pISSN 2005-8063
eISSN 2586-5854
2025. 6. 30.
Vol.17 No.2
pp. 39-48

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2025.17.2.039

Expressive voice conversion enhancing prosody and emotion consistency*

Su-Jin Koo · Hoi-Rin Kim**

School of Electrical Engineering, Korea Advanced Institute of Science & Technology (KAIST), Daejeon, Korea

Abstract

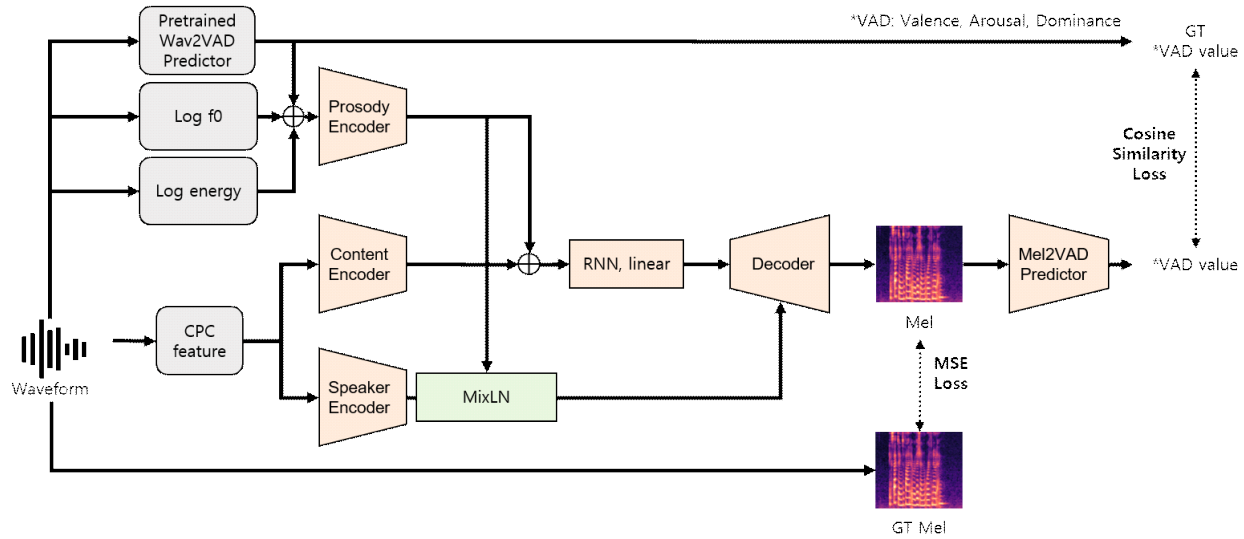
In Korean voice-conversion tasks, not only converting speaker identity but also preserving prosody and emotional consistency is essential, as intonation and rhythm are crucial in conveying meaning in the language. However, conventional voice-conversion (VC) systems focus primarily on altering speaker timbre while overlooking expressive aspects such as prosody and emotion. This limitation becomes particularly problematic in applications such as animation dubbing or emotionally expressive voice generation, where nuanced delivery is critical. Hence, we propose a novel expressive voice-conversion (EVC) model. Our model is based on the triple adaptive attention normalization–VC framework and introduces a prosody embedding that combines F0, energy, and emotional attributes represented by valence, arousal, and dominance (VAD). This embedding captures the prosodic characteristics of Korean more precisely. Furthermore, we adopt mix-layer normalization to suppress prosodic information in the speaker encoder, thereby enhancing the separation of speaker identity and prosody. To further strengthen emotional expressiveness, a dedicated VAD predictor is incorporated to guide emotion learning. Experiments conducted on Korean speech data show that our model outperforms existing EVC systems in terms of prosody preservation and emotional delivery. Notably, our model achieves an average prosody mean opinion score of 4.11, thereby indicating the generation of natural and expressive Korean speech. This study demonstrates a promising direction for improving both accuracy and expressiveness in VC systems

Keywords: expressive voice conversion, disentanglement, prosody consistency, emotion consistency

1. 서론

음색 변환(voice conversion, VC)은 발화자의 언어적 내용을 유지한 채, 음색을 다른 화자의 것으로 변환하는 기술로, 음성 생성 및 인터랙티브 콘텐츠 제작 분야에서 핵심적인 기술로 자리 잡고 있다. VC는 버추얼 유튜버, 감정 전달이 필요한 오디오 콘텐츠, 더빙 시스템 등 다양한 실제 응용 사례에서 활용 가능성이 크다. 특히 음성 합성의 자연스러움과 몰입도를 향상시키는

* This work was supported by the National Research of Foundation of Korea (No. 2021R1A2C1014044).
** hoirkim@kaist.ac.kr, Corresponding author
Received 1 Apr 2025; Revised 10 May 2025; Accepted 23 May 2025



CPC, contrastive predictive coding; MixLN, mix layer normalization.

그림 1. 제안하는 모델의 학습 과정
Figure 1. Training process of proposed model

측면에서 그 역할이 점점 확대되고 있다(Sisman et al., 2020).

VC는 텍스트에서 음성을 바로 합성하는 텍스트 음성 변환(text-to-speech, TTS)과 달리 원 화자의 실제 발화를 입력으로 사용하기 때문에, 단순한 언어 정보뿐만 아니라 발화에 담긴 운율이나 감정과 같은 표현 정보를 일정 수준 보존할 수 있다는 장점을 갖는다. 그러나 기존 VC 모델들은 주로 화자의 음색 정체성 변환에 초점을 맞추고 있어, 운율과 감정 등 표현 정보의 정밀한 재현에는 한계를 보이는 경우가 많다. 이러한 한계의 원인은 다음과 같다. 첫째, 운율 정보를 간접적으로 처리하거나 명시적으로 모델링하지 않기 때문에 변환된 음성이 단조롭고 부자연스럽게 들리는 경우가 자주 발생한다. 둘째, 감정 정보 보존이 충분하지 않아, 화자 독립적 감정(speaker-independent emotional cue)이 제대로 유지되지 못하고, 화자 종속적 감정 스타일(speaker-dependent emotional style)과 분리되지 않아 감정 전달력이 저하되는 문제가 있다. 마지막으로, 운율과 감정 정보는 밀접하게 연관되어 있음에도 이를 통합적으로 처리하는 접근이 부족하여, 결과적으로 변환된 음성에서 자연스러움과 표현력이 저하되는 현상이 나타난다.

이러한 한계는 한국어와 같은 운율적 특성이 두드러진 언어에서는 더욱 명확하게 나타난다. 한국어는 억양, 리듬, 음장(길이)이 문장의 화행(speech act)을 결정하는 데 중요한 역할을 한다. 이에 따라 같은 문장도 운율에 따라 질문, 명령, 감정 표현 등으로 전혀 다르게 해석될 수 있다. 이처럼 운율이 의미 해석에 직접적인 영향을 미치는 언어적 특성 때문에, 운율의 미세한 손실조차도 의미 왜곡이나 감정 전달력 저하로 이어질 수 있다. 그럼에도 불구하고, 많은 기존 VC 연구는 한국어와 같이 운율 민감도가 높은 언어에 대한 정밀한 분석과 실험은 상대적으로 부족한 편이다.

이러한 문제를 해결하기 위해, 최근에는 화자 독립적인 감정 정보를 보존하면서도 화자 고유의 감정 스타일을 효과적으로 변환하는 EVC(expressive voice conversion) 기술이 제안되고 있다. Du et al.(2021)은 음성 감정 인식(speech emotion recognition, SER) 모델을 활용하여 화자 정체성과 감정 스타일을 함께 모델링하는 구조를 제안하였고, Gan et al.(2022)은 자기 지도적 학습 기반 음성 표현 임베딩과 병목 특징(bottleneck features, BNF)을 활용해 언어적·준언어적 표현을 정교하게 추출하였다. 그러나 이러한 모델들은 주로 다대일(many-to-one) 또는 화자 ID 기반 다대다(many-to-many) 프레임워크에 의존하고 있어, 훈련에서 보지 않은 화자에 대한 일반화 성능이 떨어지는 한계가 있다. 이와 같은 한계를 극복하기 위해 최근에는 학습 데이터에 포함되지 않은 새로운 화자에 대해서도 음색 변환이 가능한 제로샷(zero-shot) 기반의 EVC 모델이 연구되고 있다. 예를 들어, 일부 연구에서는 데이터 증강 및 대조 학습을 활용하여 운율과 언어적 내용을 분리하는 기법을 도입하였으며, 운율 예측기를 활용하여 운율을 효과적으로 모델링하는 방법도 제안되었다(Chen et al., 2023; Deng et al., 2023). 또한, 감정과 운율 분리를 위한 mutual information loss 기반 학습(Du et al., 2022) 등이 제안되고 있다. 그러나 여전히 감정 라벨을 포함한 학습 데이터의 부족, 감정 정보의 불완전한 분리, 운율 보존의 일관성 저하 등 해결되지 않은 문제들이 존재한다.

더불어, 이처럼 EVC 모델들은 표현력 측면에서는 진전을 이루었으나, 전통적인 VC 모델들과 비교할 때 내용 보존이나 화자 유사성과 같은 핵심 성능 지표에서는 상대적으로 낮은 성능을 보이는 경향이 있다.

이에 본 연구는 감정 및 운율 정보를 정교하게 통합하는 동시에, 내용 보존과 화자 유사성 간의 균형이 유지되는 VC 모델을 기반으로 설계할 필요가 있다고 판단하였다. TriAAN-VC(triple adaptive attention normalization-VC; Park et al., 2023)는 감정 정보

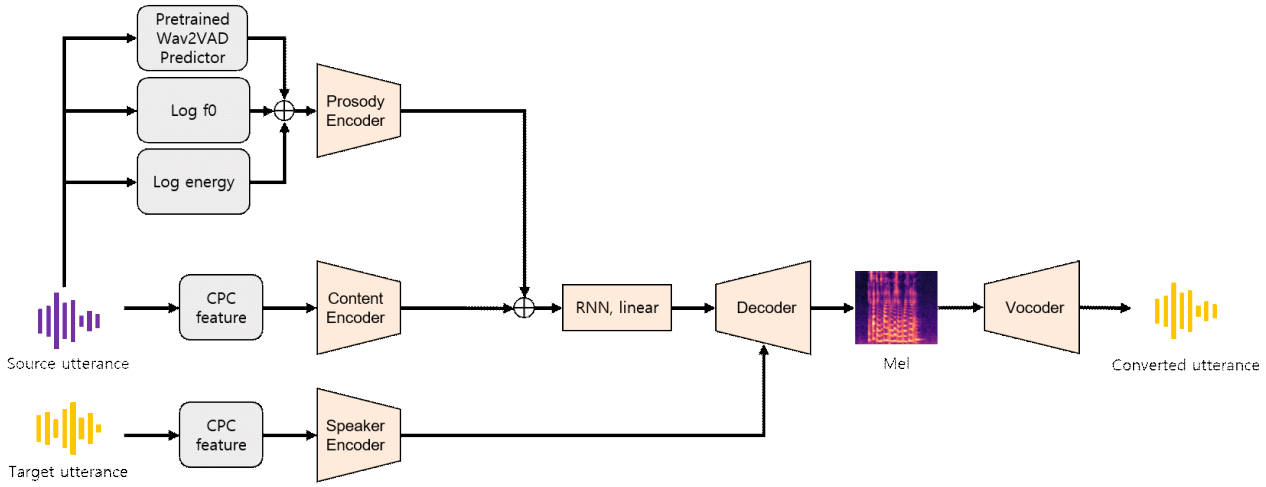


그림 2. 제안하는 모델의 추론 과정
Figure 2. Inference process of proposed model

를 명시적으로 처리하지 않지만, 전통적인 VC 모델 중에서도 화자 유사성과 내용 보존 간 균형성이 뛰어난 구조로 평가된다. 본 연구는 이를 백본으로 삼아, 운율 및 감정 표현을 보완하는 구조를 추가하였다. 이를 위해 F0, 에너지, VAD(valence, arousal, dominance)를 결합한 운율 임베딩을 도입하고, MixLN(mix layer normalization) 및 감정 예측 모듈을 추가하였다.

이러한 설계를 바탕으로, 본 논문은 다음과 같은 네 가지 측면에서 기여한다.

1. 연속적인 감정 표현인 VAD를 VC 프레임워크에 통합하여 감정 정보를 보다 세밀하게 반영하였다.
2. Mel-spectrogram으로부터 감정 상태를 회귀하는 Mel2VAD 예측기 구조를 도입하여 감정 일관성 보존에 기여하였다.
3. MixLN을 운율-화자 분리에 활용함으로써, 운율 정보의 독립성과 F0/에너지 보존 성능을 향상시켰다.
4. 위 구성 요소들을 TriAAN-VC 백본 위에 유기적으로 통합하고, 손실 함수 및 학습 스케줄을 조정하여 감정 및 운율 표현력을 강화한 VC 모델을 구현하였다.

본 논문에서는 한국어 데이터셋에 기반한 정량적·정성적 평가를 수행하였으며, 해당 접근 방식이 한국어처럼 운율에 민감한 언어에서도 자연스럽고 표현력 높은 음성을 안정적으로 생성할 수 있음을 확인하였다. 이는 기존 VC 시스템의 감정 및 운율 표현 부족 문제를 보완할 수 있음을 보여주며, 운율 중심 언어를 포함한 다양한 언어 환경으로의 확장 가능성을 뒷받침한다.

2. 제안 방법

본 연구는 TriAAN-VC(Park et al., 2023)를 백본 모델로 채택하였다. TriAAN-VC는 제로샷 음색 변환을 위한 구조로, CPC(contrastive predictive coding) 기반의 내용/화자 분리와 TriAAN을 통해 화자 정체성과 언어 내용을 효과적으로 분리하는 데 강점을 가진다. 이 구조에서 내용 인코더와 화자 인코더는 동일한 CPC feature를 입력으로 사용하지만, 구조적으로 서로 다른 정보를 학습하도록 설계되어 있다. 특히 화자 인코더에는 SSA(speaker aattention) 모듈이 포함되며, 이는 TIN(time-wise instance normalization)과 self-attention을 결합하여 채널 간 관계를 보존하고 화자의 전역 특성을 강조한다. 본 연구에서는 이 구조를 그대로 계승하여 사용하였다.

그림 1은 제안하는 모델의 전체 학습 구조를 시각적으로 나타낸 것이다. 입력 음성으로부터 네 가지 특징(CPC, Log-F0, Log-Energy, VAD)이 추출되며, 이 중 CPC는 내용 인코더와 화자 인코더에, 나머지 세 가지는 운율 인코더의 입력으로 사용된다. 운율 인코더는 운율 임베딩을 생성하고, 이는 MixLN을 통해 speaker embedding과 정규화된다. 최종적으로 내용, 화자, 운율 임베딩이 디코더로 전달되어 멜 스펙트로그램을 생성한다.

본 연구에서 기존 TriAAN-VC 구조에 새롭게 추가한 구성 요소는 다음과 같다. 첫째, 운율 인코더는 Log-F0, Log-Energy, VAD 정보를 통합하여 운율 임베딩을 생성한다. 둘째, MixLN은 화자 임베딩과 운율 임베딩 간의 간섭을 줄여 분리를 강화하는 역할을 한다. 셋째, Mel2VAD 예측기는 생성된 멜 스펙트로그램에서 감정 상태를 회귀한다. 이 중 MixLN과 Mel2VAD 예측기는 학습 과정에서만 사용되며, 추론 시에는 제외된다.

2.1. 모델

2.1.1. 특징 추출

내용 및 화자 정보의 추출을 위해 각 인코더의 입력으로 CPC 기반 임베딩을 사용하였다(van den Oord et al., 2018). CPC는 자기 지도 학습 방식으로, 과거 입력을 기반으로 미래 표현을 예측하여 시퀀스 데이터에서 의미 있는 표현을 추출하는 기법이다. 이를 통해 CPC는 음성 데이터에서 언어적 정보와 화자 특성을 효과적으로 추출한다.

또한 본 연구는 감정을 보다 정밀하게 표현하기 위해 이산적

인 라벨 대신 연속적인 값을 갖는 VAD 기반의 감정 표현 방법을 사용하였다(Islam et al., 2021). VAD는 감정을 세 가지 차원으로 나타내는 방법으로, Valence는 감정의 긍정/부정 정도를, Arousal은 감정의 강도를, Dominance는 감정의 지배성을 나타낸다. 본 연구에서는 사전 학습된 Wav2VAD 예측기를 이용하여 음성 신호로부터 추출된 감정 값을 모델의 입력으로 활용하였다.

마지막으로 운율 정보는 로그 스케일로 변환한 기본 주파수(Log-F0)와 음성 신호의 로그 에너지(Log-Energy)를 결합하여 표현하였다.

2.1.2. 운율 인코더

본 연구에서 도입한 운율 인코더는 기존 모델이 Log-F0만 활용했던 방식을 확장하여, Log-F0뿐만 아니라 VAD와 Log-Energy를 통합적으로 활용하여 운율 정보를 더 풍부하게 표현하도록 설계되었다. 운율 인코더는 입력된 특징을 Conv2D 레이어와 ReLU 활성화 함수, 배치 정규화를 포함하는 기초 컨볼루션 레이어, 총 6개의 잔차 구조를 가진 인코더 블록, 그리고 출력을 담당하는 Conv2D 레이어로 구성된다.

2.1.3. 믹스 레이어 정규화

화자 임베딩과 운율 임베딩 간의 분리를 보다 효과적으로 수행하기 위해 MixLN을 도입하였다. MixLN(Huang et al., 2022)은 기존의 LN(layer normalization) 및 CLN(conditional layer normalization)의 확장된 형태로, 화자와 운율의 특성 간 명확한 구분을 유도하는 것을 목표로 한다. 구체적으로는 운율 임베딩을 화자 임베딩과 혼합하여 미세한 변형을 유도하고, 이를 바탕으로 레이어 정규화를 적용하여 두 가지 특성 간의 간섭을 최소화한다. 결과적으로 운율 정보와 화자 정보는 더욱 효과적으로 분리되어 음성 변환의 품질을 향상시킨다.

기본적인 LN는 입력 벡터 x의 평균 μ와 분산 σ를 이용해 정규화한 후, 학습 가능한 스케일 벡터 γ와 바이어스 벡터 β를 적용하는 방식으로 정의된다.

$$LN(x) = \gamma \times \frac{x-\mu}{\sigma} + \beta \tag{1}$$

그리고 CLN은 스타일 임베딩 w에 따라 γ와 β를 조절하여 특정 스타일에 적응하도록 설계된다.

$$CLN(x,w) = \gamma(w) \times \frac{x-\mu}{\sigma} + \beta(w) \tag{2}$$

여기서 $\gamma(w)$와 $\beta(w)$는 스타일 임베딩 w를 입력으로 받아 선형변환을 통해 계산된다.

$$\gamma(w) = E_{\gamma} \times w \tag{3}$$

$$\beta(w) = E_{\delta} \times w \tag{4}$$

그에 반면, MixLN은 학습과정에서 스타일정보를 혼합하여 모델이 다양한 스타일에 적응하도록 유도한다. 구체적으로, 스타일 벡터 w를 무작위로 섞은 $\tilde{w}$를 생성하고, 두 스타일의 가중합을 통해 새로운 γ_{mix}와 β_{mix}를 계산한다.

$$\gamma_{mix}(w) = \lambda\gamma(w) + (1-\lambda)\gamma(\tilde{w}) \tag{5}$$

$$\beta_{mix}(w) = \lambda\beta(w) + (1-\lambda)\beta(\tilde{w}) \tag{6}$$

여기서 λ는 Beta 분포 $\beta(\alpha,\alpha)$에서 샘플링되며, α는 하이퍼파라미터로 설정된다. 이렇게 계산된 γ_{mix}와 β_{mix}를 사용해 입력 x를 정규화 한다.

$$MixLN(x,w) = \gamma_{mix}(w) \times \frac{x-\mu}{\sigma} + \beta_{mix}(w) \tag{7}$$

2.1.4. Mel2VAD 예측기

Mel2VAD 예측기는 생성된 멜 스펙트로그램에서 VAD 값을 예측함으로써 모델이 원본 발화의 운율과 감정을 효과적으로 보존하도록 돕는다. 예측된 VAD 값과 정답 VAD 값 간의 코사인 유사도를 기반으로 손실을 계산하여 모델의 학습을 진행한다. 이 모듈은 CNN 블록, GRU 블록, 출력 레이어로 구성된다. CNN 블록은 ReLU 활성화 함수와 BatchNorm2D를 사용하는 두 개의 Conv2D 레이어와 각 레이어 뒤에 MaxPooling을 포함한다. GRU 블록은 2층 양방향 GRU로 구성되며, 출력 레이어는 GRU의 출력을 VAD 값으로 매핑하는 완전 연결층(fully connected layer)으로 이루어진다.

2.2. 손실 함수

2.2.1. 재구성 손실

재구성 손실은 실제 멜 스펙트로그램과 예측된 멜 스펙트로그램 간의 L1손실을 사용하였으며, 이는 TriAAN-VC 백본 모델에서 채택된 방식을 따랐다. 손실 함수는 다음과 같이 정의된다.

$$L_{L1}(y,\hat{y}) = \| y - \hat{y} \|_1 \tag{8}$$

여기서 y는 정답 멜 스펙트로그램이고, $\hat{y}$는 예측된 멜 스펙트로그램이다.

2.2.2. 시암 손실

시암 손실은 실제 멜 스펙트로그램과 시간 마스킹으로 증강된 입력특징을 사용한 예측된 멜 스펙트로그램간의 L1 손실로 정의된다.

$$L_{L1}(y,\hat{y}_{siam}) = \| y - \hat{y}_{siam} \|_1 \tag{9}$$

여기서 $\hat{y}_{siam}$은 시간 마스킹된 입력 특징으로 예측된 멜 스펙트로그램을 의미한다.

2.2.3. VAD(valence, arousal, dominance) 예측 손실

VAD 예측손실은 실제 VAD와 예측된 VAD간의 코사인 유사도 손실로 정의된다. 본 연구에서는 감정 벡터의 방향 일관성을

학습하도록 유도하기 위해, 이 손실 함수를 적용하였다. vad_y가 정답 VAD이고 $vad_{\hat{y}}$가 예측된 VAD일 때 다음과 같이 계산된다.

$$L_{\cos}(vad_y, vad_{\hat{y}}) = 1 - \frac{vad_y - vad_{\hat{y}}}{\| vad_y \| \ \| vad_{\hat{y}} \|} \quad (10)$$

2.2.4. 총 손실

총 손실 함수는 학습 에포크에 따라 가중치가 다르게 적용된다. 우선, 에포크 1부터 100까지는 다음과 같이 계산된다.

$$L_{total} = \frac{L_{L1}(y,\hat{y}) + L_{L1}(y,\hat{y}_{siam})}{2} + L_{L1}(\hat{y},\hat{y}_{siam}) \quad (11)$$

그리고 에포크 101부터 500까지는 다음과 같이 계산된다.

$$L_{total} = \frac{L_{L1}(y,\hat{y}) + L_{L1}(y,\hat{y}_{siam})}{2} + L_{L1}(\hat{y},\hat{y}_{siam}) + 0.2L_{\cos}(vad_y, vad_{\hat{y}}) \quad (12)$$

이러한 학습 스케줄은 실험적으로 설정되었으며, TriAAN-VC 기반 구조에서 약 100 epoch 이후부터 VAD 예측 손실을 추가하는 것이 학습 안정성과 성능 향상 측면에서 효과적이라는 실험적 결과에 기반하였다.

2.3. 추론

추론 과정은 다음과 같다. 그림 2는 제안하는 모델의 전체 추론 흐름을 시각적으로 보여주며, 각 단계에서 어떤 정보가 어떻게 사용되는지를 나타낸다. 먼저 원 화자로부터 VAD, Log-F0, Log-Energy, 그리고 CPC 특징을 추출한다. 이렇게 추출된 특징들은 각각의 인코더를 거쳐 운율 임베딩과 내용 임베딩을 생성한다. 한편, 목표 화자에서는 CPC 특징을 활용해 화자 임베딩을 추출한다. 이 세 가지 임베딩(운율, 내용, 화자)은 디코더에 입력되어 멜 스펙트로그램을 생성하며, 마지막으로 이 멜 스펙트로그램을 보코더에 입력하여 최종 음성 파형을 얻는다.

3. 실험

3.1. 실험 개요

본 연구는 제안한 모델의 성능을 검증하기 위해 한국어 감정 음성 데이터셋을 활용하여 실험을 설계하였다. 성능 평가는 객관적 지표와 주관적 지표를 기준으로 이루어졌다. 객관적 평가는 문자 오류율(character error rate, CER), 화자 임베딩 간의 코사인 유사도(speaker embedding cosine similarity, SECS), F0 피어슨 상관계수(F0 Pearson correlation coefficient, F0 PCC), 에너지 피어슨 상관계수(energy pearson coefficient, Energy PCC)를 통해 변환 음성의 명료도, 화자 유사성, 운율 일관성을 정량적으로 분석하였다. 주관적 평가는 청취자 평가를 기반으로 음성의 자연스러움(naturalness mean opinion score, NMOS), 화자 유사성(speaker similarity mean opinion score, SMOS), 운율 유사성(prosody similarity mean opinion score, PMOS)을 측정하였다.

또한, 제안한 모델의 구성 요소별 기여도를 확인하기 위해 ablation study를 수행하였으며, 학습에 포함된 화자(seen speaker)와 포함되지 않은 화자(unseen speaker)를 구분하여 일반화 성능을 평가하였다.

비교 실험을 위해 다음의 세 가지 기존 음성 변환 모델을 베이스라인으로 설정하였다.

- StyleVC(Du et al., 2022)는 화자 정체성과 감정 스타일을 분리하는 초기 Expressive VC 모델로, mutual information loss를 통해 스타일, 화자, 내용, F0를 분리하는 구조를 갖는다. 하지만 특징들 간 상호 의존성을 완전히 제거하지 못해 감정 표현의 일관성과 음질에서 일부 한계가 있었다.
- UUVC(unified unsupervised voice conversion; Chen & Watanabe & Rudnicky, 2023)는 F0, 에너지, 발화 길이 등 운율 요소를 예측하여 변환하는 방식이며, 이산적 음성 단위를 활용해 언어 정보와 운율 정보를 분리한다. 운율 표현력이 높은 장점이 있으나, 예측 기반 운율 생성 과정에서 원 화자의 억양이 과도하게 왜곡될 가능성이 있다.
- TriAAN-VC(Park et al., 2023)는 제로샷 VC 모델로, CPC 기반 content/speaker 분리 및 TriAAN 블록을 통해 세밀한 화자 특성과 전반적인 화자 정보를 추출하고 이를 변환에 활용한다. F0 정보를 활용하긴 하지만, 감정 스타일이나 운율 정보를 명시적으로 분리하거나 학습하는 구조는 포함되어 있지 않다.

3.2. 데이터셋

본 연구에서는 AIHub에서 제공하는 한국어 감정 음성 데이터셋을 사용하였다. 이 데이터셋은 전문 성우 8명과 일반인 501명의 녹음 데이터를 포함하며, 다양한 감정 레이블이 부착되어 있다. 본 연구에서는 일반인 320명과 전문 성우 4명(총 324명)의 데이터를 학습에 사용하였다. 그리고 평가 데이터셋은 전문 성우 중 학습에 포함된 4명(seen speakers)과 포함되지 않은 4명(unseen speakers)을 각각 선정하여 구성하였다. 각 성우는 감정별로 10개의 발화를 제공하였으며, 평가 과정에서는 이 발화들을 source-target으로 매칭하여 변환 성능을 평가하였다. 이때 seen speakers와 unseen speakers는 각각 별도의 그룹 내에서 독립적으로 조합을 구성하였다.

3.3. 실험환경

입력 데이터는 16 kHz 샘플링 주파수의 오디오 신호이며, 25 ms의 윈도우 크기와 10 ms의 홉 크기로 설정된 80-bin 멜 스펙트로그램을 생성하여 모델의 입력 특징으로 사용하였다. 또한, Facebook AI에서 공개한 CPC 모델(Facebook Research, 2019)과 Hugging Face에서 공개된 Wav2VAD 예측기(3loi, 2023)를 통해 각각의 임베딩을 추출하여 모델 입력으로 활용하였다. 최적화 방법으로는 Adam 옵티마이저를 사용하였으며, 학습률은 10^{-4}으로 설정하였다. 최종 음성 합성은 ParallelWaveGAN 보코더(Yamamoto et al., 2020)를 통해 수행하였다.

3.4. 평가 지표

3.4.1. 객관적 평가 지표

변환된 음성의 전사 정확도를 평가하기 위해 CER를 측정하였다. CER은 한국어 음성 인식 모델인 Whisper- small-ko (SungBeom, 2023)을 사용하여 변환된 음성을 텍스트로 변환한 뒤, 해당 결과와 실제 정답 텍스트 간의 차이를 기반으로 계산된다. 값이 낮을수록 명료도가 높은 음성을 의미하며, 계산식은 다음과 같다.

$$CER = \frac{S + D + I}{N} \tag{13}$$

여기서 S는 치환된 문자 수, D는 삭제된 문자 수, I는 삽입된 문자 수, N은 원본 문장의 총 문자 수이다.

그리고 변환된 음성이 목표 화자와 얼마나 유사한지를 평가하기 위해 화자 검증 기반 지표인 SECS를 사용하였다. SECS는 두 음성의 화자 임베딩 간 코사인 유사도를 계산하는 방식으로, Guan et al.(2024)과 같은 음성 합성 연구에서 화자 유사성을 정량적으로 평가하는 데 널리 활용된다. 본 연구에서는 화자 임베딩 추출을 위해 Resemblyzer의 VoiceEncoder를 사용하였으며, 코사인 유사도는 다음과 같이 계산된다:

$$SECS = \frac{S_{target} \times S_{converted}}{\| S_{target} \| \, \| S_{converted} \|} \tag{14}$$

여기서 S_{target}는 목표 화자의 임베딩 벡터, $S_{converted}$는 변환된 음성의 화자 임베딩 벡터이다. 값이 1에 가까울수록 두 음성 간의 화자 유사성이 높음을 의미한다.

또한, 운율 보존도를 정량적으로 측정하기 위해 F0 PCC를 활용하였다. F0 PCC는 변환 전후의 피치(F0) 곡선 간 선형 상관관계를 나타내며, –1부터 1 사이의 값을 가진다. 본 연구에서는 변환된 음성이 원본의 억양 구조를 얼마나 잘 보존했는지를 평가하기 위해 이 지표를 사용하였다. 이는 Guo et al.(2024)을 비롯한 다양한 VC 연구에서 prosody 일관성 평가 지표로 활용되고 있으며, F0 PPC 계산 공식은 다음과 같다.

$$F0PCC = \frac{cov(F0_{source}, F0_{converted})}{\sigma_{F0_{source}} \sigma_{F0_{converted}}} \tag{15}$$

여기서 $F0_{source}$와 $F0_{converted}$는 각각 원본과 변환된 음성의 F0 벡터이다.

마지막으로 에너지 패턴 간의 상관관계를 평가하기 위해 Energy PCC를 측정하였으며, 높은 값은 더 나은 에너지 일관성을 의미한다. Energy PCC 계산 공식은 다음과 같다.

$$EnergyPCC = \frac{cov(E_{source}, E_{converted})}{\sigma_{E_{source}} \sigma_{E_{converted}}} \tag{16}$$

여기서 E_{source}와 $E_{converted}$는 각각 원본과 변환된 음성의 에너지 벡터이다.

3.4.2. 주관적 평가 지표

주관적 평가는 총 10명의 평가자가 참여하여 실시하였다. 이 중 7명은 음성 청취 및 평가 경험을 보유한 성인으로, 음성 합성 또는 음성 품질 관련 프로젝트 및 연구에 실제로 참여한 이력을 가진 전문가로 구성되었으며, 나머지 3명은 관련 경험이 없는 일반인이었다. 모든 평가자는 실험 전에 평가 목적과 MOS(mean ppinion score) 방식에 대한 안내를 받은 후, 기준 음원을 제시받고 항목별 판단 기준을 숙지하였다. 이를 통해 전문가와 일반인 간 평가 편차를 최소화하고, 평가 기준의 일관성을 확보하고자 하였다. 평가 항목은 음성의 자연스러움(NMOS), 화자 유사성(SMOS), 운율 유사성(PMOS)의 세 가지 항목으로 나누어졌다. 모든 항목은 1점(매우 나쁨)에서 5점(매우 좋음)까지의 MOS 방식으로 평가하였으며, 높은 점수는 해당 항목의 품질이 우수함을 나타낸다. 자연스러움(NMOS)은 음성의 인간적인 명료도와 자연성을 평가하며, 화자 유사성(SMOS)은 변환된 음성이 목표 화자의 음색과 얼마나 유사한지를 나타내고, 운율 유사성(PMOS)은 억양, 감정 및 리듬과 같은 운율적 특성이 원본과 얼마나 유사하게 유지되었는지 평가한다.

4. 결과

4.1. 객관적 평가 지표

표 1. 객관적 평가 결과(seen speakers)
Table 1. Objective evaluation results (seen speakers)

집단	CER (↓,%)	SECS (↑)	F0 PCC (↑)	Energy PCC(↑)
StyleVC	57.7	0.591	0.717	0.853
UUVC	**22.1**	0.652	0.708	0.968
TriAAN-VC	28.7	**0.762**	0.739	0.970
Proposed	23.7	0.751	**0.745**	**0.971**

CER, character error rate; SECS, speaker embedding cosine similarity; F0 PCC, F0 pearson correlation coefficient; Energy PCC, energy pearson coefficient; UUVC, unified unsupervised voice conversion; TriAAN-VC, triple adaptive attention normalization-voice conversion.

표 2. 객관적 평가 결과(unseen speakers)
Table 2. Objective evaluation results (unseen speakers)

집단	CER (↓,%)	SECS (↑)	F0 PCC (↑)	Energy PCC(↑)
StyleVC	55.1	0.609	0.720	0.871
UUVC	**22.1**	0.649	0.689	0.967
TriAAN-VC	27.6	**0.767**	0.742	0.971
Proposed	**22.1**	0.759	**0.747**	**0.973**

CER, character error rate; SECS, speaker embedding cosine similarity; F0 PCC, F0 pearson correlation coefficient; Energy PCC, energy pearson coefficient; UUVC, unified unsupervised voice conversion ; TriAAN-VC, triple adaptive attention normalization-voice conversion.

표 1과 표 2는 베이스라인 모델들과 제안된 모델의 객관적 평가 결과를 비교한 것이다. Seen speaker 조건에서 제안된 모델은 F0 Pearson 상관계수(0.745)와 Energy Pearson 상관계수(0.971) 모두에서 가장 높은 성능을 보이며, 운율 및 에너지 정보를 효과적으로 보존하는 것으로 나타났다. TriAAN-VC와 비교했을 때, 화자 유사도(SECS)는 0.751로 약간 낮았지만 유사한 수준이었으며, 문자 오류율(CER)은 23.7%로 훨씬 개선되었고, 운율 및 에너지 관련 지표에서도 우수한 성능을 기록하였다. UUVC와 비교하면 CER 측면에서 약간 높았으나, SECS, F0 PCC, Energy PCC 측면에서 더 나은 결과를 보이며 전반적인 품질 측면에서 우수함을 입증하였다.

Unseen Speaker 조건에서도 유사한 경향이 관찰되었다. 제안된 모델은 F0 PCC(0.747)와 Energy PCC(0.973) 항목에서 가장 높은 수치를 기록하며, 학습 과정에서 보지 못한 화자에 대해서도 운율과 에너지의 일관성을 안정적으로 유지하는 성능을 보였다. TriAAN-VC와 비교 시 SECS(0.759 vs. 0.767)는 비슷한 수준이었고, CER은 22.1%로 훨씬 개선되었다. UUVC와는 명료도에서는 유사하였으나, 운율 및 에너지 관련 지표에서는 본 모델이 꾸준히 높은 성능을 기록하였다. 이러한 결과는 제안된 접근 방식이 운율 보존 및 감정 표현의 정밀도 측면에서 유의미한 성능 개선을 달성했음을 시사한다.

4.2. 주관적 평가 지표

표 3. 주관적 평가 결과(unseen speakers)
Table 3. Subjective evaluation results (unseen speakers)

집단	NMOS(↑)	SMOS(↑)	PMOS(↑)
StyleVC	2.08	1.92	3.30
UUVC	**3.41**	2.99	3.39
TriAAN-VC	3.30	3.38	3.85
Proposed	3.40	**3.39**	**4.11**

NMOS, naturalness mean opinion score; SMOS, speaker similarity mean opinion score; PMOS, prosody similarity mean opinion score; UUVC, unified unsupervised voice conversion: TriAAN-VC, triple adaptive attention normalization-VC.

표 3은 주관적 평가에 대한 비교 결과를 보여준다.

자연스러움(NMOS) 항목에서는 UUVC가 3.41로 가장 높은 점수를 기록하며, 음성의 자연스러운 발화에 있어 약간의 우위를 보였다. 본 연구에서 제안한 모델은 3.40으로 거의 동등한 수준의 성능을 나타냈으며, TriAAN-VC(3.30)와 StyleVC(2.08)에 비해서는 확연히 높은 점수를 보였다. 이는 제안된 모델이 음성의 자연스러움을 유지하는 데 있어 충분히 경쟁력 있는 품질을 제공함을 의미한다. 화자 유사성(SMOS) 평가에서는 제안된 모델이 3.39로 가장 높은 점수를 기록하며, 화자의 정체성 보존 측면에서 우수한 성능을 입증하였다. TriAAN-VC는 3.38로 근소한 차이를 보였지만, UUVC(2.99) 및 StyleVC(1.92)와 비교하면 뚜렷한 성능 차이가 관찰되었다. 이는 제안된 모델이 unseen speaker에 대해서도 화자 일관성을 효과적으로 유지함을 보여준다. 운율 유사성(PMOS) 측면에서도 제안된 모델은 4.11로 가장 높은 점수를 나타냈으며, 이는 운율 표현의 일관성과 정밀도 측면에서 탁월한 성능을 의미한다. TriAAN-VC는 3.85, UUVC는 3.39, StyleVC는 3.30의 점수를 기록하였으며, 제안된 모델과의 성능 차이가 명확하게 드러났다. 이 결과는 제안한 접근 방식이 운율 정보 보존에 있어 뚜렷한 강점을 지님을 시사한다.

4.3. Ablation Study

표 4는 제안된 모델의 주요 구성 요소가 전체 성능에 미치는 영향을 분석한 ablation study 결과를 나타낸다. Seen speaker 조건에서는 운율 임베딩 추가 시 CER이 28.7%에서 28.5%로 소폭 개선되었으며, 이는 운율 정보가 문장 명료도 향상에 일부 기여할 수 있음을 시사한다. 이후 MixLN을 추가하면 CER이 23.9%로 크게 감소하고 F0 PCC(0.741), Energy PCC(0.971)가 함께 향상되며 운율 및 에너지 일관성 유지에 긍정적인 영향을 미쳤다. 그리고 Mel2VAD 예측기만 도입한 경우 CER은 27.0%로 감소하고 SECS는 0.765로 증가하여 화자 유사성이 향상되었다. 모든 모듈을 통합한 최종 모델은 CER 23.7%, SECS 0.751, F0 PCC 0.745, Energy PCC 0.971로 균형 잡힌 성능을 보였다.

Unseen speaker 조건에서도 유사한 경향이 확인되었다. 운율 임베딩 추가 시 CER은 27.6%로 소폭 개선되었고, MixLN 추가 시 CER이 22.3%로 크게 향상되었으며, Mel2VAD 예측기만 사용 시 SECS는 0.769로 가장 높은 화자 유사성을 보였다. 최종적으로 모든 구성 요소가 포함된 모델은 CER 22.1%, SECS 0.759, F0 PCC 0.747, Energy PCC 0.973으로 전반에 걸쳐 우수한 성능을 보였다.

5. 결론

본 연구에서는 TriAAN-VC 모델을 기반으로 운율 및 감정 정보의 보존 성능을 개선하기 위한 새로운 방법을 한국어 데이터셋을 통해 검증하였다. 기존 EVC(expressive voice conversion) 기술에서는 운율과 감정의 미세한 특성을 충분히 유지하지 못해 음성의 자연스러움과 표현력이 저하되는 한계가 존재하였다. 이를 극복하고자 운율 임베딩, MixLN, VAD 예측기 등의 추가 모듈을 도입하였다.

표 4. Ablation study 결과
Table 4. Ablation study results

Prosody embed-ding	MixLN	Mel2VAD pre-dictor	CER(↓,%)		SECS(↑)		F0 PCC(↑)		Energy PCC(↑)	
			Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen
×	×	×	28.7	27.6	0.762	0.767	0.739	0.742	0.970	0.971
◯	×	×	28.5	27.3	0.762	0.767	0.739	0.743	0.970	0.971
◯	◯	×	23.9	22.3	0.750	0.758	0.741	0.745	0.971	0.973
×	×	◯	27.0	26.1	**0.765**	**0.769**	0.740	0.743	0.971	0.972
◯	◯	◯	**23.7**	**22.1**	0.751	0.759	**0.745**	**0.747**	**0.971**	**0.973**

MixLN, mix layer normalization; CER, character error rate; SECS, speaker embedding cosine similarity; F0 PCC, F0 pearson correlation coefficient; Energy PCC, energy pearson coefficient.

운율 임베딩 모듈을 통해 기본 주파수(F0), 에너지, 그리고 VAD 값을 결합하여 보다 명확하고 풍부한 운율 정보를 표현할 수 있도록 하였으며, MixLN 기법을 활용하여 화자와 운율 특성을 효과적으로 분리함으로써 음성 변환의 일관성과 자연스러움을 개선하였다. 또한 VAD 예측기를 도입하여 음성 내의 감정 표현을 더욱 정교하게 반영함으로써 표현력을 향상시켰다.

한국어 데이터셋을 이용한 실험 결과, 운율 임베딩과 MixLN, VAD 예측기가 모두 포함된 최종 모델에서 F0 PCC와 Energy PCC가 향상되었으며, 특히 운율 일관성과 표현력이 주관적 평가 지표(PMOS)에서 뚜렷한 성능 향상을 나타냈다. 이를 통해 변환된 음성이 더욱 자연스럽고 생생한 표현력을 지녔음을 입증하였다. 본 연구의 결과는 음성 변환 기술의 활용 범위를 넓히고, 더빙, 애니메이션, 음성 합성 등의 다양한 분야에서 화자의 의도와 감정을 더욱 효과적으로 전달할 수 있는 가능성을 제시한다.

References

Chen, L. W., Watanabe, S., & Rudnicky, A. (2023, June). A unified one-shot prosody and speaker conversion system with self-supervised discrete speech units. *Proceedings of ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Rhodes Island, Greece.

Deng, Y., Tang, H., Zhang, X., Wang, J., Cheng, N., & Xiao, J. (2023, October). Pmvc: Data augmentation-based prosody modeling for expressive voice conversion. *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 184-192). Ottawa, Canada.

Du, Z., Sisman, B., Zhou, K., & Li, H. (2022). Disentanglement of emotional style and speaker identity for expressive voice conversion. arXiv. https://doi.org/10.48550/arXiv.2110.10326.

Du, Z., Sisman, B., Zhou, K., & Li, H. (2021, December). Expressive voice conversion: A joint framework for speaker identity and emotional style transfer. *Proceedings of 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 594-601). Cartagena, Colombia.

Facebook Research. (2019). CPC_audio: Contrastive predictive coding for audio representation [Machine learning model]. Retrieved from https://github.com/facebookresearch/CPC_audio

Goncalves, L., Salman, A. N., Naini, A. R., Moro-Velázquez, L., Thebaud, T., Garcia, P., Dehak, N., ... Busso, C. (2024, Jun). Odyssey 2024-speech emotion recognition challenge: Dataset, baseline framework, and results. *Proceedings of The Speaker and Language Recognition Workshop (Odyssey 2024)* (pp. 4-54). Quebec, Canada.

Guan, W., Li, Y., Li, T., Huang, H., Wang, F., Lin, J., Huang, L., ... Hong, Q. (2024, March). MM-TTS: Multi-modal prompt based style transfer for expressive text-to-speech synthesis. *AAAI Technical Track on Natural Language Processing I*, *38*(16), 18117-18125.

Guo, Y., Li, Z., Li, J., Du, C., Wang, H., Wang, S., Chen, X., & Yu, K. (2024). Vec2wav 2.0: Advancing voice conversion via discrete token vocoders. arXiv. https://doi.org/10.48550/arXiv.2409.01995.

Huang, R., Ren, Y., Liu, J., Cui, C., & Zhao, Z. (2022). Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech. arXiv. https://doi.org/10.48550/arXiv.2205.07211.

Islam, M. R., Moni, M. A., Islam, M. M., Rashed-Al-Mahfuz, M., Islam, M. S., Hasan, M. K., Hossain, M. S., ... & Ió, P. (2021). Emotion recognition from EEG signal focusing on deep learning and shallow learning techniques. *IEEE Access*, *9*, 94601-94624.

Kan-bayashi. (2019). ParallelWaveGAN: High-quality speech synthesis [Machine learning model]. Retrieved from https://github.com/kan-bayashi/ParallelWaveGAN

Park, H. J., Yang, S. W., Kim, J. S., Shin, W., & Han, S. W. (2023, June). Triaan-vc: Triple adaptive attention normalization for any-to-any voice conversion. *Proceedings of ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Rhodes Island, Greece.

Sisman, B., Yamagishi, J., King, S., & Li, H. (2020). An

overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 132-157.

SungBeom. (2023). Whisper-small-ko: Korean ASR model [Machine learning model]. Retrieved from https://huggingface.co/SungBeom/whisper-small-ko

van den Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv.https://doi.org/10.48550/arXiv.1807.03748

Veaux, C., Yamagishi, J., & MacDonald, K. (2017). CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit [Dataset]. Retrieved from Retrieved from https://doi.org/10.7488/ds/1994

Yamamoto, R., Song, E., & Kim, J. M. (2020, May). Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. *Proceedings of ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6199-6203). Barcelona, Spain.

3loi. (2023). SER-odyssey-baseline-wavLM-multi-attributes [Machine learning model]. Retrieved from https://huggingface.co/3loi/SER-Odyssey-Baseline-WavLM-Multi-Attributes

• **구수진 (Su-Jin Koo)**
한국과학기술원 전기및전자공학부 석사
대전광역시 유성구 대학로 291
Tel: 042-350-7617
Email: sujin.koo@kaist.ac.kr
관심분야: 음성합성

• **김회린 (Hoi-Rin Kim)** 교신저자
한국과학기술원 전기및전자공학부 교수
대전광역시 유성구 대학로 291
Tel: 042-350-7417
Email: hoirkim@kaist.ac.kr
관심분야: 음성인식, 음성합성, 화자인식, 패턴인식

운율 및 감정 보존을 향상시키는 음색 변환 기법*

구 수 진 · 김 회 린

한국과학기술원(KAIST) 전기및전자공학부

국문초록

한국어 음성 변환 작업에서는 음색만이 아닌 운율과 감정의 일관된 전달이 매우 중요하지만, 기존의 음색 변환(voice conversion, VC) 기술은 주로 화자의 음색만을 바꾸는 데 집중되어 있어, 한국어처럼 억양과 리듬이 의미 전달에 중요한 언어에서는 제한적인 성능을 보인다. 특히 대사 전달이 중요한 애니메이션 더빙이나 감정 표현이 중요한 콘텐츠에서 이 문제는 더욱 부각된다. 본 연구는 표현 음색 변환(expressive voice conversion, EVC) 모델을 새롭게 제안하여 이러한 문제를 해소하고자 한다. 제안된 모델은 TriAAN-VC 구조를 기반으로 하며, F0, 에너지, 그리고 정서적 차원인 VAD(valence, arousal, dominance)를 통합한 운율 임베딩을 활용하여 한국어의 운율적 특징을 정교하게 반영한다. 또한, MixLN(mix layer normalization)을 통해 화자 인코더 내 운율 정보를 효과적으로 제거하여 화자 고유 특성과 운율 사이의 간섭을 줄였다. 아울러, 감정 정보 학습을 위한 VAD 예측 모듈을 추가함으로써, 감정 표현력을 강화하였다. 한국어 화자 데이터를 대상으로 수행한 실험에서는 운율 보존 및 감정 전달 측면에서 기존 EVC 모델을 상회하는 성능을 입증하였다. 특히 주관적 평가인 PMOS(prosody mean opinion score)에서 평균 4.11점을 기록하며, 보다 자연스럽고 감정 표현이 풍부한 한국어 음성을 생성함을 확인하였다. 본 연구는 한국어 음색 변환 기술의 정밀성과 표현력을 동시에 향상시킬 수 있는 방향성을 제시한다.

핵심어: 음성 합성, 음색 변환, 운율 보존, 감정 보존

* 본 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행되었음(No. 2021R1A2C1014044)