Interspeech 2024
1-5 September 2024, Kos, Greece

Speech Formants Integration for Generalized Detection of Synthetic Speech Spoofing Attacks

Kexu Liu[1], *Yuanxin Wang*[1], *Shengchen Li*[2], *Xi Shao*[1]

[1]Nanjing University of Posts and Telecommunications, Nanjing, China
[2]School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China

kexu.liu.03@gmail.com, b21030803@njupt.edu.cn, Shengchen.li@xjtlu.edu.cn, shaoxi@njupt.edu.cn

Abstract

Existing synthetic speech detection systems struggle with high variance of performance in different spoofing attacks. This is due to the diversity of unseen synthesis algorithms, making it challenging for the system to generalize unseen spoofing attacks. To address this, we propose multi-view features with one-class learning for synthetic speech detection. The key idea is to capture *bona-fide* speech features from dynamic information of formants and XLS-R dimensions, aiming to compactly represent *bona-fide* speech in the embedding space without the need to fit various unseen spoofing attacks. To leverage multi-view features, the dynamic information of formants is integrated with XLS-R features using a parallel attention mechanism and gating modulation. Our system achieves an equal error rate (EER) of 0.39% in the ASVspoof 2019 logical access scenario, demonstrating a low performance variance of 0.069 across all 13 attacks, outperforming most mainstream single-systems.

Index Terms: ASVspoof, multi-view features, formants dynamic information, one-class learning

1. Introduction

Automatic speaker verification is a key biometric technology [1]. Speech spoofing detection enhances the reliability of speaker verification by distinguishing between genuine (*bona-fide*) and spoofed speech inputs. Recently, artificial intelligence advancements have improved speech spoofing techniques, such as text-to-speech (TTS) and voice conversion (VC) [2], [3]. To enhance the anti-spoofing speaker verification research, ASVspoof Challenges were successfully organised in 2015 [4], 2017 [5], 2019 [6], and 2021 [7].

Existing works focus on exploring the cues used to distinguish genuine speech from fake speech, especially synthetic traces of fake speech. References [8, 9, 10, 11] indicate that distinguishing features (i.e., spoofing artefacts) can be found across both spectral and temporal dimensions. Huang[12] employs the high-frequency information of speech signals to boost generalization ability against unknown attacks. Jung[13] utilizes a heterogeneous attention mechanism and a stacked node to model spoofing artefacts. However, spoofing artefacts often depend on the particular type of synthesis algorithm. In practical scenarios, the rapid advancement of artificial intelligence (AI) technology has introduced a lot of new and unseen synthesis algorithms. Consequently, fully detecting the range of diverse and unseen spoofing artefacts becomes challenging. This situation can lead to noticeable disparities in detection capabilities across different spoofing attacks.

To address these challenges, this work focuses on the feature representation of genuine speech, thereby avoiding the tricky task of exploring unknown fake speech artefacts. Moreover, it is well known that representations based on single-dimensional features often suffer from poor generalization. Building on this understanding, our approach adopts a multi-view feature construction methodology aimed at crafting features of genuine speech. This method combines two types of features: non-semantic features derived from a self-regression model and semantic features, with a specific emphasis on formants. First, for the non-semantic features, we utilize the XLS-R model [14], a large-scale model for cross-lingual speech representation learning based on wav2vec 2.0 [15]. The selection of the XLS-R model is predicated on its extensive training with genuine speech data, enabling effective and robust characterization of genuine speech. Its superior performance across various speech processing tasks [14] further underscores its aptness for our application. Secondly, the selection of formants as semantic features in the analysis of genuine speech is driven by their capacity to reflect the resonances of the vocal tract during genuine speech production [16]. This ability is crucial in speech synthesis models to simulate natural speech [17], serving our goal to accurately represent genuine speech characteristics.

Given the challenge of precisely estimating formant frequencies, inspired by [18], we employ the Gabor transform to extract dynamic information of formants from the speech formant trajectories. To integrate multi-view information, we introduce a feature fusion technique employing parallel attention

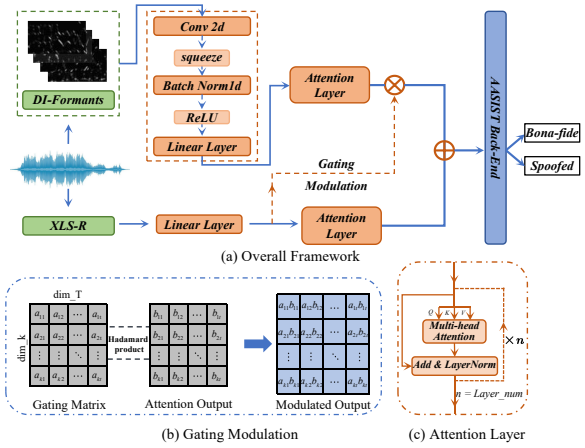


Figure 1: *(a) Overall framework of our method. The system consists of the features extraction module, the features fusion module and the AASIST Back-End. ⊗ denotes Hadamard Product in Equation (3). ⊕ denotes the operation in Equation (4). (b) Gating modulation module. (c) Attention layer module.*

10.21437/Interspeech.2024-292

streams and gating modulation for better integration of non-semantic and semantic features. Building on the multi-view feature construction of genuine speech, we adopt the one-class softmax (OC-Softmax) loss function [19]. This approach employs one-class learning to not only compact the representation of *bona-fide* speech (i.e., target class) but also ensure that spoofed speech (i.e., non-target class) is positioned outside the boundary of *bona-fide* speech, inherently generalizing to unknown spoofing attacks.

The experimental results on ASVspoof 2019 logical access (LA) dataset show that our system achieves significant and consistent improvement without the use of data augmentation as compared to most single-systems. Based on the results, we notice that *bona-fide* speech forms a compact distribution in the embedding space, in contrast to the dispersed distribution of spoofed speech. This supports our hypothesis that genuine speech has unique and identifiable features. Therefore, by focusing on simulating *bona-fide* speech rather than the specifics of unknown spoofing attacks, we lay a robust foundation for distinguishing between *bona-fide* and spoofed speech.

The rest of the paper is organized as follows: Section 2 describes our proposed method in detail. Sections 3 and 4 cover experimental design, results, and comparative analysis with other known single-systems. Section 5 concludes the study.

2. The Proposed Method

This paper introduces a novel spoofed speech detection method that combines semantic features with non-semantic features within a three-tier system architecture. Initially, we separately extract dynamic information of formants (DI-Formants) and XLS-R features, the latter sourced from a self-supervised pre-trained model. Subsequently, we construct multi-view features through parallel attention streams enhanced by gating modulation. In the final stage, the AASIST [13] classifier is employed as the back-end model. Our system architecture is detailed in Figure 1 and further explained in subsequent subsections.

2.1. Proposed Features Extraction

XLS-R Features: XLS-R [14] , a variant of Wav2vec2.0 [15], is a large-scale model for cross-lingual speech representation learning. XLS-R employs quantization to transform raw audio into discrete tokens, which are processed by a transformer network. The network learns contextual representations of these tokens, capturing subtle nuances in speech. The XLS-R based features have achieved top results in the Audio Deep Synthesis Detection (ADD) 2022 challenge [20].

In this paper, we utilize the XLS-R (300M) model.[1] Feature representation with 1024 dimensions is extracted from the XLS-R model following the example provided in the Fairseq toolkit [21]. Notably, we do not fine-tune the pre-trained model to ensure that the feature representation remains general and not overly adapted to specific known attack algorithms.

Dynamic Information of Formants (DI-Formants): This paper employs DI-Formants as supplementary genuine speech features in spoofed speech detection. Inspired by [18], DI-Formants is obtained by applying a Gabor transform to time-frequency maps with formant trajectories, reducing reliance on precise numerical estimation of formants. Initially, we utilize the Time-Varying Quasi-Closed Phase (TVQCP) analysis technique [22] for tracking formants trajectories, enabling estimation of the first three formants in speech signals. The technical details are as follows.

[1] https://huggingface.co/facebook/wav2vec2-xls-r-300m

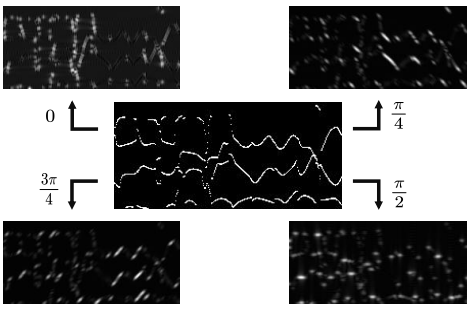


Figure 2: *Illustration of DI-Formants features extraction using 'LA_T_5769708.flac' as an example. The central time-frequency plot delineates the initial trajectories of the first three formants. Directional arrows guide to the resulting images from the Gabor transformation, with theta values in Gabor filtering denoted by* 0, $\frac{\pi}{2}$, $\frac{\pi}{4}$, *and* $\frac{3\pi}{4}$.

In our research, the time-frequency maps are binarized to accentuate the formants trajectories. Subsequently, these binary maps are convolved with a set of Gabor filters to produce Gabor feature images in four orientations: 0, $\frac{\pi}{4}$, $\frac{\pi}{2}$, and $\frac{3\pi}{4}$. These images effectively capture the dynamic information of formants in various orientations within speech signals. The number of these orientations determines the channel count for the input feature maps in our neural network. The 2D Gabor filter can be mathematically represented as:

$$G(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \cos\left(2\pi \frac{x'}{\lambda} + \psi\right) \tag{1}$$

where $x' = x\cos(\theta) + y\sin(\theta)$, $y' = -x\sin(\theta) + y\cos(\theta)$, and the parameters are defined as follows: λ is the wavelength of the sinusoidal factor, θ is the orientation of the normal to the parallel stripes of a Gabor function, ψ is the phase offset, σ is the standard deviation of the Gaussian envelope, and γ is the spatial aspect ratio. In this work, the parameters are set as follows: $\lambda = 4$, $\theta = [0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}]$, $\psi = 0$, $\sigma = \frac{\lambda}{2}$, and $\gamma = 0.5$.

As shown in Figure 2, a Gabor filter generates a strong response at specific pixel locations when its orientation aligns with the trajectory direction, and a weaker response when misaligned. Ultimately, these results form a four-channel image as input to the model.

2.2. Fusion Strategies for Features

Existing Feature Fusion Strategies: In the realm of speech processing, the fusion of features plays a pivotal role in enhancing system performance. Traditional fusion strategies like concatenation and cross-attention mechanisms have laid the groundwork [21]. Concatenation simply combines feature sets, offering a straightforward yet often effective approach. Cross-attention mechanisms, on the other hand, provide a more dynamic integration, focusing on the interplay and relevance of features from different sources.

Proposed Feature Fusion Strategy: With the aim of constructing multi-view features of genuine speech, this paper incorporates parallel attention streams, a gating modulation in the attention layer output of DI-Formants features, and learnable weights (PGL-Attention) for feature fusion. First, DI-Formants

and XLS-R features are processed separately using multi-head attention mechanisms within the encoder part of Transformer [23]. To achieve this, DI-Formants features are initially processed by a convolutional layer for channel alignment, followed by a linear transformation to match the dimension to d_k. Likewise, XLS-R features are adjusted via a linear layer to align with the d_k dimensional space. Lastly, the multi-head attention mechanism separately applies scaled dot-product attention to h distinct (Q, K, V) representations, merges these through concatenation, and then projects the result via a feed-forward layer. The detailed operation of each attention head is as follows:

$$Attention\_head_i(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \quad (2)$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ represent query, key, and value matrices, respectively, and d_k is the dimension of the keys.

Self-supervised features typically possess richer speech information and greater noise robustness. Given the susceptibility of formants information to noise interference, it is enhanced by the joint modeling of self-supervised features and formants. Central to this enhancement is a gating mechanism that modulates the attention output of the DI-Formants. This modulation is shaped by XLS-R features processed through a linear layer. It can be defined as:

$$Attention_{mod} = \sigma(Attention_1) \odot G, \quad G = WF + b \quad (3)$$

where $\odot$ represents the Hadamard Product, σ denotes the sigmoid activation function, and F denotes the XLS-R features. $Attention_1$ refers to the initial attention output derived from DI-Formants, while $\boldsymbol{G}$ represents gate value matrices, modulating the attention with the influence of XLS-R features. In this context, $\boldsymbol{W}$ is a learnable matrix, and b is the bias.

Lastly, we introduce learnable weights to merge the modulated DI-Formants and XLS-R features, optimizing these weights during training to balance the contributions of the two features:

$$F_{fused} = w_1 \cdot Attention_{mod} + w_2 \cdot Attention_{XLS-R} \quad (4)$$

where w_1 and w_2 are learnable weights bounded within [0, 1]. The architecture of our method is shown in Figure 1.

2.3. AASIST Back-End

Our classification module adapts the architecture of AASIST [13], known for its performance in audio anti-spoofing, and incorporates a novel heterogeneous stacking graph attention layer (HS-GAL) to model artefacts spanning heterogeneous temporal and spectral domains. In line with common practice, we have tailored the model to our specific needs. This customization involved replacing the sinc convolutional layer, a modification frequently adopted by predecessors to meet particular requirements. Our implementation closely follows the model architecture presented in [24].

3. Experiments

3.1. Dataset and Evaluation Metrics

A mainstream dataset, ASVspoof2019 LA [6], is used to evaluate the proposed method. This dataset is divided into three subsets: training, development, and evaluation. The evaluation set comprises 11 unseen attacks along with 2 known attacks (A16, A19). The details of this dataset are provided in Table 1.

Table 1: *Summary of the ASVspoof2019 LA dataset*

Datasets	Bona-fide	Spoofed	
	#utterance	#utterance	#attacks
Training	2580	22800	A01-A06
Development	2548	22296	A01-A06
Evaluation	7355	63882	A07-A19

This paper employs two evaluation metrics: the Equal Error Rate (EER) and the Minimum Tandem Detection Cost Function (min t-DCF) [25], which follows the impact of the Countermeasures (CM) on the reliability of the ASV system.

3.2. Training Details

Data Preparation and Loss Function: In this paper, we conduct experiments on DI-Formants features, XLS-R features, and multi-view features generated through fusion techniques. The dimension of the input features are reduced to 128. The input audio length for feature extraction is standardized to 4 seconds (64,000 samples) by either trimming or padding with repetition. The loss function employed is OC-Softmax, with hyperparameters consistent with those reported in [19].

Parameter Setting: To train our model, we utilize the Adam optimizer for updating the weights in the AASIST-backend model, setting the β_1 parameter to 0.9 and β_2 to 0.999. The epsilon is set to 10^{-8}, and the weight decay is 0.0005. Our attention mechanism consists of 3 attention layers with 8 heads each. In parallel, the parameters within the loss function are optimized using the Stochastic Gradient Descent (SGD) optimizer. The batch size for the model is set at 32, with an initial learning rate of 0.0005, which is programmed to decay by 50% every 10 epochs. All model training does not employ any form of data augmentation. All models are trained for 100 epochs on a single NVIDIA A30 GPU and then the model with the lowest validation EER is selected for evaluation. Wang[26] found that spoofing detection performance changes significantly with different random seeds. The results in this paper are averages from three runs using different random seeds.

4. Results and Discussion

4.1. Ablation Study

4.1.1. Complementary Patterns of Features

An ablation study is performed to investigate the interplay and complementary nature of non-semantic and semantic features. Specifically, DI-Formants, XLS-R features, and their fusion are tested under the same conditions. As shown in Table 2, DI-

Table 2: *Performance on the evaluation set of the ASVspoof 2019 LA Scenario. A denotes DI-Formants features and B denotes XLS-R features. M_1 denotes EER and M_2 denotes min t-DCF. When utilizing a single feature as input, none of the fusion methods are employed.*

Feature	Concatenating		Cross-Attention		**PGL-Attention**	
	M_1 (%)	M_2	M_1 (%)	M_2	M_1 (%)	M_2
A	25.08	0.7138	-	-	-	-
B	1.41	0.0378	-	-	-	-
A+B	1.01	0.0339	0.77	0.0229	**0.39**	**0.0117**

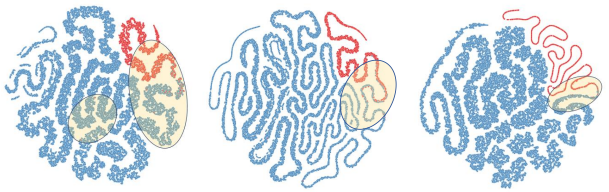
Table 3: *Performance Analysis on ASVspoof 2019 LA: Breakdown of EER (%) across 13 attacks, with pooled minimum t-DCF (P_1) and pooled EER (%) (P_2). M_1 represents the variance in performance across all 13 attack scenarios within the ASVspoof 2019 LA evaluation set, indicating the stability of model performance against different types of attacks. M_2 denotes the p-value from statistical significance tests (Wilcoxon signed-rank test) comparing the performance of the PGL-Attention method against other methods across the 13 attacks, assessing the significance of performance improvements achieved by the PGL-Attention method. The best performance figures for each column are highlighted in boldface. Asterisked (*) algorithms target known attacks (A16, A19).*

System	A07	A08	A09	A10	A11	A12	A13	A14	A15	A17	A18	A16*A19*	M_1	M_2	P_1	$P_2(\%)$
PGL-Attention	**0.06**	0.29	0.04	**0.63**	0.22	**0.12**	0.05	**0.07**	**0.18**	0.51	**0.84**	**0.12** **0.53**	**0.069**	**–**	**0.0117**	**0.39**
w/o gating	0.18	0.29	0.04	1.18	0.83	0.15	0.06	0.11	0.23	1.05	0.91	**0.12** 0.82	0.181	0.005	0.0175	0.56
Cross-Attention	0.12	0.42	0.05	1.81	0.63	0.35	**0.01**	0.36	0.77	0.85	1.28	0.24 0.67	0.265	0.001	0.0229	0.77
Concatenating	0.24	0.65	0.14	7.81	0.79	0.55	0.05	0.42	0.93	**0.48**	0.89	0.34 0.62	4.175	0.003	0.0339	1.01
AASIST [13]	0.80	0.44	**0.00**	1.06	0.31	0.91	0.10	0.14	0.65	1.52	3.40	0.72 0.62	0.778	0.001	0.0275	0.83
RawGAT-ST [27]	1.19	0.33	0.03	1.54	0.41	1.54	0.14	0.14	1.03	1.44	3.22	0.67 0.62	0.764	0.001	0.0333	1.19
Zhang *et al.* [19]	0.12	**0.18**	0.12	1.14	**0.12**	0.47	0.22	0.69	1.40	9.22	0.90	0.33 0.90	5.966	0.008	0.0590	2.19

Formants do not yield the anticipated outcomes, possibly attributed to its constrained views. Conversely, the XLS-R features demonstrate significantly superior performance, underscoring the robust representational capacity of self-supervised features. Lastly, the training-updated weights of 0.55 (w_1) and 0.68 (w_2) in Equation (4) demonstrate effective use of complementary information and improved performance.

4.1.2. Impacts of Feature Fusion Strategies

Comparison and visualization analysis: To evaluate the proposed fusion method, comparisons are made with concatenation and cross-attention techniques under identical settings. Table 3 reveals that PGL-Attention stands out, boosting performance and lowering variance against all attack types. It is important to highlight that the outlined methods encounter challenges when detecting the A18 algorithm. This difficulty can be attributed to the unique non-parallel voice conversion mechanism and the framework based on transfer learning in A18, which enable voice identity transformation effectively without requiring parallel data, thereby increasing the complexity of detection [28]. Moreover, a deeper analysis through t-distributed Stochastic Neighbor Embedding (t-SNE) visualizes embedding distributions from the evaluation set (Figure 3). The aim of multi-view features is to accurately represent genuine speech without overfitting to spoofed speech flaws, ensuring robustness to new spoofing attacks. The results of the PGL-Attention method align with our objectives, showing a dispersed spoofed speech distribution and a tightly clustered *bona-fide* speech distribution, with the smallest area of *bona-fide* and spoofed speech confusion among all methods. Compared to other baseline systems, our top-performing system achieves superior outcomes



(a) Concatenating (b) Cross-Attention (c) **PGL_Attention**

Figure 3: *Features embedding visualization of different fusion methods. Red: Bona-fide Speech. Blue: Spoofed Speech. Circled areas with positive and negative samples confusion.*

Table 4: *Comparative performance analysis of mainstream single-systems on the ASVspoof 2019 Evaluation Sets.*

System	# Param	EER (%)	min t-DCF
Ours	850K	**0.39**	**0.0117**
S^2pecNet [29]	1284k	0.77	0.0240
AASIST [13]	297K	0.83	0.0280
RawGAT-ST [27]	437K	1.06	0.0335
FFT-L-SENet [30]	1,100K	1.14	0.0368
Res-TSSDNet [31]	350K	1.64	0.0481

across multiple metrics (M_1, P_1 and P_2).

Significance Testing: To validate the superiority of our system, significance tests are conducted between our top-performing system on individual attacks and that of baseline systems. Compared to baseline systems, our top-performing system demonstrates a statistically significant advantage (p-value < 0.01).

Ablation study on gating modulation: We further investigate the impact of the gating modulation module which aim to the joint modeling of XLS-R features and DI-Formants. As shown in Table 3, the gating modulation module can improve detection performance by modulating the synergistic information.

4.2. Comparison with Other Systems

Table 4 showcases a performance comparison between our proposed method and various other recent single-systems designed for the ASVspoof 2019 LA dataset. This collection of systems spans a wide array of front-end representations and model architectures. The proposed method is the best performing of all.

5. Conclusion

In this paper, we propose a multi-view feature methodology that integrates semantic and non-semantic features into a cohesive framework, emphasizing the features of genuine speech. Our method integrates complementary features and one-class learning to capture genuine speech features and effectively cluster its distribution, avoiding fitting to spoofing artefacts. Experimental results show that our system outperforms most known single-systems without data augmentation. In particular, we demonstrate that the better fitting of the model to genuine speech distribution correlates with reduced variability in spoofed speech detection performance. In the future, addressing cross-domain genuine speech distribution fitting will be key to enhance generalization against spoofed speech in the proposed method.

6. Acknowledgements

This work is supported by the National Key Research and Development Project (No.2020AAA0106200), and the National Nature Science Foundation of China under Grants (No.61936005, No.62001038), and Gusu Innovation and Entrepreneurship Leading Talents Programme (ZXL2022472), and the project of Science and Technology Innovation Training Program of Nanjing University of Posts and Telecommunications (202310293037Z). We thank Yikang Wang and Xiaoyi Qin for their invaluable assistance and insightful discussions.

7. References

[1] K. Delac and M. Grgic, "A survey of biometric recognition methods," in *Proceedings. Elmar-2004. 46th International Symposium on Electronics in Marine*, 2004, pp. 184–193.

[2] Z. Guo, Y. Leng, Y. Wu, S. Zhao, and X. Tan, "Prompttts: Controllable text-to-speech with text descriptions," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[3] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2021.

[4] Z. Wu, T. H. Kinnunen, N. W. D. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. Interspeech*, 2015.

[5] T. H. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. W. D. Evans, J. Yamagishi, and K.-A. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Interspeech*, 2017.

[6] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. W. D. Evans, T. H. Kinnunen, and K.-A. Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," in *Pro. Interspeech*, 2019.

[7] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans *et al.*, "Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," *Proc. ASVspoof 2021 Workshop*, pp. 47–54, 2021.

[8] K. Sriskandaraja, V. Sethu, P. N. Le, and E. Ambikairajah, "Investigation of Sub-Band Discriminative Information Between Spoofed and Genuine Speech," in *Proc. Interspeech*, 2016, pp. 1710–1714.

[9] J. Yang, R. K. Das, and H. Li, "Significance of subband features for synthetic speech detection," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2160–2170, 2020.

[10] H. Tak, J. Patino, A. Nautsch, N. W. D. Evans, and M. Todisco, "An explainability study of the constant q cepstral coefficient spoofing countermeasure for automatic speaker verification," in *Odyssey 2020*. ISCA, 2020, pp. 333–340.

[11] H. Tak, J. Patino, A. Nautsch, N. Evans, and M. Todisco, "Spoofing attack detection using the non-linear fusion of sub-band classifiers," in *Proc. Interspeech*, 2020.

[12] B. Huang, S. Cui, J. Huang, and X. Kang, "Discriminative frequency information learning for end-to-end speech anti-spoofing," *IEEE Signal Processing Letters*, vol. 30, pp. 185–189, 2023.

[13] 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.

[14] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. M. Pino, A. Baevski, A. Conneau, and M. Auli, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," in *Proc. Interspeech*, 2021.

[15] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12 449–12 460.

[16] M. Kiefte, "Formants in speech perception," *Journal of the Acoustical Society of America*, vol. 140, pp. 3162–3162, 2016.

[17] P. P. Zarazaga, Z. Malisz, G. E. Henter, and L. Juvela, "Speaker-independent neural formant synthesis," *Proc. Interspeech*, 2023.

[18] X.-c. Lu, F.-p. Pan, J.-x. Yin, and W.-p. Hu, "A new formant feature and its application in mandarin vowel pronunciation quality assessment," *Journal of Central South University*, vol. 20, no. 12, pp. 3573–3581, 2013.

[19] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021.

[20] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, Wang *et al.*, "Add 2022: the first audio deep synthesis detection challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9216–9220.

[21] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, Jun. 2019, pp. 48–53.

[22] D. Gowda, S. R. Kadiri, B. Story, and P. Alku, "Time-varying quasi-closed-phase analysis for accurate formant tracking in speech signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1901–1914, 2020.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[24] H. Tak, M. Todisco, X. Wang, J. weon Jung, J. Yamagishi, and N. Evans, "Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 112–119.

[25] T. Kinnunen, H. Delgado, N. Evans, K. A. Lee, V. Vestman, A. Nautsch, M. Todisco, X. Wang, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2195–2210, 2020.

[26] X. Wang and J. Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," in *Interspeech*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:232306990

[27] H. Tak, J.-w. Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," *Proc. ASVspoof workshop*, 2021.

[28] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch *et al.*, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, 2020.

[29] P. Wen, K. Hu, W. Yue, S. Zhang, W. Zhou, and Z. Wang, "Robust Audio Anti-Spoofing with Fusion-Reconstruction Learning on Multi-Order Spectrograms," in *Proc. INTERSPEECH 2023*, 2023, pp. 271–275.

[30] Y. Zhang12, W. Wang12, and P. Zhang12, "The effect of silence and dual-band fusion in anti-spoofing system," in *Proc. Interspeech*, 2021.

[31] G. Hua, A. B. J. Teoh, and H. Zhang, "Towards end-to-end synthetic speech detection," *IEEE Signal Processing Letters*, vol. 28, pp. 1265–1269, 2021.