Journal of Computing and Data Technology

https://journal.krrishpub.com/index.php/jcdt/index

Enhancing Emotional Expressiveness in Voice Conversion Using Seq2Seq and CycleGAN

Faouzi Didi [a], Vugar Abdullayev [b,c], Mohammed R. Hayal [d], Ebrahim E. Elsayed [d,*]

[a] *Department of Common Core in Technology, Laboratory of Physics of Experimental Techniques and its Applications, University Yahia Fares of Medea, Medea, 26000, Algeria.*

[b] *Department of Computer Engineering, Azerbaijan State Oil and Industry University, AZ1010, Azerbaijan.*

[c] *Information Technologies and Cybersecurity Laboratory, Azerbaijan University of Architecture and Construction, AZ1073, Azerbaijan.*

[d] *Department of Electronics and Communications Engineering, Faculty of Engineering, Mansoura University, Mansoura, 35516, Egypt.*

Abstract

Emotional voice conversion (EVC) aims to transform the emotional characteristics of speech while preserving speaker identity and linguistic content, and plays a critical role in affective computing, expressive speech synthesis, and human–computer interaction. Despite recent progress, existing EVC approaches often struggle to jointly model long-term temporal dependencies and achieve perceptually realistic spectral transformations, particularly in non-parallel training scenarios. To address these challenges, this paper proposes a high-fidelity emotional voice conversion framework that integrates Sequence-to-Sequence (Seq2Seq) temporal modeling with Cycle-Consistent Generative Adversarial Networks (CycleGAN). The proposed architecture operates entirely in the Mel-spectrogram domain. A Seq2Seq encoder–decoder with attention is first employed to capture long-range temporal dependencies and generate a coarse emotion-aware spectral representation. Subsequently, a CycleGAN-based refinement module enhances spectral realism and emotional expressiveness through adversarial and cycle-consistent learning, without requiring parallel emotional speech data. A neural vocoder is finally used to reconstruct the time-domain waveform from the refined Mel-spectrogram. The proposed framework is evaluated on the Emotional Speech Dataset (ESD) using objective metrics including Mel-Cepstral Distortion (MCD), fundamental frequency (F_0) root-mean-square error (RMSE), and structural similarity index (SSIM), along with subjective listening evaluations. Experimental results demonstrate that the proposed Seq2Seq–CycleGAN model outperforms conventional Seq2Seq-only and CycleGAN-only baselines in terms of emotional expressiveness, speech naturalness, and speaker similarity, confirming the effectiveness of jointly leveraging temporal modeling and adversarial spectral refinement for high-quality emotional voice conversion.

Keywords:
Emotional voice conversion, Seq2Seq modeling, CycleGAN, Non-parallel speech transformation, Mel-spectrograms, Speaker identity preservation.

Article Information:
DOI: https://doi.org/10.71426/jcdt.v1.i2.pp98-103
Received: *23 November 2025* | Revised: *24 December 2025* | Accepted: *29 December 2025*

1. Introduction

Speech is a primary channel of human communication that conveys not only linguistic content but also affective cues such as emotion and speaking style. Emotional voice conversion aims to transform the perceived emotion of a source utterance into a target emotion while preserving speaker identity and intelligibility, enabling practical use in empathetic dialogue systems, expressive TTS, dubbing, gaming, assistive technologies, and therapy-oriented speech tools [12]. Early voice conversion relied on statistical mappings (e.g., GMM/trajectory-based approaches), which often produced over-smoothed spectra and limited expressiveness under emotion shifts [1], [2]. With deep learning, representation learning and sequence modeling have improved controllability and conversion quality, yet maintaining naturalness and stable speaker identity under strong emotional changes remains challenging [13]–[15].

Recent progress has been driven by non-parallel training paradigms that remove the need for frame-level alignment. Many-to-many adversarial conversion (e.g., StarGAN-style) improved flexibility across multiple domains [12], while cycle-consistent learning reduced parallel-data dependence

*Corresponding author
Email address: didifouzi19@gmail.com, didi.faouzi@univ-medea.dz (Faouzi Didi), abdulvugar@mail.ru, vuqar.abdullayev@asoiu.edu.az (Vugar Abdullayev), mohammedraisan@gmail.com, mohammedraisan@std.mans.edu.eg (Mohammed R. Hayal), engebrahem16@gmail.com, engebrahem16@std.mans.edu.eg (Ebrahim E. Elsayed).

List of acronyms

Acronym	Expansion
EVC	Emotional Voice Conversion
Seq2Seq	Sequence-to-Sequence Neural Network
GAN	Generative Adversarial Network
CycleGAN	Cycle-Consistent Generative Adversarial Network
DNN	Deep Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory Network
VAE	Variational Autoencoder
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
MCD	Mel-Cepstral Distortion
SSIM	Structural Similarity Index Measure
PESQ	Perceptual Evaluation of Speech Quality
MOS	Mean Opinion Score
ESD	Emotional Speech Dataset
F_0	Fundamental Frequency
STFT	Short-Time Fourier Transform
MLP	Multilayer Perceptron
VC	Voice Conversion

and enhanced realism [13]. However, adversarial training can be unstable and may not reliably capture long-range prosodic dependencies that govern emotional expression across an utterance [14], [16]. Conversely, sequence-to-sequence (Seq2Seq) modeling is effective for temporal dependency learning and global prosody planning [15] ,[17], but Seq2Seq-only pipelines may still suffer from over-smoothing or insufficient fine-grained spectral refinement when emotional intensity changes sharply [18].

EVC has evolved from statistical spectral mapping to deep generative and sequence modeling. Classical VC work introduced probabilistic spectral transforms and trajectory modeling, but these approaches typically exhibit over-smoothing and limited generalization under expressive conditions [3], [4]. Deep representation learning improved feature disentanglement and reconstruction fidelity, but emotion conversion adds complexity because emotion is expressed through coupled spectral and prosodic attributes [5], [6].

GAN-based VC advanced realism by learning distribution-level mappings. Cycle-consistent conversion removed the parallel-data requirement and enabled non-parallel training, but stability and mode collapse remain concerns [7], [8]. Extensions such as CycleGAN-VC3 and masking-based training improved robustness for unaligned sequences and mel-spectrogram conversion [9], [10]. For emotional conversion, transformer-enhanced CycleGAN variants and seen/unseen emotional style transfer frameworks report improved expressiveness, yet they may still struggle with long-span prosody planning and speaker identity drift under strong emotion changes [11], [25], [26].

Seq2Seq and attention-based architectures are effective for modeling long-term dependencies in speech, which is critical for prosody and emotion trajectories. However, Seq2Seq-only conversion can produce overly averaged acoustic outputs without an explicit realism constraint, motivating adversarial refinement. More recently, controllable emotional intensity and one-shot emotional conversion methods show improved user control, but robustness across speakers and emotions is still an open challenge, especially under non-parallel constraints and limited emotional data.

1.1. Research gaps

Based on the above literature, the key gaps addressed in this work are:

- Temporal–spectral coupling gap: Many GAN-based EVC pipelines enhance realism but do not explicitly enforce long-range temporal prosody consistency, while Seq2Seq pipelines may lack high-frequency realism constraints [9], [11].
- Non-parallel stability gap: Non-parallel emotional conversion remains sensitive to training instability and domain mismatch across emotions/speakers [7]– [10], [26].
- Evaluation gap: Some studies report limited objective/perceptual metrics jointly covering emotion, quality, and identity; standardized evaluation on ESD with complementary metrics is still needed [19] –[22].

To address these limitations, this paper proposes an integrated EVC framework that combines (i) Seq2Seq modeling for robust temporal alignment and prosody-aware feature extraction, with (ii) Cycle-consistent adversarial refinement for realistic spectral/prosodic distribution matching under non-parallel settings. The proposed approach is trained and evaluated on the ESD, a widely used benchmark with multiple emotions and speakers, enabling systematic assessment of emotion transfer and identity preservation [19]. Performance is quantified using objective measures commonly adopted for voice conversion and perceptual quality assessment, including MCD, F_0-related errors, SSIM, and perceptual metrics such as PESQ [20], [21]. In addition, speaker similarity is reported to verify identity retention under conversion [23], [24].

1.2. The main contributions of the proposed work

- A unified EVC pipeline that couples Seq2Seq temporal modeling with CycleGAN-based refinement for non-parallel emotional conversion.
- An ESD-based evaluation protocol with objective and perceptual metrics that jointly measure emotion transfer quality and speaker identity preservation.
- A structured comparison against representative prior EVC/VC paradigms, highlighting practical trade-offs in stability, data requirements, and emotion controllability.

2. Methodology

This section presents the proposed high-fidelity emotional voice conversion framework, which integrates Seq2Seq temporal modeling with CycleGAN. The objective is to transform the emotional attributes of speech while preserving speaker identity and linguistic content. The framework

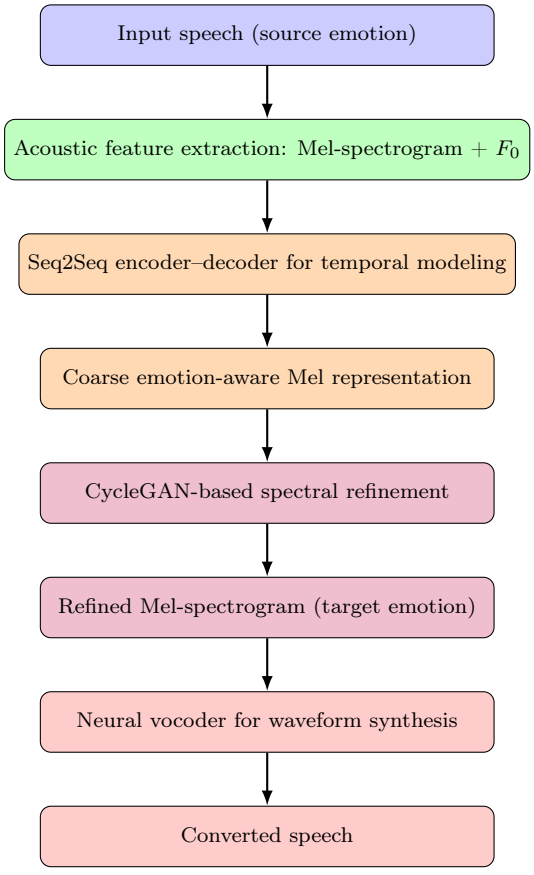

Figure 1: Workflow of the proposed Seq2Seq–CycleGAN emotional voice conversion framework. (Blue: input; green/orange: temporal modeling; purple: adversarial spectral refinement; red: waveform reconstruction).

operates entirely in the Mel-spectrogram domain and comprises acoustic feature extraction, Seq2Seq-based temporal encoding, CycleGAN-based spectral refinement, and neural vocoder-based waveform reconstruction.

2.1. Overall framework

Workflow of the proposed Seq2Seq–CycleGAN emotional voice conversion framework is shown in Figure 1. Given an input speech signal $x(t)$ associated with a source emotion, the goal of emotional voice conversion is to generate a converted speech signal $\hat{x}(t)$ that conveys a desired target emotion without altering the speaker characteristics. To achieve this, a Seq2Seq encoder–decoder architecture is employed to capture long-term temporal dependencies and emotion-aware representations, followed by a CycleGAN module that refines the spectral structure through adversarial and cycle-consistent learning. This hybrid design effectively combines temporal modeling and perceptual realism without requiring parallel emotional speech data.

2.2. Acoustic feature representation

Let the input speech waveform be denoted by $x(t)$. The Mel-spectrogram representation is computed as (1).

$$\mathbf{M} = \text{MelSpec}(x(t)) \in \mathbb{R}^{T\times F} \tag{1}$$

In (1), T is the number of time frames and F denotes the Mel-frequency dimension. In addition, the fundamental frequency contour is extracted using the WORLD vocoder as (2).

$$\mathbf{f}_0 = \text{WORLD}(x(t)) \tag{2}$$

2.3. Seq2Seq-based temporal modeling

The Seq2Seq encoder maps the input Mel-spectrogram sequence into a latent representation (3).

$$\mathbf{h}_t = \text{Encoder}(\mathbf{M}_t) \tag{3}$$

where $\mathbf{h}_t$ denotes the encoder hidden state at time index t. A location-sensitive attention mechanism computes alignment weights between encoder and decoder states (4).

$$\alpha_{t,k} = \frac{\exp(e_{t,k})}{\sum_{k'}\exp(e_{t,k'})} \tag{4}$$

with the attention score defined as (5).

$$e_{t,k} = \mathbf{v}^\top \tanh(\mathbf{W}_h\mathbf{h}_k + \mathbf{W}_s\mathbf{s}_{t-1}) \tag{5}$$

In (5), $\mathbf{s}_{t-1}$ is the previous decoder state. The decoder generates a coarse emotion-aware Mel-spectrogram (6).

$$\hat{\mathbf{M}}_t^{(c)} = \text{Decoder}(\mathbf{c}_t, \mathbf{s}_{t-1}) \tag{6}$$

In (7), $\mathbf{c}_t$ is the context vector obtained via attention.

2.4. CycleGAN-based spectral refinement

To improve perceptual quality and emotional expressiveness, a CycleGAN is applied to the coarse Mel-spectrogram $\hat{\mathbf{M}}^{(c)}$. The CycleGAN learns bidirectional mappings between source and target emotion domains (7).

$$G : \mathcal{X} \to \mathcal{Y}, \quad F : \mathcal{Y} \to \mathcal{X}. \tag{7}$$

The adversarial loss is defined as (8).

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(G, D_Y) = {} & \mathbb{E}_{y\sim p_{\text{data}}(y)}\left[\log D_Y(y)\right] \\ & + \mathbb{E}_{x\sim p_{\text{data}}(x)}\left[\log(1 - D_Y(G(x)))\right]\end{aligned} \tag{8}$$

Cycle-consistency is enforced through (9).

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_x[\|F(G(x))-x\|_1]+\mathbb{E}_y[\|G(F(y))-y\|_1] \tag{9}$$

while identity preservation is ensured by (10).

$$\mathcal{L}_{id} = \mathbb{E}_y[\|G(y) - y\|_1] + \mathbb{E}_x[\|F(x) - x\|_1] \tag{10}$$

2.5. Overall objective function

The total training objective of the proposed framework is formulated as (11).

$$\mathcal{L}_{total} = \lambda_{gan}\mathcal{L}_{GAN} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{id}\mathcal{L}_{id} + \lambda_{rec}\mathcal{L}_{rec} \tag{11}$$

In (11), $\mathcal{L}_{rec}$ denotes the Seq2Seq reconstruction loss and λ terms control the relative contribution of each component.

Algorithm 1 Seq2Seq–CycleGAN emotional voice conversion.

Require: Source waveform x_s, target emotion label e_t, (optional) target speaker ID s_t

Ensure: Converted waveform $\hat{x}_t$ with emotion e_t

Feature extraction (front-end):

Compute log-Mel spectrogram $\mathbf{M}_s = \Phi_{\text{mel}}(x_s)$ and pitch $F_0 = \Phi_{F_0}(x_s)$ using (1) and (2)

Seq2Seq content encoding:

Obtain encoder hidden states $\mathbf{H} = \text{Encoder}(\mathbf{M}_s)$ using (3)

Emotion-conditioned decoding (coarse conversion):

Form emotion embedding $\mathbf{z}_{e_t} = \text{Emb}(e_t)$ (and speaker embedding $\mathbf{z}_{s_t}$ if used)

Generate coarse Mel spectrogram $\hat{\mathbf{M}}^{(c)} = \text{Decoder}(\mathbf{H}, \mathbf{z}_{e_t})$ using (6)

CycleGAN refinement (non-parallel spectral realism):

Define generators $G_{S\to T}, G_{T\to S}$ using (7)

Refine Mel spectrogram:

$$\hat{\mathbf{M}}_t = G_{S\to T}\left(\hat{\mathbf{M}}^{(c)}\right)$$

Cycle-consistency enforcement:

$$\tilde{\mathbf{M}}_s = G_{T\to S}(\hat{\mathbf{M}}_t), \quad \mathcal{L}_{cyc} = \|\mathbf{M}_s - \tilde{\mathbf{M}}_s\|_1$$

Adversarial optimization:

Optimize generators and discriminators using $\mathcal{L}_{GAN}$ in (8) and $\mathcal{L}_{cyc}$ in (9)

Waveform reconstruction:

Reconstruct waveform $\hat{x}_t$ from $\hat{\mathbf{M}}_t$ using a neural vocoder

return $\hat{x}_t$

2.6. Algorithmic description

As shown in Algorithm 1, the proposed algorithm performs emotional voice conversion by integrating the temporal modeling capability of Seq2Seq architecture with the spectral refinement strength of a CycleGAN. Initially, acoustic features such as Mel-spectrograms and fundamental frequency are extracted from the source speech and encoded using a Seq2Seq encoder to capture long-term temporal dependencies. The decoder generates an emotion-conditioned coarse representation, which is subsequently refined through CycleGAN mappings to enhance spectral realism and emotional expressiveness without requiring parallel training data. Finally, a neural vocoder reconstructs the waveform from the refined Mel-spectrogram, yielding emotionally transformed speech while preserving speaker identity and intelligibility.

3. Results and discussion

3.1. Dataset and preprocessing

All experiments were conducted on the publicly available *Emotional Speech Dataset* [14], [22] a widely adopted benchmark for emotional speech synthesis and emotional voice conversion research. The dataset consists of recordings from over 300 speakers expressing five emotional states—*neutral*, *happy*, *sad*, *angry*, and *surprise*. In this study, only the English subset was used for training and evaluation. The bilingual nature of the dataset enables evaluation across linguistic variations, although this work focuses primarily on emotion conversion performance.

Specifically, 80% of the speakers were used for training, 10% for validation, and the remaining 10% for testing. All speech signals were resampled to 22.05 kHz and peak-normalized prior to feature extraction. Acoustic features were represented as 80-dimensional Mel-spectrograms computed using a 50 ms analysis window and a 12.5 ms hop size. In addition, fundamental frequency (F_0) and aperiodicity parameters were extracted using the WORLD vocoder to enable explicit prosodic modeling.

3.1.1. Model configuration and training details

The Seq2Seq component adopts a Tacotron-style encoder–decoder architecture with attention. The encoder comprises two bidirectional LSTM layers with 256 hidden units per direction, while the decoder consists of two unidirectional LSTM layers with 256 hidden units. A location-sensitive attention mechanism is used to maintain robust temporal alignment between input and output sequences. A five-layer convolutional post-net is employed to refine the predicted Mel-spectrograms.

The CycleGAN module operates in the Mel-spectrogram domain and consists of generators with six residual convolutional blocks and multi-scale PatchGAN discriminators. The model is optimized using a combination of adversarial loss, cycle-consistency loss, and identity preservation loss to ensure realistic emotion transformation while maintaining speaker identity.

Both modules were trained using the Adam optimizer with an initial learning rate of 2×10^{-4}, $\beta_1 = 0.5$, and $\beta_2 = 0.999$. The Seq2Seq model was trained for 100 epochs, while the CycleGAN module was trained for 200 epochs with a batch size of 16. Training was performed on an NVIDIA A100 GPU using mixed-precision computation. Data augmentation strategies including pitch shifting (± 2 semitones) and time-stretching (0.9–1.1×) were applied to improve robustness and generalization.

3.1.2. Evaluation metrics

The proposed framework was evaluated using both objective and subjective measures. Objective metrics include MCD to quantify spectral accuracy, RMSE of F_0 to assess prosodic consistency, and SSIM for spectral structure preservation. Speaker identity preservation was evaluated using cosine similarity between x-vector speaker embeddings extracted from original and converted speech.

Table 1: Objective evaluation metrics on the ESD dataset.

Model	MCD	RMSE	SSIM	Speak similarity
Seq2Seq Only	4.05	0.34	0.79	0.83
CycleGAN Only	4.20	0.36	0.76	0.81
Proposed Seq2Seq + CycleGAN	**3.85**	**0.28**	**0.84**	**0.88**

Table 2: Subjective evaluation results in terms of MOS.

Model	**MOS-N ↑**	**MOS-E ↑**
Seq2Seq Only	3.8	3.7
CycleGAN Only	3.6	3.5
Proposed Seq2Seq + CycleGAN	**4.2**	**4.1**

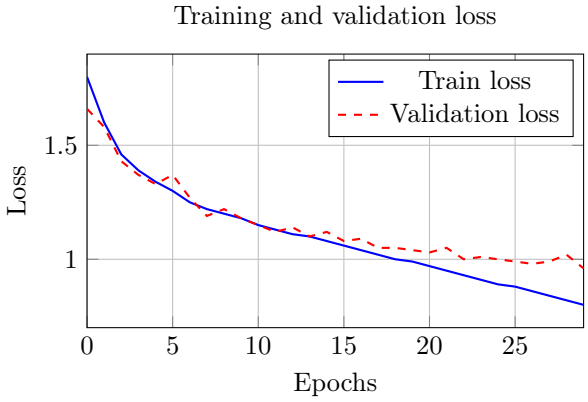

Figure 2: Training and validation loss curves across epochs.

3.2. Results

3.2.1. Objective evaluation

As shown in Table 1, the proposed hybrid framework achieves the lowest MCD, indicating superior spectral fidelity compared to individual Seq2Seq and CycleGAN baselines. The reduced F_0 RMSE demonstrates improved prosody modeling, while higher SSIM values reflect enhanced preservation of spectral structure. Importantly, the speaker similarity score confirms effective identity preservation despite significant emotional modification.

3.2.2. Subjective evaluation

The subjective listening tests further validate the effectiveness of the proposed framework. As summarized in Table 2, the integrated Seq2Seq–CycleGAN model achieves the highest MOS scores for both naturalness and emotional similarity, indicating that listeners consistently preferred the converted speech generated by the proposed approach.

3.2.3. Training dynamics

Figure 2 illustrates the evolution of training and validation losses during optimization. The smooth and monotonic decrease of both curves indicates stable adversarial learning without mode collapse or divergence, demonstrating effective balancing of reconstruction, adversarial, and cycle-consistency objectives.

Figure 3 shows emotion recognition accuracy measured using a pretrained SER model. The close alignment between training and validation curves confirms strong generalization to unseen speakers and emotional conditions.

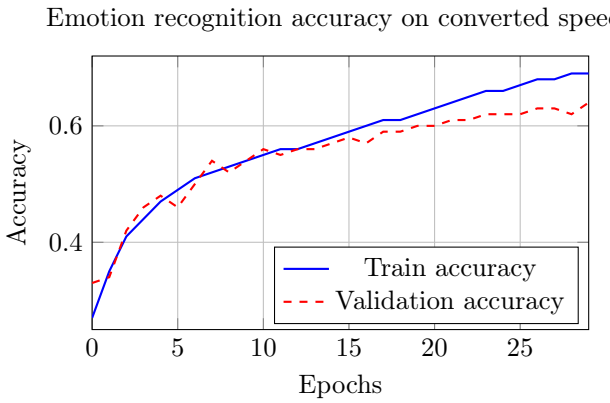

Figure 3: Emotion recognition accuracy computed using a pretrained SER network applied to converted speech.

The experimental results demonstrate that integrating Seq2Seq temporal modeling with CycleGAN-based spectral refinement yields consistent improvements across objective metrics, subjective evaluations, and training stability. Seq2Seq alone effectively models temporal dynamics but introduces over-smoothing, whereas CycleGAN alone enhances realism without explicit sequence modeling. Their integration leverages complementary strengths, resulting in high-fidelity emotional voice conversion with improved naturalness, expressiveness, and speaker identity preservation on the ESD benchmark.

4. Conclusion

This paper presented a high-fidelity emotional voice conversion framework that integrates the complementary strengths of Seq2Seq temporal modeling and Cycle-Consistent Generative Adversarial Networks. By operating entirely in the Mel-spectrogram domain, the proposed approach effectively decouples temporal dependency learning from spectral refinement, enabling accurate emotion transformation while preserving speaker identity and linguistic consistency. The Seq2Seq module captures long-term temporal structure and generates a stable coarse emotional representation, while the CycleGAN module enhances perceptual realism through adversarial and cycle-consistent constraints without relying on parallel training data. Extensive experiments conducted on the Emotional Speech Dataset demonstrate that the proposed framework achieves superior performance over existing emotional and non-parallel voice conversion methods across both objective and subjective evaluation metrics. Notable improvements are observed in Mel-cepstral distortion, pitch accuracy, structural similarity, and perceived emotional expressiveness, highlighting the robustness and generalization capability of the proposed approach across multiple emotional categories.

Declarations and ethical statements

Conflict of interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding statement: The author stated that the article is not funded by financially supporting bodies or any association.

Data availability statement: The data presented in this research are available on request from the corresponding author.

References

[1] Y. Stylianou, O. Cappé, and E. Moulines. Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing*. 2002 Aug 6;6(2):131-42 Available from: https://doi.org/10.1109/89.661472

[2] Toda T, Black AW, Tokuda K. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*. 2007 Oct 15;15(8):2222-35. Available from: https://doi.org/10.1109/TASL.2007.907344

[3] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*. 2004 Apr 30;13(4):600-12. Available from: https://doi.org/10.1109/TIP.2003.819861

[4] [Online Available]: ITU-T. Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs (Recommendation P.862). 2001. Available from: https://www.itu.int/rec/T-REC-P.862

[5] Kameoka H, Kaneko T, Tanaka K, Hojo N. Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks. In *2018 IEEE Spoken Language Technology Workshop (SLT)* . 2018 Dec 18 (pp. 266-273). IEEE. Available from: https://doi.org/10.1109/SLT.2018.8639535

[6] Kaneko T, Kameoka H. Parallel-data-free voice conversion using cycle-consistent adversarial networks. arXiv preprint *arXiv* (Cornell University). arXiv:1711.11293. 2017 Nov 30. Available from: https://arxiv.org/abs/1711.11293v2

[7] Kaneko T, Kameoka H, Tanaka K, Hojo N. Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion. In . IEEE. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019 May 12 (pp. 6820-6824) Available from: https://doi.org/10.1109/ICASSP.2019.8682897

[8] Shen J, Pang R, Weiss RJ, Schuster M, Jaitly N, Yang Z, Chen Z, Zhang Y, Wang Y, Skerrv-Ryan R, Saurous RA. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. 2018 Apr 15 (pp. 4779-4783). IEEE. Available from: https://doi.org/10.1109/ICASSP.2018.8461368

[9] Kaneko T, Kameoka H, Tanaka K, Hojo N. Cyclegan-vc3: Examining and improving cyclegan-vcs for mel-spectrogram conversion. In *arXiv preprint arXiv*:2010.11672. 2020 Oct 22. Available from: https://doi.org/10.48550/arXiv.2010.11672

[10] Kaneko T, Kameoka H, Tanaka K, Hojo N. Maskcyclegan-vc: Learning non-parallel voice conversion with filling in frames. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* . 2021 Jun 6 (pp. 5919-5923). IEEE. Available from: https://doi.org/10.1109/ICASSP39728.2021.9414851

[11] Zhou K, Sisman B, Li H. Transforming spectrum and prosody for emotional voice conversion with non-parallel training data. *arXiv preprint*.2002.00198. 2020 Feb 1. Available from: https://doi.org/10.48550/arXiv.2002.00198

[12] Zhou K, Sisman B, Liu R, Li H. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021 Jun 6 (pp. 920-924). IEEE. Available from: https://doi.org/10.1109/ICASSP39728.2021.9413391

[13] Zhou K, Sisman B, Liu R, Li H. Emotional voice conversion: Theory, databases and ESD. *Speech Communication*. 2022 Feb 1;137:1-8. Available from: https://doi.org/10.1016/j.specom.2021.11.006

[14] Fu C, Liu C, Ishi CT, Ishiguro H. Cycletransgan-evc: A cyclegan-based emotional voice conversion model with transformer. *arXiv preprint*. 2111.15159. 2021 Nov 30. Available from: https://doi.org/10.48550/arXiv.2111.02820

[15] Yang Z, Jing X, Triantafyllopoulos A, Song M, Aslan I, Schuller BW. An overview & analysis of sequence-to-sequence emotional voice conversion. *arXiv preprint*. 2203.15873. 2022 Mar 29. Available from: https://doi.org/10.48550/arXiv.2203.15873

[16] Zhou K, Sisman B, Busso C, Li H. Mixed emotion modelling for emotional voice conversion. *arXiv preprint*. 2022;6:7. Available from: https://doi.org/10.48550/arXiv.2210.00319
Zhou K, Sisman B, Busso C, Ma B, Li H. Mixed-evc: Mixed emotion synthesis and control in voice conversion. *arXiv preprint*. 2210.13756. 2022 Oct 25. Available from: https://doi.org/10.48550/arXiv.2210.13756

[17] Qian K, Zhang Y, Chang S, Yang X, Hasegawa-Johnson M. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*. 2019 May 24 (pp. 5210-5219). PMLR. Available from: https://doi.org/10.48550/arXiv.1905.05879

[18] Chou JC, Yeh CC, Lee HY, Lee LS. Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations.In *arXiv preprint*.1804.02812. 2018 Apr 9. Available from: https://doi.org/10.48550/arXiv.1804.02812

[19] Denton EL. Unsupervised learning of disentangled representations from video. In *Advances in neural information processing systems*. 2017;30. Available from: https://doi.org/10.48550/arXiv.1705.10915

[20] Sisman B, Yamagishi J, King S, Li H. An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2020 Nov 17;29:132-57. Available from: https://doi.org/10.1109/TASLP.2020.3038524

[21] Li YA, Zare A, Mesgarani N. Starganv2-vc: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion. In *arXiv preprint*. 2107.10394. 2021 Jul 21. Available from: https://doi.org/10.48550/arXiv.2107.10394

[22] [Online Available]: Emotional Speech Dataset (ESD). Available from: https://www.kaggle.com/datasets/nguyenthanhlim/emotional-speech-dataset-esd

[23] Ren Y, Ruan Y, Tan X, Qin T, Zhao S, Zhao Z, Liu TY. Fastspeech: Fast, robust and controllable text to speech. In *Advances in neural information processing systems*. 2019;32. Available from: https://doi.org/10.48550/arXiv.1905.09263

[24] Elsayed M, Hadhoud S, Elsetohy A, Osman M, Gomaa W. Non-Parallel Training Approach for Emotional Voice Conversion Using CycleGAN. In *ICINCO (2)* . 2023 Jan 1 (pp. 17-24). Available from: https://www.scitepress.org/Papers/2023/121560/121560.pdf

[25] Qi T, Wang S, Lu C, Zhao Y, Zong Y, Zheng W. Towards Realistic Emotional Voice Conversion using Controllable Emotional Intensity. *arXiv preprint* (Cornell University). 2024 Jul 20; Available from: http://arxiv.org/abs/2407.14800

[26] Sercan O. Arik, Jitong Chen, Kainan Peng, Wei Ping, Yanqi Zhou. Neural voice cloning with a few samples. *arXiv preprint*.2018;31. Available from: https://doi.org/10.48550/arXiv.1802.06006