Voice Conversion-Driven Speaker Anti-Spoofing Framework Using Attention-Based Seq2Seq Modeling

Govind Khandelwal
AI and Machine Learning, IIT Jammu
2023pai9236@iitjammu.ac.in

Rantu Buragohain
Electrical Engineering, IIT Jammu
2021ree1027@iitjammu.ac.in

Karan Nathwani
Electrical Engineering, IIT Jammu
karan.nathwani@iitjammu.ac.in

***Abstract*—Recent advances in voice conversion (VC) have enabled highly realistic transformation of a speaker's voice while preserving linguistic content. Although useful for many applications, this capability poses security risks by facilitating spoofing attacks on automatic speaker verification (ASV) systems. In this work, we propose a novel framework for the VC-driven anti-spoofing framework that jointly models spoof generation and detection in a unified attack–defense paradigm. High-quality spoofed speech is generated using an attention-based Sequence-to-Sequence (Seq2Seq) VC model and combined with genuine speech to build a dataset for training robust spoofing countermeasures capable of detecting subtle acoustic artifacts introduced by modern VC systems. For detection, a pretrained YAMNet-based feature extractor with fully connected layers is used for binary classification of genuine and spoofed speech. The framework follows a Generative Adversarial Network (GAN)-like co-design strategy, iteratively improving both spoof generation and detection models to enhance robustness against VC-based attacks. Experimental evaluation on publicly available speech datasets demonstrates the effectiveness of the proposed approach is capable of detecting high-quality VC attacks.**

***Index Terms*—Voice conversion (VC), automatic speaker verification (ASV), voice anti-spoofing, voice conversion and identification system.**

I. Introduction

Voice conversion (VC) modifies the paralinguistic attributes of speech while preserving its linguistic content and can be used for speaker identity modification, assistive speech systems, emotion transfer and accent adaptation. Traditional VC methods, such as Gaussian Mixture Model (GMM)–based approaches [1], require parallel time-aligned speech data and provide limited prosody modeling. Recently, deep learning (DL)-based approaches such as feed-forward networks (FFNs), recurrent neural networks (RNNs) [2], variational autoencoders, and adversarial learning frameworks [3] have significantly improved the quality and flexibility of conversion. While these advances have also heightened the threat of voice spoofing [4]. Most existing anti-spoofing methods have been proposed over the past decade [5] [6] for earlier spoofing techniques, and often fail to detect very natural-sounding speech produced by modern VC models. Motivated by this gap, we propose a novel framework that exploits advanced VC to improve spoofing detection, enable reliable detection of artificially converted speech even when it closely resembles the target speaker.

Recent feature extraction methods such as RawNet [5], HuBERT [7], and VGGish [8] achieve strong spoofing detection performance, but are often trained on datasets that lack artifacts from modern VC systems, limiting their generalization to new VC-based attacks. To address this, we propose a VC-driven anti-spoofing framework that integrates spoofed speech generation and detection within a unified pipeline. The approach uses pre-trained HuBERT encoders for robust feature extraction and a modified Griffin−Lim algorithm to improve speech reconstruction by refining phase information while preserving the magnitude spectrum. The key contributions of this work are summarized as follows.

- We propose a generative adversarial–inspired framework consisting of a VC model for generating high-quality spoofed speech and an anti-spoofing model for distinguishing genuine and spoofed speech.
- We design a Seq2Seq VC architecture with an attention mechanism and a pretrained HuBERT encoder for robust feature extraction and effective sequence alignment.
- We incorporate a *modified Griffin–Lim* algorithm to improve speech reconstruction by fixing the magnitude spectrum and iteratively refining phase information.
- We develop an *anti-spoofing system* that combines feature extraction with a neural network classifier trained on both genuine speech and VC-generated spoofed samples.
- We evaluate the proposed framework using *YAMNet* [9] and compare it with several state-of-the-art (SOTA) feature extractors, including modified group delay function [10], RawNet [5], HuBERT [7], and VGGish [8]. Extensive experiments and comparative analyses demonstrate the effectiveness of the proposed approach.

The remainder of this paper is organized as follows: Section II reviews the related work. Section III presents the proposed methodology. Section IV describes the experimental methodology. Section V reports and discusses the results, and Section VI concludes the paper.

II. Related Works

VC research has evolved substantially from early statistical mapping methods to modern end-to-end neural models. Early approaches predominantly used frame-based statistical models, like GMMs [1] and Nonlinear Frequency Mapping (NFM)

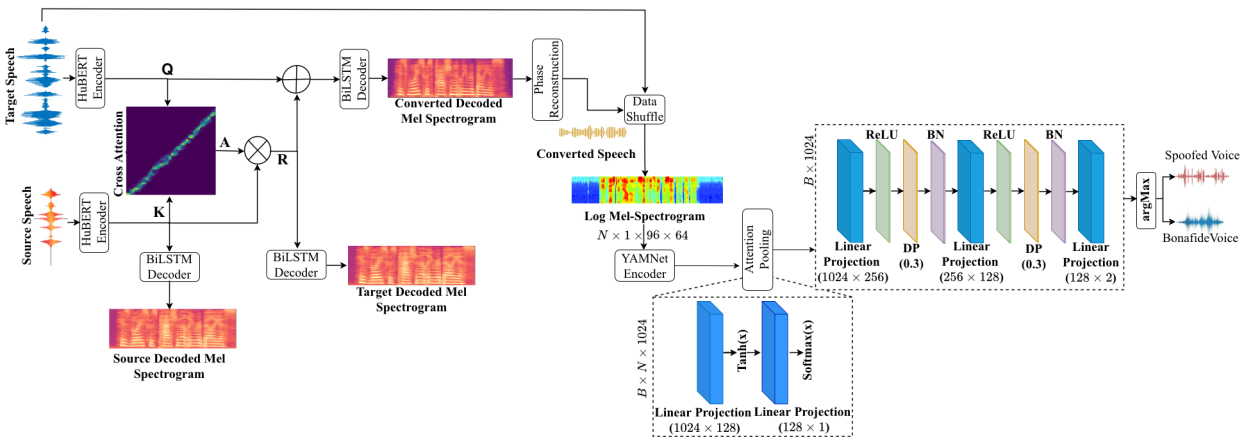

Fig. 1. Proposed model of voice conversion and identifying real and converted voice.

[11], necessitating time-aligned parallel speech data from both source and target speakers. While effective in controlled environments, these approaches exhibit limited flexibility and often fail to model prosodic variations and duration changes accurately. Subsequent advances led to the development of neural network-based VC models, including Restricted Boltzmann Machines (RBMs) [12] and generative adversarial network (GAN) frameworks such as CycleGAN [13] and StarGAN [14]. GAN-based approaches provide parallel-data-free VC and significantly improve naturalness and speaker resemblance. However, they typically require large training datasets and may struggle to convert speech duration effectively.

Recently, attention-based `Seq2Seq` architectures have emerged as a highly successful paradigm for VC, facilitating end-to-end mapping of variable-length acoustic sequences while concurrently modeling spectral features and speech duration [15]. Models like ATTS2S-VC framework exhibit enhanced conversion quality and temporal alignment relative to conventional GMM-based approaches. Simultaneously, anti-spoofing methods have progressed from handcrafted spectral and prosodic features with classical classifiers [16] to more advanced methodologies. However, these conventional methods often encounter difficulties in identifying speech produced by contemporary neural VC and text-to-speech (TTS) models, whose outputs display remarkably lifelike attributes. This difficulty underscores the necessity for good anti-spoofing methodologies adept at accurately detecting sophisticated generated speech.

III. Proposed Methodology

Fig. 1 depicts the outline of the proposed framework, which integrates VC and voice spoofing detection system.

A. Voice Conversion Model

Let the source and target acoustic feature sequences be denoted by

$$X = [x_1, \ldots, x_I], \qquad Y = [y_1, \ldots, y_J]. \tag{1}$$

Both sequences are passed through pretrained HuBERT encoders to obtain contextualized speech embeddings:

$$K = HuBERTencoder(X), Q = HuBERTencoder(Y). \tag{2}$$

where, $K = [k_1, \ldots, k_I]$, $Q = [q_1, \ldots, q_J]$, $k_i, q_j \in \mathbb{R}^{768}$.

1) Attention-Based Alignment: Source and target speech often differ in length and temporal structure. In order to accurately predict the output sequence, we adopt an attention mechanism to compute soft alignments between source embeddings K and target embeddings Q. At each time frame of the embeddings Q, the attention mechanism gives a probability distribution that describes the relationship between the given time frame feature q_j and the embeddings K. Consequently, the attention matrix A can be written as

$$e_{ij} = f_{FFNN}(k_i, q_j) \tag{3}$$

The scores are normalized using the softmax function:

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_i exp(e_{ij})} \tag{4}$$

where f_{FFNN} indicates a function described by feed-forward NNs and a_{ij} is an element (i, j) of the attention matrix A.

2) Context Vector Construction: Obtain long-range temporal dependencies R between the source and target sequences by multiplying attention matrix A with source embeddings K: $R = AK$. To form the fused representation used for conversion, we combine R with target embeddings: $Z = R + Q$. The system utilizes BiLSTM decoders to reconstruct the mel-spectrogram from the context vector representation. The detailed parameters of the VC model are shown in Table I.

B. Voice Spoofing Detection

For spoofing detection, a pretrained `YAMNet` model is employed as a feature extractor to obtain rich audio embeddings from raw speech. The pretrained `YAMNet` network generates compact, fixed-length embeddings that effectively condense

TABLE I
VOICE CONVERTER MODEL PARAMETERS.

Layer (type:depth-idx)	Output Shape	#Params
VoiceConvertor	[1, 262, 128]	–
DecoderBiLSTM: 1-1	[1, 262, 128]	–
LSTM: 2-1	[1, 262, 1536]	23,617,536
Linear: 2-2	[1, 262, 128]	196,736
DecoderBiLSTM: 1-2	[1, 232, 128]	(recursive)
LSTM: 2-3	[1, 232, 1536]	(recursive)
Linear: 2-4	[1, 232, 128]	(recursive)
DecoderBiLSTM: 1-3	[1, 232, 128]	–
LSTM: 2-5	[1, 232, 1536]	23,617,536
Linear: 2-6	[1, 232, 128]	196,736

complex acoustic information into a low-dimensional representation. These embeddings capture speaker-specific characteristics and discriminative spectral properties that are essential for distinguishing genuine speech from spoofed or manipulated audio.

Before processing audio with YAMNet, several preprocessing steps are applied. Since the model expects a fixed input format, each audio sample is first **resampled to 16 kHz**, ensuring compatibility with the network architecture and consistency across samples. The resampled waveform is then converted into a **96×64 log-mel spectrogram**, which provides a time–frequency representation of the signal. YAMNet processes this sequence of spectrogram patches and produces a corresponding sequence of **1024-D embedding vectors**.

To aggregate temporal information, an `attention pooling` layer computes a weighted global embeddings. This resulting pooled representation is then fed into two fully connected layers with Rectified Linear Unit (ReLU) activation, batch normalization (BN), and dropout (DP). The final classifier outputs logits determine whether the input speech signal is bonafide or spoofed. The detailed parameters of the YAMNet-based voice spoofing detection system are shown in Table II.

TABLE II
YAMNET-BASED CLASSIFIER PARAMETERS.

Layer (type)	Output Shape	#Params
Linear-1	[-1, 30, 128]	131,200
Tanh-2	[-1, 30, 128]	0
Linear-3	[-1, 30, 1]	129
Linear-4	[-1, 256]	262,400
ReLU-5	[-1, 256]	0
Dropout-6	[-1, 256]	0
BatchNorm1d-7	[-1, 256]	512
Linear-8	[-1, 128]	32,896
ReLU-9	[-1, 128]	0
Dropout-10	[-1, 128]	0
BatchNorm1d-11	[-1, 128]	256
Linear-12	[-1, 2]	258

IV. EXPERIMENTAL METHODOLOGY

A. Dataset

We used the CMU Arctic database [17] consisting of utterances by two female speakers **clb** and **slt** and two male speakers **ksp** and **rms**. We trained the VC model for **clb** as source speaker and **slt** as target speaker and generated converted speech data, similarly we trained the model for **ksp** as source speaker and **rms** as target speaker and generated corresponding converted speech data. In the **Voice Spoofing Detection** model, we used 300 sentences from a target speaker (**slt**) along with the corresponding 300 converted speech samples The genuine target speech samples were labeled as "Actual Voice", while the converted speech samples were labeled as "Converted Voice." The combined dataset was randomly shuffled and then fed into the model for training. After training, the model's performance was evaluated using an unseen test set consisting of 200 sentences from another target speaker (**rms**) and the corresponding 200 converted speech samples generated by the same voice conversion model. Additionally, we evaluated the proposed model on the Chula Spoofed Speech (CSS) [18] dataset, where spoofed speech samples are generated using two-stage TTS systems. Specifically, FastPitch combined with UnivNet–generated speech was employed as a spoofed input for evaluation.

B. Experiment Setup

1) Pre-processing: We adopt a standardized pre-processing pipeline to extract content features for both the source and target speakers prior to VC. For VC, raw audio waveforms are loaded using librosa library and resampled to 16 kHz to match HuBERT's expected input format. First, we utilize the pre-trained HuBERT Base model from the torchaudio library as the primary feature extractor. The HuBERT encoder operates directly on raw audio waveforms and produces high-level content embeddings that are invariant to speaker identity, making it suitable for VC tasks.

For the anti-spoofing subsystem, we employ the pre-trained `YAMNet` model. Similarly, all audio signals are resampled to 16 kHz and transformed into 96×64 log-mel spectrograms, as required by `YAMNet`. This process includes mel-filterbank computation, logarithmic amplitude scaling, and temporal framing. The resulting spectrogram are subsequently used to generate compact and perceptually meaningful embeddings for spoofing detection.

2) Model Details and Training Paradigm: The VC model comprises two pre-trained HuBERT encoders, three BiLSTM decoders, and an attention-based alignment mechanism. During training, we minimize the $L1$ loss between the mel-spectrogram of the converted speech and the target speech.

$$\mathcal{L}_{\text{Seq2Seq}} = \left\| \hat{Y} - Y \right\| \tag{5}$$

This loss encourages the converted speech to closely match the acoustic characteristics of the target speaker, improving perceptual similarity and conversion quality. To stabilize training and preserve linguistic content, context preservation losses are introduced for both source and target speech:

$$\mathcal{L}_{\text{source}} = \left\| \tilde{X} - X \right\|, \quad \mathcal{L}_{\text{target}} = \left\| \tilde{Y} - Y \right\| \tag{6}$$

Because speech progresses monotonically over time, the attention matrix should be approximately diagonal. To accelerate

the training of the attention mechanism, a guided attention loss is introduced. In most speech processing tasks, the alignment between source and target sequences is expected to be nearly monotonic and time-incremental. It is therefore reasonable to assume that the i-th time frame of the source speech aligns approximately linearly with the j-th time frame of the target speech, i.e., $i \approx \alpha j$, where $\alpha \approx \frac{I}{J}$. Under this assumption, the attention matrix A should be nearly diagonal. To enforce this structure, a penalty matrix G is defined as

$$g_{ij} = 1 - \exp\left(-\frac{\left(\frac{i}{I} - \frac{j}{J}\right)^2}{2\sigma_g^2}\right). \tag{7}$$

The guided attention loss $\mathcal{L}_{ga}$ is defined as

$$\mathcal{L}_{\text{ga}} = \|G \odot A\| \tag{8}$$

The overall training objective is defined as:

$$\mathcal{L}_{proposed} = \mathcal{L}_{Seq2Seq} + \lambda_{ga}\mathcal{L}_{ga} + \lambda_{cp}\left(\mathcal{L}_{source} + \mathcal{L}_{target}\right). \tag{9}$$

The hyperparameters are set to σ_g=0.4, λ_{ga}=10,000, and λ_{cp}=10. The model is trained with a batch size of 32, a learning rate of $1e-5$, and 128 epochs.

The spoofing detection model uses a pre-trained `YAMNet` network as a feature extractor, followed by an attention-based pooling layer and two fully connected layers. The attention pooling learns frame-level importance weights to produce a fixed-length global representation, emphasizing regions that are most indicative of spoofing artifacts such as synthesis distortions, replay noise, or vocoder inconsistencies. The aggregated pooled 1024-dimensional representation is then passed through a classifier consisting of two fully connected layers (1024×256 and 256×128), each followed by ReLU activation, DP (0.3), and BN. Finally, a linear layer of size 128×2 produces the output logits. The model is trained using cross-entropy loss, with predictions obtained via `argMax` over the logits. Optimization is performed using the Adam optimizer with a learning rate of 0.001, weight decay of $1e{-}4$, and a cosine annealing learning rate scheduler. The dataset was divided into 30% for training, 35% for evaluation, and 35% for testing. This split was chosen to allow for a rigorous validation and an unbiased performance evaluation. Given the sufficient dataset size, 30% of the data was adequate for model training, while the larger validation and test sets provided a more reliable evaluation of generalization on unseen data.

3) Post Processing: During experimentation, several key limitations of the VC model were observed. The system operates on mel-spectrograms, which preserve magnitude information but discard phase information. Consequently, waveform reconstruction from the transformed mel-spectrogram often introduces artifacts, manifested as dark regions in the spectrogram and reduced speech intelligibility.

To address this issue, we introduce a phase reconstruction module based on a modified `Griffin−Lim` algorithm. The proposed approach fixes the magnitude of the converted mel-spectrogram and iteratively refines only the phase information to obtain a consistent time-frequency representation. By constraining the magnitude and optimizing the phase, the reconstructed waveform more closely matches the phase characteristics of the target speech. This post-processing step significantly reduces noise and artifacts, resulting in smoother mel-spectrograms and improved perceptual quality of the synthesized speech.

4) Computational Details: The experiments were conducted on an Apple M4 Pro CPU machine with 24 GB RAM, using Python 3.9.6 and PyTorch version 2.2.2. The proposed VC method required approximately 50 hours for training and about 20 minutes for testing. In contrast, the voice spoofing detection model required 40.25 minutes for training and 17.17 minutes for evaluation.

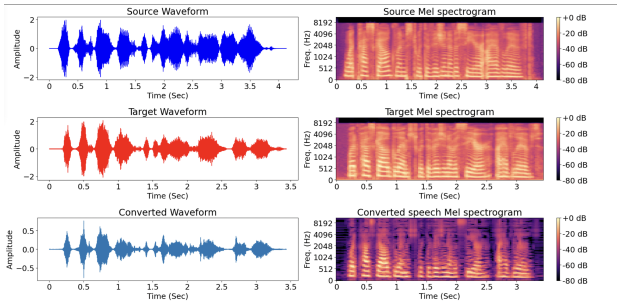

Fig. 2. Source, target, and converted speech mel-spectrograms and corresponding waveforms generated by the proposed VC model, shown after phase reconstruction.

V. Results and Discussion

A. Discussion

Fig. 2 depicts the source, target, and converted speech waveforms along with their corresponding mel-spectrograms obtained using the pretrained HuBERT-based VC model and the modified `Griffin−Lim` phase reconstruction. The converted speech exhibits spectro-temporal patterns and waveform characteristics closely matching those of the target speaker, confirming effective speaker characteristic transfer while preserving linguistic content. Fig. 3 presents the confusion matrices comparing the SOTA models with the proposed YAMNet framework. The diagonal elements represent correctly classified instances, while the off-diagonal elements indicate misclassifications, revealing class-wise confusion patterns.

B. Comparative Analysis

For comparison, the following SOTA techniques are evaluated: Modified GDF [10], RawNet [5], HuBERT [7], and VGGish [8]. As shown in Table III, the modified GDF method achieves a significantly lower accuracy $\eta \leftarrow$61.57%, mainly due to its high sensitivity to noise, spectral zeros, and phase estimation errors. Conversely, DL-based feature extractors have significantly better performance: RawNet [5], HuBERT [7], and VGGish [8] attain accuracies of 87.71%, 89.05%, and 88.45%, respectively, and the proposed YAMNet framework surpasses all SOTAs with an accuracy of 96.29%. Additionally, we have evaluated the proposed framework on the Chula

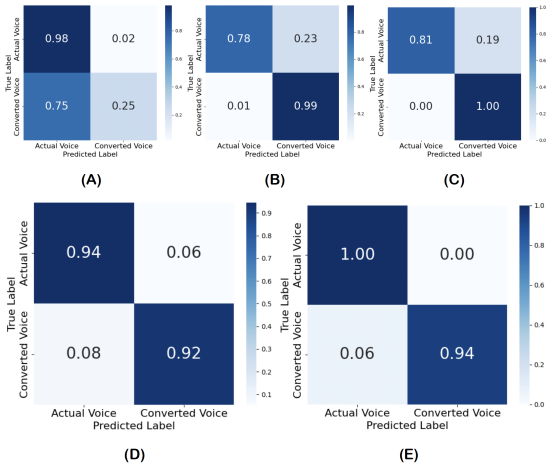

Fig. 3. Confusion matrices of (A) Modified Group Delayed function, (B) RawNet, (C) HuBERT, (D) VGGish, and (E) the proposed YAMNet-based voice spoofing detection method.

Spoofed Speech (CSS) dataset, in which synthetic speech samples are generated using five high-quality Text-to-Speech (TTS) systems. The resulting spoofed speech is highly natural and perceptually convincing, making it extremely difficult for human listeners to distinguish it from genuine speech. As shown in Table IV, the modified GDF [10], RawNet [5], HuBERT [7], and VGGish [8] attain accuracies of 52.77%, 47.22%, 61.11%, and 52.77%, respectively, and the proposed YAMNet framework achieved better accuracy than SOTAs, which is 75.00%. The proposed framework provides near-optimal classification performance with minimal training duration and compact model dimensions, underscoring it's efficacy and efficiency in speech spoofing detection.

TABLE III
COMPARATIVE ANALYSIS WITH SOTAS USING SPOOFED SPEECH GENERATED BY THE VC MODEL.

Methods	**Performance Metrics (%)**				**Training Time (s)**	**#Params**
	η	***Se***	***Sp***	***F1-Score***		
Modified GDF [10]	61.57	30.44	92.77	42.29	36.64	42.1K
RawNet [5]	87.71	98.00	99.89	89.62	66.19	11.3M
VGGish [8]	88.45	81.48	95.43	85.91	378.16	83.6K
HuBERT [7]	89.05	96.80	81.31	90.29	549.42	329K
YAMNet (Proposed)	**96.29**	**92.58**	**100.00**	**95.96**	**64.92**	**428K**

TABLE IV
COMPARATIVE ANALYSIS USING CHULA SPOOFED SPEECH (CSS) DATASET [18].

Methods	**Performance Metrics (%)**				**Training Time (s)**	**#Params**
	η	***Se***	***Sp***	***F1-Score***		
RawNet [5]	47.22	38.88	55.55	42.22	1.35	11.3M
Modified GDF [10]	52.77	11.11	94.44	19.04	3.22	42.1K
HuBERT [7]	52.77	16.66	88.88	26.08	9.91	329K
VGGish [8]	61.11	44.44	77.77	53.33	11.57	83.6K
YAMNet (Proposed)	**75.00**	**77.77**	**72.22**	**75.67**	**1.17**	**428K**

VI. CONCLUSION AND FUTURE WORK

We improved a `Seq2Seq`-based voice conversion (VC) framework by integrating an attention mechanism and a pre-trained HuBERT encoder, as well as a modified `Griffin--Lim` algorithm for phase reconstruction. Experimental findings demonstrate that the proposed framework significantly outperforms the SOTAs, attaining higher conversion quality with fewer training epochs. The converted speech was later used to train several voice spoofing detection systems. The proposed YAMNet-based framework achieved a validation accuracy of 96.29% with only 428K trainable parameters, which is suitable for resource-constrained Automatic Speaker Verification (ASV) systems. Future work will include validating the framework on larger and more diverse datasets, extending it to multi-speaker and cross-lingual VC, evaluating robustness under real-time conditions and exploring neural vocoders, phase-aware reconstruction, and model compression for improved speech naturalness and real-time deployment.

ACKNOWLEDGEMENT

This work is financially supported by TIH IIT Guwahati under Project No. TIH/TD/0425.

REFERENCES

[1] L.-H. Chen, Z.-H. Ling, W. Guo, and L.-R. Dai, "GMM-based voice conversion with explicit modelling on feature transform," in *7th Int. Symp. on Chinese Spoken Lang. Proc.*, pp. 364–368, IEEE, 2010.
[2] Nakashika et al., "Voice Conversion Using RNN Pre-Trained by Recurrent Temporal Restricted Boltzmann Machines," *IEEE/ACM Trans. on Audio, Speech, and Language Proc.*, vol. 23, no. 3, pp. 580–587, 2015.
[3] S. Dhar, N. D. Jana, and S. Das, "An adaptive-learning-based generative adversarial network for one-to-one voice conversion," *IEEE Trans. on Artificial Intelligence*, vol. 4, no. 1, pp. 92–106, 2022.
[4] Kamel et al., "A Survey of Threats Against Voice Authentication and Anti-Spoofing Systems," *arXiv preprint arXiv:2508.16843*, 2025.
[5] Tak et al., "End-to-end anti-spoofing with rawnet2," in *Int. Conf. on Acou., Speech and Sig. Proc.*, pp. 6369–6373, IEEE, 2021.
[6] Hershey et al., "VGGish: A VGG-like Audio Classification Model," in *Proc. ICASSP*, 2017.
[7] Hsu et al., "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Trans. on Audio, Speech, and Language Proc.*, 2021.
[8] Shaoyong et al., "Cross-modal supervised learning for better acoustic representations," *arXiv preprint arXiv:1911.07917*, 2019.
[9] Shahzad et al., "Enhancing Voice Spoofing Detection: A Hybrid Approach With VGGish-LSTM Model for Improved Security in Automatic Speaker Verification Systems," *IEEE Access*, 2025.
[10] Murthy et al., "The modified group delay function and its application to phoneme recognition," in *Int. Conf. on Acou., Speech, and Signal Proc. Proceedings.(ICASSP'03).*, vol. 1, pp. I–68, IEEE, 2003.
[11] R. Aihara, T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary," in *Int. Conf. on Acou., Speech and Signal Proc. (ICASSP)*, pp. 7894–7898, 2014.
[12] K.-S. Lee, "Restricted boltzmann machine-based voice conversion for nonparallel corpus," *IEEE Sig. Proc. Lett.*, vol. 24, no. 8, pp. 1103–1107, 2017.
[13] C. Fu, C. Liu, C. T. Ishi, and H. Ishiguro, "Cycletransgan-evc: A cyclegan-based emotional voice conversion model with transformer," *arXiv preprint:2111.15159*, 2021.
[14] S. Sakamoto, A. Taniguchi, T. Taniguchi, and H. Kameoka, "Stargan-vc+ asr: Stargan-based non-parallel voice conversion regularized by automatic speech recognition," *arXiv preprint arXiv:2108.04395*, 2021.
[15] Tanaka et al., "AttS2S-VC: Sequence-to-sequence voice conversion with attention and context preservation mechanisms," in *IEEE Int. Conf. on Acou., Speech and Sig. Proc. (ICASSP)*, pp. 6805–6809, IEEE, 2019.
[16] Evans et al., *Anti-spoofing, Voice Conversion*, pp. 1–10. 09 2014.
[17] "Festvox: CMU_ARCTIC Databases."
[18] Slscu, "GitHub - SLSCU/CSS: An official repository of CSS: Chula Spoofed Speech (CSS) dataset — a large-scale Thai spoofed speech dataset.."