Bridging the Age Gap: Towards Detecting Neural Audio Codec Synthesized Elderly Speech Deepfake

Anonymous submission to Interspeech 2026

Abstract

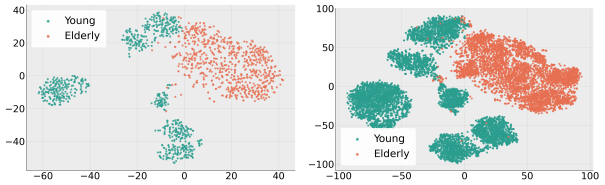
In this study, we introduce the Elderly CodecFake Detection (ECFD) task and release the Elderly-CodecFake (ECF) dataset in English and Chinese. We show that state-of-the-art CF detectors trained on previous benchmark CF datasets generalize poorly to elderly speech, revealing a critical vulnerability. We further hypothesize and demonstrate that multimodal foundation models (FMs) such as LanguageBind (LB) and ImageBind (IB) are more effective for ECFD due to their exposure to elderly content during cross-modal pretraining. Motivated by prior evidence that fusion of FMs enhances downstream performance, we explore fusion of FMs for ECFD. To this end, we propose **BONSAI**, a novel framework that employs Jensen–Shannon Divergence as the fusion mechanism. **BONSAI** with the fusion of LB and IB achieves an average EER (%) of 1.66 and outperforms individual FMs as well as competitive SOTA baselines, establishing a new benchmark for the ECFD task.

Index Terms: CodecFake Detection, Speech Deepfake Detection, Elderly Speech, Multimodal Foundation Models

1. Introduction

In recent years, the boundary between genuine and speech deepfakes has become increasingly blurred. Current text-to-speech (TTS) and voice-conversion (VC) models can generate speech utterances in nearly human-level realism. While these capabilities support valuable applications in assistive communication, human–computer interaction, and entertainment industry, they also enable serious misuse. They can be exploited for impersonation fraud, misinformation campaigns, and unauthorized replication of personal identities. Recognizing this threat, the research community has established dedicated benchmarks to drive the development of reliable countermeasures [1, 2, 3, 4, 5]. Early detection approaches relied on handcrafted features combined with traditional machine learning classifiers [6, 7, 8]. Succeeding them, researchers explored various deep learning approaches for detection of speech deepfakes leading to notable improvements [9, 10, 11]. More recently, usage of large-scale speech foundation models (PTMs) have gained focus in the community; architectures such as Wav2vec2 and WavLM, trained on massive unlabeled corpora, have demonstrated strong transferability to speech deepfake detection [12, 13, 14, 15, 16, 17, 18].

However, prior studies have largely focused on detecting speech deepfakes generated via TTS, VC, or traditional vocoder-based models. With the rapid advancement of audio language models (ALMs), a new class of deepfakes has emerged, demanding dedicated countermeasures. These ALMs rely on neural audio codecs (NACs) for both encoding and synthesis (For example, AudioLM is built on Sounstream as NAC backbone [19]), giving rise to the term CodecFakes (CFs). The first investigations into CF detection were conducted by Wu et al. [20] and Lu et al. [21], who established initial benchmarks and detection strategies for this emerging threat. They showed that models trained on existing vocoder-based datasets fail when evaluated for detection of CFs. As such various researchers have come up with various approaches [22, 23, 24]. Nevertheless, existing resources for CF detection are predominantly built on younger adult speakers, overlooking other demographics. This leaves older adults particularly vulnerable, as elderly speech exhibits distinct vocal traits—such as increased breathiness, reduced pitch stability, and irregular temporal patterns—that differ markedly from younger populations [25] (Figure 1 (a)). Reflecting these differences, related fields such as speech emotion recognition have established dedicated benchmarks and challenges targeting elderly speech [26, 27]. In contrast, research in CF detection has yet to address this demographic, leaving a critical gap in both datasets and methodologies.



(a) (b)

Figure 1: *t-SNE Plots for Real Speech (a) and CF Speech (b) samples for TIS Corpus[28]; Clear separation of Young and Elderly Speech can be observed in both Real and CF scenarios*

To address this limitation, we introduce Elderly CodecFake Detection (ECFD) task (Figure 1 shows the support for the need of ECFD task) and present the Elderly-CodecFake (ECF) dataset, comprising authentic and NAC-generated utterances from older speakers across diverse age bands, NAC backbones, and two languages—English and Chinese. We first evaluate state-of-the-art (SOTA) CF detectors trained on existing datasets [21, 20] and observe substantial performance degradation on NAC-generated elderly voices, revealing poor cross-demographic generalization. Motivated by the success of foundation models (FMs) for speech deepfake detection [14, 18], we also explore FMs for ECFD. To our end, *we hypothesize that multimodal FMs such as LanguageBind (LB) and ImageBind (IB)—are better suited for ECFD, as their cross-modal pretraining improve contextual understanding and implicitly capture age-related cues through exposure to visual contexts such as faces and scenes involving elderly individuals*. The effectiveness of leveraging multimodal FMs due to their cross-modality information prior for audio-centric tasks has also been demonstrated in related domains, such as non-verbal

human vocalization emotion recognition [29]. To validate our hypothesis, we conduct a systematic comparison between leading SOTA speech and multimodal FMs, and the results shows multimodal FMs as champion. Further, inspired by prior evidence that fusion of FMs improves performance for speech deepfake detection [30], we explore such fusion for ECFD and propose **BONSAI** (**B**ridging Fusi**ON** via Jen**S**en–Sh**A**nnon D**I**vergence), which employs Jensen-Shannon divergence as the fusion mechanism. **BONSAI**, applied to LB and IB, achieves the best performance, outperforming individual FMs and competitive SOTA baselines, establishing a new benchmark for ECFD. Our work lays the foundation for CF detection in underrepresented demographics.

To summarize, the main contributions are as follows: (i) We formalize the novel task of ECFD and introduce the ECF dataset. (ii) We benchmark SOTA CF detectors trained on previous benchmark CF datasets and demonstrate a substantial performance degradation on elderly speech, revealing poor cross-demographic robustness. (iii) We show that multimodal FMs are best suited for ECFD by comparing SOTA speech and multimodal fmS. (iv) We propose a novel framework, **BONSAI**, which leverages Jensen-Shannon Divergence for fusion of FMs. **BONSAI** with LB and IB, achieves the best performance with 1.66 as average EER (%), outperforming individual FMs and competitive SOTA baselines. *We release few samples[1] and the full dataset along with the code will be released after the double-blind review process. The dataset will be released through controlled-access to prevent misuse and for research purposes only*.

2. Elderly Codecfake Dataset

This section first describes the sources of real elderly speech data, followed by an overview of the NACs employed in this study. All data are obtained from publicly available corpora. Finally, we detail the pipeline used to generate ECF dataset.

Real Elderly Speech Source: **SeniorTalk** (E1) [31]: It is a Mandarin conversational speech corpus specifically designed for super-aged seniors (Age 75 to 85), containing 55.53 hours of spontaneous dialogues from 202 speakers across 16 provinces. By prioritizing natural interaction rather than read speech, SeniorTalk offers a realistic benchmark for developing age-inclusive voice technologies. The dataset is splitted to training, validation, and test sets. **TIS Corpora** (E2) [28]: The dataset comprises 1152 utterances from 96 speakers in English across diverse demographics, including younger (18–45) and older (60+) adults from White, Black, and South Asian backgrounds. It was designed to support inclusive speech technology by incorporating speakers from multiple age groups and racial identities.

Neural Audio Codecs: We follow previous works on CF detection works [21, 20] for selection of NACs and employ publicly released, reproducible NACs that reflect current SOTA generation pipelines. **Descript Audio Codec (DAC)** [32]: We utilize DAC models operating at sampling rates of 16 kHz, 24 kHz, and 44 kHz. **EnCodec** [33]: We adopt the 24 kHz and 48 kHz variants in our experiments. **SoundStream** [34]: The 16 kHz configuration is employed. **Speech Tokenizer** [35]: We use the 16 kHz model. **FunCodec** [36]: The officially released 16 kHz version is incorporated. **AudioDec** [37]: Models at 28 kHz and 48 kHz are used. **SNAC** [38]: Multiple configurations at 24 kHz, 32 kHz, and 44 kHz are employed. **MIMI** [39]: We use the 24 kHz version. Considering all configurations, a total of fourteen NAC variants are used in this work.

[1] https://anonymous.4open.science/w/ElderlyCodecFake-BF43/

ECF Dataset Generation Process: To build the ECF dataset, we adopt a structured pipeline inspired by following Wu et al. [20] and Lu et al. [21], one of the foundational CF detection work. The procedure converts real elderly speech into NAC-generated counterparts using multiple NACs that constitute the core of modern ALM systems. We start from publicly available elderly speech datasets described above, where each original recording is treated as a real reference sample. For creating synthetic counterparts, every input speech sample undergoes a NAC encoding–decoding process. First, it is is transformed into a discrete latent sequence through the pre-trained encoder of a NAC, and this representation is subsequently reconstructed using the corresponding decoder. The resulting synthetic speech preserves linguistic content and speaker identity but contains NAC-induced distortions unique to each individual NAC, yielding realistic yet synthetic elderly speech. This procedure is applied across all fourteen NAC variants, producing parallel data such that each real speech utterance has a one-to-one synthetic counterpart for every codec type. For the SeniorTalk dataset, we follow the official data split. CF samples for the training and validation sets are generated using SNAC, DAC, EnCodec, SoundStream, SpeechTokenizer, and FunCodec (including their variants). For the test set, CF samples are generated using AudioDec, SNAC, and Mimi (including their variants). The TIS corpus contains speech samples from twelve elderly speakers. As no official split is provided, we perform a speaker-independent partition: eight speakers for training, two for validation, and two for testing. CF samples for all splits of TIS corpus are generated following the same procedure used for SeniorTalk. SeniorTalk and TIS contain 60029 and 720 real elderly utterances, respectively. In total, this results in 60749 real elderly speech utterances across both datasets. After generating CF data using fourteen NAC variants, the resulting dataset comprises 850486 elderly CF speech samples.

3. Methodology

This section presents the FMs employed in our study, followed by the downstream modeling approaches. We then detail the proposed framework, **BONSAI**.

3.1. Foundation Models

The FMs considered are SOTA in their respective benchmarks.

Mutlimodal FMs: We select LanguageBind (LB) [40] and ImageBind (IB) [41] as multimodal FMs. IB maps diverse modalities (image, audio, text, IMU, depth, thermal) into a shared image-centric embedding space via an InfoNCE objective, achieving strong cross-modal generalization without explicit paired supervision. LB similarly aligns modalities such as video, depth, audio, and infrared to a frozen language encoder using contrastive learning.

Speech FMs: We employ Wav2vec2 [42], WavLM [43], and Whisper [44] as SOTA speech FMs. Wav2vec2 is included due to its demonstrated effectiveness in CF detection, particularly when combined with AASIST as a downstream model [21]. We additionally incorporate WavLM, which has achieved strong performance across multiple speech processing tasks in the SUPERB benchmark and in speech deepfake detection [15]. Finally, we consider Whisper, a multilingual FM trained on 96 languages, unlike Wav2vec2 and WavLM which are primarily English-centric. Whisper has recently shown leading performance in speech deepfake detection [30]. Wav2vec2 is a self-supervised model trained using a contrastive learning objective, while WavLM is also

self-supervised but optimized through masked prediction and speech denoising tasks. In contrast, Whisper is pre-trained using a supervised multi-task learning framework. For all three FMs, we utilize their base variants.

All audio samples are resampled to 16 kHz prior to feature extraction to ensure compatibility across FMs. We extract representations from the last hidden layer of each frozen FM by applying average pooling. For multimodal FMs, only the audio encoder branch is utilized. The resulting embedding dimensionalities are 768 for LB, Wav2vec2, WavLM; 1024 for IB; and 512 for the Whisper encoder.

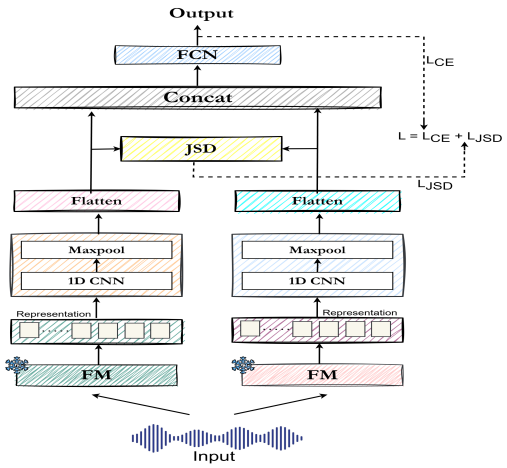


Figure 2: *Proposed Framework:* **BONSAI***; JSD stands for Jensen-Shannon Divergence*

3.2. Modeling

Here, we detail the downstream modeling with individual FMs followed by discussion of the novel proposed approach for FMs fusion, **BONSAI**.

Downstream Modeling with Individual FMs: For downstream modeling, we adopt two widely used architectures for speech deepfake detection: AASIST [11] and a convolutional neural network (CNN) [30]. AASIST is a graph neural network–based architecture. In contrast, the CNN model consists of a1D-convolutional layer followed by max-pooling layer with a fully connected layer as output classifier. For both AASIST and CNN, we strictly follow the architectural configurations and implementation details described in Lu et al. [21] and Phukan et al. [30], respectively, to ensure reproducibility and fair comparison.

BONSAI: The modeling architecture of **BONSAI** is illustrated in Figure 2. **BONSAI** leverages Jensen–Shannon Divergence (JSD) as a novel fusion mechanism for aligning FM representations. While JSD is primarily employed in multimodal learning for distribution matching and regularization [45, 46], we repurpose it here as a novel loss function to explicitly align heterogeneous FM representations in a unified representation space. JSD is a symmetric and bounded measure of similarity between probability distributions which quantifies how individual distributions deviate from their shared mean distribution. We employ JSD because different FMs produce representations with distinct distributional characteristics due to differences in their pretraining objectives. Direct feature concatenation or linear fusion may therefore fail to account for distributional mismatch. By minimizing JSD between projected FMs representational space, **BONSAI** explicitly encourages distributional alignment in a stable and symmetric manner. The detailed modeling flow of **BONSAI** is given below: Representations from each FM are first passed through 1D-CNN layer with filter size 3 and we use 32 filters followed by maxpooling identical to those used with the CNN model mentioned in Downstream Modeling with Individual FMs. Then, the resulting features are flattened and projected to a shared dimensional space to ensure dimensional consistency and reduce computational overhead. Let e_a and e_b denote the projected flattened representations from two FMs. These vectors are normalized using softmax to obtain probability distributions p and q. The JSD alignment loss is defined as: $\mathcal{L}_{JSD} = \frac{1}{2}KL(p \parallel m) + \frac{1}{2}KL(q \parallel m)$ where $m = \frac{1}{2}(p+q)$ and $KL(\cdot)$ denotes the Kullback-Leibler (KL) divergence. Minimizing $\mathcal{L}_{JSD}$ encourages the two FM representations to capture complementary yet consistent information. Following this, the features are passed to a fully connected network (FCN) consisting of a dense layer with 120 neurons, followed by an output layer with a softmax activation function. The final training objective jointly optimizes classification ($\mathcal{L}_{CE}$) and alignment: $\mathcal{L} = \lambda\mathcal{L}_{CE} + (1-\lambda)\mathcal{L}_{JSD}$ where λ controls the balance between cross-entropy supervision and JSD-alignment. The number of trainable parameters in **BONSAI** ranges from 3.8M to 4.02M, depending on the dimensionality of the underlying FM representations.

Model	E1	E2	
		Young	Elderly
AASIST	30.18	14.07	27.45
Wav2vec2-AASIST	**28.32**	**12.89**	**25.76**

Table 1: *EER (↓) in %. Trained on Lu et al. [21] and evaluated on ECFD test sets; E1: SeniorTalk, E2: TIS Corpus; The abbreviations are kept same for Table 2 and Table 3*

4. Experiments

4.1. Training Details

We train the models by combining the training sets from SeniorTalk and TIS, while validation and testing are performed separately on the respective validation and test splits of each individual dataset. All models are trained for 20 epochs using a learning rate of 1e-3 and a batch size of 32. We employ the Adam optimizer with cross-entropy loss for classification. For experiments involving **BONSAI**, the alignment weight λ is set to 0.65, selected based on preliminary validation experiments. Dropout is employed during training to reduce overfitting, while class weighting is applied to handle class imbalance.

4.2. Experimental Results

We use Equal Error Rate (EER) as the evaluation metric, following prior work on speech deepfake detection and CF detection [30, 20, 21].

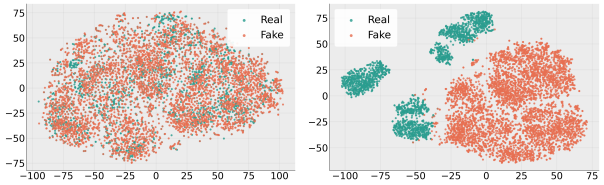
Training on previous benchmark CF dataset [21] and testing on ECFD (zero-shot): We train AASIST [20] and Wav2Vec2 with AASIST as the downstream classifier (Wav2Vec2-AASIST) [21] on the CF dataset introduced by Lu et al. [21], which contains samples in both English and Chinese. These models represent SOTA CF detection approaches and are subsequently evaluated on the ECFD test set. The results are presented in Table 1. Furthermore, to analyze potential age-related performance bias, we generate CF samples for the

younger speech subset of the E2: TIS corpus using the same neural audio codecs (NACs) employed to construct the ECFD test set. This enables us to assess whether models trained on the Lu et al. [21] CF dataset exhibit differential performance when evaluated on younger versus elderly speech samples. The results reveal clear cross-age performance degradation: the models achieve substantially lower EER on the younger subset, while the EER nearly doubles on the elderly subset. Notably, although the younger speech subset also constitutes an out-of-distribution condition relative to the training data, the models retain comparatively strong performance on younger speech than on elderly speech. This indicates that the observed degradation cannot be attributed solely to distribution mismatch, but is further attributed towards by age-related acoustic differences. Among the evaluated models, Wav2Vec2-AASIST demonstrates superior performance, likely due to its FM backbone, which provides better representations for CF detection. We also

Model	E1	E2 (Elderly)	Avg
End to End: AASIST			
AASIST	14.54	13.66	14.10
Downstream: AASIST			
Wav2vec2	11.76	11.02	11.39
WavLM	11.34	10.66	11.00
Whisper	10.12	9.86	9.99
IB	6.53	5.79	6.16
LB	**6.48**	**5.21**	**5.85**
Downstream: CNN			
Wav2vec2	11.02	10.29	10.66
WavLM	10.67	9.13	9.90
Whisper	8.46	8.14	8.30
IB	**5.41**	**5.26**	**5.34**
LB	**4.81**	**4.30**	**4.56**

Table 2: *Evaluation scores (EER (↓) in %) for training and evaluation on ECFD; Avg represents the average of EER across E1 and E2 (Elderly)*

In-domain Training and Evaluation on ECFD: Table 2 reports the in-domain results on the ECFD dataset. We first evaluate the AASIST baseline [20], as it demonstrated better performance than prior baselines in Table 1. We then evaluate different FMs using both AASIST and CNN as downstream classifiers. The results show that multimodal FMs consistently outperform speech-only FMs, thereby supporting our hypothesis. To further analyze this behavior, we perform a qualitative analysis using t-SNE visualizations of the raw FM representations from LB and Wav2Vec2. The t-SNE plots reveal clearer separation and better clustering between real and fake classes for LB compared to Wav2Vec2. Furthermore, the CNN achieves stronger performance than AASIST while being more lightweight, whereas AASIST appears more prone to overfitting in this setting. In Table 3, we investigate whether combining FMs can further improve ECFD performance. We use simple concatenation as a baseline, employing the same architecture and training protocol as **BONSAI**, except without the JSD-based alignment loss. Overall, **BONSAI** consistently outperforms concatenation across all FM pairs, demonstrating that explicit representation alignment through JSD enables more effective exploitation of complementary information. The performance gains are modest for speech-only FM pairs but become more pronounced when fusing multimodal FMs or combining speech and multimodal FMs. Notably, combinations involving LB and IB achieve the strongest performance, indicating that multimodal FMs provide highly complementary representations for ECFD. We also evaluated KL divergence as an alternative alignment objective, given that JSD is derived from KL divergence. However, KL divergence showed less stable convergence and achieved inferior performance than **BONSAI**, with results comparable to simple concatenation. Consequently, we present concatenation as the baseline and omit KL divergence results due to space constraints.



(a) (b)

Figure 3: *t-SNE Plots (a) Wav2vec2 (b) LanguageBind*

Pairs	Concatenation			BONSAI		
	E1	E2 (Elderly)	Avg	E1	E2 (Elderly)	Avg
W2v2 + WL	9.36	9.01	9.18	7.54	6.97	7.26
W2v2 + Wh	7.17	6.63	6.90	6.24	5.64	5.94
W2v2 + IB	5.20	4.79	5.00	4.87	3.22	4.05
W2v2 + LB	**4.47**	4.22	4.35	4.00	3.76	3.88
WL + Wh	7.94	7.35	7.65	5.68	5.17	5.43
WL + IB	5.12	4.72	4.92	4.36	4.09	4.23
WL + LB	4.86	4.43	4.65	3.92	3.56	3.74
Wh + IB	4.51	**4.02**	**4.27**	**3.23**	**2.95**	**3.09**
Wh + LB	**3.78**	**3.14**	**3.46**	**2.78**	**2.36**	**2.57**
IB + LB	**3.01**	**2.50**	**2.76**	**1.80**	**1.51**	**1.66**

Table 3: *Evaluation scores (EER (↓) in %) for training and evaluation on ECFD; Avg represents the average of EER across E1 and E2 (Elderly); W2v2: Wav2vec2, WL: WavLM; Wh: Whisper*

Comparison to SOTA: AASIST and Wav2vec2-AASIST represent SOTA architectures for CF detection [21, 20]. From Table 1 and Table 3, we observe that **BONSAI** with the fusion of LB and IB achieves the lowest average EER of 1.66% (1.80% on E1 and 1.51% on E2 (Elderly)), thereby establishing a new SOTA for the ECFD task.

5. Conclusion

In this work, we introduced the ECFD task and the ECF dataset comprising English and Chinese speech. Our analysis demonstrated that existing SOTA CF detection models trained on prior benchmark datasets generalize poorly to elderly speech, revealing a critical robustness gap. Furthermore, we showed that multimodal FMs, such as LB and IB, provide more effective representations for ECFD. Building on this insight and motivated by the complementary nature of FMs, we proposed **BONSAI**, a novel fusion framework that leverages Jensen–Shannon Divergence for representation alignment. Experimental results demonstrated that **BONSAI** with LB and IB achieves an average EER of 1.66%, outperforming individual FMs and competitive baselines, and establishing a new benchmark for the ECFD task. In future work, we plan to extend the ECF dataset to additional languages to further improve the robustness and generalizability of ECFD systems.

6. Generative AI Use Disclosure

AI Assistants were utilized exclusively to enhance grammatical accuracy, clarity, and the overall readability of the manuscript. These tools did not contribute to the development of scientific concepts, data analysis, generation of results, or interpretation of findings. The authors assume full responsibility for the accuracy and integrity of the work.

7. References

[1] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *INTERSPEECH 2015 16th Annual Conference of the International Speech Communication Association*. International Speech Communication Association, 2015, pp. 2037–2041.

[2] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," 2017.

[3] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[4] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch *et al.*, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2023.

[5] X. Wang, H. Delgado, H. Tak, J.-w. Jung, H.-j. Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T. Kinnunen *et al.*, "Asvspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale," *arXiv preprint arXiv:2408.08739*, 2024.

[6] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech." in *Interspeech*, 2015, pp. 2062–2066.

[7] ——, "Cochlear filter and instantaneous frequency based features for spoofed speech detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 618–631, 2016.

[8] B. Balamurali, K. E. Lin, S. Lui, J.-M. Chen, and D. Herremans, "Toward robust audio spoofing detection: A detailed comparison of traditional and learned features," *IEEE Access*, vol. 7, pp. 84 229–84 241, 2019.

[9] X. Tian, X. Xiao, E. S. Chng, and H. Li, "Spoofing speech detection using temporal convolutional neural network," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2016, pp. 1–6.

[10] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A light convolutional gru-rnn deep feature extractor for asv spoofing detection," in *Proc. Interspeech*, vol. 2019, 2019, pp. 1068–1072.

[11] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2022, pp. 6367–6371.

[12] J. M. Martín-Doñas and A. Álvarez, "The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9241–9245.

[13] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," in *The Speaker and Language Recognition Workshop (Odyssey 2022)*. ISCA, 2022.

[14] P. Kawa, M. Plata, M. Czuba, P. Szymański, and P. Syga, "Improved DeepFake Detection Using Whisper Features," in *Interspeech 2023*, 2023, pp. 4009–4013.

[15] Y. Guo, H. Huang, X. Chen, H. Zhao, and Y. Wang, "Audio deepfake detection with self-supervised wavlm and multi-fusion attentive classifier," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 702–12 706.

[16] A. Pimentel, Y. Zhu, H. R. Guimarães, and T. H. Falk, "Efficient audio deepfake detection using wavlm with early exiting," in *2024 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2024, pp. 1–6.

[17] H. M. Tran, D. Lolive, D. Guennec, A. Sini, A. Delhay, and P.-F. Marteau, "Leveraging SSL Speech Features and Mamba for Enhanced DeepFake Detection," in *Interspeech 2025*, 2025, pp. 5323–5327.

[18] H. M. Tran, D. Lolive, A. Sini, A. Delhay, P.-F. Marteau, and D. Guennec, "Multi-level ssl feature gating for audio deepfake detection," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 11 766–11 775.

[19] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi *et al.*, "Audiolm: a language modeling approach to audio generation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 31, pp. 2523–2533, 2023.

[20] H. Wu, Y. Tseng, and H. yi Lee, "Codecfake: Enhancing anti-spoofing models against deepfake audios from codec-based speech synthesis systems," in *Interspeech 2024*, 2024, pp. 1770–1774.

[21] Y. Lu, Y. Xie, R. Fu, Z. Wen, J. Tao, Z. Wang, X. Qi, X. Liu, Y. Li, Y. Liu, X. Wang, and S. Shi, "Codecfake: An initial dataset for detecting llm-based deepfake audio," in *Interspeech 2024*, 2024, pp. 1390–1394.

[22] X. Chen, J. Du, H. Wu, L. Zhang, I. Lin, I. Chiu, W. Ren, Y. Tseng, Y. Tsao, J.-S. R. Jang *et al.*, "Codecfake+: A large-scale neural audio codec-based deepfake speech dataset," *arXiv preprint arXiv:2501.08238*, 2025.

[23] J. Cui, B. Yu, Q. Wang, F. Meng, and J. Lu, "Whiadd: Semantic-acoustic fusion for robust audio deepfake detection," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 11 610–11 618.

[24] Y. Xie, Y. Lu, R. Fu, Z. Wen, Z. Wang, J. Tao, X. Qi, X. Wang, Y. Liu, H. Cheng *et al.*, "The codecfake dataset and countermeasures for the universally detection of deepfake audio," *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[25] S. Rojas, E. Kefalianos, and A. Vogel, "How does our voice change as we age? a systematic review and meta-analysis of acoustic and perceptual voice data from healthy adults over 50 years of age," *Journal of Speech, Language, and Hearing Research*, vol. 63, no. 2, pp. 533–551, 2020.

[26] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, H. Baumeister, A. D. MacIntyre, and S. Hantke, "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing Masks," in *Interspeech 2020*, 2020, pp. 2042–2046.

[27] G. Soğancıoğlu, O. Verkholyak, H. Kaya, D. Fedotov, T. Cadée, A. A. Salah, and A. Karpov, "Is Everything Fine, Grandma? Acoustic and Linguistic Modeling for Robust Elderly Speech Emotion Recognition," in *Interspeech 2020*, 2020, pp. 2097–2101.

[28] C. Maltezou-Papastylianou, R. Scherer, and S. Paulmann, "Human voices communicating trustworthy intent: A demographically diverse speech audio dataset," *Scientific Data*, vol. 12, no. 1, p. 921, 2025.

[29] O. C. Phukan, M. M. Akhtar, S. R. Behera, S. Kalita, A. B. Buduru, R. Sharma, S. M. Prasanna *et al.*, "Strong alone, stronger together: Synergizing modality-binding foundation models with optimal transport for non-verbal emotion recognition," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[30] O. C. Phukan, G. S. Kashyap, A. B. Buduru, and R. Sharma, "Heterogeneity over homogeneity: Investigating multilingual speech pre-trained models for detecting audio deepfake," *arXiv preprint arXiv:2404.00809*, 2024.

[31] Y. Chen, H. Wang, shiyao wang, J. Chen, J. He, J. Zhou, X. Yang, Y. Wang, Y. Lin, and Y. Qin, "Seniortalk: A chinese conversation dataset with rich annotations for super-aged seniors," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. [Online]. Available: https://openreview.net/forum?id=QzWPUEuMKU

[32] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved rvqgan," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[33] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[34] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[35] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, "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

[36] Z. Du, S. Zhang, K. Hu, and S. Zheng, "Funcodec: A fundamental, reproducible and integrable open-source toolkit for neural speech codec," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 591–595.

[37] Y.-C. Wu, I. D. Gebru, D. Marković, and A. Richard, "Audiodec: An open-source streaming high-fidelity neural audio codec," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[38] H. Siuzdak, F. Grötschla, and L. A. Lanzendörfer, "Snac: Multi-scale neural audio codec," in *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024.

[39] A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave, and N. Zeghidour, "Moshi: a speech-text foundation model for real-time dialogue," *arXiv preprint arXiv:2410.00037*, 2024.

[40] B. Zhu, B. Lin, M. Ning, Y. Yan, J. Cui, W. HongFa, Y. Pang, W. Jiang, J. Zhang, Z. Li, C. W. Zhang, Z. Li, W. Liu, and L. Yuan, "Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=QmZKc7UZCy

[41] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 15 180–15 190.

[42] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "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.

[43] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "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.

[44] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[45] F. Xiao, "Multi-sensor data fusion based on a generalised belief divergence measure," *arXiv preprint arXiv:1806.01563*, 2018.

[46] T. Sutter, I. Daunhawer, and J. Vogt, "Multimodal generative learning utilizing jensen-shannon-divergence," *Advances in neural information processing systems*, vol. 33, pp. 6100–6110, 2020.