Deepfake Audio Detection Using Spectrogram-based Feature and Ensemble of Deep Learning Models

Lam Pham[1*], Phat Lam[2*], Truong Nguyen[3], Huyen Nguyen[4], Alexander Schindler[5]

Abstract— **In this paper, we propose a deep-learning-based system for the task of deepfake audio detection. This work is a part of the proposed toolchain for speech analysis in EUCINF (EUropean Cyber and INFormation) project, which is an European project with multiple partners in Europe. In particular, the raw input audio is first transformed into various spectrograms using three transformation methods of Short-time Fourier Transform (STFT), Constant-Q Transform (CQT), Wavelet Transform (WT) combined with different auditory-based filters of Mel, Gammatone, linear filters (LF), and discrete cosine transform (DCT). Given the spectrograms, we evaluate a wide range of classification models based on three deep learning approaches. The first approach is to train the spectrograms using our proposed baseline models of CNN-based model (CNN-baseline), RNN-based model (RNN-baseline), C-RNN model (C-RNN baseline). Meanwhile, the second approach is to apply the transfer learning from computer vision models such as ResNet-18, MobileNet-V3, EfficientNet-B0, DenseNet-121, SuffleNet-V2, Swint, Convnext-Tiny, GoogLeNet, MNASsnet, and Reg-Net. In the third approach, we leverage the state-of-the-art audio pre-trained models of Whisper, Seamless, Speechbrain, and Pyannote to extract audio embeddings from the input spectrograms. Then, the audio embeddings are explored by a Multilayer perceptron (MLP) model to detect fake or real audio samples. Finally, high-performance deep learning models from these approaches are fused to achieve the best performance. We evaluated our proposed models on ASVspoof 2019 benchmark dataset. Our best ensemble model achieved an Equal Error Rate (EER) of 0.03, which is highly competitive to top-performing systems in the ASVspoofing 2019 challenge. Experimental results also highlight the potential of selective spectrograms and deep learning approaches to enhance model performance on the task of audio deepfake detection.**

Items— **deepfake audio, spectrogram, feature extraction, classification model.**

I. INTRODUCTION

Sound-based applications represent a revolutionary paradigm in the rapidly evolving landscape of Internet of Sound (IoS) technology, where audio signals serve as the primary medium for data transmission, control, and interaction among interconnected devices [1], [2]. A voice-activated module in an IoS system, such as smart home devices, voice banking, home automation systems, and virtual assistants, relies on recognizing the user's voice to activate critical functions and generally involves confidential information. However, with the advancement of deep learning technologies, the emergence of spoofing speech attacks, commonly referred to as 'Deepfake', has become more prevalent. These attacks involve various AI-based speech synthesis techniques (e.g. Speech to Text [3], Voice Conversion [3], Scene Fake [4], Emotion Fake [5]), posing significant threats to the integrity and authenticity of voice-activated systems. Consequently, detecting audio deepfakes (DAD) has become a crucial area of research, drawing considerable attention from the research community. Several benchmark datasets and following challenges such as ASVspoof [6], Audio Deep synthesis Detection (ADD) [7] have been proposed, which facilitates the creation of various systems and techniques to handle this task.

Existing DAD research can be divided into two kinds: pipeline solutions (consisting of a front-end feature extractor and a back-end classifier) and end-to-end solutions [8]. The top-performing DAD systems proposed in the ASVspoof and ADD competitions are mainly score-level fusion systems [8]. However, there is a lack of comprehensive evaluation to present how individual spectrograms and classifiers affect overall performance, which is crucial for further research motivation and research direction. Other successful systems utilize deep features through supervised embedding methods, such as DNNs [9] and RNNs [10]. Despite their effectiveness, these embeddings are trained on specific datasets and may encounter overfitting issue and susceptibility to adversarial attacks. This reduces the model's ability to generalize to new, unseen data, particularly when the dataset is not sufficiently large or diverse. Meanwhile, other approaches that can manage generalization and domain adaptation, such as transfer learning and leveraging embeddings from large pre-trained audio models, have not been extensively explored. To tackle these mentioned limitations, we propose an ensemble of deep-learning-based models for the audio deepfake detection task, which is achieved via a comprehensive analysis in terms of multiple spectrogram-based features and deep-learning approaches. Our key contributions can be highlighted as:

- Evaluated the efficacy of different spectrograms in combination with auditory filters to the detection performance.
- Evaluated a wide range of architectures leveraging both transfer learning and end-to-end networks.
- Explored the performance of audio embeddings extracted from state-of-the-art pre-trained models (e.g. Whisper, Speechbrain, Pyannote) on deepfake detection.
- Proposed an ensemble model via selective spectrograms

L. Pham and A. Schindler are with Austrian Institute of Technology, Vienna, Austria

P. Lam and T. Nguyen are with HCM University of Technology, Ho Chi Minh city, Vietnam

H. Nguyen is with Tokyo University of Agriculture and Technology, Tokyo, Japan

(*) Main and equal contribution into the paper.

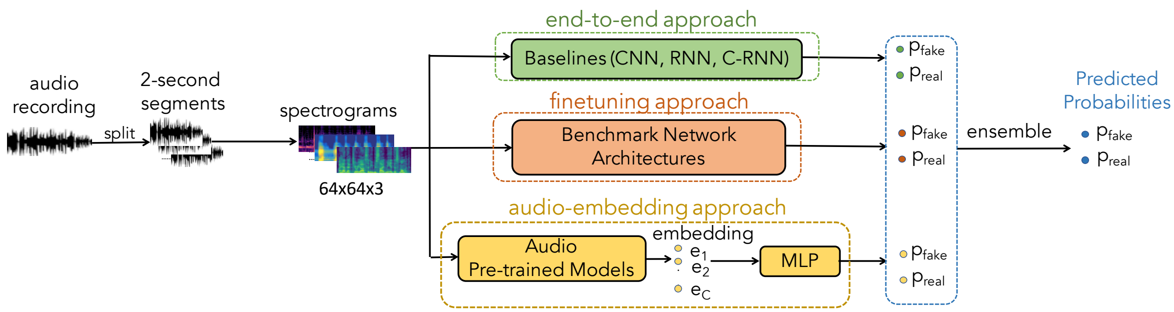

Fig. 1. The high-level architecture of deepfake audio detection system

and models from experiments that further improves deepfake audio detection system performance.

II. Proposed Deep-Learning-Based Systems

The high-level architecture of the proposed deep-learning-based system for audio deepfake detection, which is denoted in Fig. 1, comprises two main parts: front-end spectrogram-based feature extraction and back-end deep learning model for classification. In particular, the raw input audio recordings are first split into two-second segments. This segment length generally provides sufficient context to capture important features and allows faster training and inference for applications that require real-time detection. Next, the two-second audio segments are transformed into spectrograms. Finally, the spectrograms are explored by back-end deep learning models to detect real or fake audio segments.

There are three deep-learning-based approaches proposed in this paper. The first approach is shown in the upper part of Fig. 1. In this approach, referred to as the end-to-end approach, proposed models train spectrograms as input data. In the second approach as shown in the middle part in Fig. 1, referred to as the finetuning approach, we fine-tune benchmark network architectures, which are popularly used in the computer vision domain. Regarding the third approach as shown in the lower part in Fig. 1, we leverage state-of-the-art pre-trained models, which were trained on large audio datasets in advance. Then, we feed spectrograms input into these audio pre-trained models to obtain audio embeddings. A Multilayer Perceptron (MLP) is then proposed to train the audio embeddings as input for classifying into either real or fake classes. We refer this approach to as the audio-embedding approach. Finally, individual and high-performance models from three approaches are selected and fused to achieve the best performance.

A. Spectrogram-based feature extraction

Fig. 2 presents how 6 different spectrograms are generated in this paper. In particular, these spectrograms are generated from three transformation methods including Short-time Fourier Transform (STFT), Constant-Q Transform (CQT), and Wavelet Transform (WT). Inspired by [11], [12], each type of spectrogram focuses on different perspectives on frequency content and may capture artifacts in deepfake audio from different angles. The combination of these spectrograms allows classification models to learn a broader range of features and patterns, potentially improving its ability to generalize and detect deepfakes [11]. Additionally, we also establish different auditory-based filters: Mel and Gammatone filters focus on subtle variations relevant to human auditory perception and the linear filter (LF) isolates specific frequency bands. Incorporating these filters alongside pre-defined spectrograms enriches the available features and further enhances the detection system's robustness to variations. As we use the same settings of the window length, the hop length, and the filter number with 1024, 512, and 64 for all spectrograms, the generated spectrograms present the same tensor shape of $64{\times}64$. Then, DCT is applied to spectrograms across the temporal dimension. Finally, we apply the first and the second-order derivative to these spectrograms, generate three-dimensional tensor of $64{\times}64{\times}3$ (i.e. the original spectrogram, the first-order derivative, and the second-order derivative are concatenated across the third dimension).

B. End-to-end deep learning approach

Regarding the end-to-end deep learning approach, we propose three baseline models of CNN-based model, RNN-based model, and C-RNN-based model, which are referred to as the CNN baseline, RNN baseline, and C-RNN baseline, respectively. The detailed configuration of these baselines is presented in Table I. CNNs are the most common architecture for this task, which can effectively capture and learn spectral features within local frequency bands such as harmonic structures, formants, pitch variations, high-frequency artifacts, etc. On the other hand, RNNs focus on detecting natural sequential patterns that can be disrupted in synthetic audio [13] (e.g. temporal coherence, prosodic features such as rhythm, stress, and intonation). Consequently, the use of C-RNN baseline is based on the idea of combining both spectral features and temporal features for distinguishing characteristics of deepfake audio.

C. Transfer learning approach

Additionally, we also evaluate a wide range of benchmark network architectures in the computer vision domain such as ResNet-18, MobileNet-V3, EfficientNet-B0, DenseNet-121, SuffleNet-V2, Swint, Convnext-Tiny, GoogLeNet, MNASsnet, RegNet. In particular, these networks were trained on the ImageNet1K dataset [14] in advance. Their pre-trained weights can capture rich and generalized features of pattern recognition in images that can be potentially adapted to identifying patterns in spectrograms via parameter fine-tuning.

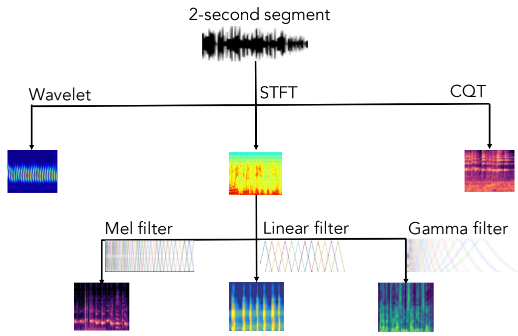

Fig. 2. Generate spectrograms using different spectrogram transformation methods and auditory filter models

TABLE I

THE CNN, RNN, AND C-RNN BASELINE NETWORK ARCHITECTURES

Models	Configuration
CNN baseline	**3** $\times$ {Conv(32/64/128)-ReLU-AP-Dropout(0.2)} **1** $\times$ {Dense(256)-ReLU-Dropout(0.2)} **1** $\times$ {Dense(2)-Softmax}
RNN baseline	**2** $\times$ {BiLSTM(128/64)-ReLU-Dropout(0.2)} **1** $\times$ {Dense(256)-ReLU-Dropout(0.2)} **1** $\times$ {Dense(2)-Softmax}
C-RNN baseline	**3** $\times$ {Conv(32/64/128)-ReLU-AP-Dropout(0.2)} **2** $\times$ {BiLSTM(128/64)-ReLU-Dropout(0.2)} **1** $\times$ {Dense(256)-ReLU-Dropout(0.2)} **1** $\times$ {Dense(2)-Softmax}

In this approach, the final dense layer of these mentioned networks is modified to match the binary classification task of deepfake audio detection.

D. Audio-embedding deep learning approach

In the audio-embedding deep learning approach, we leverage the state-of-the-art audio pre-trained models of Whisper [15], Seamless [16], Speechbrain [17], and Pyanote [18], [19]. These pre-trained models are utilized for their ability to capture robust and high-level feature representations of genuine speakers in practice such as pitch, tone, accent, and intonation from their diverse training data. This capability is crucial for distinguishing between real and fake audio. Therefore, the spectrogram inputs are first fed into these pre-trained models to obtain audio embeddings. Given the audio embeddings, we propose a Multilayer Perceptron (MLP), as shown in Table II, to detect real or fake audio.

E. Ensemble of models

As an individual model works on two-second audio segment, the predicted probability of an entire audio recording is computed by averaging of predicted probabilities over all two-second segments in that recording. Consider $\boldsymbol{p}^{(n)} = [p_1^{(n)}, p_2^{(n)}, ..., p_C^{(n)}]$, with C being the category number of the n-th out of N two-second segments in one audio recording, the probability of an entire audio recording is calculated by the average classification probability which denoted as $\bar{\boldsymbol{p}} = [\bar{p}_1, \bar{p}_2, ..., \bar{p}_C]$ where:

$$\bar{p}_c = \frac{1}{N}\sum_{n=1}^{N} p_c^{(n)} \quad for \quad 1 \leq c \leq C \tag{1}$$

TABLE II

THE AUDIO PRE-TRAINED MODELS AND THE MULTILAYER PERCEPTRON

Models	Using License	Embedding size/configuration
Whisper [15]	MIT	512
SpeechBrain [17]	Apache2-0	192
SeamLess [16]	MIT	1024
Pyannote [18], [19]	MIT	512
MLP	Our proposal	**1** $\times$ {Dense(128)-ReLU } **1** $\times$ {Dense(2)-Softmax }

To combine the results from individual models, we propose a MEAN fusion. In particular, we first conduct experiments on the individual models, then obtain the predicted probability as $\hat{\mathbf{p}}_s = (\bar{p}_{s_1}, \bar{p}_{s_2}, ..., \bar{p}_{s_C})$ where C is the category number and the s-th out of S individual models evaluated. Next, the predicted probability after MEAN fusion $(\hat{p}_1, \hat{p}_2, ..., \hat{p}_C)$ is obtained by:

$$\hat{p_c} = \frac{1}{S}\sum_{s=1}^{S} \hat{p}_{s_c} \quad for \quad 1 \leq c \leq C \tag{2}$$

Finally, the predicted label $\hat{y}$ for an entire audio sample is determined as:

$$\hat{y} = \text{argmax}(\hat{p}_1, \hat{p}_2, ..., \hat{p}_C) \tag{3}$$

III. EXPERIMENTS AND RESULTS

A. Datasets and evaluation metrics

We evaluate the proposed models on the Logic Access dataset of ASVspoofing 2019 challenge. The Logic Access dataset comprises three subsets (fake sample/real sample) of 'Train' (22800/2580), 'Development' (22296/2548), and 'Evaluation' (63882/7355), in which fake audio samples were generated from 19 AI-based generative systems. The models are trained on 'Train' subset, then evaluated on 'Development' subset. Finally, the models are tested on the 'Evaluation' subset to report the final result.

We obey the ASVspoofing 2019 challenge, then use the Equal Error Rate (ERR) as the main metric for evaluating proposed models. We also report the Accuracy, F1 score and AUC score to compare the performance among proposed models.

B. Results and discussion

Evaluation of spectrogram inputs: Considering the efficacy of feature extraction among proposed spectrogram inputs (i.e., systems from A1 to A6), STFT outperforms other spectrograms (models such as A1, A4, A6 achieve the best ERR score of 0.08 while the combination of STFT & LF obtains slightly better accuracy and F1 score of 0.88 and 0.9 respectively). This result suggests that STFT is more appropriate for identifying deepfake artifacts due to its uniform resolution in time and frequency [20] while the interpretable features extracted from linearly filtered signals are suitable for isolating specific frequency bands in classification algorithms.

Multiple deep learning approaches: Regarding end-to-end deep learning approach (from A1 to B2), both RNN and C-RNN approaches obtain ERR scores of 0.14 and 0.17, significantly worse than using only CNN with the best score

TABLE III
PERFORMANCE COMPARISON AMONG DEEP LEARNING MODELS AND ENSEMBLE OF HIGH-PERFORMANCE MODELS ON LOGIC ACCESS EVALUATION SUBSET IN ASVSPOOFING 2019

Systems	**Spectrograms**	**Models**	**Acc**	**F1**	**AUC**	**ERR**
A1	STFT	CNN	**0.87**	**0.89**	**0.96**	**0.08**
A2	CQT	CNN	0.89	0.90	0.92	0.14
A3	WT	CNN	0.84	0.86	0.89	0.17
A4	STFT & LF	CNN	**0.88**	**0.90**	**0.96**	**0.08**
A5	STFT & MEL	CNN	0.86	0.88	0.95	0.11
A6	STFT & GAM	CNN	**0.85**	**0.87**	**0.96**	**0.08**
B1	STFT & LF	RNN	0.92	0.91	0.88	0.17
B2	STFT & LF	CRNN	0.88	0.90	0.96	0.14
C1	STFT & LF	ResNet-18	0.49	0.58	0.51	0.47
C2	STFT & LF	MobileNet-V3	0.59	0.67	0.52	0.48
C3	STFT & LF	EfficientNet-B0	0.52	0.61	0.51	0.48
C4	STFT & LF	DenseNet-121	0.58	0.66	0.51	0.48
C5	STFT & LF	ShuffleNet-V2	0.64	0.71	0.53	0.48
C6	STFT & LF	Swin_T	**0.84**	**0.87**	**0.94**	**0.09**
C7	STFT & LF	ConvNeXt-Tiny	**0.88**	**0.90**	**0.96**	**0.075**
C8	STFT & LF	GoogLeNet	0.53	0.62	0.51	0.47
C9	STFT & LF	MNASNet	0.62	0.70	0.54	0.47
C10	STFT & LF	RegNet	0.50	0.60	0.50	0.48
D1	STFT & LF	Whisper+MLP	**0.85**	**0.88**	**0.95**	**0.10**
D2	STFT & LF	Speechbrain+MLP	0.77	0.81	0.81	0.25
D3	STFT & LF	Seamless+MLP	0.86	0.88	0.87	0.20
D4	STFT & LF	Pyannote+MLP	0.64	0.71	0.78	0.27
A1 + A2	STFT, CQT	CNN	**0.91**	**0.92**	**0.98**	**0.06**
A1 + A3	STFT, WT	CNN	0.88	0.90	0.96	0.09
A1 + A2 + A3	STFT, CQT, WT	CNN	0.90	0.92	0.98	0.07
A4 + A5	STFT&LF, STFT&MEL	CNN	0.88	0.90	0.97	0.08
A4 + A6	STFT&LF, STFT&GAM	CNN	0.87	0.89	**0.98**	**0.065**
A4 + A5 + A6	STFT& LF, STFT&MEL, STFT&GAM	CNN	0.88	0.90	0.98	0.069
A4 + C6	STFT&LF	CNN, Swint_T	0.87	0.89	0.96	0.078
A4 + C7	STFT&LF	CNN, ConvNeXt-Tiny	0.88	0.90	**0.97**	**0.07**
A4 + C6 + C7	STFT&LF	CNN, ConvNeXt-Tiny, Swint_T	0.88	0.89	0.97	0.072
A2 + A4 + A6 + C7 + D1	**CQT, STFT&LF, STFT&GAM**	**CNN, ConvNeXt-Tiny, Whisper**	**0.90**	**0.91**	**0.994**	**0.03**

of 0.08. This indicates the specific patterns indicative of deepfake audio might not be primarily temporal but rather frequency in the spectrogram representation. In the finetuning and audio embedding-based approaches (C1 to C10 and D1 to D4), Swint, Convnext-Tiny and Whisper stand out as the best systems with competitive EER scores of 0.09, 0.0075 and 0.10 respectively. This suggests the potential of these approaches when choosing the appropriate networks for enhancement.

Ensembles: The experimental results presented in Table III underscore the significant effectiveness of ensemble techniques in detecting audio deepfakes. Specifically, the combination of STFT and CQT spectrograms (A1+A2) achieves a score of 0.06, marking an improvement of 0.02 compared to best systems utilizing single spectrograms. Similarly, ensembles of models show slight enhancements such as the combination of CNN and ConvNeXt-Tiny which helps to reduce the ERR compared to individual models. These findings suggest that diverse feature extraction via ensembling multiple spectrograms substantially enhances overall performance compared to evaluating a wide range of models on a single spectrogram. Importantly, the ensemble of both spectrograms and models demonstrates significant improvement. Our best-performing system (A2, A4, A6, C7, and D1) achieves an ERR score and AUC of 0.03 and 0.994 respectively, placing in the top-3 in terms of EER score in the ASVspoof 2019 challenge [6]. These results highlight the strength of ensemble technique by leveraging multiple spectrograms for feature extraction and deep learning models for pattern recognition.

IV. CONCLUSION

This paper has evaluated the efficacy of a wide range of spectrograms and deep learning approaches for deepfake audio detection. By establishing the ensemble of selective spectrograms and models, our best system achieves the EER score of 0.03 on LA dataset of ASVspoofing 2019 challenge, which is very competitive to state-of-the-art systems. Additionally, our comprehensive evaluation also indicate the potential of certain types of spectrogram (e.g. STFT) and deep learning approaches (e.g. CNN-based, fine-tuning pre-trained models), which can provide initial guidance for further improvement of deepfake audio detection.

ACKNOWLEDGMENTS

EUCINF project is co-funded by European Union under grant agreement N°101121418. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] Luca Turchet et al., "The internet of sounds: Convergent trends, insights, and future directions," *IEEE Internet of Things Journal*, vol. 10, no. 13, pp. 11264–11292, 2023.

[2] Luca Turchet et al., "The internet of audio things: State of the art, vision, and challenges," *IEEE internet of things journal*, vol. 7, no. 10, pp. 10233–10249, 2020.

[3] Zhizheng Wu et al., "Spoofing and countermeasures for speaker verification: A survey," *speech communication*, vol. 66, pp. 130–153, 2015.

[4] Jiangyan Yi et al., "Scenefake: An initial dataset and benchmarks for scene fake audio detection," *Pattern Recognition*, vol. 152, pp. 110468, 2024.

[5] Yan Zhao et al., "Emofake: An initial dataset for emotion fake audio detection," 2023.

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

[7] Jiangyan Yi et al., "Add 2022: the first audio deep synthesis detection challenge," in *Proc. ICASSP*, 2022, pp. 9216–9220.

[8] Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao, "Audio deepfake detection: A survey," *arXiv preprint arXiv:2308.14970*, 2023.

[9] Nanxin Chen et al., "Robust deep feature for spoofing detection — the SJTU system for ASVspoof 2015 challenge," in *Proc. Interspeech*, 2015, pp. 2097–2101.

[10] Alejandro Gomez-Alanis et al., "A light convolutional gru-rnn deep feature extractor for asv spoofing detection," in *Proc. Interspeech*, 2019, pp. 1068–1072.

[11] Lam Pham, Khoa Tran, Dat Ngo, Hieu Tang, Son Phan, and Alexander Schindler, "Wider or deeper neural network architecture for acoustic scene classification with mismatched recording devices," in *Proceedings of the 4th ACM International Conference on Multimedia in Asia*, 2022.

[12] Lam Pham, Dat Ngo, Dusan Salovic, Anahid Jalali, Alexander Schindler, Phu X. Nguyen, Khoa Tran, and Hai Canh Vu, "Lightweight deep neural networks for acoustic scene classification and an effective visualization for presenting sound scene contexts," *Applied Acoustics*, vol. 211, pp. 109489, 2023.

[13] Akash Chintha et al., "Recurrent convolutional structures for audio spoof and video deepfake detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 1024–1037, 2020.

[14] Jia Deng et al., "Imagenet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009, pp. 248–255.

[15] Alec Radford et al., "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023, pp. 28492–28518.

[16] Barrault Loïc et al., "Seamless: Multilingual expressive and streaming speech translation," *arXiv preprint arXiv:2312.05187*, 2023.

[17] Mirco Ravanelli et al., "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[18] Alexis Plaquet and Hervé Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization," in *Proc. INTERSPEECH*, 2023.

[19] Hervé Bredin, "pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in *Proc. INTERSPEECH*, 2023.

[20] Daniel Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 2, pp. 236–243, 1984.