), Poor(C), Very Poor(D)의 5단계 품질 등급으로 분류하여 직관적으로 품질을 평가한다. 본 연구에서 제안하는 ESQ 메트릭은 데이터 증강 과정에서 원본 감정 발화의 핵심 특징들이 얼마나 잘 보존되었는지 정량적으로 측정할 수 있으며, 품질 기준을 만족하는 데이터만 선별하는 필터링 도구로 활용할 수 있다. 또한 32개 세부 지표를 통해서 손실된 특징 정보 진단을 통해서 증강 모델의 성능 저하의 원인을 사전 파악하여 개선 방향을 추적하는 피드백 도구로 활용될 수 있다.

본 논문에서는 제안한 ESQ 메트릭의 7개 그룹 및 32개 세부 지표에 대한 당위성 검증을 위해, 서로 다른 원리와 특성을 가진 세 가지 음성 증강 기법으로 생성된 데이터셋을 활용하여 비교 분석을 수행한다. 비교 분석을 위한 모델은 VTLP(Vocal Tract Length Perturbation)[17], StarGAN-VC(Star Generative Adversarial Network for Voice Conversion)[18], CycleGAN-VC(CycleGAN for Voice Conversion)[19]이며, 각 모델의 선정 근거는 다음과 같다. VTLP는 주파수 축 선형 변환을 통한 전통적 신호 처리 기반 증강 기법으로, 시간 영역 특성과 운율 구조를 원본 그대로 보존하는 특성을 가진다. 이러한 특성은 ESQ의 Time Domain, Frequency Domain, Prosodic 그룹 지표들이 데이터 증강 시, 원본의 고유 특성을 보존하는지를 평가 검증하기에 적합하며, 딥러닝 기반 모델과의 비교를 위한 기준선(baseline)으로 활용된다. StarGAN-VC는 다중 도메인(multi-domain) 학습을 통해 여러 감정 클래스 간 변환을 수행하는 생성적 적대 신경망 기반 모델로, 단일 모델로 다대다(many-to-many) 감정 변환이 가능하다. 이러한 특성은 ESQ의 Emotional Consistency와 Clustering Quality 그룹 지표들이 복수 감정 간 변환 시 감정 특성 보존 및 클래스 분리도 적정성 평가 검증을 위해 활용된다. CycleGAN-VC는 cycle consistency loss를 통해 언어 정보와 화자 정보를 보존하면서 감정만 선택적으로 변환하는 쌍을 이루지 않는 (unpaired) 학습 기반 모델이다. 이러한 특성은 ESQ의 Distribution Similarity와 Perceptual 그룹 지표들이 원본과 증강 데이터 간 통계적 분포 유사성 및 지각적 품질 적절성 평가 검증을 위해 활용된다. 따라서 전통적 신호 처리(VTLP), 다중 도메인 생성 모델(StarGAN-VC), 순환 일관성 기반 변환 모델(CycleGAN-VC)이라는 서로 다른 증강 원리를 가진 세 모델을 통해, ESQ 메트릭이 다양한 증강 기법의 특성을 다차원적으로 변별하고 각 기법의 강점과 약점을 정량적으로 분석할 수 있음을 입증한다.

II. Proposed Method

본 논문에서는 증강된 감정 오디오 데이터의 유효성을 검증하기 위해서 감정 오디오에 특화된 품질 평가 방법을 제안한다. 제안한 감정 오디오 증강 기법의 품질 평가 지표는 크게 음성학, 신호 처리, 그리고 지각적 측면을 다양하게 고려하였으며, 본 연구에서는 Time Domain, Frequency Domain, Prosodic, Perceptual Quality, Emotional Consistency, Distribution Similarity, Clustering Quality의 7개 그룹으로 정의하였으며, 각 그룹에 대한 세부 지표는 총 32개의 평가 지표로 정의한다.

본 논문에서는 감정 발화 오디오에 특화된 품질을 정량적으로 평가하기 위하여, 품질 지표를 7개 그룹 지표로 세분화하였다. 각 그룹은 감정 오디오의 특성을 다각도로 반영할 수 있도록 설계되었으며, 세부 특징은 정량적으로 계산된 후 그룹별 품질 점수로 산출된다. 예를 들어, 시간 도메인(Time Domain) 그룹 지표는 RMS energy, zero-crossing rate, duration, temporal dynamics, attack time, decay characteristics 등의 세부 지표를 포함하고, 운율(Prosodic) 그룹 지표는 pitch mean, pitch variance, tempo consistency, intonation pattern 등을 통해 감정적 표현력을 측정한다. 이와 같은 방식으로 각 7개 그룹 지표는 고유한 특성을 정량화하여 최종 품질 점수 산출에 기여한다. 최종 종합 품질 점수를 도출하기 위해, 본 연구에서는 Table 1과 같이 7개 그룹 지표별 특성의 중요도를 반영한 가중치를 설정하였다. 감정 표현과 직결되는 감정 일관성(Emotional Consistency) 지표는 감정 인식 분야에서 감정의 분류 및 감정 표현을 위한 유의미한 특징 정보를 포함한다는 결과[20]를 기반으로 감정 상태를 표현하는 핵심 요소로써 가장 높은 21%의 가중치를 할당하고, 음성의 음질과 관련된

주파수 영역(Frequency Domain) 과 지각적 품질(Perceptual Quality) 지표에는 각각 20%의 가중치를 할당하였다. 주파수 영역 특징은 음색. 음질의 음향적 품질의 기초를 반영하며, 음성의 스펙트럼 특성은 감정적 색채(밝기, 거칠기)에 영향을 준다. 지각적 품질 지표는 청취자가 지각하는 음질 및 청감의 특성을 반영[21-22]한다는 점에서 중요한 의미를 갖는다. 시간 영역(Time Domain)은 에너지 변화, 제로크로싱 레이트, 시간적 패턴 등이 감정 인식 및 표현과 연관[23]이 있는 보조적 지표로써, 16%의 가중치를 할당하였다. 분포 유사성(Distribution Similarity)은 신호의 통계적 유사성. 데이터 분포에 대한 특징 정보 지표로써 감정적 품질과 간접적 상관관계를 가지며[24], 보조 검증용으로써 12%의 가중치를 할당하였다. 운율(Prosodic)은 감정 표현 및 인식과 연관이 있는 발화의 억양. 피치. 리듬 정보를 반영[25]하는 지표로써, 10%의 가중치를 할당하였다. 마지막으로 클러스터링 품질(Clustering Quality)은 고차원 특징 공간에서 감정 정보가 임베딩 공간 내에서 클러스터 형태로 존재하며[26], 증강 데이터가 원본 데이터와 동일한 감정 클래스의 분포를 형성하는지를 확인하는 보조적 역할을 수행하는 지표로써, 1%의 가중치를 할당하였다.

이와 같이, 그룹별 가중치를 차등 적용함으로써 본 연구에서 제안한 ESQ 메트릭은 단일 지표에 의존하지 않고, 감정적·음향적·통계적 관점이 종합적으로 반영된 품질 점수를 제공한다. 이를 통해 증강된 감정 발화 데이터가 원본의 감정적 특성을 얼마나 효과적으로 보존하는지를 정량적으로 측정할 수 있다.

본 논문에서 제안하는 감정 오디오에 특화된 품질 평가 방법의 프레임워크는 Fig. 1과 같다. 프레임워크는 총 5개 단계로 구성되어 있으며, 첫 번째 단계에서 증강된 오디오의 품질을 측정하기 위해서 원본 오디오()와 증강된 오디오()가 입력된다. 두 번째 단계에서 각 원본 오디오와 증강된 오디오에 대해서 Time Domain, Frequency Domain, Prosodic, Perceptual Quality, Emotional Consistency, Distribution Similarity, Clustering Quality의 7개 그룹의 32개의 세부 특징 정보를 검출한다. 세 번째 단계에서 검출된 원본 오디오 특징 정보와 증강된 오디오 특징 정보 간의 32개의 유사도 가 계산된다. 네 번째 단계는 그룹별 가중치 평균을 계산하여, 7개 그룹의 점수를 하나의 점수로 통합 수행한다. 마지막 단계에서 Excellent, Very Good, Good, Poor, Very Poor의 5개 품질 등급으로 출력된다.

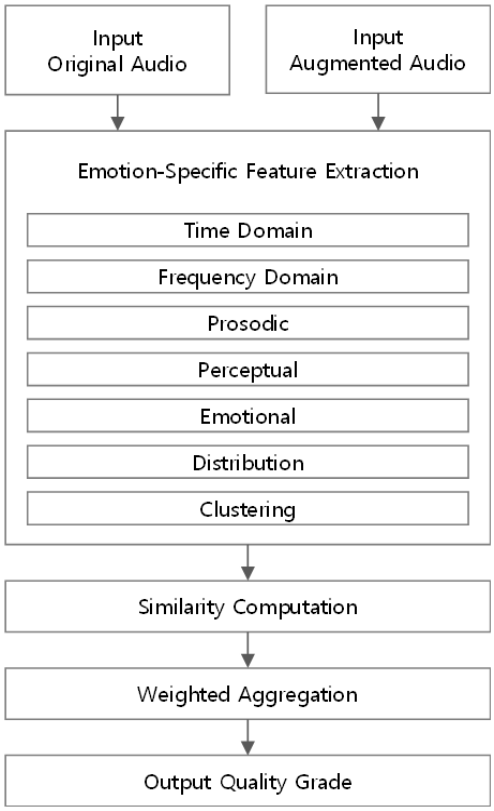

Fig. 1. Proposed Emotion-Specific Audio Quality Assessment Framework

1. Time-Domain Features

시간 영역 특징(Time-Domain Feature)은 시간 축에서 오디오 신호의 특성을 분석하는 지표로써, 원본과 증강 오디오의 기본적인 시간 특성이 얼마나 유사한지를 측정한다. 본 지표는 RMS 에너지(RMS energy), 영점 교차율(ZCR, Zero Crossing Rate), 지속 시간(duration), 시간적 변화(temporal dynamics), 어택 시간(attack time), 감쇠 특성(delay characteristics)으로 총 6개의 세부 지표로 구성된다. 시간 영역 특징의 각 세부 지표는 수식(1)~ 수식(6)과 같으며, 각 수식에서 x는 원본(original) 오디오 신호를 의미하며, $\hat{x}$는 증강(augmented) 오디오 신호를 의미한다. 각 세부 지표의 유사도 점수는 S로 표기하며, 0에서 1사이의 값을 갖는다. 유사도 점수는 1에 가까울수록 원본과 증강 신호가 유사함을 의미하고, 0에 가까울수록 상이함을 의미한다. N은 오디오 신호의 총 샘플 수를 의미하며, T는 총 프레임 수를 의미한다.

RMS 에너지[27] 유사도는 수식 (1)과 같다. 수식 (1)에서 $x(n)$은 시간 인덱스 n에서의 신호 진폭 값을 의미한다. RMS 에너지는 오디오 신호의 전체적인 음량과 세기를 나타내며, 감정 표현에서 중요한 지표이다.

$$S_{RMS}=1-\min\left(1,\frac{|RMS(x)-RMS(\hat{x})|}{RMS(x)}\right),$$
$$RMS(x)=\sqrt{\frac{1}{N}\sum_{n=1}^{N}x^2(n)} \quad (1)$$

영점 교차율[28] 유사도는 수식 (2)와 같다. 수식 (2)에서 $sign(\ )$은 부호함수를 의미하며, x가 0보다 클 경우, 1을 반환하고 x가 0보다 같거나 작을 경우 -1을 반환한다. 영점 교차율은 신호가 0을 교차하는 빈도로, 음성의 유성음(voiced)과 무성음(unvoiced) 특성과 관련된다. 낮은 ZCR은 주기적인 유성음일 가능성이 높으며, 높은 ZCR은 잡음성 무성음일 가능성이 높다. 감정에 따라 발화의 유성음과 무성음의 비율과 발음 특성이 달라지기 때문에 감정 별 특성을 반영하는 중요한 지표이다.

$$S_{ZCR}=1-\min\left(1,\frac{|ZCR(x)-ZCR(\hat{x})|}{\max(ZCR(x),ZCR(\hat{x}))}\right),$$
$$ZCR(x)=\frac{1}{N-1}\times\sum_{n=1}^{N-1}\left(\frac{|sign(x(n+1)-sign(x(n)))|}{2}\right), \quad (2)$$
$$sign(x(n))=\begin{cases}1, & \text{if } x(n)>0\\ -1, & \text{if } x(n)\le 0\end{cases}$$

지속 시간[29] 유사도는 수식 (3)과 같다. 수식 (3)에서 D는 신호 지속 시간을 의미하며, 단위는 초(second)이다. N은 각 원본과 증강된 오디오 신호의 총 샘플 수를 의미한다. fs는 샘플링 주파수(sampling rate)를 의미한다. 지속 시간은 감정 표현에서 발화 속도(speed rate)와 휴지기(pause)는 중요한 요소이다. 감정에 따라, 흥분 또는 분노와 같은 감정 상태에서는 빠른 발화 속도로 짧은 지속 시간을 보이는 경향이 있으며, 차분하거나 슬픈 감정 상태에서는 느린 발화 속도와 긴 휴지기를 보이는 경향이 있다.

$$S_D=1-\min\left(1,\frac{|D(x)-D(\hat{x})|}{\max(D(x),D(\hat{x}))}\right),$$
$$D(x)=N/fs \quad (3)$$

시간적 변화[30] 유사도는 수식 (4)와 같으며, $TD(x)$는 시간적 변화율을 의미한다. 수식 (4)에서 σ는 표준편차, μ는 평균, △RMS는 프레임별 RMS 에너지 벡터이다. 시간적 변화는 감정 표현의 역동성과 강도를 나타내는 지표이다. 감정에 따라, 흥분이나 분노과 같은 고각성(high arousal) 감정은 에너지가 급격하고 불규칙하게 변화하고, 강조(emphasis) 또는 악센트(accent) 표현 시 특정 음절에서 에너지가 급증하여 높은 TD 값을 보이며, 슬픔과 같은 저각성(low arousal) 감정은 에너지 변화가 완만하고 안정적이어서 낮은 TD값을 보이는 경향이 있다.

$$S_{TD}=1-\min\left(1,\frac{|TD(x)-TD(\hat{x})|}{\max(TD,TD(\hat{x}))}\right),$$
$$TD(x)=\frac{\sigma(\triangle RMS(x))}{\mu(RMS(x))} \quad (4)$$

어택 시간[31] 유사도는 수식 (5)와 같다. 수식 (5)에서 n_onset은 첫 번째 음향 $onset$이 발생하는 샘플 인덱스를 의미하며, fs는 샘플링 주파수를 의미한다. n_onset 수식에서 $E_local(n)$은 샘플 n 주변의 국소 에너지, $E_\max$는 전체 신호의 최대 에너지, θ는 $onset$ 검출 임계값을 의미한다. $\arg\min$은 조건을 만족하는 가장 작은 n값을 반환한다. 본 논문에서는 θ를 0.1(10%)을 적용하였다. E_local 수식에서 w는 국소 윈도우의 반경(radius)을 의미하며, 본 논문에서는 400샘플(25ms at 16kHz)을 사용하였으며, $2w+1$은 윈도우 크기를 의미한다. 어택 시간은 오디오 신호가 시작되는 시점의 특성을 나타내며, 음성의 발화 개시(speech onset) 패턴을 반영한다. 감정에 따라, 흥분과 분노와 같은 감정은 갑작스럽고 강한 발화 시작으로 인해 짧은 어택 시간을 보이며, 슬픔과 같은 감정은 부드럽고 점진적인 발화 시작으로 인해서 상대적으로 긴 어택 시간을 보이는 경향을 보인다.

$$S_{AT}=1-\min\left(1,\frac{|AT(x)-AT(\hat{x})|}{\max(AT(x),AT(\hat{x}))}\right),$$
$$AT(x)=n\_onset/fs,$$
$$E\_local(n)=\frac{1}{2w+1}\sum_{i=n-w}^{n+w}x^2(i), \quad (5)$$
$$E\_\max=\max E\_local(n),\ n=w+1,\ldots,N-w$$

감쇠 특성[32] 유사도는 수식 (6)과 같으며, DC는 각 오디오의 에너지 감쇠 비율(decay ratio)을 의미한다. 수식 (6)에서 N은 전체 샘플 수를 의미하며, N_half는 전체 샘플 수의 절반을 의미한다. E_first_half는 전반부 신호(0~50%) 구간의 평균 에너지를 의미하며, E_Second_half는 후반부 신호(50~100%) 구간의 평균 에너지를 의미한

다. 감쇠 특성은 음성의 에너지가 시간에 따라 감쇠 하는 패턴을 나타내며, 감정별로 다른 감쇠 패턴 특성을 나타낸다. 높은 유사도 값은 원본의 에너지 감쇠 패턴이 잘 보존되었음을 의미한다.

$$
\begin{aligned}
&S_{DC} = 1 - \min\left(1, \frac{|DC(x) - DC(\hat{x})|}{\max(DC(x), DC(\hat{x}))}\right),\\
&DC(x) = \\
&\frac{E\_first\_half(x) - E\_second\_half(x)}{E\_first\_half(x)} \times 100\\
&E\_first\_half = \frac{1}{N\_half} \times \sum_{n=1}^{N\_half} |x(n)|^2\\
&E\_second\_half = \frac{1}{N\_half} \times \sum_{n=N\_half+1}^{N} |x(n)|^2
\end{aligned}
\tag{6}
$$

2. Frequency-Domain Features

주파수 영역 특징은 주파수 축에서 오디오 신호의 스펙트럼 특성을 분석하는 지표로써, 음성의 운율적 특성을 분석한다. 주파수 영역 특징 지표는 스펙트럼 중심(spectral centroid), 스펙트럼 대역폭(spectral bandwidth), 스펙트럼 롤-오프(spectral roll-off), 스펙트럼 대비(spectral contrast), MFCC 유사성(MFCC similarity), 하모닉 비율(harmonic ratio)으로 총 6개의 세부 지표로 구성된다. 주파수 영역 특징의 각 세부 지표는 수식 (7) ~ 수식 (12)와 같다.

스펙트럼 중심[33] 유사도는 수식 (7)과 같으며, $f(k)$는 k번째 주파수 빈의 주파수 값(Hz)을 의미한다. $|X(k)|^2$은 k번째 주파수 빈의 스펙트럼 크기를 의미하며, K는 총 주파수 빈 개수를 의미한다. 스펙트럼 중심은 주파수 분포의 무게중심으로 음색의 밝기를 나타낸다. 따라서 감정에 따라 다른 주파수 분포를 갖기 때문에 감정별 특성을 검출할 때 중요한 지표이다.

스펙트럼 대역폭[33] 유사도는 수식 (8)과 같다. 수식 (8)에서 k는 주파수 빈 인덱스를 의미하며, N는 총 주파수 빈 개수로써 FFT 크기의 절반을 의미한다. $S(k)$는 k번째 주파수 빈에서의 스펙트럼 크기(magnitude)를 의미한다. SC는 스펙트럼 중심값으로 수식 (7)에서 계산된 값을 의미한다. $(k-SC)^2$은 각 주파수 빈이 스펙트럼 중심으로부터 떨어진 제곱의 값을 나타낸다. 스펙트럼 대역폭은 주파수 분포의 퍼짐 정도를 나타내며, 음성의 풍부함과 관련된 특징 지표이다. 감정 발화에서 대역폭은 감정의 강도 및 표현 방식과 관련이 있다. 예를 들어, 흥분된 감정은 넓은 대역폭을 나타내며, 차분한 감정은 상대적으로 좁은 대역폭을 나타내는 경향이 있다.

$$
\begin{aligned}
&S_{SC} = 1 - \min(1, |SC(x) - SC(\hat{x})| / SC(x)),\\
&SC = \frac{\sum_{k=1}^{K} f(k) \times |X(k)|^2}{\sum_{k=1}^{K} |X(k)|^2},\\
&f(k) = \frac{k \times fs}{2K}
\end{aligned}
\tag{7}
$$

$$
\begin{aligned}
&S_{SB} = 1 - \min\left(1, \frac{\| SB(x) \bullet SB(\hat{x}) \|}{\max(SB(x), SB(\hat{x}))}\right),\\
&SB = \sqrt{\frac{\sum_{k=1}^{N} S(k) \times (k - SC)^2}{\sum_{k=1}^{N} S(k)}}
\end{aligned}
\tag{8}
$$

스펙트럼 롤오프[34] 유사도는 수식 (9)와 같다. 수식 (9)에서 k는 주파수 빈 개수(FFT 크기의 절반)를 의미하며, $S(i)$는 i번째 주파수 빈의 스펙트럼 크기(magnitude)를 의미한다. R은 롤-오프 임계 값을 의미하며, $\arg\min$은 조건을 만족하는 가장 작은 k값을 반환한다. 본 논문에서는 R은 85%를 적용하였다. 즉, 스펙트럼 롤-오프는 전체 스펙트럼 에너지의 85%가 집중되는 주파수 경계를 의미하며, 신호의 고주파 성분 분포를 측정한다. 낮은 롤-오프 값은 에너지가 주로 저주파 대역에 집중되고 있음을 의미하며, 높은 롤-오프 값은 고주파 성분이 많이 포함되어 밝고 날카로운 음색을 의미한다.

$$
\begin{aligned}
&S_{SR} = 1 - \min\left(1, \frac{|SR(x) - SR(\hat{x})|}{\max(SR(x), SR(\hat{x}))}\right),\\
&SR(x) = \arg\min_k \left\{ \sum_{i=1}^{k} S(i) \geq R \times \sum_{i=1}^{N} S(i) \right\}
\end{aligned}
\tag{9}
$$

스펙트럼 대비[35] 유사도는 수식 (10)과 같으며, B는 서브 밴드(sub-band)의 개수를 의미한다. $Peak_b$는 b번째 서브 밴드의 피크 에너지, $Valley_b$는 b번째 서브 밴드의 밸리 에너지를 의미하며, $\log_{10}$은 상용로그를 의미한다. 본 논문에서는 B는 7개 서브 밴드를 사용하였다. 스펙트럼 대비는 각 주파수 대역에서 에너지 피크와 밸리 간의 차이

를 측정하는 지표이다. 높은 대비 값은 해당 대역의 에너지가 불균일하고 변화가 큰 것을 의미하며, 낮은 대비 값은 에너지가 고르게 분포되어 있음을 의미한다. 각 서브 밴드별로 독립적으로 계산하고 평균을 취함으로써 전 대역에 걸친 스펙트럼 구조의 복잡성과 명료도를 정량화한다. 감정 발화에서 스펙트럼 대비는 음성의 명료도(clarity)와 선명도(sharpness)를 반영한다. 감정에 따라, 흥분이나 분노와 같은 감정은 강하고 명확한 발성으로 인해 높은 대비를 보이는 경향이 있으며, 슬픔과 같은 감정은 부드럽고 균일한 발성으로 낮은 대비를 보이는 경향이 있다.

$$S_{ST} = 1 - \min\left(1, \frac{|ST(x) - ST(\hat{x})|}{\max(ST(x), ST(\hat{x}))}\right),$$
$$ST(x) = \frac{1}{B}\sum_{b=1}^{B} \log_{10} \frac{Peak\_b(x)}{Valley\_b(x)} \quad (10)$$

MFCC[36] 유사도는 수식 (11)과 같다. 수식 (11)에서 $\cos ine_distance$은 코사인 거리 함수를 의미한다. 수식 (11)에서 $\bullet$는 벡터 내적을 의미하며, $\| \bullet \|$은 벡터의 크기 (norm)을 의미한다. T는 총 프레임 수, $MFCC_t(x)$는 t번째 프레임의 MFCC 특징벡터를 의미한다. $MFCC\_t(x)$ 수식에서 M은 MFCC의 계수, c_i는 i번째 MFCC 계수를 의미한다. 본 논문에서는 M을 13으로 설정하였다. MFCC는 인간의 청각 인지 특성을 반영한 멜 스케일(Mel scale) 기반의 켑스트럼(cepstrum) 특징으로, 음성 신호의 스펙트럼 포락선(spectral envelope)을 효과적으로 표현한다. 멜 스케일은 저주파에서는 선형적으로, 고주파에서는 로그로 변환하여 인간의 주파수 지각 특성을 모델링한다. 감정 발화에서 MFCC는 화자의 성도(vocal tract) 특성과 발성 방식을 반영한다.

하모닉 비율[37] 유사도는 수식 (12)와 같다. $MFCC$는 평균 MFCC 벡터를 의미한다. 수식 (12)에서 $E_harmonic$은 조화음 성분 에너지를 의미하며, E_total은 전체 신호의 에너지를 의미한다. N은 총 샘플수를 의미하며, $x(n)$은 시간 n에서의 신호 값을 의미한다. 조화음 성분 에너지는 기본 주파수와 조화음 성분의 에너지 합으로 계산된다. K는 조화음의 개수, E_k는 k번째 조화음 성분의 에너지를 의미한다. 본 논문에서는 10개의 조화음을 적용하였다. 조화음 비율은 주기적인 조화음 성분과 비주기적인 잡음 성분의 비율을 나타내는 지표이다. 높은 값(0.7~1.0)은 유성음(voiced sound)일 확률이 높으며, 낮은 값(0~0.3)은 무성음(unvoiced sound)이거나 잡음 성분이 많음을 의미한다. 또한 중간 값(0.3~0.7)은 혼합 음성 또는 전환 구간일 확률이 높다.

$$S_{MFCC} = 1 - cosine\_distance(MFCC(x) - MFCC(\hat{x})),$$
$$\cos ine\_distance(u,v) = 1 - \frac{MFCC(x) \bullet MFCC(\hat{x})}{\| MFCC(x) \| \times \| MFCC(\hat{x}) \|}, \quad (11)$$
$$MFCC(x) = \frac{1}{T}\sum_{t=1}^{T} MFCC\_t(x),$$
$$MFCC\_t(x) = [c_1, c_2, ..., c_M]$$

$$S_{HR} = 1 - \min\left(1, \frac{|HR(x) - HR(\hat{x})|}{\max(HR(x) - HR(\hat{x}))}\right),$$
$$HR(x) = E\_harmonic(x) / E\_total(x),$$
$$E\_total(x) = \sum_{n=1}^{N} x^2(n), \quad (12)$$
$$E\_harmonic(x) = \sum_{k=1}^{K} E_k$$

3. Prosodic Features

운율 특징은 음성의 운율적 특성을 분석하는 지표로써, 감정 인식에 핵심적인 운율 특성의 보존 정도를 측정한다. 운율 특징 지표는 평균 피치(pitch mean), 피치 분산(pitch variance), 템포 일관성(temp consistency), 억양 패턴(intonation pattern)으로 총 4개의 세부 지표로 구성된다.

평균 피치[38] 유사도는 수식 (13)과 같다. 수식 (13)에서 PM는 평균 피치 주파수(Hz)를 의미한다. T_{voiced}는 유성음(voiced) 프레임의 개수, PM_t은 t번째 유성음 프레임의 기본 주파수를 의미한다. 무성음(unvoiced) 프레임은 평균 계산에서 제외된다. 유성음과 무성음의 판별은 자기상관 피크 값으로 결정된다. $\tau_{\max}$는 자기상관 최댓값의 지연을 의미한다. 피치 범위는 80~400 Hz로 제한되며, 자기상관 피크가 0.3 이상인 프레임만 유성음으로 판별하여 평균 계산에 포함된다.

$$S_{PM} = 1 - \min\left(1, \frac{|PM(x) - PM(\hat{x})|}{\max(PM(x), PM(\hat{x}))}\right),$$
$$PM(x) = \frac{1}{T_{voiced}} \times \sum_{t=1}^{T_{voiced}} (PM_t), \quad (13)$$
$$PM_t = f_s / \tau_{\max}$$

피치 분산[39] 유사도는 수식 (14)와 같다. 수식 (14)에서 $PV(x)$는 신호의 피치 변동 계수(coefficient of variation of pitch)를 의미한다. 변동 계수는 표준편차를 평균으로 나는 상대적 변동성 지표로써 무차원(dimensionless) 값이다. σ_PV는 피치 표준편차, PV_mean은 평균 피치를 의미한다. T_voiced는 유성음 프레임 개수를 의미하고, PV_t 는 t번째 유성음 프레임의 기본 주파수를 의미한다. 피치 변동 계수는 발화 내에서 피치가 얼마나 변동하는지 나타내는 정규화된 지표이다.

$$
\begin{aligned}
S_{PV} &= 1 - \min\left(1, \frac{|PV(x) - PV(\hat{x})|}{\max(PV(x), PV(\hat{x}))}\right), \\
PV(x) &= \frac{\sigma\_PV(x)}{PV\_mean(x)}, \\
\sigma\_PV(x) &= \sqrt{\frac{\sum_{t=1}^{T\_voiced} (PV_t - PV\_mean)^2}{T\_voiced}},
\end{aligned}
\tag{14}
$$

템포 일관성[40] 유사도는 수식 (15)와 같으며, TC는 템포 일관성(tempo consistency)을 계산하는 함수이다. CV_{BI}는 비트 간 간격의 변동 계수이며, BI 는 비트 간 간격(Beat Interval)을 의미한다. 비트는 에너지 포락선의 지역 최댓값 중 임계값($E_{\max} \times 0.3$) 이상인 피크로 검출된다. $E_{\max}$는 최대 에너지를 의미한다. CV_{BI}에서 σ_{BI}는 비트 간 간격의 표준편차, μ_{BI}는 평균 비트 간 간격을 의미한다. 템포 일관성은 발화 리듬의 규칙성을 나타낸다. TC가 1에 가까울수록 규칙적인 리듬을 의미하며, 0에 가까울수록 불규칙한 리듬을 의미한다.

$$
\begin{aligned}
S_{TC} &= 1 - \min\left(1, \frac{|TC(x) - TC(\hat{x})|}{\max(TC(x), TC(\hat{x}))}\right), \\
TC(x) &= \frac{1}{1 + CV_{BI}(x)} \\
CV_{BI}(x) &= \frac{\sigma_{BI}(x)}{\mu_{BI}(x)}
\end{aligned}
\tag{15}
$$

억양 패턴[41] 유사도는 수식 (16)과 같으며, $IC(x)$는 신호의 억양 복잡도(Intonation Complexity)를 계산하는 함수이다. $\mu_|\triangle f0|$는 피치 1차 차분의 절댓값 평균을 의미하며, $\mu_|\triangle^2 f0|$는 피치 2차 차분의 절댓값 평균을 의미한다. c는 정규화 상수로써, 본 논문에서는 50Hz를 사용하였다. $\triangle f0$는 인접한 유성음 프레임 간의 피치 변화량을 의미하며, T_{voiced}는 유성음 프레임 개수를 의미한다. $\triangle^2 f0$는 피치 변화율의 변화로써 피치 가속도(pitch acceleration)를 의미한다. 각 1차, 2차 차분의 절댓값 평균을 사용하는 이유는 피치의 상승/하강 방향과 무관하게 변화의 크기만을 측정하기 위함이다. 감정 발화에서 억양 복잡도는 감정 표현의 다양성과 강도를 반영한다.

$$
\begin{aligned}
S_{IC} &= 1 - \min\left(1, \frac{|IC(x) - IC(\hat{x})|}{\max(IC(x), IC(\hat{x}))}\right), \\
IC(x) &= \frac{\sqrt{\mu\_|\triangle f0| + \mu\_|\triangle^2 f0|}}{c}, \\
\mu\_|\triangle f0| &= \frac{\sum_{t=1}^{T_{voiced}-1} |\triangle f0_t|}{T_{vooiced-1}}, \\
\mu\_|\triangle^2 f0| &= \frac{\sum_{t=1}^{T_{voiced}-2} |\triangle^2 f0_t|}{T_{vooiced-2}}
\end{aligned}
\tag{16}
$$

4. Perceptual Features

지각적 품질(Perceptual Quality) 특징은 인간의 청각적 지각에 기반한 품질 측정 지표로써, 사람이 실제로 느끼는 주관적 음질의 유사성을 측정한다. 지각적 품질 특징 지표는 에너지 유사성(energy similarity), 포먼트 코사인 유사성(formant cosine similarity), 음량 유사성(loudness similarity), 거칠기 측정(roughness measure), 선명도 지수(clarity index)로 총 5개의 세부 지표로 구성된다.

에너지 유사도[42]는 수식 (17)과 같다. 수식 (17)에서 $E(x)$는 신호의 총 에너지를 의미한다. N은 신호의 총 샘플 수, $x(n)$은 시간 n에서의 신호 진폭 값을 의미한다. 총 에너지는 신호의 전체적인 음향 수준과 세기를 나타내는 지표로써, 인간이 인지하는 소리의 크기를 정량화한다. 강조(emphasis)한 악센트(accent)가 많은 발화일수록 에너지 피크로 인해 총 에너지가 증가한다.

$$
\begin{aligned}
S_E &= 1 - \min\left(\frac{|E(x) - E(\hat{x})|}{\max(E(x), E(\hat{x}))}\right), \\
E(x) &= \sum_{n=1}^{N} x^2(n)
\end{aligned}
\tag{17}
$$

포먼트 코사인 유사도[43]는 수식 (18)과 같으며, $FC(x)$는 신호의 스펙트럼 벡터를 의미한다. S_{FC}는 코사

인 유사도를 계산하는 형태로, 값의 범위는 0~1 사이로, 1에 가까울수록 유사도가 높음을 의미한다. T는 총 프레임 수를 의미하고, $|FFT_t(x)|$는 t번째 프레임의 FFT 크기 스펙트럼을 의미한다. 각 프레임의 스펙트럼을 평균하여 발화 전체의 대표 스펙트럼을 계산한다. 포먼트는 음성의 모음 특성을 결정하는 주파수 영역으로, 음성의 명료도와 인식률을 위한 특징을 검출한다. 각 감정 별로 다른 포먼트 패턴을 보인다.

$$S_{FC} = \frac{FC(x) \cdot FC(\hat{x})}{\| FC(x) \| \cdot \| FC(\hat{x}) \|}, \quad FC(x) = \frac{1}{T}\sum_{t=1}^{T} |FFT_t(x)| \tag{18}$$

음량 유사도[44]는 수식 (19)와 같다. 수식 (19)에서 $L(x)$는 신호의 음량(loudness)을 의미하며, N은 총 샘플 수를 의미한다. $x(n)$은 시간 n에서의 신호 값, P_ref는 기준 레벨(reference level)을 의미한다. 본 논문에서는 기준 레벨을 디지털 신호 표준 기준인 1.0을 사용하였으며, 단위는 데시벨(dB)이다. 음량 유사도는 인지되는 음량 수준의 보존 정도를 측정하며, 높은 값일수록 청각적으로 인지되는 음량이 유지됨을 의미한다. 감정 발화에서 음량은 감정의 강도(intensity)와 각성도(arousal)를 반영하는 핵심 지표로써, 감정별로 다른 음량 특성을 갖는다. 예를 들어, 분노는 큰 음량의 특성을 보이고 슬픔은 작은 음량의 특성을 보인다.

$$S_L = 1 - \min\left(1, \frac{|L(x) - L(\hat{x})|}{\max(L(x), L(\hat{x}))}\right), \quad L(x) = 10 \times \left(\log_{10}\left|\frac{\sum_{n=1}^{N} x^2(n)}{P\_ref^2}\right|\right) \tag{19}$$

거칠기 측정 유사도[45]는 수식 (20)과 같으며, $RM(x)$는 신호의 거칠기(roughness)를 의미하며, N은 총 샘플 수를 의미한다. $\triangle x(n)$은 인접 샘플 간의 차이(1차 미분)를 의미하며, $\mu_\triangle$은 차분의 평균을 의미한다. 거칠기는 신호의 순간적 변화 정도를 나타내며, 음성의 매끄러움이나 거친 음을 측정한다. 차분하고 안정적인 발화는 낮은 거칠기를 보이며, 흥분되거나 거친 음성은 높은 거칠기를 보이는 경향이 있다.

$$S_{RM} = 1 - \min\left(1, \frac{|RM(x) - RM(\hat{x})|}{\max(RM(x), RM(\hat{x}))}\right), \quad RM(x) = \sqrt{\frac{\sum_{n=1}^{N-1}(\triangle x(n) - \mu\_\triangle)^2}{N-1}} \tag{20}$$

선명도 지수 유사도[46]는 수식 (21)과 같으며, $CI(x)$는 신호의 선명도(clarity)를 의미하며, N은 총 샘플 수를 의미한다. $|x(n)|$은 시간 n에서의 신호 절댓값을 의미하며, $\max|x(n)|$은 신호의 최대 절댓값을 의미한다. 선명도 지수는 신호의 진폭 균일성 정도를 나타내며 1에 가까울수록 신호가 균일하고, 0에 가까울수록 피크가 두드러지는 특징을 갖는다. 높은 CI는 균일한 진폭 분포와 안정적 발성을 의미하며, 낮은 CI는 두드러진 피크와 역동적 에너지 분포를 의미한다.

$$S_{CI} = 1 - \min\left(1, \frac{|CI(x) - CI(\hat{x})|}{\max(CI(x), CI(\hat{x}))}\right), \quad CI(x) = \frac{\sum_{n=1}^{N} |x(n)|}{N \times max_n(|x(n)|)}, \quad \max(x(n)) = \max\{|x(1)|, |x(2)|, ..., |x(N)|\} \tag{21}$$

5. Emotional Consistency

감정 일관성 특징은 감정적 특성의 보존 정도를 측정하는 지표로써, 원본의 감정적 특성이 증강 후에도 그대로 유지되는지 측정한다. 감정 일관성 특징은 감정 임베딩 유사성(emotion embedding similarity), 각성도 일관성(arousal consistency), 감정가 일관성(valence consistency)으로 총 3개의 세부 지표로 구성된다.

감정 임베딩 유사도[47]는 수식 (22)와 같다. $E(x)$는 신호의 감정 임베딩 벡터를 의미하며, 감정 임베딩 벡터는 스펙트럼 특징(F_spec), 운율 특징(F_pros), 시간 특징(F_temp) 이 벡터 연결(vector concatenation)로 통합 구성된다. 수식 (22)에서 T는 오디오의 총 프레임 수를 의미한다. $E(x)$는 45차원 감정 임베딩 벡터로써, 스펙트럼 특징은 $F_spec = [\mu_MFCC; \sigma_MFCC; \triangle\mu_MFCC]$으로 MFCC의 평균, 표준편차 델타 평균의 39차원 벡터로 구성된다. 운율 특징은 $F_pros = [\mu_f0; \sigma_f0; \mu_E; \sigma_E]$으로 피치와 에너지의 평균 및 표준편차의 4차원 벡터로 구성된다. 시간 특징은 $F_temp = [\mu_ZCR; \mu_flux]$으로 ZCR과 스펙트럼의 평균의 2차원 벡터로 구성된다. $d_\cos$는 코사인 거

리를 의미한다. 감정 임베딩은 음색, 운율, 시간 특성을 통합하여 감정의 다차원적 특성을 검출한다.

$$
\begin{aligned}
&S_{EE} = 1 - d\_\cos(E(x), E(\hat{x})),\\
&d\_\cos(.,.) = 1 - \frac{E(x) \cdot E(\hat{x})}{\| E(x) \| \cdot \| E(\hat{x}) \|}\\
&E(x) = [F\_spec(x); F\_pros(x); F\_temp(x)]
\end{aligned}
\tag{22}
$$

각성도 일관성 유사도[48]는 수식 (23)과 같으며, $A(x)$는 신호의 각성도 지표(Arousal Index)를 의미하며, 에너지 동적 특성으로부터 계산된다. E_{std}는 프레임별 에너지의 표준편차로써 에너지의 시간적 변동성을 나타내고, E_{range}는 에너지 범위로써 에너지의 변화폭을 나타낸다. ZCR_{mean}은 평균 영점 교차율로써 신호의 잡음성과 활동성을 의미한다. w_1, w_2, w_3은 각 특징의 가중치를 의미한다. 본 논문에서는 각 0.5, 0.3, 0.2를 적용하였다. 감정별 각성도는 고각성 감정일수록 에너지가 크고 급격히 변화하며 발성이 불안정한 특성을 보이며, 저각성 감정은 에너지가 작고 완만하게 변화하여 발성이 안정적인 특성을 보인다.

감정가 일관성 유사도[49]는 수식 (24)와 같다. $V(x)$는 신호의 감정가 지수를 의미하며, w_{SC}와 w_{SR}은 스펙트럼 중심 V_{SC}와 스펙트럼 롤-오프 V_{SR}의 가중치를 의미한다. 본 논문에서는 가중치를 각 0.6, 0.4로 설정하였다. SC_{ref}는 기준 스펙트럼 중심을 의미하며, 본 논문에서는 2000Hz를 설정 하였다. SR_{ref}는 기준 스펙트럼 롤-오프로써 본 논문에서는 4000Hz를 설정하였다. 감정가 일관성은 감정의 긍정과 부정 정도를 나타내며, 스펙트럼 밝기로 측정된다. 높은 주파수 성분이 많을수록 긍정으로 인지되며, 낮은 주파수 성분이 많을수록 부정으로 인지되는 특성을 보인다.

$$
\begin{aligned}
&S_A = 1 - \min\left(1, \frac{|A(x) - A(\hat{x})|}{\max(A(x), A(\hat{x}))}\right),\\
&A(x) = w_1.E_{std}(x) + w_2.E_{range}(x) + w_3.ZCR_{mean}(\\
&E_{std}(x) = \frac{\sigma(E_{frame})}{\mu(E_{frame})},\\
&E_{range}(x) = \frac{E_{\max} - E_{\min}}{E_{\max}},\\
&ZCR_{mean} = \frac{1}{T}\sum_{n=n_{\star t}}^{n_{end}} x^2(n)
\end{aligned}
\tag{23}
$$

$$
\begin{aligned}
&S_V = 1 - \min\left(\frac{|V(x) - V(\hat{x})|}{V_{range}}\right),\\
&V(x) = w_{SC} \times V_{SC}(x) + w_{SR} \times V_{SR}(x),\\
&V_{SC}(x) = \frac{SC(x) - SC_{ref}}{SC_{ref}},\\
&V_{SR}(x) = \frac{SR(x) - SR_{ref}}{SR_{ref}}
\end{aligned}
\tag{24}
$$

6. Distribution Similarity

분포 유사성 지표는 신호의 통계적 분포 특성을 비교하는 지표로써, 신호의 통계적 분포가 얼마나 유사한지 정량적으로 측정한다. 분포 유사성 지표는 코사인 유사성(cosine similarity), KL 발산 유사성(KL diverence similarity), 바서슈타인 유사성(Wasserstein similarity), JS 발산(JS divergence), 바타차리야 유사성(Bhattacharyya similarity)으로 총 5개의 세부 지표로 구성된다.

코사인 유사도[50]는 수식 (25)와 같다. 코사인 유사도는 두 신호 간의 방향적 유사도를 측정하며, 높은 값은 신호의 전체적인 패턴이 유사함을 의미한다. 수식 (25)에서 $F(x)$는 오디오 신호의 통계적 특징 벡터를 의미하며, F_{TD}는 시간 영역 특징의 대푯값으로써, 평균 시간적 변화율, 어택 시간, 감쇠 비율 등의 6차원 벡터로 구성된다. F_{FD}는 주파수 영역 특징의 대푯값으로써, 평균 스펙트럼 중심, 평균 스펙트럼 대역폭, 평균 스펙트럼 롤-오프, 평균 하모닉 비율 등의 19차원으로 구성된다. F_{PR}은 운율 특징의 대푯값으로써, 템포, 억양 등의 4차원 벡터로 구성되며, 모든 특징을 결합한 29차원의 벡터로 구성된다. 코사인 유사도는 특징 벡터의 방향적 유사성을 측정함으로써 1에 가까울수록 특징 패턴의 유사도가 높음을 의미한다.

$$
\begin{aligned}
&S_{CS} = \frac{F(x) \cdot F(\hat{x})}{\| F(x) \| \cdot \| F(\hat{x}) \|},\\
&F(x) = [F_{TD}(x); F_{FD}(x); F_{PR}(x)]
\end{aligned}
\tag{25}
$$

KL 발산 유사도[51]는 수식 (26)과 같다. 수식 (26)에서 D_KL은 쿨백-라이블러 발산(Kullback-Leibler Divergence)을 의미하며, H는 신호의 정규화된 히스토그램 분포를 의미한다. M은 히스토그램의 총 빈의 개수를 의미하며, $H(x)[i]$와 $H(\hat{x})[i]$는 각 i번째 빈에서 정규화된 확률값을 의미하며, ln은 자연로그를 의미한다. KL 발

산 유사도는 두 신호의 확률 분포 차이를 측정하는 지표로써 원본 분포를 기준으로 증강 분포가 얼마나 다른지를 정량화한다. 높은 유사도 값은 원본의 확률 분포 특성이 잘 보존되었음을 의미한다.

$$S_{KL} = 1 - \min\left(1, \frac{D\_KL(H(x) \| H(\hat{x}))}{\log(M)}\right),$$
$$D\_KL(H(x) \| H(\hat{x})) = \sum_{i=1}^{M} H(x)[i] \times \ln \frac{H(x)[i]}{H(\hat{x})[i]} \quad (26)$$

바서슈타인 유사도[52]는 수식 (27)과 같다. 수식 (27)에서 W_1은 바서슈타인 거리(Wasserstein distance)를 의미하며, H는 신호의 정규화된 히스토그램 분포를 의미한다. $\sigma_combined$는 결합 표준편차를 의미한다. M은 히스토그램의 총 빈의 개수를 의미하며, bin_i는 i번째 빈의 중심값을 의미한다. $H(x)[i]$와 $H(\hat{x})[i]$는 각 i번째 빈에서 정규화된 확률값을 의미한다. σ은 신호의 진폭 표준편차를 의미한다. 바서슈타인 유사도는 두 분포를 일치시키는 데 필요한 최소 이동 비용을 측정하는 지표로써 분포 간의 기하학적 거리를 고려하여 신호의 전체적인 분포 형태 유사성을 평가한다. 높은 유사도 값은 원본 신호의 진폭 분포 형태가 잘 보존되었음을 의미한다.

$$S_{WT} = 1 - \min\left(1, \frac{W_1(H(x), H(\hat{x}))}{\sigma\_combined}\right),$$
$$W_1 = \sum_{i=1}^{M} |bin\_i| \cdot |H(x)[i] - H(\hat{x}[i])| \quad (27)$$
$$\sigma\_combined = \sqrt{\sigma^2(x) + \sigma^2(\hat{x})}$$

JS 발산 유사도[53]는 수식 (28)과 같다. 수식 (28)에서 D_JS는 젠슨-섀넌 발산(Jensen-Shannon divergence)을 의미하며, H는 신호의 정규화된 히스토그램 분포를 의미한다. D_KL은 수식(26)에서 정의한 쿨백-라이블러 발산을 의미한다. M은 두 분포의 평균 분포를 의미하며, 각 빈에서 원본과 증강 분포의 확률값을 평균한 값이다. JS 발산은 KL 발산의 대칭적 버전으로, 두 분포 간의 유사성을 안정적으로 측정하는 지표이다. KL 발산과 달리 두 분포를 대칭적으로 비교하여 어느 분포를 기준으로 하든 동일한 결과를 제공한다. 높은 분포 값일수록 원본의 통계적 분포 특성이 양방향으로 잘 보존이 되었음을 의미한다.

$$S_{JS} = 1 - \sqrt{(D\_JS(H(x) \| H(\hat{x})))},$$
$$D\_JS(H(x) \| H(\hat{x})) = 0.5 \times D\_KL(H(x) \| M) + 0.5 \times D\_KL(H(\hat{x}) \| M), \quad (28)$$
$$M = \frac{H(x) + H(\hat{x})}{2}$$

바타차리야 유사도[54]는 수식 (29)과 같다. S_BC는 바타차리야 계수(Bhattacharyya coefficient)를 의미하며, H는 신호의 정규화된 히스토그램 분포를 의미한다. M은 히스토그램의 총 빈의 개수를 의미하며, $H(x)[i]$와 $H(\hat{x})[i]$는 각 i번째 빈에서 정규화된 확률값을 의미한다. 바타차리야 계수는 두 확률 분포의 겹치는 정도를 측정하는 지표이다. 높은 유사도 값은 원본 신호의 확률 분포가 전반적으로 잘 보존되었음을 의미한다.

$$S\_BC = \sum_{i=1}^{M} \sqrt{H(x)[i] \times H(\hat{x})[i]} \quad (29)$$

7. Clustering Quality

클러스터링 품질은 특징 공간에서 유사성을 측정하는 지표로써, 특징 벡터 공간에서 원본과 증강 파일의 구조적 유사성을 검증한다. 클러스터링 품질 지표는 실루엣 점수(silhoutte score), 데이비스-볼딘 점수(Davies-Bouldin score), 칼린스키-하라바즈 점수(Calinski-Harabasz score)로 총 3개의 세부 지표로 구성된다.

실루엣 점수[55]는 수식 (30)과 같다. 수식 (30)에서 F는 오디오의 특징 벡터를 의미한다. d는 특징 벡터의 차원수를 의미하며, $F_i(x)$는 특징 벡터 $F(x)$의 i번째 원소를 의미한다. 실루엣 점수는 고차원 특징 공간에서 두 오디오 샘플이 얼마나 같은 방향을 가리키는지 여부를 측정한다. 코사인 유사도가 1에 가까울수록 두 특징 벡터가 거의 같은 방향을 향하고 있음을 의미하기 때문에 높은 유사도를 나타낸다. 따라서 증강 데이터가 원본의 핵심 특성이 얼마나 보존되었는지를 표현할 수 있는 지표 중 하나이다.

$$S_{SL} = \cos(F(x), F(\hat{x})),$$
$$\cos(F(x), F(\hat{x})) = \frac{\sum_{i=1}^{d} (F\_i(x) \times F\_i(\hat{x}))}{\sqrt{\left(\sum_{i=1}^{d} F\_i^2(x)\right)} \times \sqrt{\sum_{i=1}^{d} F\_i^2(} \quad (30)$$

데이비스-볼딘 점수[56]는 수식 (31)과 같다. 수식 (31)에서 $d_normalized$는 정규화된 벡터 간의 유클리드 거리를 의미한다. d는 특징 벡터의 차원 수를 의미하며, $F_i(x)$은 단위 벡터로 정규화된 특징 벡터의 i번째 원소를 의미한다. 데이비스-볼딘 점수는 두 오디오 특징 벡터를 단위 벡터로 정규화한 후, 실제 공간에서 거리를 측정한다. 벡터의 크기 차이를 제거하고 순수한 패턴의 유사성만을 평가한다. 거리가 작을수록 두 오디오가 비슷한 패턴을 가지고 있음을 의미하며, 증강된 데이터가 원본의 구조적 특징을 잘 유지했음을 의미한다.

$$
\begin{aligned}
&S_{DB} = 1 - d\_normalized\big(F(x), F(\hat{x})\big),\\
&d\_normalized\big(F(x), F(\hat{x})\big) =\\
&\parallel F\_norm(x) - F\_norm(\hat{x}) \parallel,\\
&F\_norm(x) = \frac{F(x)}{\parallel F(x) \parallel}
\end{aligned}
\tag{31}
$$

칼린스키-하라바즈 점수[57]는 수식 (32)와 같다. 수식 (32)에서 p은 피어슨 상관계수를 의미하며, -1에서 1 사이의 값을 갖는다. d는 특징 벡터의 차원 수를 의미하며, $F_i(x)$는 특징 벡터 $F(x)$의 i번째 원소를 의미한다. μ_x와 $\mu_\hat{x}$는 각 원본과 증강 특징 벡터의 평균값을 의미한다. 칼린스키-하라비스 점수는 두 오디오 특징 벡터 간의 선형 상관관계를 측정한다. 피어슨 상관계수를 기반으로 두 신호가 얼마나 비례적으로 변화하는지 평가한다. 높은 상관계수는 증강 과정에서 원본의 변화 패턴이 일관되게 유지되었음을 의미하며, 감정적 특성 보존 여부와 연관이 있는 측정 지표이다.

$$
\begin{aligned}
&S_{CH} = \frac{(1 + p(F(x), F(\hat{x})))}{2},\\
&p\big(F(x), F(\hat{x})\big) =\\
&\frac{\sum_{i=1}^{d}\big((F\_i(x) - \mu\_x)\big(F\_i(\hat{x}) - \mu\_\hat{x}\big)\big)}{\sqrt{\sum_{i=1}^{d}(F\_i(x) - \mu\_x)^2} \times \sqrt{\sum_{i=1}^{d}\big(F\_i(x) - \mu\_\hat{x}\big)^2}},\\
&\mu\_x = \frac{\sum_{i=1}^{d} F\_i(x)}{d}
\end{aligned}
\tag{32}
$$

8. Quality Score

최종 품질 점수는 수식 (33)과 같다. 수식 (33)에서 Q_i는 각 품질 평가 그룹별 정규화된 점수를 의미하며, G는 그룹의 개수로써 본 논문에서는 7개의 그룹으로 구성되어 있다. Q_i는 각 그룹의 점수로써 i번째 그룹의 점수를 의미한다. M_ij는 i번째 그룹의 j번째 지표의 점수를 의미한다. 각 그룹 별 상대적 중요성을 반영한 가중치를 Table 1과 같이 설정하였으며, W_i로 표현되며, i번째 가중치를 의미한다. 수식 (33)에 의해서 계산된 최종 품질 점수는 정규화되어 0~1 사잇값으로 표현되며, 1에 가까울수록 품질이 좋음을 의미하며, 0에 가까울수록 품질이 낮음을 의미한다.

$$
\begin{aligned}
&Q_{Score} = \sum_{i=1}^{G} W\_i \times Q\_i,\\
&Q\_i = \frac{1}{n\_i} \times \sum_{j=1}^{n\_i} M\_ij
\end{aligned}
\tag{33}
$$

본 논문에서는 Table 2와 같이 사전 정의된 임계 값(threshold value)을 적용하여 Excellent, Very Good, Good, Poor, Very Poor의 5단계로 품질 등급을 정의한다. 첫 번째 Excellent(S) 품질 등급은 품질 점수가 0.85 이상인 경우를 의미한다. 원본 대비 거의 완벽하게 품질이 보존되었음을 의미하며, 모든 감정적 특성이 원본과 매우 유사하게 유지된 고품질 증강 결과를 의미한다. Excellent 등급의 데이터는 학습에 활용할 때 원본 데이터와 동등한 수준의 성능 향상을 기대할 수 있는 품질 등급을 의미한다. 두 번째 Very Good(A) 품질 등급은 품질 점수가 0.7 이상이고, 0.85 이하인 경우로 정의하였다. 본 등급은 우수한 감정적 특징 보존 품질을 의미하며, 원본의 감정 표현력이 높은 수준으로 유지됨으로써 실제 응용에서 원본과 거의 구분되지 않는 품질 상태를 의미한다. 품질 저하가 미미함으로 감정 인식 성능에 미치는 부정적 영향이 최소화됨으로써 학습 데이터로 활용하기에 매우 적합한 품질 등급을 의미한다. 세 번째 Good(B) 품질 등급은 품질 점수가 0.55 이상이고, 0.7 이하인 경우로 정의하였다. 본 등급은 대부분의 감정적 특징이 양호한 수준으로 보존된 품질 상태를 의미한다. 일부 미세한 품질 저하가 존재할 수 있지만 감정의 핵심 표현 요소는 적절히 유지되어 전반적인 감정 인식에는 큰 영향을 미치지 않는 수준을 의미한다. 학습 데이터 확장 목적으로 충분히 활용이 가능한 품질을 의미한다. 네 번째 Poor(C) 등급은 육안 또는 청각적

Table 1. Detailed Characteristic Items and Importance Weights for Seven Groups of Proposed Emotion-Specific Audio Quality Metrics

No	Main Quality Measure	Detail Quality Measure	Detail Feature Description	Weight
1	Emotional Consistency	emotion embedding similarity	MFCC based Similarity	21%
		arousal consistency	Degree of Emotional Activation	
		valence consistency	Degree of Positivity/Negativity	
2	Frequency Domain	spectral centroid	Brightness of Tone	20%
		spectral bandwidth	Frequency Dispersion	
		spectral rolloff	High-Frequency Distribution	
		spectral contrast	Differences between Frequency Bands	
		mfcc similarity	Voice Feature	
		harmonic ratio	Speech to Noise Ratio	
3	Perceptual Quality	energy similarity	Voice Volume	20%
		formant cosine similarity	Voice Resonance	
		loudness similarity	Perceptual Loudness	
		roughness measure	Smoothness of Sound Quality	
		clarity index	Voice Clarity	
4	Time Domain	rms energy	Signal Strength	16%
		zero crossing rate	Signal Noise or Percussive	
		duration	Signal Length	
		temporal dynamics	Signal Dynamic Changes	
		attack time	Sound Start Characteristics	
		decay characteristics	Sound End Characteristics	
5	Distribution Similarity	cosine similarity	Vector Direction Similarity	12%
		kl divergence similarity	Probability Distribution Difference	
		wasserstein similarity	Distribution Distance	
		js divergence	Symmetric Distribution Difference	
		bhattacharyya similarity	Degree of Distribution Overlap	
6	Prosodic	pitch mean	Fundamental Frequency (f_0)	10%
		pitch variance	Degree of Pitch Conversion	
		tempo consistency	Rhythm Stability	
		intonation pattern	Pitch Transformation Complexity	
7	Clustering Quality	silhouette score	Cluster Quality	1%
		davies bouldin score	Cluster Separation	
		calinski harabasz score	Cluster Density	

Table 2. Quality Grade Classification System for Emotion Audio Data Augmentation

Quality		Score Range	Quality Characteristics
Excellent	'S'	$0.85 \leq Q_{score}$	• High-fidelity preservation • Complete retention of emotional attributes • Negligible perceptual differences • Consistently strong performance across evaluation metrics
Very Good	'A'	$0.70 \leq Q_{score} < 0.85$	• Minimal quality degradation • Preservation of emotion recognition accuracy • Only subtle differences observed • Generally favorable performance across most metrics
Good	'B'	$0.55 \leq Q_{score} < 0.70$	• Moderate quality preservation • Largely retained emotional information • Perceptually acceptable variations in quality • Quality degradation observed in certain metrics
Poor	'D'	$0.40 \leq Q_{score} < 0.55$	• Noticeable quality degradation • Partial loss of emotional information • Perceptually discernible degradation • Quality deterioration across multiple metrics
Very Poor	'D'	$Q_{score} < 0.40$	• Substantial quality degradation • Potential for emotional confusion • Clearly perceptible quality differences • Poor performance across multiple metrics

으로 품질 저하가 감지되어, 사람이 원본과의 차이를 인지할 수 있을 정도의 품질을 의미하며, 품질 점수가 0.4 이상 0.55 이하인 경우로 정의하였다. 기본적인 음성 정보는 유지가 되었지만, 감정적 특징의 손상이 존재하기 때문에 본 등급의 데이터는 신중한 선별 후에 제한적으로 활용이 가능한 품질 등급을 의미한다. 마지막으로 Very Poor(D) 품질 등급은 심각한 품질 저하로 인해 감정적 특징이 크게 손상되어 감정 인식 성능에 부정적인 영향을 미칠 가능성이 높으며, 원본과의 명확한 차이로 인해 개선이 필요한 품질 단계로써 품질 점수가 0.4 이하인 경우를 의미한다. 본 품질 등급의 데이터는 학습에 활용하지 않거나, 추가적인 품질 개선 과정이 필수적인 품질 상태임을 의미한다.

III. Experimental Results

1. Experimental Environment

본 논문에서는 감정 발화 오디오의 품질을 검증하기 위한 실험을 수행하기 위해 Emo-DB(Emotional Speech Database) 공개 데이터 셋을 사용하였다. Emo-DB는 남성 화자 5명과 여성 화자 5명, 총 10명의 20-30대 전문 화자가 녹음에 참여하여 제작된 데이터 셋으로, 총 535개의 발화로 구성되어 있다. 데이터는 분노(W, Wut), 지루함(L, Langeweile), 불안(A, Angst), 행복(F, Freude), 슬픔(T, Trauer), 역겨움(E, Ekel), 중립(N, Neutral)의 7가지 감정 클래스를 포함한다. 본 논문에서는 Emo-DB의 각 발화를 2초 단위로 분할(segment)하여 실험 데이터로 사용하였으며, 2초보다 짧은 발화는 학습에서 제외하였다. 본 논문에서 사용한 Emo-DB의 각 클래스 별 원본 발화 파일 수, 총 발화 길이(total duration) 및 2초 단위로 분할 했을 때 분할된 파일 개수는 Table 3과 같다.

Table 3. Basic information about the audio files by class in Emo-DB used in the experiment

Class	Num of Origin Wav Files	Total Duration	Segments (2 sec)
F	71	180.62	52
N	79	186.38	51
W	127	335.38	105
T	62	251.28	89
A	69	154.1	41
L	81	225.1	72
E	46	154.23	54

오디오 데이터 증강에는 MetricGAN 모델을 적용하였다. MetricGAN은 본래 음성 신호의 품질을 개선하기 위해 제안된 모델로, 전통적인 GAN과 달리 판별자(D, Discriminator)가 단순히 생성 신호의 진위를 구분하지 않고, PESQ 또는 STOI와 같은 객관적 음질 평가 지표의 점수를 예측하는 회귀 모델로 구성되어 있다. 이러한 특성으로 인해 사용자는 목표 품질 점수(target score)를 지정함으로써, 생성자(G, Generator)가 해당 품질 수준을 만족하도록 오디오를 생성·개선할 수 있다. 본 연구에서는 MetricGAN의 이러한 특성을 활용하여, STOI 지표를 기준으로 목표 점수를 다섯 단계(0.35, 0.50, 0.65, 0.75, 0.85)로 설정하였다. 이후 각 감정 클래스별로 해당 목표 점수에 대응하는 5단계 품질의 증강 데이터를 생성하였다. 특히 5단계로 구분된 증강 데이터 셋을 구축한 이유는, 제안하는 ESQ 메트릭이 감정 발화 오디오의 특성을 기반으로 품질 차이를 정량적으로 구분하고 점수화할 수 있는지를 검증하기 위함이다. 즉, 단계 별로 점진적으로 향상된 품질을 가진 데이터를 구성함으로써, 제안한 감정 오디오 품질 지표가 기존의 일반 음성 품질 지표(PESQ, STOI, SNR)에 비해 감정 발화의 세부 품질 변화를 얼마나 정확하게 정량화할 수 있는지를 실험적으로 검증할 수 있도록 설계하였다.

2. Quality Verification of Augmented Data

본 논문에서 제안한 ESQ 메트릭의 유효성을 검증하기 위해 Emo 데이터 셋을 5단계 품질 레벨로 증강 수행했으며, 본 실험 결과표에서는 Lv. 1, Lv. 3, Lv. 5의 증강 데이터에 대해 7개 메트릭 그룹의 성능 결과를 비교하였다. Table 4 ~ Table 10은 7개 메트릭 각 메트릭 그룹의 품질 레벨별 점수를 보여준다. Table 11은 각 7개 메트릭 그룹의 품질 레벨별 평균 점수를 기반으로 레벨 간 향상률을 보여준다. 실험 결과, 전체적으로 모든 그룹에서 품질 레벨이 높아질수록 품질 점수가 상승하는 일관된 경향을 확인 할 수 있었다. 또한 전체 평균을 기준으로 Lv. 1에서 Lv. 3으로의 향상률을 69.87%, Lv. 3에서 Lv. 5로의 향상률은 11.68%의 결과를 보이며, 최종적으로 Lv. 1 대비 Lv. 5로의 전체 평균 향상률은 90.75%임을 볼 수 있다. 향상률 결과를 보면, 대부분 초기 단계(Lv. 1→Lv. 3)에서 더 큰 향상 폭을 보이고 후기 단계(Lv. 3→Lv. 5)에서는 상대적으로 완만한 향상률을 보인다. 본 결과는 오디오 데이터를 수행할 때 사용한 MetricGAN의 주목적이 품질 개선에 초점이 맞춰져 있기 때문에 초기에 급격한 개선이 이루어지다가 고품질 단계에서는 점진적으로 수렴하는 일반적인 패턴과 일치하는 것으로 분석된다.

Table 4. Performance Comparison Results of Four Detailed Indicators in the 'Prosodic' Group of ESQ Metrics by Quality Level

	Lv.	F	N	W	T	A	L	E
Pitch Mean	1	0.460	0.092	0.477	0.262	0.339	0.444	0.372
	3	0.679	0.831	0.784	0.820	0.770	0.819	0.816
	5	0.799	0.797	0.781	0.804	0.772	0.827	0.840
Pitch Variance	1	0.980	0.000	0.678	0.867	0.725	0.741	0.948
	3	0.874	0.934	0.947	0.822	0.864	0.959	0.703
	5	0.981	0.938	0.976	0.826	0.848	0.997	0.903
Tempo Consistency	1	0.442	0.430	0.613	0.605	0.511	0.602	0.567
	3	0.548	0.553	0.536	0.705	0.499	0.502	0.631
	5	0.544	0.508	0.532	0.656	0.510	0.520	0.535
Intonation Pattern	1	0.890	0.000	0.814	0.952	0.773	0.812	0.899
	3	0.891	0.895	0.958	0.913	0.779	0.923	0.744
	5	0.940	0.912	0.979	0.908	0.767	0.954	0.889
Avg.	1	0.693	0.130	0.646	0.672	0.587	0.650	0.697
	3	0.748	0.803	0.806	0.815	0.728	0.801	0.723
	5	0.816	0.788	0.817	0.799	0.724	0.825	0.792

Table 5. Performance Comparison Results of Three Detailed Indicators in the 'Emotional Consistency' Group of ESQ Metrics by Quality Level

	Lv.	F	N	W	T	A	L	E
Emotion Embedding Similarity	1	0.904	0.798	0.897	0.900	0.851	0.846	0.894
	3	0.945	0.911	0.954	0.969	0.940	0.929	0.930
	5	0.965	0.971	0.961	0.974	0.958	0.959	0.969
Arousal Consistency	1	0.113	0.797	0.038	0.090	0.070	0.123	0.084
	3	0.618	0.723	0.756	0.676	0.692	0.772	0.696
	5	0.740	0.715	0.759	0.677	0.783	0.772	0.769
Valence Consistency	1	0.182	0.129	0.078	0.034	0.309	0.115	0.203
	3	0.449	0.202	0.657	0.428	0.522	0.306	0.400
	5	0.815	0.718	0.807	0.676	0.844	0.671	0.887
Avg.	1	0.399	0.575	0.338	0.341	0.410	0.362	0.394
	3	0.671	0.612	0.789	0.691	0.718	0.669	0.676
	5	0.840	0.801	0.843	0.775	0.862	0.801	0.875

Table 6. Performance Comparison Results of Six Detailed Indicators in the 'Frequency Domain'' Group of ESQ Metrics by Quality Level

	Lv.	F	N	W	T	A	L	E
Spectral Centroid	1	0.094	0.065	0.035	0.014	0.217	0.050	0.074
	3	0.329	0.181	0.473	0.378	0.435	0.354	0.279
	5	0.646	0.566	0.627	0.524	0.695	0.534	0.743
Spectral Bandwidth	1	0.310	0.311	0.342	0.263	0.340	0.326	0.424
	3	0.453	0.269	0.610	0.421	0.408	0.299	0.425
	5	0.783	0.639	0.811	0.660	0.814	0.606	0.868
Spectral Roll-off	1	0.143	0.103	0.100	0.031	0.216	0.100	0.183
	3	0.292	0.117	0.452	0.262	0.321	0.152	0.266
	5	0.650	0.527	0.645	0.501	0.679	0.492	0.753
Spectral Contrast	1	0.094	0.000	0.101	0.107	0.074	0.118	0.084
	3	0.513	0.586	0.572	0.703	0.589	0.603	0.248
	5	0.635	0.660	0.665	0.511	0.665	0.594	0.633
MFCC similarity	1	0.904	0.798	0.897	0.900	0.851	0.846	0.894
	3	0.945	0.911	0.954	0.969	0.940	0.929	0.930
	5	0.965	0.971	0.961	0.974	0.958	0.959	0.969
Harmonic Ratio	1	0.021	0.110	0.020	0.390	0.226	0.304	0.477
	3	0.178	0.578	0.466	0.626	0.549	0.657	0.309
	5	0.446	0.601	0.428	0.642	0.574	0.669	0.530
Avg.	1	0.261	0.231	0.249	0.284	0.321	0.291	0.356
	3	0.452	0.440	0.588	0.560	0.540	0.499	0.410
	5	0.688	0.661	0.689	0.635	0.731	0.642	0.749

Table 7. Performance Comparison Results of Five Detailed Indicators in the 'Perceptual' Group of ESQ Metrics by Quality Level

	Lv.	F	N	W	T	A	L	E
Energy Similarity	1	0.091	0.000	0.043	0.062	0.092	0.062	0.058
	3	0.466	0.500	0.520	0.502	0.518	0.547	0.563
	5	0.509	0.502	0.507	0.493	0.541	0.557	0.560
Formant Cosine Similarity	1	0.133	0.004	0.172	0.229	0.168	0.118	0.372
	3	0.413	0.561	0.585	0.484	0.534	0.529	0.502
	5	0.599	0.591	0.612	0.489	0.545	0.573	0.613
Loudness Similarity	1	0.289	0.000	0.172	0.178	0.252	0.237	0.144
	3	0.686	0.697	0.707	0.704	0.720	0.734	0.746
	5	0.706	0.705	0.699	0.699	0.741	0.754	0.751
Roughness Measure	1	0.759	0.000	0.472	0.462	0.458	0.732	0.242
	3	0.082	0.577	0.669	0.589	0.681	0.732	0.471
	5	0.690	0.554	0.686	0.533	0.620	0.697	0.643
Clarity Index	1	0.284	0.000	0.258	0.223	0.300	0.246	0.154
	3	0.668	0.654	0.671	0.652	0.659	0.699	0.702
	5	0.661	0.667	0.671	0.640	0.673	0.702	0.693
Avg.	1	0.311	0.001	0.223	0.231	0.254	0.279	0.194
	3	0.463	0.598	0.631	0.586	0.622	0.648	0.597
	5	0.633	0.604	0.635	0.571	0.624	0.657	0.652

Table 8. Performance Comparison Results of Six Detailed Indicators in the 'Time Domain' Group of ESQ Metrics by Quality Level

	Lv.	F	N	W	T	A	L	E
RMS Energe	1	0.259	0.000	0.195	0.209	0.279	0.233	0.158
	3	0.708	0.672	0.734	0.677	0.713	0.721	0.733
	5	0.704	0.705	0.736	0.658	0.745	0.720	0.715
Zero Crossing Rate	1	0.0	0.036	0.000	0.000	0.028	0.000	0.000
	3	0.053	0.031	0.127	0.267	0.291	0.489	0.040
	5	0.613	0.540	0.367	0.496	0.720	0.569	0.790
Duration	1	0.562	0.496	0.563	0.675	0.450	0.634	0.725
	3	0.704	0.638	0.691	0.649	0.660	0.688	0.725
	5	0.687	0.665	0.673	0.653	0.657	0.680	0.728
Temporal Dynamics	1	0.113	0.797	0.038	0.090	0.070	0.123	0.084
	3	0.618	0.723	0.756	0.676	0.692	0.772	0.696
	5	0.740	0.715	0.759	0.677	0.783	0.772	0.769
Attack Time	1	0.651	0.695	0.601	0.261	0.561	0.574	0.344
	3	0.737	0.739	0.651	0.447	0.640	0.631	0.547
	5	0.661	0.660	0.637	0.394	0.661	0.636	0.561
Decay Characteristics	1	0.214	0.000	0.196	0.291	0.352	0.712	0.321
	3	0.245	0.484	0.437	0.312	0.367	0.644	0.294
	5	0.332	0.546	0.460	0.315	0.409	0.648	0.315
Avg.	1	0.300	0.337	0.266	0.254	0.290	0.380	0.272
	3	0.511	0.548	0.566	0.505	0.561	0.658	0.506
	5	0.623	0.638	0.605	0.532	0.663	0.671	0.646

Table 9. Performance Comparison Results of Five Detailed Indicators in the 'Distribution Similarity' Group of ESQ Metrics by Quality Level

	Lv.	F	N	W	T	A	L	E
Energy Similarity	1	0.002	0.001	0.001	0.004	0.000	0.001	0.002
	3	0.007	0.014	0.007	0.011	0.013	0.014	0.006
	5	0.015	0.018	0.009	0.012	0.018	0.015	0.015
Formant Cosine Similarity	1	0.288	0.843	0.240	0.257	0.257	0.224	0.266
	3	0.880	0.866	0.892	0.823	0.862	0.835	0.871
	5	0.911	0.903	0.912	0.838	0.902	0.891	0.910
Loudness Similarity	1	0.652	0.185	0.609	0.551	0.556	0.573	0.506
	3	0.873	0.866	0.886	0.865	0.863	0.864	0.884
	5	0.891	0.888	0.896	0.866	0.891	0.894	0.895
Roughness Measure	1	0.884	0.933	0.857	0.834	0.824	0.854	0.814
	3	0.968	0.972	0.976	0.963	0.965	0.968	0.976
	5	0.981	0.979	0.983	0.962	0.977	0.978	0.981
Clarity Index	1	0.834	0.837	0.778	0.741	0.721	0.787	0.700
	3	0.965	0.968	0.975	0.958	0.962	0.964	0.974
	5	0.980	0.977	0.982	0.957	0.976	0.976	0.979
Avg.	1	0.532	0.560	0.497	0.477	0.472	0.488	0.458
	3	0.739	0.737	0.747	0.724	0.733	0.729	0.742
	5	0.755	0.753	0.756	0.727	0.753	0.751	0.756

Table 10. Performance Comparison Results of Three Detailed Indicators in the 'Clustering' Group of ESQ Metrics by Quality Level

	Lv.	F	N	W	T	A	L	E
Emotion Embedding Similarity	1	0.997	0.996	0.989	0.989	0.994	0.993	0.993
	3	0.998	0.997	0.999	0.996	0.998	0.995	0.998
	5	0.999	0.998	0.999	0.997	0.999	0.998	0.999
Arousal Consistency	1	0.962	0.955	0.931	0.929	0.947	0.943	0.943
	3	0.967	0.961	0.975	0.957	0.967	0.949	0.968
	5	0.976	0.974	0.978	0.967	0.979	0.968	0.978
Valence Consistency	1	0.998	0.998	0.994	0.994	0.997	0.996	0.996
	3	0.999	0.998	0.999	0.998	0.999	0.997	0.999
	5	0.999	0.999	0.999	0.999	0.999	0.999	0.999
Avg.	1	0.986	0.983	0.971	0.971	0.980	0.977	0.978
	3	0.988	0.985	0.991	0.983	0.988	0.980	0.988
	5	0.991	0.990	0.992	0.987	0.992	0.988	0.992

Table 11. Average Performance Comparison Results of Seven ESQ Metric Groups by Quality Level

Group	Lv.	F	N	W	T	A	L	E	Avg.	QG
Prosodic	1	0.693	0.130	0.646	0.672	0.587	0.650	0.697	0.582	B
	3	0.748	0.803	0.806	0.815	0.728	0.801	0.723	0.775	A
	5	0.816	0.788	0.817	0.799	0.724	0.825	0.792	0.794	A
Emotional Consistency	1	0.399	0.575	0.338	0.341	0.410	0.362	0.394	0.403	C
	3	0.671	0.612	0.789	0.691	0.718	0.669	0.676	0.689	B
	5	0.840	0.801	0.843	0.775	0.862	0.801	0.875	0.828	A
Frequency Domain	1	0.261	0.231	0.249	0.284	0.321	0.291	0.356	0.285	D
	3	0.452	0.440	0.588	0.560	0.540	0.499	0.410	0.498	C
	5	0.688	0.661	0.689	0.635	0.731	0.642	0.749	0.685	B
Perceptual	1	0.311	0.001	0.223	0.231	0.254	0.279	0.194	0.213	D
	3	0.463	0.598	0.631	0.586	0.622	0.648	0.597	0.592	B
	5	0.633	0.604	0.635	0.571	0.624	0.657	0.652	0.625	B
Time Domain	1	0.300	0.337	0.266	0.254	0.290	0.380	0.272	0.300	D
	3	0.511	0.548	0.566	0.505	0.561	0.658	0.506	0.551	B
	5	0.623	0.638	0.605	0.532	0.663	0.671	0.646	0.626	B
Distribution Similarity	1	0.532	0.560	0.497	0.477	0.472	0.488	0.458	0.498	C
	3	0.739	0.737	0.747	0.724	0.733	0.729	0.742	0.736	A
	5	0.755	0.753	0.756	0.727	0.753	0.751	0.756	0.750	A
Clustering	1	0.986	0.983	0.971	0.971	0.980	0.977	0.978	0.978	S
	3	0.988	0.985	0.991	0.983	0.988	0.980	0.988	0.986	S
	5	0.991	0.990	0.992	0.987	0.992	0.988	0.992	0.991	S

Table 12. Improvement Rate Comparison Results of ESQ Metric Groups by Quality Level

Group	Lv.1	Lv.3	Lv.5	Improvement Rate (%)		
				Lv.1→3	Lv.3→5	Lv.1→5
Prosodic	0.582	0.775	0.794	▲ 33.16	▲ 2.45	▲ 36.43
Emotional Consistency	0.403	0.689	0.828	▲ 0.93	▲ 20.17	▲ 105.46
Frequency Domain	0.286	0.498	0.685	▲ 74.74	▲ 37.55	▲ 140.35
Perceptual	0.213	0.592	0.625	▲ 177.93	▲ 5.57	▲ 193.43
Time Domain	0.300	0.551	0.626	▲ 83.67	▲ 13.61	▲ 108.67
Distribution Similarity	0.498	0.736	0.750	▲ 47.97	▲ 1.90	▲ 50.60
Clustering	0.978	0.986	0.991	▲ 0.82	▲ 0.51	▲1.33
Avg.	0.466	0.690	0.757	▲ 69.87	▲ 11.68	▲ 90.75

Table 13. Setting group-specific weights for ESQ metric

Group	Lv.1 Avg.	Lv.5 Avg.	$\triangle = Lv.5 - Lv.1$	$\triangle / \sum \triangle$	Proposed Weight (%)
Prosodic	0.582	0.794	0.212	0.104	10
Emotional Consistency	0.403	0.828	0.425	0.208	21
Frequency Domain	0.286	0.685	0.399	0.196	20
Perceptual	0.213	0.625	0.412	0.202	20
Time Domain	0.300	0.626	0.326	0.160	16
Distribution Similarity	0.498	0.75	0.252	0.124	12
Clustering	0.978	0.991	0.013	0.006	1

Table 14. Comparison Results Between Conventional Quality Measurement Methods (PESQ, STOI, SNR) and ESQ by Quality Level

Level	Target STOI	ESQ	PESQ	STOI	SNR
1	0.35	0.470	4.98	0.718	21.3
2	0.50	0.588	4.99	0.732	22.2
3	0.65	0.675	4.99	0.731	22.4
4	0.75	0.702	5.00	0.747	22.7
5	0.85	0.738	5.00	0.836	22.8

Table 12는 제안한 ESQ 메트릭의 그룹별 평균 품질 평가 점수와 품질 단계별 개선율을 비교한 결과를 보여준다. Table 12의 결과에서 향상률을 기반으로, 지각적 품질과 시간 도메인 항목이 레벨 1에서 레벨 3단계인 초기 개선에서 향상률이 높으며, 레벨 3에서 레벨 5단계는 대부분의 항목에서 상대적으로 낮은 향상률을 보이며, 품질이 수렴되는 현상이 관찰된다. 클러스터링 품질과 운율 항목은 이미 높은 기준점에서 시작하여 안정적인 품질을 유지한다. 분석 결과, 7개 그룹에서 모두 품질 레벨 증가에 따른 일관된 향상 패턴을 보이는 것은 제안한 ESQ 메트릭의 내적 일관성(internal consistency)과 구성 타당성(construct validity)이 높음을 입증한다. 각 그룹이 서로 다른 음향학적 측면을 평가하고 있음에도 불구하고 동일한 품질 변화를 보이는 것은 ESQ 메트릭이 감정 오디오의 품질을 종합적이고 신뢰성 있게 측정하고 있음을 의미한다. Fig. 2는 본 논문에서 제안한 증강된 감정 오디오 품질 측정 방법에 대한 정확도를 검증하기 위해서 Emo-DB 학습 데이터 셋을 MetricGAN 모델을 이용하여 클래스 별로 5개 레벨 품질로 증강한 데이터의 스펙트로그램 결과를 예시로 보여준다. Emo-DB 학습 데이터 셋은 총 7개의 감정으로 이루어져 있으며, Fig. 2에서는 예시로 A(불안) 클래스, E(역겨움) 클래스, F(행복) 클래스에 대한 원본 및 1~5 레벨로 증강된 데이터 결과를 비교한다. 각 클래스의 레벨별 증강된 결과를 보면, 레벨 1의 경우 원본과 비교하여 주요 특성 정보의 손실이 큰 것을 육안으로 확인 할 수 있으며 레벨이 증가할수록 주요 특성 정보가 단계적으로 뚜렷해지는 것을 확인할 수 있다. 특히 마지막 품질 레벨 5는 원본과의 유사도가 가장 높은 것을 확인할 수 있으며, 각 클래스별로 레벨이 높아질수록 생성된 오디오의 품질이 높아지는 것을 정성적으로 확인할 수 있다.

본 논문에서 제안하는 ESQ 메트릭의 최종 품질 점수를 계산하기 위해서 각 7개 그룹에 대한 가중치를 감정 품질 변화의 실측 개선 폭(△)을 기반하여 Table. 13과 같이 산정하였다. Table. 12에서 계산된 품질 레벨 5(Lv. 5)와 레벨 1(Lv. 1)의 차이를 $\triangle = Lv.5 - Lv.1$로 계산하고 이를 전체 합으로 정규화하여 $w = \triangle / \sum \triangle$로 정의하였다. 분석 결과, 감정 일관성이 0.425, 지각적 품질이 0.412로 그룹의 개선 폭이 가장 큼으로써 감정 품질 향상에 대한 기여도가 높게 나타났으며, 주파수 도메인이 0.399로 유사한 비중을 보였다. 시간 도메인과 분포 유사도는 각 0.326, 0.252로 다음으로 높은 기여도를 보였다. 반면 클러스터링 품질의 경우, Lv. 1부터 높은 값을 유지함으로써 품질 변화량이 작고, 감정 보존 변화에 대한 영향성이 상대적으로 작은 것으로 확인되었다. 따라서 본 기여도에 따라서 감정 일관성 그룹이 21%, 주파수 도메인과 운율 그룹의 가중치가 각 20%, 시간 도메인과 분포 유사도 그룹이 각 16%, 12%의 가중치가 적용되었으며, 클러스터링 품질 그룹은 1%의 가중치를 적용하여 각 그룹의 감정 품질 개선 기여도를 실 측정 데이터에 근거하여 정량적으로 반영 적용하였다.

Table 14는 본 논문에서 제안한 품질 측정 방법(ESQ, Emotion-Specific Quality)을 이용하여 각 레벨 별로 측정된 품질 점수와 기존 품질 측정 방법인 PESQ, STOI, SNR의 품질 점수와 비교한 결과를 보여준다. Table 14의 각 품질 점수는 Emo 데이터 셋의 각 7개 감정 클래스의 5

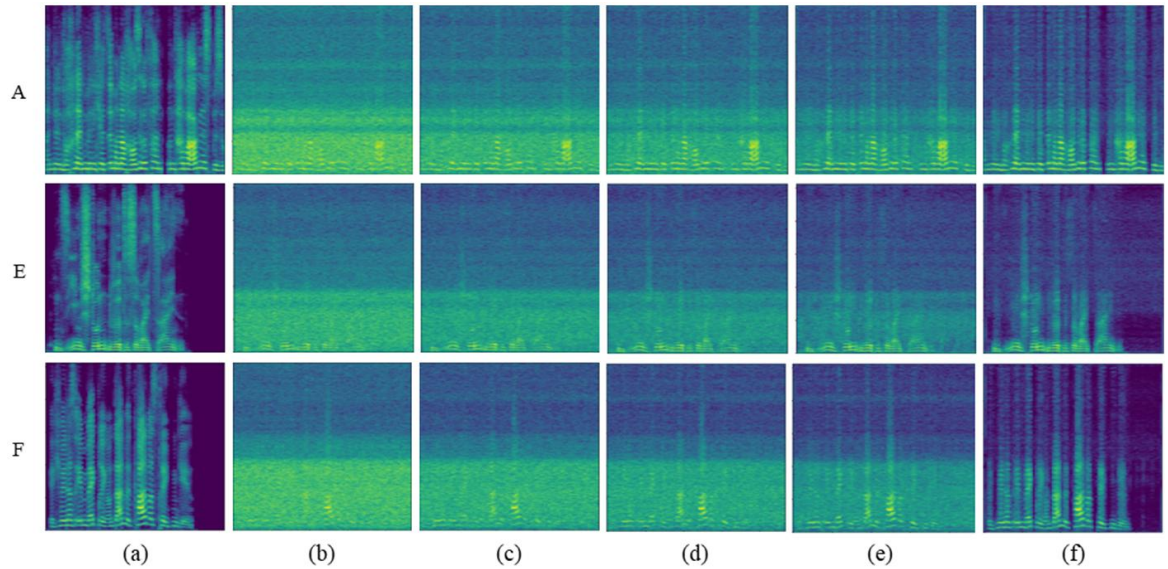

Fig. 2. Spectrogram Comparison of Multi-level Quality-Enhanced Audio Sample Classes (A: Anxiety, E: Disgust, F: Happiness) from Emo Dataset (a) Original Audio, (b) Level 1, (c) Level 2, (d) Level 3, (e) Level 4, (f) Level 5

단계 레벨 증강 데이터로 측정하였으며, 모든 클래스의 레벨 별 품질 점수의 평균 점수이다. PESQ 평가 지표 결과는 레벨 1의 품질 평균 점수는 4.98이고, 레벨 2와 레벨 3의 품질 평균 점수가 4.99로 동일하다. 레벨 4와 레벨 5의 품질 평균 점수는 5.0으로 동일한 결과를 보인다. 레벨별로 격차가 거의 존재하지 않으며, Fig. 2에서와 같이 레벨별로 품질의 수준이 다름이 존재하지만, 레벨 간 동일한 점수가 존재하는 것을 볼 수 있다. 특히 PESQ의 경우, 품질의 범위를 0~5.0으로 설정했을 때의 결과로써 레벨 4와 레벨 5의 경우 5.0의 품질 점수가 나온 것은 원본과의 품질 차이가 거의 없음을 의미하지만 Fig. 2를 통해서 실제 품질 결과는 차이가 존재함을 확인하였다.

STOI 평가 지표는 품질 평가의 범위가 0~1.0으로 점수가 1에 가까울수록 품질의 명료도가 높음을 의미한다. 레벨 별 STOI 평가 점수 비교 결과, 레벨 1은 0.718이고 레벨 2는 0.732로 약 0.014 정도 향상된 것을 볼 수 있다. 하지만 레벨 3이 0.731로 레벨 2보다 오히려 0.001 낮은 점수를 보인다. 레벨 4는 레벨 3에 비해 0.016 높아진 결과를 보이며, 레벨 5는 레벨 4에 비해 0.089 높은 품질을 보임으로써 레벨이 높아질수록 품질 점수가 높아지는 경향은 있지만 품질 레벨 구간별 편차가 일정하지 않으며, 레벨 간의 차이가 존재하지 않은 결과도 볼 수 있다. SNR 품질 지표의 경우, 품질 점수가 20dB 이상이면 비교적 양호한 품질을 의미하고 30dB 이상인 경우 우수한 품질임을 의미한다. 실험 결과 레벨 단계가 높아질수록 SNR 점수가 높아지는 경향이 있지만 레벨 1의 품질 점수가 21.3이고, 레벨 5의 품질 점수가 22.9로 품질 점수 차가 1.5로써 레벨 간 차가 매우 작아서 레벨 별 품질 형평성이 낮음을 볼 수 있다. 제안한 방법인 ESQ의 품질 지표는 0~1.0 범위에서 1에 가까울수록 품질이 높음을 의미한다. 실험 결과, Level 1의 품질 점수는 0.470였으며 Level 5의 품질 점수는 0.738으로, 두 구간 사이의 품질 차이가 0.268로 확인되었다. 또한 각 레벨 간 품질 격차는 Level 1에서 Level 2간은 0.118, Level 2에서 Level 3간은 0.087, Level 3에서 Level 4간은 0.027, Level 4에서 Level 5간은 0.036의 차이를 보임으로써, 전반적으로 일정하고 안정적으로 품질이 증가하는 경향을 보임을 확인할 수 있다.

따라서 제안된 방법은 측정된 품질 점수에 따라서 증강된 데이터의 품질 및 특징 정보의 복원 정도를 역으로 유추가 가능함을 볼 수 있다. 이와 같은 결과는 기존의 객관적 음성 품질 지표(PESQ, STOI, SNR)는 주로 전송 음질이나 잡음 환경에서의 신호 보존성을 평가하도록 설계되었기 때문에 감정 발화 오디오 증강에서 나타나는 감정적 단서와 발화 특성의 품질 차이를 정밀하게 반영할 수 없는 단점이 존재한다. 반면 제안한 ESQ 지표는 감정 발화 오디오의 특성에 최적화되어 레벨별 품질 차이를 안정적으로 측정할 수 있음을 확인할 수 있다.

Table 15는 기존의 발화 음성 증강 모델인 VTLP, StarGAN-VC, CycleGAN-VC를 이용하여 증강한 데이터셋을 ESQ 메트릭으로 품질을 측정한 실험 결과로 7개 그룹의 32개 세부 지표의 결과를 보여준다. 실험 결과, 첫 번째로 감정 일관성 그룹에서 emotion embedding similarity 지표는 모든 모델이 0.97 이상으로 고품질을 보였으며, VTLP 모델이 0.983으로 가장 우수한 결과를 보였다. valence consistency 지표에서는 CycleGAN-VC 모델이 0.859로 감정가 표현이 가장 높았으며, arousal consistency는 CycleGAN-VC와 StarGAN-VC 모델이 각 0.764, 0.763으로 거의 동등한 결과를 보였다. 감정 일관성 그룹의 총점은 CycleGAN-VC 모델이 0.866으로 가장 높은 점수로 감정 특성 보존이 우수함을 볼 수 있었으며, 그 다음으로 VTLP 모델이 0.855로 감정 특성 변화 없이 원본을 유지한 모델이 높은 점수를 보였으며, StarGAN-VC가 0.833으로 가장 낮은 결과를 보였다. 두 번째, 주파수 영역 그룹에서 MFCC similarity 지표는 VTLP 모델이 0.984로 가장 좋은 결과로 스펙트럼 포락선이 보존된 것을 알 수 있으며, 나머지 모델도 0.970, 0.971로 높은 결과를 보였다. Spectral centroid, spectral bandwidth, spectral roll-off, spectral contrast는 모든 모델이 0.5 초반대의 결과를 보였다. harmonic ratio는 VTLP 모델이 0.523으로 가장 높았으며, StarGAN-VC 모델이 0.504로 그다음으로 높고, CycleGAN-VC 모델이 0.441로 가장 낮은 점수를 보였다. 주파수 영역 그룹의 총점은 VTLP 모델이 0.608로 가장 좋은 결과를 보였다. VTLP 모델은 주파수 축을 선형 변환함으로써 모든 주파수 특성이 다른 모델에 비해서 일관성 있게 유지하고 있음을 볼 수 있다. 반면에 StarGAN-VC와 CycleGAN-VC 모델이 MFCC는 높지만, harmonic ratio가 낮은 결과를 보이는 이유는 딥러닝 모델이 전체적으로 스펙트럼 형태는 유사하게 생성하지만, 미세한 배음 구조는 재현하지 못하고 있음을 의미한다. 세 번째, 지각적 품질 그룹에서 formant cosine similarity 지표는 CycleGAN-VC 모델이 0.942로 가장 높은 결과를 보임으로써 화자 특성을 가장 잘 유지하고 있음을 볼 수 있다. roughness 지표는 StarGAN-VC 모델이 0.894로 가장 높은 결과로써 부드러운 음성을 생성하고 있음을 볼 수 있다. clarity index 지표는 VTLP 모델이 0.427로 상

대적으로 높은 결과를 보임으로써 명료도가 가장 높음을 볼 수 있다. 지각적 특성은 CycleGAN-VC 모델이 최종 점수 0.685로 가장 높은 점수를 보이며, formant similarity와 energy similarity 지표에서 높은 점수를 통해 화자의 고유한 성도 특성을 보존하면서 감정에 맞는 음색 변화와 에너지 조절을 학습함으로써 동일한 화자가 다른 감정을 표현하는 것이 자연스러우며, 지각적으로 우수한 감정 음성을 생성한 것으로 볼 수 있다. 네 번째, 시간 도메인 그룹에서 duration 지표는 모든 모델에서 0.9 이상으로 높은 결과를 보임으로써, 모든 증강 모델이 원본 발화의 시간적 길이를 효과적으로 보존했음을 의미한다. duration을 제외한 나머지 모든 지표에서 VTLP 모델이 높은 점수를 보였다. VTLP 모델은 주파수 축의 워핑(warping)만을 수행하기 때문에 시간 영역의 파형 구조가 거의 그대로 보존되기 때문에 다른 모델에 비해서 상대적으로 보존률이 높은 결과를 보였다. 반면에 decay characteristics 지표는 모든 모델에서 낮은 성능을 보인다. decay characteristics는 발화의 미세한 뉘앙스를 전달하는 특징 정보로, 현재 증강 기법들이 미세 특징 정보를 재현하지 못하는 결과를 볼 수 있으며, 공통적으로 향후 개선의 필요한 부분임을 확인할 수 있다. 다섯 번째, 분포 유사도 그룹에서 wasserstein similarity 지표는 모든 모델에서 0.998로 동일하게 높은 성능을 보였으며, KL divergence similarity 지표의 경우, CycleGAN-VC 모델이 0.949로 가장 좋은 성능을 보임으로써 원본 데이터의 통계적 분포를 효과적으로 학습하고 재현함을 볼 수 있다. VTLP 모델은 cosine similarity 지표에서는 0.901로 가장 높은 유사도를 보였지만, KL divergence와 JS divergence에서 각 0.794, 0.686으로 전반적으로 분포 유사도가 낮은 결과를 보인다. 이는 원본 데이터의 특징 방향성은 유지하지만, 클래스 간 확률적 분포 특성의 재현이 낮음을 확인 할 수 있다. 본 지표를 통해서 GAN기반 방법은 확률적 분포 학습에 우수하지만 VTLP는 특징 방향성 보존에 강점을 보인다. 여섯 번째로 운율 그룹에서 VTLP 모델은 ptich mean 지표가 0.742, pitch variance 지표가 0.727로 피치의 변동성이 유지되고 보존이 높은 것을 볼 수 있으며, 총 점수가 StarGAN-VC 모델에 비해 9.2%, CycleGAN-VC 모델에 비해 8.4% 높은 결과를 볼 수 있다. 반면에 StarGAN-VC 모델은 intonation pattern이 0.509로 억양 패턴이 다른 모델에 비해 손실된 것을 볼 수 있으며, CycleGAN-VC 모델은 pitch variance 지표가 0.510으로써 운율 정보가 많이 손실된 것을 볼 수 있다. 마지막으로 클러스터링 품질 그룹은 모든 모델이 높은 점수를 보임으로써 증강 데이터가 감정 클래스 경계를 보존하는 특성을 가진 것을 볼 수 있다. 본 지표 결과를 분석한 결과, VTLP 모델의 경우, 운율과 시간 특성은 우수하지만, 분포 매칭은 상대적으로 낮은 특성을 확인할 수 있었다. StarGAN-EVC 모델은 분포 학습은 우수하지만, 운율 정보에서 손실이 높은 것을 볼 수 있다. 또한 StarGAN-EVC 모델은 VTLP와 CycleGAN-VC모델에 비해서 거의 모든 지표에서 가장 낮은 결과를 보였다. CycleGAN-VC 모델은 감정 일관성과 지각 품질 지표에서는 성능이 가장 높았으며, 미세 운율 정보에서는 손실이 있음을 확인할 수 있었다.

Table 16은 Table 15의 결과를 기반으로 각 모델 별 핵심 특징을 종합 요약한 정보를 보여준다. ESQ 메트릭을 기반으로 측정한 결과, VTLP 모델이 약 0.706으로 증강된 데이터의 품질 점수가 딥러닝 기반 모델보다 높은 것을 볼 수 있으며, StarGAN-VC 모델이 0.688로 가장 낮은 품질 점수의 결과를 볼 수 있다. 본 결과를 통해 최신 딥러닝 모델이 반드시 최고 성능을 보이지 않으며 전통적인 기법이 여전히 경쟁력 높은 품질을 제공함을 확인 할 수 있다. 또한 ESQ 메트릭을 이용하여 각 증강 모델에 대한 해당 그룹 점수와 세부 지표 점수를 통해서 모델의 약점이 되는 부분과 강점인 특징 정보를 검출함으로써 심층 분석을 수행할 수 있다. 예를 들어, CycleGAN-VC 모델이 감정 일관성 그룹에서 가장 높은 품질 점수를 보였기 때문에 감정 표현이 중요한 용도의 경우에 추천 적용할 수 있다. 따라서 Table 15와 Table 16의 결과 분석을 통해서 제안한 ESQ 메트릭이 기존의 단일 차원의 품질 측정이 아니라, 다차원의 품질 측정을 통해서 각 세부 지표에 대한 강점과 약점을 분석할 수 있는 장점을 보유함을 확인할 수 있었다.

또한 본 논문에서는 제안한 ESQ 메트릭의 유효성을 검증하기 위해서 세 가지 음성 증강 모델 VTLP, StarGAN-VC, CycleGAN-VC에 대한 객관적 품질 평가와 실제 감정 인식 성능 간의 상관관계를 분석 수행하였다. Table 17은 세 가지 증강 모델에 대한 기존 품질 메트릭 PESQ, STOI, SNR과 제안한 ESQ의 평가 결과를 보여준다. 실험 결과, VTLP 모델은 모든 메트릭에서 가장 우수한 성능을 보였으며, 이는 고전 신호 처리 기반 기법의 특성상 원본 음성의 특성을 잘 보존하기 때문으로 분석된다. 딥러닝 기반 모델인 StarGAN-VC와 CycleGAN-VC의 비교 결과 기존 메트릭에서는 두 기법이 거의 동일한 성능을 보였지만, ESQ는 CycleGAN-VC가 StarGAN-VC보다 약 1.31% 높은 품질로 측정되었다.

품질 메트릭 결과를 검증하기 위해서 세 가지 증강 모델로 증강한 데이터 셋을 이용하여, LSTM, CNN_1D, CNN_2D, ResNet_18, ResNet_50, Transformer 6가지 분류 모델을 사용하여 분류 정확도 평가를 수행하였으며 Table 18에 결과를 보여준다. 실험 결과, VTLP 모델이 0.563∓0.040으로 가장 높은 분류 정확도를 보였으며, 본 결과는 기존 품질 메트릭 및 ESQ 결과와 일치한다. 또한 CycleGAN-VC 모델의 분류 정확도가 0.487∓0.143, StarGAN-VC 모델의 분류 정확도가 0.332∓0.054로 약 46.7% 높은 성능을 보였다. 본 결과는 기존 품질 메트릭 PESQ, STOI, SNR에서는 0~0.009%로 품질 차이를 거의 측정하지 못했지만, 제안한 ESQ 메트릭은 약 1.31%의 차이로 두 모델 간의 품질 차이를 측정할 수 있음을 확인할 수 있다. 따라서 본 실험을 통해서 ESQ의 평가 순위와 실제 감정 분류 정확도와 일치하고 있음을 볼 수 있으며, ESQ 메트릭은 32개 세부 지표를 통해서 다차원적 품질 분석을 통해 각 증강 기법의 강점과 약점을 구체적으로 분석할 수 있는 장점이 존재함을 검증하였다. 제안한 ESQ 메트릭은 감정적 특성이 중요한 응용 분야에서 증강 데이터의 품질을 평가하는 효과적인 도구로 활용될 수 있을 것으로 기대한다.

Table 15. The results of measuring the quality of the data set augmented with the existing VTLP, StarGAN-VC, and CycleGAN-VC models using the ESQ metric

No.	Main Quality Measure	Detail Quality Measure	VTLP	StarGAN-VC	CycleGAN-VC
			2013	2018	2018
1	Emotional Consistency	emotion embedding similarity	0.983	0.970	0.974
		arousal consistency	0.744	0.763	0.764
		valence consistency	0.839	0.767	0.859
		Avg.	0.855	0.833	0.866
2	Frequency Domain	spectral centroid	0.527	0.500	0.505
		spectral bandwidth	0.538	0.506	0.515
		spectral rolloff	0.528	0.501	0.506
		spectral contrast	0.548	0.545	0.549
		mfcc similarity	0.984	0.970	0.971
		harmonic ratio	0.523	0.504	0.441
		Avg.	0.608	0.588	0.581
3	Perceptual	energy similarity	0.744	0.763	0.764
		formant cosine similarity	0.913	0.878	0.942
		loudness similarity	0.597	0.578	0.569
		roughness measure	0.655	0.894	0.832
		clarity index	0.427	0.210	0.319
		Avg.	0.667	0.664	0.685
4	Time Domain	rms energy	0.629	0.601	0.592
		zero crossing rate	0.546	0.517	0.527
		duration	0.908	0.911	0.914
		temporal dynamics	0.627	0.589	0.604
		attack time	0.571	0.480	0.516
		decay characteristics	0.219	0.213	0.190
		Avg.	0.583	0.552	0.557
5	Distribution Similarity	cosine similarity	0.901	0.869	0.884
		kl divergence similarity	0.794	0.942	0.949
		wasserstein similarity	0.998	0.998	0.998
		js divergence	0.686	0.659	0.676
		bhattacharyya similarity	0.888	0.870	0.883
		Avg.	0.853	0.868	0.878
6	Prosodic	pitch mean	0.742	0.765	0.685
		pitch variance	0.727	0.607	0.510
		tempo consistency	0.597	0.544	0.621
		intonation pattern	0.580	0.509	0.497
		Avg.	0.662	0.606	0.578
7	Clustering Quality	silhouette score	0.998	0.998	0.998
		davies bouldin score	0.920	0.914	0.924
		calinski harabasz score	0.955	0.950	0.957
		Avg.	0.958	0.954	0.960
ESQ Score			0.706	0.688	0.697

Table 16. Summary of key features of the existing VTLP, StarGAN-VC, and CycleGAN-VC models

	VTLP	StarGAN-VC	CycleGAN-VC
ESQ Score	0.706±0.062	0.688±0.068	0.697±0.063
Rank	1	3	2
Strengths	Prosodic, Time Domain, Frequency Domain	-	Emotional, Distribution, Perceptual
Weaknesses	Distribution	Prosodic, Time Domain	Prosodic, Frequency
Features	Stable quality based on traditional signal prosessing	Balanced performance but weak in some measure	Excellent in emotional consistency and distribution similarity
Bast Use	Prosody focus	Balanced needs	Emotion focus

Table 17. Comparison of augmented quality assessment results between existing metrics and the proposed ESQ metric

	PESQ	STOI	SNR(dB)	ESQ
VTLP	2.138±1.170	0.901±0.060	0.36±3.97	0.706±0.062
StarGAN-VC	1.070±0.148	0.350±0.148	-13.16±2.22	0.688±0.068
CycleGAN-VC	1.070±0.164	0.353±0.060	-13.15±2.21	0.697±0.063

Table 18. Comparison of emotion recognition classification model accuracy based on speech augmentation model

	VTLP	StarGAN-VC	CycleGAN-VC
LSTM	0.6335	0.4163	0.2194
CNN_1D	0.5294	0.3529	0.4137
CNN_2D	0.5701	0.3710	0.6619
ResNet_18	0.5747	0.3122	0.4928
ResNet_50	0.5068	0.2715	0.5324
Transformer	0.5656	0.2670	0.6043
Avg. ∓ Std.	0.563∓0.040	0.332∓0.054	0.487∓0.143

IV. Conclusions

딥러닝 기반 모델의 성능은 학습 데이터의 품질과 양에 크게 좌우되며, 특히 음성·오디오 처리 분야에서는 데이터 획득의 제약으로 인해 여전히 충분한 규모와 다양성을 확보하기 어려운 문제가 존재한다. 이러한 문제점을 극복하는 방법으로 다양한 데이터 증강 기법이 제안되고 있으며, 실제 학습에 활용하기 전에 증강된 데이터의 품질을 객관적으로 검증하는 것은 필수적인 과정이다. 그러나 기존의 대표적인 음성 품질 지표인 PESQ, STOI, SNR은 주로 지각적 음질, 명료도, 신호 대 잡음비에 초점을 두고 있어, 감정 음성과 같이 고차원적이고 복합적인 특성을 갖는 오디오 데이터 품질을 분석할 때 어려움이 존재한다. 실제 본 논문의 실험 결과에서도, 기존 지표들은 증강 레벨에 따른 품질 차이를 명확하게 반영하지 못하거나, 경우에 따라 역전된 결과를 보이는 등 데이터 품질 변화를 일관되게 반영하지 못한 결과를 볼 수 있었다. 따라서 본 논문에서는 기존 품질 측정 지표의 단점을 보완하기 위해서 감정 오디오에 특화된 새로운 품질 측정 지표를 제안하였다. 본 품질 측정 지표는 Time Domain, Frequency Domain, Prosodic, Perceptual, Emotional Consistency, Distribution Similarity, Clustering Quality 등 총 7개 그룹 지표에 대한 음성 특성 항목을 기반으로 품질을 평가하며, 각 항목 별 세부 항목에 대한 점수뿐만 아니라 종합 품질 점수를 함께 제공한다. 이러한 다차원적 평가 구조를 통해 증강된 데이터가 원본 데이터의 시간적·주파수적 특성은 물론, 감정적 일관성과 분포적 유사성까지 유지하는지를 정량적으로 검증할 수 있다. 실험 결과, 제안한 품질 측정 지표는 기존 품질 지표 대비 레벨별 품질 차이를 더 안정적이고 의미 있게 반영하는 것을 확인하였다. 특히, 레벨이 증가함에 따라 품질 점수가 일정한 격차로 향상되는 결과를 보여, 증강 데이터가 단계적으로 개선됨을 설명할 수 있었다. 이는 감정 기반 오디오 증강의 유효성 평가에 있어 기존 지표와 같이, 단일 차원의 품질 측정이 아닌,

다차원의 품질 측정을 수행함으로써 감정 특성의 보존 여부를 확인할 수 있으며, 각 세부 지표에 대한 강점과 약점을 명확하게 파악함으로써 증강 모델의 개선 방향을 도출할 수 있도록 상세한 분석 결과를 제공한다. 또한 실제 학습 데이터 셋 구축 과정에서 증강 기법을 선택하고 최적화하는 데 있어 중요한 기준으로 활용될 수 있음을 시사한다. 본 논문에서 제안한 감정 오디오 특화 품질 메트릭은 향후 음성 감정 인식, 대화형 AI 학습, 멀티모달 감정 분석 등 다양한 응용 분야에서 데이터 증강 검증의 신뢰도를 높이고, 결과적으로 딥러닝 기반 감정 오디오 처리의 성능 향상에 기여할 수 있을 것으로 기대된다.

REFERENCES

[1] S. Zeng, “Audio-visual affect recognition,” IEEE Transaction on Multimedia, Vol 9, No. 2, pp. 424-428, 2007. DOI : 10.48550/arXiv.2208.00344

[2] J. Jeong, S. Mondol, Y. Kim, S. Lee, “An Effective Learning Method for Automatic Speech Recognition,” in Korean CI Patients’ Speech. Electronics, Vol. 10, No. 7, 807, 2021. DOI : 10.3390/electronics10070807

[3] M. Faghani, H. Rezaee-Dehsorkh, N. Ravanshad, H. Aminzadeh, “Ultra-Low-Power Voice Activity Detection System Using Level-Crossing Sampling,” Electronics, Vol. 12, No. 4, 795, 2023. DOI : 10.3390/electronics12040795

[4] B. T. Atmaja and A. Sasou, “Effects of Data Augmentations on Speech Emotion Recognition,” Sensors, 22(16), 5941, 2022. DOI : 10.3390/s22165941

[5] A.W. Rix, A.W, J. G. Beerends, M. P. Hollier, A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 749-752, 2001. DOI : 10.1109/ICASSP.2001.941023

[6] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, “An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech,” IEEE Trans. Audio Speech Lang. Process. Vol. 19, No. 7, pp. 2125-2136, 2011. DOI : 10.1109/TASL.2011.2114881

[7] Y. Hu, P. C. Loizou, “Evaluation of Objective Quality Measures for Speech Enhancement,” IEEE Trans. Audio Speech Lang. Process, Vol. 16, pp. 229-238, 2008. DOI : 10.1109/TASL.2007.911054

[8] K. Shruti, P. Anurag, F. Tiago, "Task-specific speech enhancement and data augmentation for improved multimodal emotion recognition under noisy conditions," Frontiers in Computer Science, Vol. 5, 10.3389, 2023. DOI : 10.3389/fcomp.2023.1039261

[9] A. Zgank, G. Donaj and D. Vlaj, "Speech Quality Assessment and Emotions - Effect on the PESQ Metric," 2024 ELEKTRO (ELEKTRO), pp. 1-4, 2024. DOI : 10.1109/ELEKTRO60337.2024.10556949

[10] S. G. Leem, D. Fulford, J. P. Onnela, D. Gard and C. Busso, "Selective Acoustic Feature Enhancement for Speech Emotion Recognition With Noisy Speech," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 917-929, 2024. DOI : 10.1109/TASLP.2023.3340603

[11] S. Wang, J. Li, L. hao, H. Liu, L. Zhu, X. Zhu, “Speech Enhancement Performance Based on the MANNER Network Using Feature Fusion,” Electronics, Vol. 12, No. 8, 1768, 2023. DOI : 10.3390/electronics12081768

[12] Y. -L. Wu, C. -C. Lee, “MetricAug: A Distortion Metric-Lead Augmentation Strategy for Training Noise-Robust Speech Emotion Recognition,” Proc. INTERSPEECH, pp. 3587-3591, 2023. DOI : 10.21437/Interspeech.2023-819

[13] C. C. Lo, S. W. Fu, W. C. Huang, X, Wang, J. Yar, “MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion,” INTERSPEECH 2019, pp. 1541-1545, 2019. DOI : 10.21437/Interspeech.2019-2003

[14] G. Mittag, B. Naderi, A. Chehadi, S. Moller, “NISQA: A Deep Neural Network for Multidimensional Non-Intrusive Speech Quality Assessment,” INTERSPEECH 2019, pp. 2127-2131, 2021. DOI : 10.21437/Interspeech.2021-299

[15] F. Burkhardt, W. F. Sendlmeier, “Verification of acoustical correlates of emotional speech using formant-synthesis,” Proc. ITRW on Speech and Emotion, pp. 151-156, 2000.

[16] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, S. S. Narayanan, “The INTERSPEECH 2010 paralinguistic challenge,” Proc. Interspeech pp. 2794-2797, 2010. DOI : 10.21437/Interspeech.2010-739

[17] N. Jaitly and G. E. Hinton, “Vocal tract length perturbation (vtlp) improves speech recognition,” in Int. Conf. Mach. Learn. (ICML) Workshop on Deep Learn. Audio, Speech, Lang. Process., 2013. https://www.cs.toronto.edu/~hinton/absps/perturb.pdf

[18] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," in Proc. IEEE SLT Workshop, pp. 266-273, 2018. https://doi.org/10.48550/arXiv.1806.02169

[19] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks," 2018 26th European Signal Processing Conference (EUSIPCO), pp. 2100-2104, 2018. doi: 10.23919/EUSIPCO.2018.8553236

[20] M. Pervaiz, T. Khan,"Speech Emotion Recognition with Distilled Prosodic and Linguistic Affect Representations," International Journal of Advanced Computer Science and Applications(IJACSA), Vol. 7, No. 8, pp. 84-90, 2016. DOI:10.14569/IJACSA.2016.070813

[21] D. Bitouk, R. Verma, A. Nenkova, "Class-Level Spectral Features for Emotion Recognition,", Speech Communication, Vol. 52, Issues 7-8, pp. 613-625, 2010. DOI: 10.1016/j.specom.2010.02.010

[22] C. Wang, Y. Ren, N. Zhang, F. Cui,"Speech Emotion Recognition Based on Multi-feature and Multi-lingual Fusion" Multimedia Tools and Applications, Vol. 81. No. 4, pp. 4897-4907, 2022. DOI:10.1007/s11042-021-10553-4

[23] Y. Wang, C. Lu, Y. Zong, H. Lian, Y. Zhao, S. Li, "Time-Frequency Transformer: A Novel Time Frequency Joint Learning Method for Speech Emotion Recognition," Communications in Computer and Information Science, pp. 415-427, 2023. DOI:10.1007/978-981-99-8138-0_33

[24] Y. Zhao, J. Wang, Y. Zong, W. Zheng, H. Lian, L. Zhao, "Deep implicit distribution alignment networks for cross-corpus speech emotion recognition," EEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5, 2023. DOI: 10.1109/ICASSP49357.2023.10095388

[25] N. Gasteiger, J. Lim, M. Hellou, B. Macdonald, "A Scoping Review of the Literature On Prosodic Elements in Human-Robot/Agent Interaction," International Journal of Social Robotics, Vol. 16, No. 4, pp. 1-12, 2022. DOI:10.1007/s12369-022-00913-x

[26] I. R. Ulgen, Z. Du, C. Busso, B. Sisman, "Revealing emotional clusters in speaker embeddings: a contrastive learning strategy for speech emotion recognition," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 12081-12085, 2024. DOI: 10.1109/ICASSP48485.2024.10447060

[27] K. Bhangale and M. Kothandaraman, "Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network," Electronics, Vol. 12, No. 4, 839, 2023. DOI: 10.3390/electronics12040839

[28] R. Bachu, S. Kopparthi, B. Adapa, and B. Barkana, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal," in American Society for Engineering Education (ASEE) Zone Conference Proceedings, pp. 1-7, 2008. DOI:10.18260/1-2-1153-53698

[29] S. Abdulkadir, S. Hassan, and V. Harpale, "Speech emotion recognition based on multi-feature speed rate and LSTM," Neurocomputing, Vol. 551, 2024. DOI: 10.1016/j.neucom.2024.127883

[30] C. Jiang, P. Fu, H. Tao, P. Lei, and L. Zhao, "Parallelized Convolutional Recurrent Neural Network with Spectral Features for Speech Emotion Recognition," IEEE Access, Vol. 7, pp. 90368-90377, 2019. DOI: 10.1109/ACCESS.2019.2927384

[31] S. Cunningham, "Supervised machine learning for audio emotion recognition," Personal and Ubiquitous Computing, Vol. 25, pp. 637-650, 2021. DOI: 10.1007/s00779-020-01389-0

[32] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in International Conference on Acoustics, Speech, and Signal Processing, pp. 381-384, 1990. DOI: 10.1109/ICASSP.1990.115702

[33] M. Alhussein, G. Muhammad, M. S. Hossain, and S. U. Amin, "Speech Emotion Recognition Using Attention Model," International Journal of Environmental Research and Public Health, Vol. 20, No. 6, 5140, 2023. DOI: 10.3390/ijerph20065140

[34] K. Bhangale and M. Kothandaraman, "Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network," Electronics, Vol. 12, No. 4, 839, 2023. DOI: 10.3390/electronics12040839

[35] M. A. Khan, H. Ashraf, M. I. Javed, M. A. Iqbal, and A. H. Dar, "Enhanced speech emotion understanding using advanced attention-centric convolutional networks," Biomedical Signal Processing and Control, Vol. 103, 2025. DOI: 10.1016/j.bspc.2025.107271

[36] A. Davis, M. Alhussein, K. Haider, M. Iqbal, and M. Muhammad, "Real-time speech emotion recognition using deep learning and data augmentation," Artificial Intelligence Review, Vol. 58, 2025. DOI: 10.1007/s10462-024-11065-x

[37] H. Zhiyan and W. Jian, "Speech Emotion Recognition Based on Improved Masking EMD and Convolutional Recurrent Neural Network," Frontiers in Psychology, Vol. 13, 1075624, 2022. DOI: 10.3389/fpsyg.2022.1075624

[38] T. Bänziger and K. R. Scherer, "The role of intonation in emotional expressions," Speech Communication, Vol. 46, No. 3-4, pp. 252-267, 2005. DOI: 10.1016/j.specom.2005.02.003

[39] E. Rodero, "Intonation and emotion: Influence of pitch levels and contour type on creating emotions," Music Perception, Vol. 28, No. 1, pp. 73-83, 2010. DOI: 10.1525/mp.2010.28.1.73

[40] S. Lee, S. Yildirim, A. Kazemzadeh, and S. Narayanan, "An articulatory study of emotional speech production," in Proc. Interspeech, pp. 497-500, 2005.

[41] A. Ortony and T. J. Turner, "What's basic about basic emotions?," Psychological Review, Vol. 97, No. 3, pp. 315-331, 1990. DOI: 10.1037/0033-295X.97.3.315

[42] P. A. Abhang, B. W. Gawali, and S. C. Mehrotra, "Introduction to EEG- and Speech-Based Emotion Recognition," Academic Press, 2016.

[43] J. Park, Y. Kim, and E. Bulyko, "Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence," Information Sciences, Vol. 566, pp. 299-315, 2021. DOI: 10.1016/j.ins.2021.02.049

[44] P. Larrouy-Maestri, D. Poeppel, and M. D. Pell, "The Sound of Emotional Prosody: Nearly 3 Decades of Research and Future Directions," Perspectives on Psychological Science, Vol. 19, No. 1, pp. 3-29, 2024. DOI: 10.1177/17456916231217722

[45] R. G. Slyh, W. T. Nelson, and E. G. Hansen, "Analysis of mrate, shimmer, jitter, and F0 contour features across stress and speaking style in the SUSAS database," in Proc. ICASSP, 1999. DOI:

10.1109/ICASSP.1999.758345

[46] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," Speech Communication, Vol. 48, No. 9, pp. 1162-1181, 2006. DOI: 10.1016/j.specom.2006.04.003

[47] S. G. Leem, D. Fulford, J. P. Onnela, D. Gard, and C. Busso, "Selective Acoustic Feature Enhancement for Speech Emotion Recognition With Noisy Speech," IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 32, pp. 917-929, 2024. DOI: 10.1109/TASLP.2023.3340603

[48] J. A. Russell, "A circumplex model of affect," Journal of Personality and Social Psychology, Vol. 39, No. 6, pp. 1161-1178, 1980. DOI: 10.1037/h0077714

[49] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," Development and Psychopathology, Vol. 17, No. 3, pp. 715-734, 2005. DOI: 10.1017/S0954579405050340

[50] A. Davis, M. Alhussein, K. Haider, M. Iqbal, and M. Muhammad, "Real-time speech emotion recognition using deep learning and data augmentation," Artificial Intelligence Review, Vol. 58, 2025. DOI: 10.1007/s10462-024-11065-x

[51] S. Kullback and R. A. Leibler, "On information and sufficiency," The Annals of Mathematical Statistics, Vol. 22, No. 1, pp. 79-86, 1951. DOI: 10.1214/aoms/1177729694

[52] C. Villani, "Optimal Transport: Old and New," Springer-Verlag Berlin Heidelberg, 2009. DOI: 10.1007/978-3-540-71050-9

[53] J. Lin, "Divergence measures based on the Shannon entropy," IEEE Transactions on Information Theory, Vol. 37, No. 1, pp. 145-151, 1991. DOI: 10.1109/18.61115

[54] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," Bulletin of the Calcutta Mathematical Society, Vol. 35, pp. 99-109, 1943. https://www.jstor.org/stable/25047882

[55] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," Journal of Computational and Applied Mathematics, Vol. 20, pp. 53-65, 1987. DOI: 10.1016/0377-0427(87)90125-7

[56] D. L. Davies and D. W. Bouldin, "A cluster separation measure," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-1, No. 2, pp. 224-227, 1979. DOI: 10.1109/TPAMI.1979.4766909

[57] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," Communications in Statistics - Theory and Methods, Vol. 3, No. 1, pp. 1-27, 1974. DOI: 10.1080/03610927408827101

Authors

Do Kyung Shin received the B.S. in computer science and engineering from Baekseok University, Republic of Korea, in 2006, and the M.S. and Ph.D in computer science and engineering from Hanyang University, Republic of Korea, in 2008 and 2015, respectively. Dr. Shin has been working as a Chief Researcher Engineer at LIG Nex1 since 2015. Her major research interests include computer vision, image processing, speech emotion recognition, intelligent target identification, deep learning, and data augmentation.

Young Dae Kim received the B.S. in electronics engineering from Ajou University, Republic of Korea, in 2002. Mr. Kim has been working as a Chief Researcher Engineer at LIG Nex1 since 2002. His major research interests are in the development of naval combat system utilizing artificial intelligence.