AI

MDPI

Article

Advancements in End-to-End Audio Style Transformation: A Differentiable Approach for Voice Conversion and Musical Style Transfer

Shashwat Aggarwal, Shashwat Uttam, Sameer Garg, Shubham Garg, Kopal Jain and Swati Aggarwal *

Shashwat Aggarwal; shashwata.co@nsit.net.in; Netaji Subhas University of Technology, New Delhi, India
Shashwat Uttam; shashwatu.co@nsit.net.in; Netaji Subhas University of Technology, New Delhi, India
Sameer Garg; sameerg.co@nsit.net.in; Netaji Subhas University of Technology, New Delhi, India
Shubham Garg; shubhamg.co@nsit.net.in; Netaji Subhas University of Technology, New Delhi, India
Kopal Jain; Department of Electrical Engineering, Indian Institute of Technology, Kharagpur, India; kopaljain@kgpian.iitkgp.ac.in
Swati Aggarwal *; swati.aggarwal@himolde.no; Molde University College, Molde, Norway, Netaji Subhas University of Technology[#], New Delhi, Indiahttps://orcid.org/0000-0001-5986-6915
* Correspondence: author
[#] past affiliation where most of the work was carried out

Abstract: We introduce a fully differentiable end-to-end audio transformation network designed to convert the style of one audio sample to another. This method offers three significant advantages: (a) it operates without the need for parallel utterances, transcriptions, or time alignment processes; (b) it utilizes a global conditioning mechanism, making it vocabulary agnostic and capable of transforming audio styles regardless of the target identity; and (c) it performs one-shot audio transformations without intermediate phonetic representations, thus eliminating the necessity for phonetic alignments and speaker-independent ASR networks. We assess our method against existing approaches in voice conversion and musical style transfer tasks. Subjective evaluations demonstrate the superiority of our approach. The network employs an encoder-decoder architecture that integrates neural network models known for their explainability in Natural Language Processing tasks.

Keywords: Voice Conversion; Musical Style Transfer; Audio Transformations; End-to-End Audio Pipeline

1. Introduction

Audio transformation deals with the transformation of syntactic, acoustic, and semantic variations of one audio to another. It includes multiple applications such as voice conversion, timbre transfer, speaker morphing, emotion transformation, etc. [1]. One of the most widely studied applications of it is Voice Conversion (VC). Voice conversion deals with the transformation of paralinguistic features of the source audio with that of the target while preserving the linguistic features. Many approaches for VC have been developed over the years.

Most of the early VC approaches have focused upon statistical methods based on Gaussian Mixture Models (GMM) to convert voice from source to target speaker [2-3]. It has also been approached with feed-forward Deep Neural Networks [4] and an exemplar-based framework using non-negative matrix factorization [5-6]. Despite producing good results, these approaches often used complex feature pipelines consisting of domain-specific features and require parallel time-aligned source and target speech data, which is difficult and expensive to collect.

Recently there have been some approaches such as [7-9] that overcome the requirement for parallel time-aligned data by using an attribute label along with the acoustic features to perform local conditioning to convert an attribute of source speech (e.g. speaker identity) to target attribute. In general, though the quality of the converted audio obtained with non-parallel methods is usually limited compared with that of audio obtained through statistical methods using parallel data, these can eliminate the need for parallel data which is costly to obtain. However, these approaches still suffer from the limitation of being training vocabulary dependent. These approaches because of the use of local conditioning mechanisms can only convert the voice to a target speaker which was present during the training phase.

There have been some attempts such as [10-11] which overcome the aforementioned limitation and perform voice conversion for any arbitrary speaker. These approaches use automatic speech recognition (ASR) systems to convert the input source speech to intermediate phonetic representations which are further synthesized as output target speech using text-to-speech systems. Although these systems can perform any-to-any voice conversion, they have some downsides to offer such as the performance of such methods is heavily dependent upon the accuracy of the ASR system used. Secondly, these approaches rely on intermediate phonetic transcriptions to train or finetune the ASR system used which are usually hard to obtain, thus decreasing the portability of such systems to newer languages or datasets [12]. Lastly, these systems are primarily applicable only to the application of voice conversion.

This paper seeks to address some of the limitations highlighted in the aforementioned studies. Our approach is robust to the training vocabulary, enabling one-shot audio transformations without relying on intermediate phonetic representations or automatic speech recognition (ASR) systems. By directly operating on acoustic features such as spectrograms or Mel-frequency cepstral coefficients (MFCCs), our method circumvents the need for domain-specific, complex feature engineering pipelines. We evaluate the effectiveness of our approach on two challenging tasks: (a) voice conversion and (b) musical style transfer. Our results are benchmarked against three existing methodologies, demonstrating the efficacy and versatility of our proposed solution. This paper introduces a novel, fully differentiable, end-to-end audio transformation framework with the following key contributions:

1. Vocabulary-Agnostic Framework: Unlike traditional methods, our approach is robust to the training vocabulary, enabling one-shot audio transformations for unseen speakers or musical instruments.
2. Removal of Phonetic Representations and ASR Dependency: By bypassing the need for intermediate phonetic representations or automatic speech recognition (ASR) systems, our method improves generalizability and reduces reliance on resource-intensive processes.
3. Simplified Feature Engineering: By directly utilizing acoustic features like spectrograms and MFCCs, our approach avoids complex, domain-specific feature pipelines, making it adaptable across datasets and tasks.

This work can find applications where audio transformation is made accessible to everyone. Like in voice conversion, the proposed method would allow easy voice changes without needing special datasets or complex alignments, which are usually hard to get and use. This would facilitate for example, creation of personalized voice assistants, dubbing for movies, and tools for people with speech difficulties. In music, the proposed method would help change the style of songs, allowing musicians and producers to experiment with new genres and creative ideas. Also, this could be finding applicability in entertainment industry, like making video games and virtual environments more realistic with better voice and sound changes. By removing the need for specialized tools, the proposed method can lead to improving user experience by making audio technologies more accessible and personalized.

2. Related Works

A flexible framework for spectral conversion (SC) was proposed [8] to address the limitations of requiring aligned corpora for training. Traditional SC frameworks often rely on parallel corpora, phonetic alignments, or explicit frame-wise correspondence to learn conversion functions or synthesize target spectra. However, these dependencies significantly restrict the practicality of SC applications due to the limited availability of parallel corpora. To overcome this, the proposed framework leverages a variational auto-encoder (VAE) to enable training with non-parallel corpora. The framework incorporates an encoder to extract speaker-independent phonetic representations and a decoder to reconstruct the designated speaker's voice, eliminating the need for parallel corpora or phonetic alignments.

[10] focused on achieving voice conversion (VC) across arbitrary speakers, referred to as any-to-any VC, using just a single target-speaker utterance. Two systems are explored: (1) the i-vector-based VC (IVC) system and (2) the speaker-encoder-based VC (SEVC) system. Both approaches utilize Phonetic Posterior Grams as speaker-independent linguistic features extracted from speech samples. A multi-speaker deep bidirectional long short-term memory (DBLSTM) model is trained in both systems to perform VC, with additional inputs encoding speaker identities. In the IVC system, the speaker identity for a new target speaker is represented using i-vectors, whereas in the SEVC system, it is represented by speaker embeddings predicted by a separately trained model. Experimental results demonstrate the effectiveness of both systems in enabling any-to-any VC with a single target-speaker utterance, with the IVC system outperforming the SEVC system in terms of speech quality and similarity to the target speaker's genuine voice.

Previous research [13] has also explored the use of pseudo-recurrent structures, such as self-attention mechanisms and quasi-recurrent neural networks, to design efficient text-to-speech (TTS) acoustic models. These models demonstrated remarkable advancements, achieving a synthesis speedup of 11.2 times on CPU and 3.3 times on GPU when compared to traditional recurrent baseline models. Despite these improvements in speed, the quality of the synthetic speech was maintained at levels comparable to the original recurrent models, making the approach competitive with state-of-the-art vocoder-based statistical parametric speech synthesis systems. Additionally, another study [14] introduced a fully end-to-end neural network capable of learning to translate speech spectrograms into target spectrograms of another language, effectively mapping content across languages in a consistent canonical voice. This advancement addresses the challenge of speech-to-speech translation (S2ST), a field critical for breaking down linguistic barriers and fostering communication among individuals who do not share a common language.

Recent advancements in voice conversion systems have predominantly focused on modifying spectral parameters, such as the spectral envelope. An approach from [15] extends this by incorporating prosodic features, specifically Wavelet modelling of the F0 contour, to enhance voice quality and naturalness. Accent conversion (AC) modifies a non-native speaker's accent to resemble a native accent while preserving their vocal timbre. [16] enhances AC applicability and quality by employing an end-to-end text-to-speech system trained on native speech to generate native references, eliminating the need for reference speech during conversion. The system leverages reference encoders to integrate multi-source information, combining acoustic features from native references and linguistic data with conventional phonetic posterior grams (PPGs).

[17] introduces a Sparse Anchor-Based Representation (SABR) algorithm for exemplar selection in native-to-nonnative voice conversion (VC). Utilizing phoneme labels and clustering, the algorithm addresses poor time alignment commonly found in such conversions. Foreign Accent Conversion (FAC) traditionally relies on native reference utterances or speaker-specific systems, limiting scalability. To overcome these constraints, a novel FAC system [18] generalizes to unseen non-native (L2) speakers without requiring native (L1) references. This many-to-many approach allows native-accented synthesis while preserving the speaker's identity.

A method combining time-frequency filtering and CycleGAN-based conditional adversarial networks [19] to enhance the perceived quality of separated sources. Predominant pitch tracks are extracted using a pitch estimation algorithm, with binary masks generated for each track and its harmonics. A CycleGAN-based network refines the spectrogram images to improve perceptual quality, and the enhanced spectrogram is reconstructed into audio using the inverse short-time Fourier transform.

Although machine learning models demonstrate exceptional predictive capabilities, they are often criticized for their opaque nature, often referred to as "black boxes" [20]. This lack of transparency presents significant challenges in understanding the underlying mechanisms of these models and evaluating the reliability of their predictions [21]. Interpretability in machine learning refers to the extent to which humans can comprehend and articulate the decision-making processes and behaviors of these complex models [22]. Explainable Artificial Intelligence (XAI) has emerged as a solution to enhance the interpretability of machine learning models. It employs two primary approaches: intrinsic explainability, which involves designing models inherently interpretable, and post-hoc explainability, which provides insights into the decision-making process after the model has been trained [23].

To delve deeper into the interpretability of Long Short-Term Memory (LSTM) networks, one study [24] analyzed their performance using n-gram models, finding that LSTMs excel in tasks requiring long-range reasoning. Another study [25] introduced a novel interpretation framework inspired by principles of computational theory. Furthermore, researchers in [26] developed an interpretable variant of recurrent neural networks (RNNs) known as SISTA-RNN. This architecture is grounded in the sequential iterative soft-thresholding algorithm and leverages the concept of deep unfolding [27]. Additionally, a new explainable convolutional neural network (XCNN) was proposed in [28] as an end-to-end framework aimed at enhancing interpretability. A separate investigation [29] explored the use of fine-grained information to explain the decisions made by encoder-decoder networks utilizing CNNs and LSTMs.

Attention mechanisms, introduced as part of modern deep learning frameworks, have been a topic of ongoing debate. While some studies argue that attention weights can serve as reliable indicators of feature importance and provide meaningful explanations [31], others contend that the distributions of attention weights lack inherent interpretability and require further processing to yield insights [31-32]. To address these conflicting perspectives, a study [33] conducted a manual analysis of attention mechanisms across various natural language processing (NLP) tasks. The findings demonstrated that attention weights can indeed be interpretable and are correlated with measures of feature importance that encapsulate linguistic attributes.

In musical style transfer, advancements in neural architectures, such as generative adversarial networks (GANs), have facilitated domain-specific tasks like instrument recognition and adaptation [33]. These methods have been applied to datasets like IRMAS, demonstrating their potential to transform musical attributes effectively. Despite these advancements, the limitations of parallel data dependency, vocabulary constraints, and reliance on complex feature pipelines remain largely unaddressed. This work aims to bridge these gaps by proposing a fully differentiable, end-to-end framework that eliminates parallel data requirements, is vocabulary agnostic, and operates directly on acoustic features.

The traditional sequence-to-sequence (seq2seq) learning framework encodes a source sequence into a fixed-length vector in a single step, which often limits its ability to effectively model the structural correspondence between source and target sequences. To address this limitation, rather than relying on linearly weighted attention mechanisms, a recurrent neural network (RNN)-based approach, termed cyclic sequence-to-sequence (Cseq2seq), was proposed in [34]. Key observations include: (1) Cseq2seq effectively learns source-target correspondences without requiring explicit attention mechanisms, and (2) the encoder and decoder can share RNN parameters without compromising performance.

3. Method

An encoder-decoder based architecture, along with a reference encoder, has been used to reconstruct the input acoustic feature sequence during the training phase and perform audio style transform by conditioning the input source audio sequence with the target-specific style embeddings computed from the reference encoder during the testing phase. A GAN based fine tuning scheme similar to [35] has also been employed, to get rid of any noisy artifacts and improve upon the naturalness of the generated audio. The network architecture for the method is shown in Figure 1 and explained below.

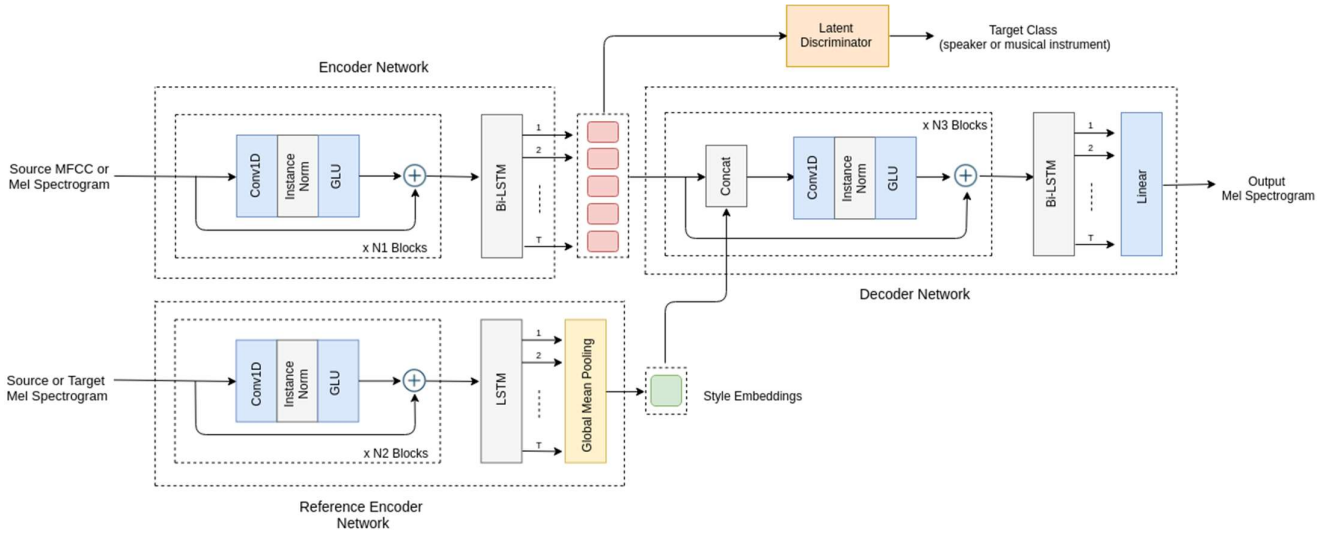

Figure 1. Overview of the method. The encoder network takes acoustic features of the source audio as input. The reference encoder takes the Mel-spectrogram of the source audio during training and of the target class during the testing phase. The decoder network combines the outputs of encoder and reference encoder networks to reconstruct or transform audio. A latent discriminator based adversarial training scheme is employed to learn target independent encoded representations.

3.1. Encoder/Decoder Networks

This architecture incorporates a hybrid approach that combines one-dimensional (1D) convolutional layers integrated with gated linear units (GLUs) [36] and bidirectional Long Short-Term Memory (LSTM) networks [37] to construct the encoder and decoder components. The 1D convolutional layers, augmented by GLUs, are instrumental in capturing the spectral relationships within the sequences of input acoustic features. The bidirectional LSTMs model the temporal characteristics of those acoustic sequences. Inspired by recent works [38-39], residual connections and instance normalization in are also included in the encoder and decoder networks. These inclusions help in stabilizing the training process and the generation of high-resolution output audio sequences.

3.2. Reference Encoder

To remove the training vocabulary dependence and the requirement for intermediate phonetic representations, the authors have trained a reference encoder jointly with the encoder-decoder networks. The reference encoder is trained to capture target specific style embeddings, where target corresponds to a speaker or a musical instrument in our case.

The reference encoder is designed similar to the encoder network, with the main difference being the use of unidirectional LSTMs instead of bidirectional LSTMs. A global mean pooling layer has also been added on top of the unidirectional LSTMs to capture the global style specific features from the input audio while ignoring the local phonetic specific features. The global mean pooling layer ensures that the learned style embeddings are independent of local features such as phonetic content.

Before training the reference encoder jointly with the encoder-decoder network, it has first been pre-trained on a simple classification task to predict the target audio class from the input acoustic feature sequences. This pre-training ensures that the reference encoder can learn a mapping from the global style specific features of the input audio sequence to a fixed length vector, which we denote as audio style embeddings. These style embeddings are then further fine-tuned by jointly training the reference encoder with the encoder-decoder networks. These target specific style embeddings provide global conditioning and help in transforming the audio from source to target class.

3.3. Training Process

During the training process, the authors have utilized the acoustic features of the ground truth audio, specifically the Mel-Frequency Cepstral Coefficients (MFCCs) and Mel Spectrograms, as inputs to the encoder and reference encoder networks. The reference encoder is designed to compress these input acoustic features into a fixed-length vector representation, referred to as style embeddings. These style embeddings capture the stylistic attributes of the audio and are subsequently concatenated with the latent representation generated by the encoder network. This combined representation is then fed into the decoder, which reconstructs the input acoustic features corresponding to the original audio sequence, ensuring that the stylistic and temporal details are effectively preserved.

We use a combination of Mean Absolute Error (MAE) and Pearson Correlation Coefficient $r_{yy'}$ as our reconstruction loss function as given in (1). Here, $r_{yy'}$ is defined as the Pearson Correlation Coefficient between the predicted output y' and the ground truth y, calculated as:

$$r_{yy'} = \frac{Cov(y, y')}{\sigma_y \sigma_{y'}}$$

where Cov(y,y') denotes the covariance between y and y' , and σ_y and $\sigma_{y'}$ represent their respective standard deviations. The value of $r_{yy'}$ ranges from −1 to 1, with 1 indicating a perfect positive correlation. To maximize $r_{yy'}$, we minimize the negative of its value in the loss function.

$$L_{rec}(\theta) = \sum_{i=1}^{n} \left| |y_i - y_i'| \right| - r_{yy} \quad (1)$$

3.4. Latent Discriminator

A latent discriminator-based adversarial training scheme is employed to ensure that the encoder learns target class-independent latent representations. An auxiliary classifier acts as the discriminator, tasked with predicting the target class c from the encoded representation z of an input audio utterance. The discriminator's loss is defined as:

$$L_{lat}(\theta) = -\mathrm{E}[log(\mathrm{P}(\mathrm{y}|enc(x))] \quad (2)$$

where *P(y|enc(x))* represents the predicted probability of the target class y given *encoding of* x, and E denotes the expectation over the latent distribution *p(enc(x))*. In contrast, the encoder is trained adversarial to maximize the discriminator's uncertainty, with the encoder's loss given as:

$$L_{ae}(\theta) = L_{rec}(\theta) - \beta L_{lat}(\theta) \quad (3)$$

This adversarial interplay ensures the encoded representations are class-invariant, enabling effective audio transformations that preserve target-independent features. Here, β is a hyperparameter that controls the relative weight of the adversarial loss term in the encoder's overall loss function. Adjusting β helps balance the encoder's focus on disentangling target-specific features while maintaining effective representation learning.

The authors devised the latent discriminator using a bank of gated convolutional layers along with instance normalization and dropout layers. The discriminator takes the encoded latent representations of an acoustic feature sequence as input and predicts the probability distribution over the target class. This latent discriminator based adversarial training scheme is essential since it enforces a regularization over the encoded latent representations and ensures that the learned representations are target class independent.

3.5. WGAN Based Fine Tuning

An adversarial based fine-tuning scheme is also applied to remove any noisy artifacts and buzzy sound effects present in the generated audio and to improve upon its naturalness. Given the well-known challenges associated with training Generative Adversarial Networks (GANs), the authors adopt a more stable variant, the Wasserstein GAN with Gradient Penalty (WGAN-GP) [40]. This approach is not only easier to train but also exhibits improved convergence behaviour. In the proposed framework, the decoder acts as the generator under this fine-tuning strategy. For the discriminator, we design a network comprising a series of two-dimensional (2D) convolutional layers, enabling it to differentiate between genuine acoustic feature sequences and those synthesized by the model. The discriminator outputs a scalar value that represents the "realness" of an input feature sequence x; a higher scalar value indicates a higher likelihood that x is real.

The discriminator is trained to maximize the adversarial loss by correctly identifying real and generated feature sequences. Conversely, the generator (decoder) is optimized to deceive the discriminator by minimizing a combination of the adversarial loss and the reconstruction loss. This dual-objective setup ensures that the generator not only produces realistic acoustic features but also preserves the fidelity of the original input, facilitating high-quality audio transformation.

3.6. Process of Conversion

During the inference phase, audio transformation can be achieved by feeding the acoustic features of the target audio whose style is to be transferred as an input to the reference encoder while feeding the acoustic features of the source audio as input to the base encoder. The output from the decoder is the transformed audio sequence with local phonetic specific features from the source audio and global style specific features from the target audio respectively.

4. Experiments

4.1. Datasets

The authors assess the effectiveness of our proposed method on two distinct audio transformation tasks: voice conversion and musical style transfer. For the voice conversion task, we leverage two datasets: CMU Arctic [41] and L2 Arctic [42]. The CMU Arctic dataset comprises approximately 1,150 utterances spoken by seven speakers, representing a mix of US English and various other accents. Complementing this, the L2 Arctic dataset extends the CMU Arctic by including recordings from twenty non-native English speakers whose first languages (L1s) include Hindi, Korean, Mandarin, Spanish, and Arabic. Notably, all speakers in the L2 Arctic dataset narrate the same set of utterances as recorded in the original CMU Arctic dataset, enabling consistent comparisons. These datasets maintain consistent utterances across speakers, allowing for controlled experiments and fair benchmarking of audio transformation tasks, even if they do not explicitly include vocalized voice features.

For the musical style transfer task, the authors utilize the IRMAS dataset [43], which is specifically designed for instrument recognition in musical audio. This dataset provides a diverse range of recordings, making it suitable for evaluating the ability of the proposed method to adapt and transform musical styles effectively. These datasets collectively ensure a comprehensive evaluation of the approach across diverse audio transformation challenges. It consists of musical audio excerpts of ten different musical instruments, such as cello, acoustic guitar, piano, etc. A subset of 12 speakers, six females and six males, across six different nationalities, i.e., English, Hindi, Korean, Mandarin, Spanish, and Arabic respectively, was selected for voice conversion task. While for musical style transfer, a subset of 6 musical instruments, namely piano, saxophone, violin, flute, trumpet, and acoustic guitar respectively, is selected. The dataset is randomly split into training and testing sets in a 5:1 split ratio for each task.

4.2. Audio Formats

For input acoustic features, the authors employ Mel-Frequency Cepstral Coefficients (MFCCs) and Mel Spectrograms in the case of voice conversion tasks, while for musical style transfer, we utilize Mel Spectrograms exclusively. These acoustic features are chosen for their ability to effectively capture the spectral and temporal characteristics of audio, which are essential for accurate transformation. All acoustic features are computed using the parameters specified in Table 1, ensuring consistency and reproducibility across the evaluations. These features serve as the foundation for the model's encoding and transformation processes, enabling robust and high-quality results for both tasks. Audio is synthesized from the predicted Mel spectrograms using Griffin-Lim algorithm [44].

Table 1 outlines the final configuration of parameters used for acoustic feature extraction, including sample rate, frame length, frame shift, number of FFT points, number of Mel bands, and number of MFCCs. These parameters were selected through a series of preliminary experiments designed to balance computational efficiency, fidelity of representation, and suitability for the target audio transformation tasks. The sample rate of 16,000 Hz was chosen as it provides sufficient frequency resolution to capture human speech and musical instruments without incurring excessive computational overhead. The values of frame length and frame shift reflect a trade-off between temporal resolution and frequency detail. A frame length of 50 ms ensures adequate frequency resolution for capturing harmonic structures, while the 12.5 ms shift reduces redundancy without sacrificing temporal dynamics.

The high FFT resolution was adopted to accurately represent fine spectral details, essential for both voice and music transformations. The configuration of 128 Mel bands balances the need for detailed spectral representation with computational efficiency, enabling the model to learn meaningful representations of diverse audio styles, and the 40 MFCCs capture the most critical features for speech and audio analysis while minimizing redundant information, making them suitable for both tasks evaluated in this work. The final parameter configuration was derived iteratively, guided by empirical evaluations of model performance on validation data. The chosen parameters consistently produced high-quality audio transformations across seen and unseen targets, validating their effectiveness in meeting the project goals.

Table 1. Parameters used for computation of Acoustic Features.

Parameter	Value
Sample Rate	16000
Frame Length	50 ms
Frame Shift	12.5 ms
n-FFTs	2048
# Mels	128
# MFCCs	40

4.3. Training Details

The authors train the network using the Adam optimizer, configured with a learning rate (lr) of 0.001, along with momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. These parameters were chosen to obtain the best results in terms of accuracy and Mean Opinion Score. A batch size of 32 is employed to balance computational efficiency and convergence stability. The training process spans 100 epochs, with an initial 30-epoch pre-training phase dedicated to the reference encoder. This pre-training ensures that the reference encoder effectively captures style embeddings before the full network is fine-tuned, enhancing the overall performance and stability of the model. Finally, a 20 epoch GAN based fine tuning was performed.

4.4. Baselines

The evaluation of our proposed method against eight existing systems aims to ensure a comprehensive comparison across a variety of approaches and their capabilities. The selected baselines represent a diverse set of established and state-of-the-art methods. The Conditional Variational Autoencoder (CVAE) [8] and Conditional Sequence-to-Sequence Network (CSeq2Seq) [29] employ local conditioning mechanisms for audio transformations and serve as representative approaches for methods utilizing explicit conditioning on attributes. Their inclusion allows us to benchmark against widely recognized frameworks that share structural similarities with our approach.

The Any-to-Any Voice Conversion Network [10] uses an SI-ASR system for intermediate phonetic transcriptions, enabling comparisons with approaches dependent on explicit intermediate representations. It highlights the advantages of bypassing such dependencies in our framework. By incorporating prosodic features such as F0 contour modeling, a deep neural network (DNN) system [15] demonstrates the impact of prosody on audio quality and naturalness. It provides a contrast to our focus on spectral features. The fifth system [16] leverages reference encoders to integrate multi-source information, combining acoustic features from native references and linguistic data with conventional phonetic posterior grams (PPGs).

The Sparse Anchor-Based Representation (SABR) Algorithm [17] addresses time-alignment issues in native-to-nonnative voice conversion. Comparing against it emphasizes our method's ability to perform robust transformations without requiring specialized alignment strategies. The Foreign Accent Conversion (FAC) System [18] is a many-to-many approach that generalizes to unseen non-native speakers without native references, directly aligning with our goal of robust transformations for unseen targets.

CycleGAN-Based Networks [19] incorporates adversarial networks for perceptual quality enhancement, and demonstrates the effectiveness of adversarial techniques in audio transformations, enabling us to benchmark our WGAN-based fine-tuning approach.

These systems were selected not only for their methodological diversity but also for their relevance to the tasks of voice conversion and musical style transfer. Each one represents a unique approach to handling challenges such as parallel data dependency, phonetic alignment, or generalization across unseen targets. Comparing our results with these systems provides a holistic view of the strengths and limitations of our method in the broader context of existing solutions.

4.5. Evaluation Metrics

To evaluate the performance of our method with the baselines, the authors compute the mean opinion score (MOS), the higher, the better. The score has been computed for both seen (speakers or musical instruments present in the training set) and unseen (speakers or musical instruments not present in the training set) targets. In addition to MOS, for the voice conversion task, the authors also evaluate the naturalness of the generated audio for four cases:

- Intra-Gender voice conversions.
- Inter-Gender voice conversions.
- Intra-Nationality voice conversions.
- Inter-Nationality voice conversions.

5. Results and Discussion

In Table 2, the subjective evaluations (MOS) of all the baselines and the proposed method are reported for both the tasks. To evaluate the audio quality, the Mean Opinion Score (MOS) is calculated, following a standard methodology. Human evaluators rate the audio generated by our method and baseline models on a 5-point numerical scale, where 1 corresponds to "bad," 2 to "poor," 3 to "fair," 4 to "good," and 5 to "excellent." Each audio sample from the experiments was assessed by five human raters with normal hearing.

These results demonstrate that the proposed method gives better results than seven systems for audio transformations targeting previously seen identities and achieves competitive performance with one system across both seen and unseen target identities. The baselines relying on an intermediate Automatic Speech Recognition (ASR) system have significant drawbacks. Specifically, its dependence on phonetic transcriptions limits portability to new datasets, as obtaining such transcriptions is resource-intensive. The proposed method relies solely on easily extractable acoustic features, making it adaptable to any dataset transformations without requiring intermediate ASR systems.

This approach captures fundamental phonetic properties, as well as the identity-specific nuances of speakers or instruments. This enables the model to apply these attributes to unseen words, pitches, target speakers, or musical instruments with minimal degradation in audio quality. This flexibility highlights the robustness of our method compared to conventional systems. In addition to subjective MOS evaluations, we present a visual analysis of the outputs in Figure 2, where examples of spectra generated by our approach are shown.

It provides a visual analysis of the audio transformations achieved by the proposed method, showcasing the MFCC and spectrogram plots for the source audio, target audio, and the generated audio. The MFCC plots illustrate how the model captures the spectral envelope of the source audio while effectively adapting it to match the stylistic attributes of the target audio. This is particularly evident in the transformed spectrograms, where the harmonic structures and energy distributions align closely with the target audio while retaining key phonetic features from the source.

To further validate the representations encoded by the reference encoder, the learned style embeddings for voice conversion and musical style transfer tasks have been analyzed. These embeddings, shown in Figure 3, highlight the ability of the encoder to preserve identity-related characteristics across diverse audio transformation scenarios. The style embeddings are visualized using the t-SNE algorithm with perplexity = 30 and number of iterations = 300 respectively. The t-SNE plots show that the reference encoder is able to cluster sounds belonging to same target identity classes together, thus confirming that the reference encoder can encode the global style specific features and the target identity.

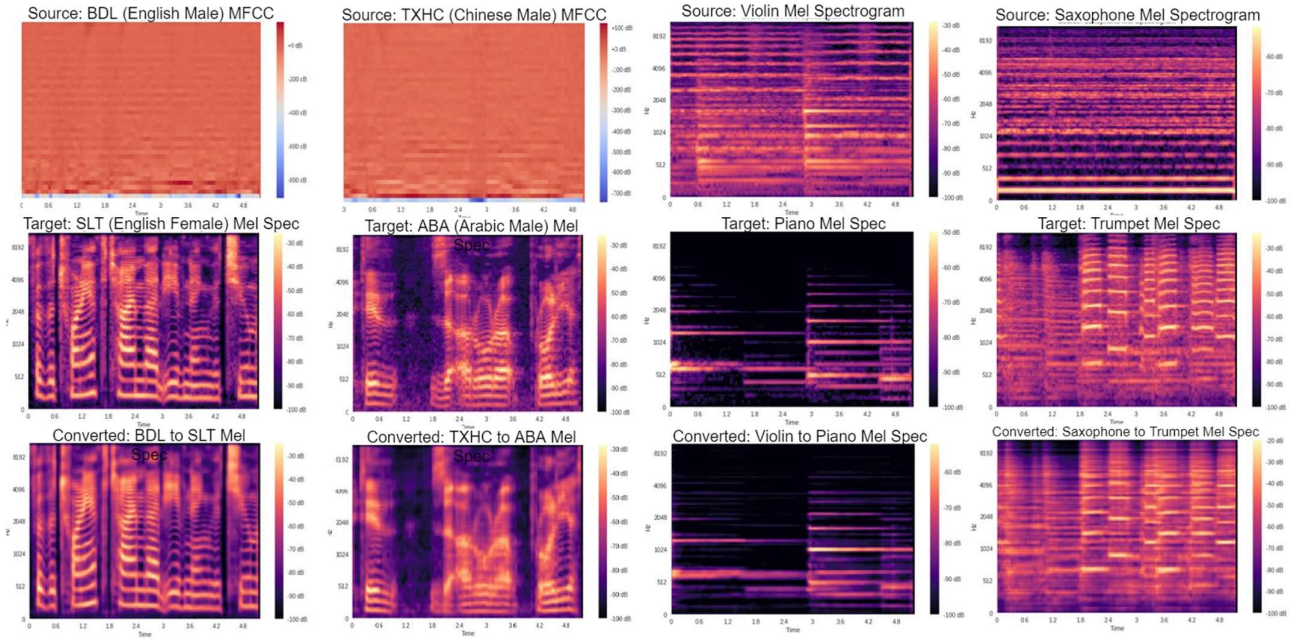

Figure 2. MFCC and Spectrogram plots for source audio, target audio and generated audio, for Voice Conversion and Musical Style Transfer.

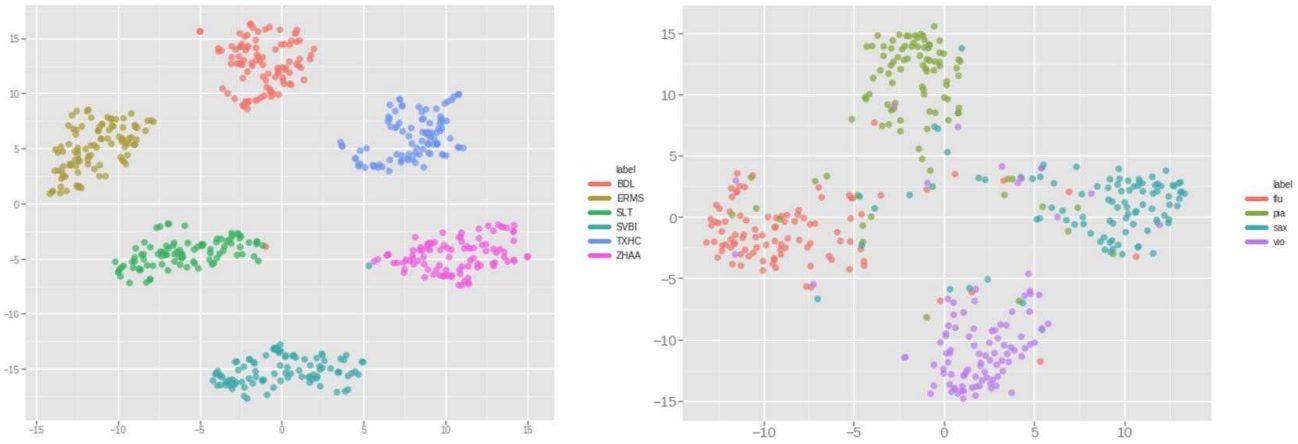

Figure 3. Learned Style Embeddings. We visualize the learned style embeddings using two-dimensional t-SNE plots for six random speakers (three females and three males) on left and for four random musical instruments on right.

Table 2. Mean opinion score (MOS) for both seen and unseen targets on voice conversion and musical style transfer tasks. Higher MOS is better.

Method	**MOS**	
Seen Target	**Voice Conversion**	**Musical Style Transfer**
Ground Truth	4.53	4.07
FAC	3.03	-
DNN	3.09	-
SABR	3.18	-
CVAE	3.31	3.08
C-Seq2Seq	3.50	3.26
PPG sequence	3.52	-
MSVC	3.77	-
CycleGAN	-	3.64
Our Method	3.72	3.50
Unseen Target	**Voice Conversion**	**Musical Style Transfer**
DNN	2.72	-
MSVC	3.51	-
Our Method	3.44	3.36

Figure 4 shows the MOS for naturalness, calculated on both seen and unseen speakers to evaluate our method over cross-nationality and cross-gender audio transformations. The results indicate that the proposed method is able to generate intelligible and natural speech across gender as well as nationality. While these results indicate the robustness of the method in generating natural-sounding outputs, it is important to note that the primary focus of this work is on vocal voice transformation rather than general speech synthesis.

The intelligibility and naturalness of the transformed audio serve as evidence of the model's effectiveness in preserving the stylistic and phonetic nuances essential to high-quality voice transformation. These findings reinforce the framework's suitability for applications requiring nuanced changes to vocal characteristics while maintaining overall audio quality.

Finally, to ensure that the latent representations from the encoder are independent of the target identity after the latent discriminator based adversarial training, the authors train a target class verification system that takes the latent representations from the encoder as input to predict the target class identity. The verification accuracies for both the tasks, with and without the latent adversarial training are reported in Table 3. The drop-in verification accuracies after latent discriminator based adversarial training confirm that the encoder is able to learn latent representations which are independent of the target identity.

While the proposed method eliminates dependencies on parallel data and phonetic alignments, it relies on high-quality acoustic feature extraction (e.g., MFCCs and Mel spectrograms) to achieve optimal performance. Additionally, the adversarial training phase, while enhancing robustness, introduces computational overhead, which may limit the framework's applicability in real-time scenarios.

The method's reliance on Griffin-Lim reconstruction for audio synthesis, although effective, introduces occasional artifacts, impacting the naturalness of the transformed audio in certain cases. Furthermore, while competitive with ASR-based systems, the proposed approach's quality lags slightly behind in some unseen target scenarios, underscoring the challenges in generalizing across significantly diverse datasets.

The model processes acoustic features like MFCCs and Mel spectrograms to capture essential phonetic characteristics such as formants and harmonics. Future enhancements may involve using metrics like Mel Cepstral Distortion to measure spectral distance between original and transformed audio, reflecting how well phonetic structures are preserved [45]. The model encodes speaker-specific vocal traits or instrument-specific tonal qualities into fixed-dimensional embeddings, which can be evaluated using metrics like Speaker Verification Accuracy to ensure the generated audio retains the identity of the target [46].

This framework demonstrates generalization beyond training data by generating audio for unseen scenarios—such as new speakers—without compromising style or intelligibility. Metrics like Phonetic Transcription Error Rate can assess how accurately the transformed audio aligns with intended phonetic content [47]. The model produces audio transformations with high naturalness and fidelity, ensuring that the output sounds realistic. Signal-to-Noise Ratio can quantify the clarity of the transformed audio by comparing signal strength to background noise. Future improvements can leverage these metrics to enhance objective evaluations and refine the model's performance.

Explainable AI (ExAI) in natural language processing (NLP) predominantly emphasizes deciphering the internal mechanisms of underlying models rather than providing insights into specific classification outputs. A comprehensive review [48] consolidates progress in various aspects of interpretability, including the behavior of word embeddings, the internal dynamics of RNNs and transformers, the rationale behind model decisions, and the array of visualization techniques employed. The review also underscores the interconnected nature of these interpretability methods, shedding light on how they complement and build upon one another.

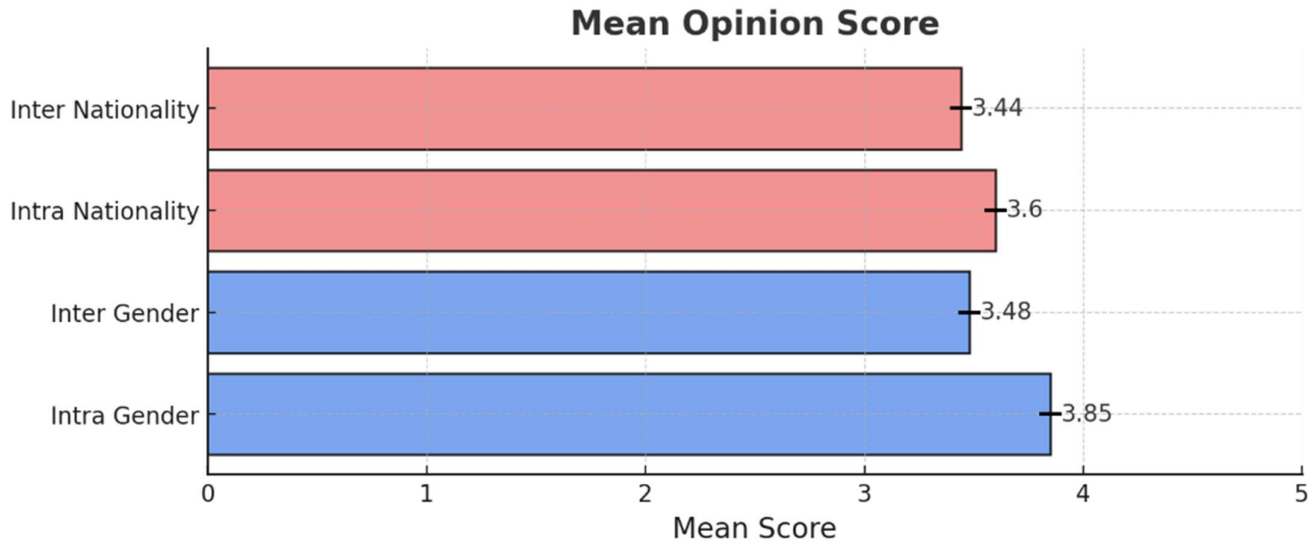

Figure 4. MOS on naturalness. Computed for the following cases, (a) Inter-Nationality, (b) Intra-Nationality, (c) Inter Gender, and (d) Intra Gender

Among NLP architectures, Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs) exhibit relatively higher inherent interpretability due to their structured operation and transparent feature extraction processes. However, attention-based models, despite their effectiveness, require more rigorous and targeted analysis to achieve complete transparency. This calls for the development of advanced techniques to elucidate how these models prioritize and weigh different inputs during decision-making.

Table 3. Target verification accuracies before and after the latent adversarial training. L.A.: Latent Adversarial.

Task	Acc. w/o L.A. Training (%)	Acc. w/ L.A. Training (%)
Voice Conversion	93.2	57.15
Musical Style Transfer	86.34	69.1

6. Conclusions

The proposed fully differentiable, end-to-end audio transformation framework offers several impactful implications and opportunities for both future research and practical applications. By removing the need for parallel, time-aligned data and intermediate phonetic representations, this approach reduces the burden of data collection and improves scalability across diverse datasets and languages. Its vocabulary-agnostic design further enhances its versatility, enabling audio transformations for previously unseen speakers, musical instruments, and styles.

Optimizing the framework for lower latency and computational efficiency could facilitate real-time applications, such as live voice modulation or musical improvisation tools. Additionally, fine-tuning the model for specialized domains like healthcare or education (e.g., accent conversion for language learners) could expand its applicability. However, ethical concerns, such as potential misuse of voice conversion for identity spoofing or spreading misinformation, must be systematically addressed.

The evaluation framework employed in this study combined objective metrics with subjective insights, recognizing that aspects of audio quality are influenced by human perception. Future evaluations should include detailed statistical analyses of subjective ratings—such as standard deviation, interquartile range, or confidence intervals—to better quantify variability and consistency in evaluators' opinions.

This would provide a clearer understanding of subjective assessments and strengthen the robustness of conclusions. Additional metrics, such as listener agreement rates or score breakdowns by demographic groups, could further validate the model's reliability across diverse contexts. In summary, this study provides a comprehensive and scientifically rigorous analysis of audio transformation techniques.

Acknowledgments: We would like to thank students of the Computer Department, NSUT for participating in subjective experiments.

References

1. Helena Liz-López, Mamadou Keita, Abdelmalik Taleb-Ahmed, Abdenour Hadid, Javier Huertas-Tato, and David Camacho. 2023. Generation and detection of manipulated multimodal audiovisual content: Advances, trends and open challenges. (August 2023).
2. Yannis Stylianou, Olivier Cappe, and Eric Moulines. 1998. Continuous Probabilistic Transform for Voice Conversion. IEEE Transactions on Speech and Audio Processing 6, 2 (1998), 131–142.
3. Tomoki Toda, Alan W. Black, and Keiichi Tokuda. 2007. Voice Conversion based on Maximum Likelihood Estimation of Spectral Parameter Trajectory. IEEE Transactions on Audio, Speech and Language Processing 15, 8 (2007), 2222–2235.
4. Seyed H. Mohammadi and Alexander Kain. 2014. Voice Conversion using Deep Neural Networks with Speaker-Independent Pre-training. In 2014 IEEE Spoken Language Technology Workshop (SLT). 19–23.
5. Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. 2013. Exemplar-based Voice Conversion using Sparse Representation in Noisy Environments. IEICE Transactions on Information and Systems E96-A, 10 (2013), 1946–1953.
6. Zhizheng Wu, Tuomas Virtanen, Eng S. Chng, and Haizhou Li. 2014. Exemplar-based Sparse Representation with Residual Compensation for Voice Conversion. IEEE/ACM Transactions on Audio, Speech and Language Processing 22, 10 (2014), 1506–1521.
7. Albert Haque, Michelle Guo, and Prateek Verma. 2018. Conditional End-to-End Audio Transforms. In Proc. Iterspeech. 2295–2299.
8. Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang. 2016. Voice conversion from non-parallel corpora using Variational Auto-encoder. Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC).
9. Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. 2018. StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks. arXiv preprint arXiv:1806.02169 (2018).
10. Songxiang Liu, Jinghua Zhong, Lifa Sun, Xixin Wu, Xunying Liu, and Helen Meng. 2018. Voice Conversion Across Arbitrary Speakers Based on a Single Target-Speaker Utterance. In Proc. Annual Conference of the International Speech Communication Association (Interspeech).
11. Feng-Long Xie, Frank K. Soong, and Haifeng Li. 2016. A KL divergence and DNN-based approach to voice conversion without parallel training sentences. In Proc. Interspeech.
12. Natalia Bogach, Elena Boitsova, Sergey Chernonog, Anton Lamtev, Maria Lesnichaya, Iurii Lezhenin, Andrey Novopashenny, Roman Svechnikov, Daria Tsikach, Konstantin Vasiliev, Evgeny Pyshkin, and John Blake. 2021. Speech Processing for Language Learning: A Practical Approach to Computer-Assisted Pronunciation Teaching. 10, 3 (January 2021).
13. Santiago Pascual. 2020. Efficient, end-to-end and self-supervised methods for speech processing and generation.
14. Ye Jia, Ron Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. 2019. Direct speech-to-speech translation with a sequence-to-sequence model. ISCA.
15. Gonzales MG, Lucas CR, Bayona MG, De Leon FA. Voice conversion of Philippine spoken languages using Deep Neural Networks. 2020 IEEE 8th Conference on Systems, Process and Control (ICSPC). Published online December 11, 2020:118-121. doi:10.1109/icspc50992.2020.9305801
16. Li W, Tang B, Yin X, et al. Improving Accent Conversion with Reference Encoder and End-To-End Text-To-Speech. *arXiv (Cornell University)*. Published online January 1, 2020. doi:10.48550/arxiv.2005.09271
17. Liberatore C, Gutierrez-Osuna R. An Exemplar Selection Algorithm for Native-Nonnative Voice Conversion. Interspeech 2022. Published online August 27, 2021:841-845. doi:10.21437/interspeech.2021-1740
18. Quamer W, Das A, Levis J, Chukharev-Hudilainen E, Gutierrez-Osuna R. Zero-Shot Foreign Accent Conversion without a Native Reference. *Interspeech 2022*. Published online September 16, 2022:4920-4924. doi:10.21437/interspeech.2022-10664
19. Reghunath LC, Rajan R. Predominant audio source separation in polyphonic music. *EURASIP Journal on Audio Speech and Music Processing*. 2023;2023(1). doi:10.1186/s13636-023-00316-4

20. Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, and Amir Hussain. 2023. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cognitive Computation (August 2023).
21. Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). 6, (September 2018).
22. Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. Electronics 8, 8 (July 2019), 832.
23. Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Javier Del Ser, Adrien Bennetot, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. 58, (June 2020).
24. Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and Understanding Recurrent Networks.
25. Bo-Jian Hou and Zhi-Hua Zhou. 2020. Learning With Interpretable Structure From Gated RNN. 31, 7 (February 2020).
26. Scott Wisdom, Thomas Powers, James Pitton, and Les Atlas. 2016. Interpretable Recurrent Neural Networks Using Sequential Sparse Recovery.
27. John R. Hershey, Jonathan Le Roux, and Felix Weninger. 2014. Deep Unfolding: Model-Based Inspiration of Novel Deep Architectures.
28. Amirhossein Tavanaei. 2020. Embedded Encoder-Decoder in Convolutional Networks Towards Explainable AI.
29. Hui Liu, Qingyu Yin, and William Yang Wang. 2019. Towards Explainable NLP: A Generative Explanation Framework for Text Classification. Association for Computational Linguistics
30. Martin Tutek and Jan Snajder. 2020. Staying True to Your Word: (How) Can Attention Become Explanation? Proceedings of the 5th Workshop on Representation Learning for NLP (2020), 131-142.
31. Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2020. On Identifiability in Transformers.
32. Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. Association for Computational Linguistics.
33. Juan Bosch, Jordi Janer, Ferdinand Fuhrmann, and Perfecto Herrera. 2012. A Comparison of Sound Segregation Techniques for Predominant Instrument Recognition in Musical Audio Signals. In ISMIR. 559–564.
34. Biao Zhang, Deyi Xiong, and Jinsong Su. 2016. Cseq2seq: Cyclic Sequence-to-Sequence Learning. arXiv preprint arXiv:1607.08725 (2016).
35. Ju chieh Chou, Cheng chieh Yeh, Hung yi Lee, and Lin shan Lee. 2018. Multi-target Voice Conversion without Parallel Data by Adversarially learning Disentangled Audio Representations. arXiv preprint arXiv:1804.02812 (2018).
36. Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language Modeling with Gated Convolutional Networks. In Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org. 933–941.
37. Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv preprint arXiv:1508.01991 (2015).
38. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In Proc. CVPR. 770–778.
39. Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for Real-time Style Transfer and Super-Resolution. In Proc. European Conference on Computer Vision. Springer, Cham. 694–711.
40. Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang. 2017. Voice Conversion from Unaligned Corpora using Variational Autoencoding Wasserstein Generative Adversarial Networks. In Proc. Annual Conference of the International Speech Communication Association (Interspeech). 3364–3368.
41. John Kominek and Alan Black. 2004. The CMU Arctic Speech Databases. In Fifth ISCA workshop on speech synthesis.
42. Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Khudilaynen, John Levis, and Ricardo Gutierrez-Osuna. 2018. L2-ARCTIC: A Non-Native English Speech Corpus. Perception Sensing Instrumentation Lab (2018).
43. Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention Interpretability Across NLP Tasks.
44. D. Griffin and Jae Lim. 1984. Signal estimation from modified short-time fourier transform. Acoustics, Speech and Signal Processing, IEEE Transactions 32, 2 (1984), 236–243.
45. Kubichek RF. Mel-cepstral distance measure for objective speech quality assessment. *Pacific Rim Conference on Communications, Computers and Signal Processing*. Published online May 19, 1993. doi:https://doi.org/10.1109/pacrim.1993.407206
46. Kinnunen T, Karpov E, Franti P. Real-time speaker identification and verification. *IEEE Transactions on Audio, Speech and Language Processing*. 2006;14(1):277-288. doi:https://doi.org/10.1109/tsa.2005.853206
47. Cucchiarini C. Phonetic transcription: a methodological and empirical study. *Handlenet*. Published online 2024. doi:https://doi.org/9090066993
48. Julia El Zini and Mariette Awad. 2022. On the Explainability of Natural Language Processing Deep Models. 55, 5 (July 2022).