Towards Detecting Neural Audio Codec Synthesized Heart Sounds

Anonymous submission to Interspeech 2026

Abstract

In this paper, we introduce Synthetic Heart Sound Detection (SHAC), a task aimed at identifying phonocardiograms (PCGs) synthesized using neural audio codecs (NACs). To facilitate research in this direction, we release **CARDIOFAKE**, the first benchmark dataset for SHAC containing both real and codec-synthesized PCGs. We benchmark spectral representations (MFCC, LFCC) and self-supervised learning (SSL) representations (e.g., WavLM) for the task. Furthermore, we propose **GROOT**, a fusion framework that integrates spectral and SSL features for leveraging their complementary behavior. Experiments show that **GROOT**, combining MFCC and WavLM, achieves state-of-the-art performance, outperforming individual representations and competitive baselines.

Index Terms: Synthetic Heart Sound Detection, Phonocardiograms, Neural Audio Codecs

1. Introduction & Background

Spoofing attack detection (SAD) is widely regarded as a core safeguard for biometric systems, and has been systematically explored in speech [1] and facial recognition [2]. In the speech domain, extensive research has investigated replay and voice conversion-based attacks, leading to standardized evaluations such as the ASVspoof challenges [3, 4]. Through standardized protocols, these benchmarks have supported the development of effective countermeasures [5, 6]. Work on face recognition has followed a similar path, from early texture-based methods to modern deep learning approaches [7, 8, 9]. Parallel advances have been made also in fingerprint and iris recognition [10]. Collectively, these efforts underscore how community-driven efforts have catalyzed rapid progress in spoofing detection across established biometric systems. Despite continued progress in spoofing attack detection, the threat remains substantial. Conventional biometric modalities—including facial images, fingerprints, speech-have consistently been shown to be vulnerable to ever-growing advancement in sophisticated spoofing techniques. In contrast, phonocardiograms (PCGs), or heart sounds, have been regarded as a promising biometric modality, offering uniqueness, inherent liveness, and a natural resilience against traditional spoofing methods [11]. Unlike fingerprints, faces, or voices, heart sounds are directly tied to physiological processes, making them intrinsically difficult to forge. As such, subsequent research proposed a range of strategies for building biometric systems based on heart sounds. They have used wavelet-based features [12], cepstral features [13] with classical ML algorithms to modern day transformer-based architectures [14].

However, the very perception of heart sounds as a non-invasive biometric modality renders their vulnerability particularly concerning. With rapid advances in neural audio codec (NAC)–based synthesis, adversaries can now generate synthetic heart sounds that are perceptually indistinguishable from genuine recordings, posing a direct threat similar to those observed in synthetic speech generation [15, 16]. To confront this risk, we present the first systematic investigation into the vulnerability of heart sound biometrics under NAC-driven spoofing attacks—a new research direction for secure biometric authentication. As part of this effort for spoofing detection of such attacks-we coin the novel task of Synthetic Heart Sound Detection (SHAC) and release the first benchmark dataset, **CARDIOFAKE** (**FAKE** Phono**CARDIO**grams) comprising both real and codec-synthesized heart sounds. We conduct a comprehensive evaluation of spectral features (MFCC, LFCC) and self-supervised learning (SSL)–based representations (e.g., WavLM). *We hypothesize that these two classes of features are complementary: spectral features are highly sensitive to NAC-induced distortions at the acoustic level, while SSL representations capture broader temporal structure and variability in heart sounds*. Building on this hypothesis, we introduce **GROOT** (Fusion via **GR**ammian **O**ptimal Transp**O**r**T**), a novel fusion framework that integrates spectral and SSL representations through novel grammian optimal transport. By combining MFCC with WavLM, **GROOT** achieves state-of-the-art (SOTA) performance, surpassing both individual representations and competitive baselines. Our benchmark and baselines establish a foundation for future research on robust countermeasures against this emerging class of attacks.

Key contributions of this work are threefold:

- We coin the novel task of SHAC as a new research direction for secure biometric authentication.
- We release the first benchmark dataset for SHAC, **CARDIOFAKE**, containing both real and codec-synthesized heart sounds, and conduct a comprehensive evaluation of spectral and SSL-based representations, providing critical insights into their strengths and limitations.
- We propose **GROOT**, a novel framework that integrates spectral and SSL representations to exploit their complementary strengths. At its core, **GROOT** employs a novel grammian optimal transport mechanism. **GROOT** achieves SOTA performance, outperforming individual representations and competitive baselines, thereby setting a strong foundation for future countermeasures against this emerging class of spoofing attacks.

We will release the dataset, codes and models curated for this work after the review process.

2. CARDIOFAKE Dataset

This section outlines the resources and methodology employed in creating the **CARDIOFAKE** dataset, including the heart sounds corpora, the NACs used for synthesis, and the overall pipeline for producing the artificial samples.

2.1. Heart Sound Dataset

For synthesizing heart sounds, we employ CirCor DigiScope dataset [17], which is openly accessible via PhysioNet [18]. Our work concentrates on the open-access portion of the dataset, comprising recordings from 963 patients. Each patient record is labeled with one of three categories: Present, Absent, or Unknown. Altogether, the collection provides 3,163 phonocardiogram recordings, with durations spanning 5 to 65 seconds.

2.2. Neural Audio Codecs

We utilize the NACs used by Lu et al. [15] and Wu et al. [16], focusing on SOTA and publicly available codecs that adversaries might use for generation of synthetic heart sounds. The NACs leveraged are given as follows:

Descript Audio Codec (DAC) [19]: It is a high-fidelity VQ-GAN–based model that leverages residual vector quantization (RVQ) along with adversarial and multi-scale spectral losses; we employ its 16kHz variant.

Encodec[20]: It is a real-time convolutional encoder–decoder codec with RVQ, integrates time/frequency reconstruction losses and spectrogram adversarial objectives; we use the 24kHz version.

Soundstream [21]: It is a NAC tailored for low-bitrate speech compression, employing an encoder–decoder framework with RVQ and multi-scale STFT discriminators to balance fidelity and compression efficiency, supporting bitrates from 3–18 kbps. We adopt its 16kHz variant.

Speech Tokenizer [22]: It serves as a unified audio tokenizer bridging semantic and acoustic cues. Built on an encoder–decoder with RVQ, it hierarchically disentangles content and paralinguistic information across layers, producing composite token sequences. We employ its default 16kHz configuration.

FunCodec [23]: It incorporates RVQ with semantic augmentation and adversarial training to yield compact, expressive representations; we use the openly released LibriTTS trained and bilingual 16kHz version.

AudioDec [24]: It adopts an autoencoder framework with a two-stage training strategy—metric losses for convergence followed by decoder-only adversarial fine-tuning for fidelity; we employ the 28kHz variant.

SNAC [25]: SNAC extends RVQ with hierarchical quantizers across temporal scales, depthwise convolutions, noise injection, and local attention; we use 24kHz version in our study.

2.3. CARDIOFAKE Generation Pipeline

We design a controlled pipeline inspired by prior work on NAC synthesized deepfakes [16] for building **CARDIOFAKE**. We start with CirCor DigiScope dataset, where each utterance serves as a real reference. Synthetic samples are created through a NAC synthesis–resynthesis loop: the original waveform is first encoded into a discrete latent representation by a pre-trained NAC encoder and then reconstructed by its decoder, yielding a synthetic version. This process preserves the underlying cardiac acoustic patterns while introducing subtle codec-induced artifacts, resulting in realistic yet synthetic heart sounds. We apply this approach across 7 NACs, producing parallel datasets in which each real utterance has a one-to-one synthetic counterpart per codec. So, in total, we have 3163 real heart sounds and 22141 synthetic heart sounds. Two evaluation settings are defined: seen, where test utterances are generated using the same NACs as in training (SNAC, DAC, EnCodec, Soundstream, Speech Tokenizer), and unseen, where test utterances are synthesized with different codecs (FunCodec, AudioDec) to evaluate generalization.

3. Methodology

3.1. Feature Extraction

We employ MFCC[1] and and linear-frequency cepstral coefficients (LFCC[2]) as spectral representations. We extract 14-dimensional LFCC and 40-dimensional MFCC after average pooling. We employ different SOTA SSL representations as they have shown effectiveness for heart sound classfication tasks [26]. We use Wav2vec2[3] [27] pre-trained on 960 hours of Librispeech. Its self-supervised framework learns contextualized representations by masking speech frames and predicting quantized latent targets. Lastly, we use SOTA models in SUPERB i.e. Unispeech-SAT[4] [28] and WavLM[5] [29]. Unispeech-SAT extends self-supervised learning by incorporating speaker-aware objectives.This enables the model to disentangle speaker identity from linguistic content, whereas WavLM incorporates masked prediction with denoising objectives. We resample all the inputs to 16 kHz and extract representations by average pooling over the final hidden layer of each frozen SSL model. We extract 768 dimension representation for all the SSL models.

3.2. Individual Representations Modeling

We follow previous work and experiment with two downstream modeling architectures that shown SOTA results on audio deepfake detection with SSL representations [30]. Firstly, we apply a fully connected network (FCN) with two dense layers of 180 and 60 neurons to the extracted representations. Secondly, we use CNN, we place a 1D-CNN layer with 32 filters on top of the representations, followed by a max-pooling layer, flatten and a FCN with the same details as the FCN above. The output layer of both models uses a sigmoid activation function.

3.3. GROOT

We propose **GROOT** for the fusion of representations and the architecture is shown in Figure 1. The extracted representations are first passed through a 1D-CNN block with the same design as used in the individual CNN models. The output is then flattened and linearly projected to a 120-dimensional vector, where the dimensionality reduction is introduced mainly to reduce computational cost. These projected features are then fed into the fusion module, which employs the proposed novel grammian optimal transport (Gram-OT) for aligning the representations. Vanilla optimal transport (OT) has been widely adopted

[1] https://librosa.org/doc/main/generated/librosa.feature.mfcc.html

[2] https://spafe.readthedocs.io/en/latest/features/lfcc.html

[3] https://huggingface.co/facebook/wav2vec2-base

[4] https://huggingface.co/microsoft/unispeech-sat-base

[5] https://huggingface.co/microsoft/wavlm-base

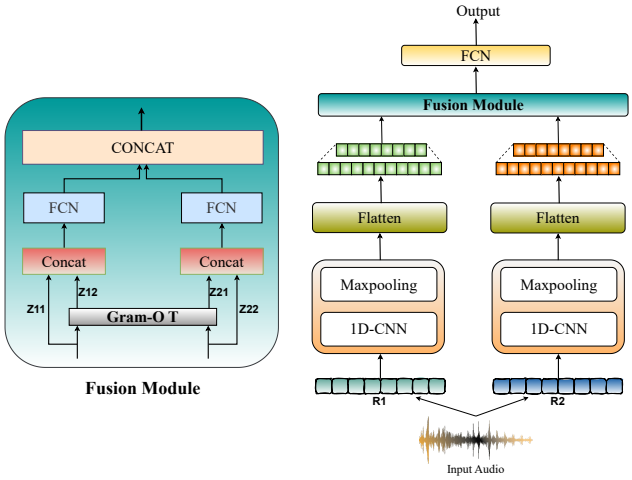


Figure 1: *Proposed Framework:* ***GROOT****: R1 and R2 represent input features from two branches. Z_{11} and Z_{22} denote features from the respective FCN branches, while Z_{12} and Z_{21} denote transported features*

for representation alignment [31, 32], but it directly compares raw features, making it sensitive to scaling and noisy variations. In contrast, Gram-OT compares representations through their gram matrices, which capture correlations between features and reflect global relational patterns across the representation space. This enables Gram-OT to preserve meaningful characteristics such as rhythm while being more robust to noise, distortions, and variability across inputs. Let us consider the feature vectors of two representations after flattening are R_1 and R_2. We first compute their gram matrices:

$$G_{R_1} = R_1 R_1^\top \quad G_{R_2} = R_2 R_2^\top$$

We then construct a cost matrix based on the frobenius distance between the two gram matrices:

$$M = \frac{\|G_{R_1} - G_{R_2}\|_F}{\max\limits_{(R_1,R_2)} \|G_{R_1} - G_{R_2}\|_F}$$

Here, frobenius distance is used in place of euclidean distance used in vanilla OT, as it provides the natural generalization of euclidean distance from vectors to matrices, making it well-suited for comparing gram matrices. To align the features, we apply the sinkhorn algorithm on this cost to obtain the optimal transport plan Γ

$$\Gamma = \text{Sinkhorn}(M)$$

Using Γ, we transport the feature spaces of the two representations into one another:

$$R_2 \to R_1 = \Gamma \cdot R_2 \quad R_1 \to R_2 = \Gamma^\top \cdot R_1$$

Finally, the transported features are concatenated with their corresponding original representations to form the fused representations F_1 and F_2:

$$F_1 = \text{Concat}(R_2 \to R_1, R_1) \quad F_2 = \text{Concat}(R_1 \to R_2, R_2)$$

The fused representations F_1 and F_2 are first passed in parallel through FCN with a dense layer of 80 neurons each, and their outputs are subsequently concatenated. This concatenated vector is then processed by another FCN consisting of two dense layers with 120 and 30 neurons, respectively, followed by the final output layer with a sigmoid activation function for binary classification.

4. Experiments

4.1. Training and Hyperparameter Details

All models are trained for 50 epochs with a batch size of 32, using the Adam optimizer and binary cross-entropy loss. To mitigate overfitting, dropout regularization is applied. We keep a learning rate of 1e-3 for the experiments. We also use class-weightage during training to handle the class-imbalance.

4.2. Experimental Results

We begin by evaluating whether NAC-synthesized heart sounds retain patient-specific identity, thereby assessing the credibility of such spoofing attacks. To this end, we perform a closed-set user identification experiment on subjects from the real heart sound corpus. Each patient is treated as a distinct class, and a supervised classifier is trained under four train–test regimes: Real→Real, Real→Fake, Fake→Real, and Fake→Fake (where Real = authentic heart sound, Fake = NAC-synthesized heart sound). The classifier achieves 89.11% accuracy in the Real→Real setting, confirming that genuine heart sounds reliably encode identity cues. Importantly, the Real→Fake regime still reaches 86.29%, indicating that synthetic reconstructions preserve most of the patient-specific information. Moreover, models trained on synthetic data perform even better: 95.07% in Fake→Fake and 93.08% in Fake→Real. These results demonstrate that NAC-based synthesis preserves discriminative identity cues and the higher accuracy is due to the more samples in the synthetic set. In summary, NAC-generated heart sounds constitute highly identity-preserving deepfakes, posing a critical challenge for biometric systems that may struggle to differentiate authentic from synthetic.

PTM's	FCN		CNN	
	ACC ↑	EER ↓	ACC ↑	EER ↓
Seen				
LF	76.99	15.19	79.02	14.96
MF	77.82	15.04	81.56	12.55
W2V	83.62	**12.13**	86.65	10.37
UNS	80.30	12.30	82.81	11.59
WAL	**84.54**	12.51	**87.72**	**9.45**
Unseen				
LF	72.45	18.93	73.99	18.08
MF	74.93	17.60	78.74	16.91
W2V	79.56	16.03	83.61	13.74
UNS	74.02	18.07	78.47	18.69
WAL	**80.54**	**15.01**	**84.02**	**13.39**

Table 1: *Accuracy (ACC) and Equal Error Rate (EER) for Seen and Unseen conditions; Abbreviation used: LFCC (LF), MFCC (MF), Wav2vec2 (W2V), Unispeech-SAT (UNS), WavLM (WAL). All scores are in %. Abbreviations are consistent in Table 2.*

Table 1 presents the results of spectral and SSL features with FCN and CNN downstream models under both seen and unseen evaluation conditions. Overall, CNN-based downstream models consistently outperform their FCN counterparts. Among

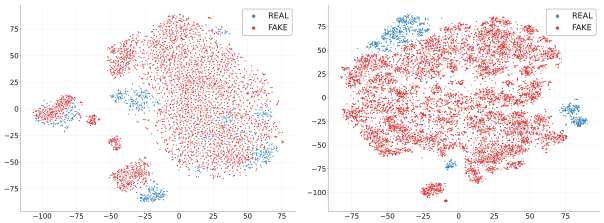
individual representations, WavLM with CNN emerges as the strongest performer, surpassing Wav2vec2 and Unispeech-SAT across both conditions. Furthermore, SSL features consistently outperform spectral counterparts (MFCC and LFCC), highlighting their effectiveness. Table 2 reports the results of representation fusion. We compare against two baselines: simple concatenation and optimal transport (OT) [31]. For a fair comparison, we retain the same architecture as **GROOT** for OT, differing only in the computation of the gram matrix. Similarly, all training settings are kept identical to those used in **GROOT** for both concatenation and OT.We observe that fusion of representations through **GROOT** consistently achieves the best overall performance. Moreover, a clear trend emerges: heterogeneous fusion of spectral and SSL representations outperforms homogeneous fusion, thereby validating our hypothesis on the complementarity of these feature classes. We observe the best performance with fusion of MFCC and WavLM through **GROOT**. These findings not only establish strong baselines for this novel task but also open new avenues for future research.

Fusion	Concat		OT		GROOT	
	ACC ↑	EER ↓	ACC ↑	EER ↓	ACC ↑	EER ↓
Seen						
LF + MF	80.05	11.82	82.41	10.91	84.61	8.35
LF + W2V	86.18	8.72	88.93	8.10	90.50	7.42
LF + UNS	81.12	12.00	84.50	11.18	87.09	9.63
LF + WAL	86.32	7.18	88.36	7.03	91.77	6.06
MF + W2V	86.57	7.20	88.82	6.87	90.83	6.14
MF + UNS	84.00	11.23	85.88	11.22	87.90	9.72
MF + WAL	**87.70**	**7.40**	**89.07**	**6.86**	**93.20**	**5.86**
W2V + UNS	86.99	9.87	87.79	9.06	89.00	8.61
W2V + WAL	86.26	8.32	88.92	7.82	91.60	6.20
UNS + WAL	85.01	10.54	86.23	8.92	88.74	7.55
Unseen						
LF + MF	79.28	16.70	81.87	15.22	83.70	13.99
LF + W2V	84.11	12.34	84.25	10.83	86.00	9.87
LF + UNS	79.05	17.98	80.47	16.00	82.39	14.80
LF + WAL	81.22	13.27	83.01	12.87	85.31	10.98
MF + W2V	81.00	**12.08**	84.72	**10.34**	85.78	10.70
MF + UNS	80.72	14.89	81.51	13.70	83.04	11.49
MF + WAL	**84.33**	13.11	**84.99**	12.06	**86.10**	**9.75**
W2V + UNS	83.11	13.38	84.20	12.40	85.81	11.13
W2V + WAL	83.97	12.90	84.09	11.29	85.70	10.00
UNS + WAL	81.02	15.80	82.68	14.10	84.48	12.51

Table 2: *Evaluation scores for fusion of features; All scores are in %; OT stands for Optimal Transport*

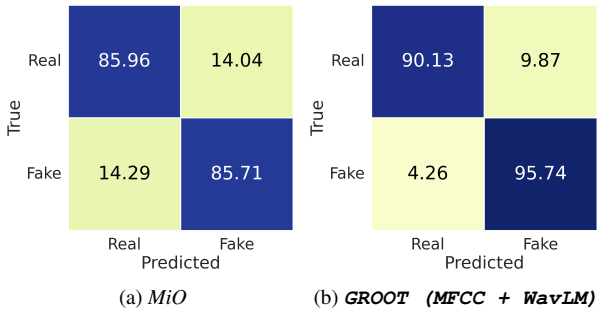
4.3. Comparison to SOTA

As this is the first work addressing SHAC, there are currently no task-specific SOTA models for direct comparison. Therefore, we compare our best-performing method, **GROOT** with MFCC + WavLM (See Table 2), against strong baselines from general audio deepfake detection, namely AASIST [33] and MiO [34]. AASIST is a graph neural network-based architecture, while MiO performs outer-product fusion of SSL representations, making them competitive baselines for this task. We train them following the training configuration for **GROOT** for fair comparison. Our proposed **GROOT** achieves 93.20% accuracy and 5.86% EER in the Seen setting, and 86.10% accuracy with 9.75% EER in the Unseen setting. In comparison, AASIST achieves 85.15% accuracy and 14.91% EER (Seen) and 73.13% accuracy with 16.43% EER (Unseen). MiO shows slight improvement over AASIST with 86.98% accuracy and 12.34% EER (Seen) and 75.89% accuracy with 14.09% EER (Unseen).We further visualize the learned representations using t-SNE plots from the penultimate layers of MiO and **GROOT** (MFCC + WavLM) in Fig. 2. The visualization shows clearer separation and tighter clustering between real and fake heart sounds for **GROOT**, indicating more discriminative representations. Additionally, the confusion matrices for both models are shown in Fig. 3, where **GROOT** demonstrates fewer misclassifications compared to MiO. These observations further confirm that **GROOT** significantly outperforms strong audio deepfake baselines, establishing it as a strong baseline for the SHAC task.



(a) *MiO* (b) ***GROOT (MFCC + WavLM)***

Figure 2: *t-SNE plots*



(a) *MiO* (b) ***GROOT (MFCC + WavLM)***

Figure 3: *Confusion Matrices*

5. Conclusion

In summary, this work introduces the novel task of SHAC, highlighting the emerging risk posed by NACs. To facilitate research in this direction, we release **CARDIOFAKE**, the first benchmark dataset for SHAC, containing both real and codec-synthesized heart sounds. We perform extensive evalaution of both spectral (MFCC, LFCC) and SSL representations (e.g., WavLM) for SHAC. Finally, we present **GROOT**, a novel framework that effectively integrates SSL and spectral features by exploiting their complementary behavior, setting a SOTA for SHAC by outperforming both individual representations and strong competitive baselines.

6. Generative AI Use Disclosure

AI-assisted tools were used only to enhance grammar, clarity, and overall presentation of the manuscript. These tools were not involved in developing the scientific ideas, conducting data analysis, generating results, or interpreting the findings. The authors take full responsibility for the accuracy, validity, and integrity of the work.

7. References

[1] J. Sanchez *et al.*, "Toward a universal synthetic speech spoofing detection using phase information," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 810–820, 2015.

[2] L. Li *et al.*, "Face recognition under spoofing attacks: countermeasures and research directions," *Iet Biometrics*, vol. 7, no. 1, pp. 3–14, 2018.

[3] Z. Wu *et al.*, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *INTERSPEECH 2015*, 2015, pp. 2037–2041.

[4] M. Todisco *et al.*, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," *arXiv preprint arXiv:1904.05441*, 2019.

[5] G. Lavrentyeva *et al.*, "Audio replay attack detection with deep learning frameworks." in *Interspeech*, 2017, pp. 82–86.

[6] H. Delgado *et al.*, "Asvspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," *arXiv preprint arXiv:2109.00535*, 2021.

[7] J. Määttä *et al.*, "Face spoofing detection from single images using micro-texture analysis," in *2011 international joint conference on Biometrics (IJCB)*. IEEE, 2011, pp. 1–7.

[8] D. Wen *et al.*, "Face spoof detection with image distortion analysis," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 746–761, 2015.

[9] Z. Boulkenafet *et al.*, "Face anti-spoofing based on color texture analysis," in *2015 IEEE international conference on image processing (ICIP)*. IEEE, 2015, pp. 2636–2640.

[10] D. Menotti *et al.*, "Deep representations for iris, face, and fingerprint spoofing detection," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 864–879, 2015.

[11] K. Phua *et al.*, "Heart sound as a biometric," *Pattern recognition*, vol. 41, no. 3, pp. 906–919, 2008.

[12] G. Gautam and D. Kumar, "Biometric system from heart sound using wavelet based feature set," in *2013 International Conference on Communication and Signal Processing*, 2013, pp. 551–555.

[13] S. Verma *et al.*, "Analysis of heart sound as biometric using mfcc & linear svm classifier," *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, vol. 3, no. 1, pp. 6626–6633, 2014.

[14] Y. Cao *et al.*, "Enabling passive user authentication via heart sounds on in-ear microphones," *IEEE Transactions on Dependable and Secure Computing*, vol. 22, no. 2, pp. 1195–1209, 2025.

[15] Y. Lu *et al.*, "Codecfake: An initial dataset for detecting llm-based deepfake audio," in *Interspeech 2024*, 2024, pp. 1390–1394.

[16] H. Wu *et al.*, "Codecfake: Enhancing anti-spoofing models against deepfake audios from codec-based speech synthesis systems," in *Interspeech 2024*, 2024, pp. 1770–1774.

[17] J. Oliveira *et al.*, "The circor digiscope dataset: from murmur detection to murmur classification," *IEEE journal of biomedical and health informatics*, vol. 26, no. 6, pp. 2524–2535, 2021.

[18] A. L. Goldberger *et al.*, "Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals," *Circulation [Online]*, vol. 101, no. 23, pp. e215–e220, 2000.

[19] R. Kumar *et al.*, "High-fidelity audio compression with improved rvqgan," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[20] A. Défossez *et al.*, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[21] N. Zeghidour *et al.*, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[22] X. Zhang *et al.*, "Speechtokenizer: Unified speech tokenizer for speech language models," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=AF9Q8Vip84

[23] Z. Du *et al.*, "Funcodec: A fundamental, reproducible and integrable open-source toolkit for neural speech codec," in *ICASSP 2024-2024*. IEEE, 2024, pp. 591–595.

[24] Y.-C. Wu *et al.*, "Audiodec: An open-source streaming high-fidelity neural audio codec," in *ICASSP 2023*. IEEE, 2023, pp. 1–5.

[25] H. Siuzdak *et al.*, "Snac: Multi-scale neural audio codec," in *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024.

[26] D. S. Panah *et al.*, "Exploring wav2vec 2.0 model for heart murmur detection," in *2023 (EUSIPCO)*. IEEE, 2023, pp. 1010–1014.

[27] A. Baevski *et al.*, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[28] S. Chen *et al.*, "Unispeech-sat: Universal speech representation learning with speaker aware pre-training," *ICASSP 2022*, pp. 6152–6156, 2021.

[29] ——, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[30] O. Chetia Phukan *et al.*, "Heterogeneity over homogeneity: Investigating multilingual speech pre-trained models for detecting audio deepfake," in *Findings: NAACL 2024*, Jun. 2024, pp. 2496–2506.

[31] S. Pramanick *et al.*, "Multimodal learning using optimal transport for sarcasm and humor detection," in *Proceedings of WACV*, 2022, pp. 3930–3940.

[32] K. Rho *et al.*, "Lavcap: Llm-based audio-visual captioning using optimal transport," in *ICASSP 2025*. IEEE, 2025, pp. 1–5.

[33] J.-w. Jung *et al.*, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP 2022*. IEEE, 2022, pp. 6367–6371.

[34] O. C. Phukan, G. Kashyap, A. B. Buduru, and R. Sharma, "Heterogeneity over homogeneity: Investigating multilingual speech pre-trained models for detecting audio deepfake," in *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024, pp. 2496–2506.