



Towards Source Attribution of Singing Voice Deepfake with Multimodal Foundation Models

Orchid Chetia Phukan^{*1}, Girish^{*1,2}, Mohd Mujtaba Akhtar^{*1,3}, Swarup Ranjan Behera⁴, Priyabrata Mallick⁴, Pailla Balakrishna Reddy⁵, Arun Balaji Buduru¹, Rajesh Sharma^{6,7}

¹IIIT-Delhi, India, ²UPES, India, ³V.B.S.P.U, India, ⁴Independent Researcher, India, ⁵Reliance AI, India, ⁶University of Tartu, Estonia, ⁷Plaksha University, India

Correspondence: orchidp@iiitd.ac.in

Abstract

In this work, we introduce the task of singing voice deepfake source attribution (SVDSA). We hypothesize that multimodal foundation models (MMFMs) such as ImageBind, LanguageBind will be most effective for SVDSA as they are better equipped for capturing subtle source-specific characteristics—such as unique timbre, pitch manipulation, or synthesis artifacts of each singing voice deepfake source due to their cross-modality pre-training. Our experiments with MMFMs, speech foundation models and music foundation models verify the hypothesis that MMFMs are the most effective for SVDSA. Furthermore, inspired from related research, we also explore fusion of foundation models (FMs) for improved SVDSA. To this end, we propose a novel framework, **COFFE** which employs Chernoff Distance as novel loss function for effective fusion of FMs. Through **COFFE** with the symphony of MMFMs, we attain the topmost performance in comparison to all the individual FMs and baseline fusion methods.

Index Terms: Source Attribution, Singing Voice Deepfake, Deepfake Detection

1. Introduction

"Imagine discovering a new song by your favorite artist, only to learn they never recorded it." With generative technologies advancing at an unprecedented pace, this scenario is no longer hypothetical. Singing voice deepfakes (SVDs) have evolved to a level where they can convincingly mimic an artist's vocal timbre, seamlessly intertwining *speech articulation* with *musical tonality* [1]. While these innovations open new frontiers in creative expression, they also introduce profound challenges related to authenticity, intellectual property rights, and the ethical deployment of AI in music generation. As deepfake synthesis becomes increasingly sophisticated, the challenge extends beyond mere detection - understanding the provenance of synthetic audio is now imperative. Tracing the origins of SVDs has become crucial for safeguarding artistic integrity and mitigating the risks of misuse, yet this remains an uncharted problem in singing voice deepfake forensics.

While Singing Voice Deepfake Detection (SVDD) has seen notable advancements [1, 2, 3], the equally critical challenge of source attribution - identifying which model or method generated a deepfake - remains largely unexplored. In this work, we introduce the task of singing voice deepfake source attribution (SVDSA). Unlike conventional deepfake detection, which merely classifies an audio clip as real or fake, source attribution seeks to trace its origin, revealing the generative process behind its creation [4, 5]. Source attribution of deepfake speech has

captured significant attention in the research community in recent years [6] in contrast to SVDSA. Muller et al. [7] used an RNN-based approach for characterizing seen and unseen speech deepfake source signatures. Further, Klein et al. [8] and [9] has shown the potential of using state-of-the-art (SOTA) speech foundation models (SFMs) such as wav2vec2, Whisper for speech deepfake source attribution (SDSA). These foundation models (FMs) provides sufficient performance benefit as well as take away the need of training models from scratch. As such these FMs have not only captured attention for SDSA but also for speech deepfake detection [10], SVDD [11] and so on.

So, as the primary research in SVDSA, we explore various FMs and we hypothesize that multimodal foundation models (MMFMs), such as ImageBind (IB) and LanguageBind (LB) will be most effective for SVDSA as they are particularly well-suited for capturing subtle, source-specific characteristics—such as unique timbre, pitch variations, and synthesis artifacts—present in singing voice deepfake sources. This advantage arises from their cross-modality pretraining, which enables them to learn rich, complementary representations by leveraging diverse contextual and acoustic information. To validate our hypothesis, we perform a large-scale comparison of MMFMs, SFMs, and music foundation models (MFMs). We consider SFMs and MFMs in our study as research on SVDD has shown their efficacy [1, 3]. Further, motivated by previous research in various related tasks such as speech deepfake detection [10], speech recognition [12] as well as SVDD [11, 13] where fusion of FMs have shown improved performance due to the emergence of complementary behavior of the FMs, we also explore the fusion of FMs for SVDSA. To our end, we propose a novel framework, **COFFE** (Fusion using ChernOFF DistanceE), for effective fusion of FMs. It leverages Chernoff Distance as a novel loss function for aligning the FMs to a joint feature space. With **COFFE** through the fusion of LB and IB, we obtain the topmost performance in comparison to all the individual FMs, baseline fusion techniques and setting SOTA in benchmark SVD dataset for future research in SVDSA.

To summarize, the main contributions are as follows:

- We introduce SVDSA, pioneering the task of tracing the generative origins of synthetic singing voices.
- We demonstrate the effectiveness of MMFMs, which outperform unimodal SFMs and MFMs for SVDSA due to their multimodal pre-training.
- We propose a novel framework, **COFFE** for fusion of FMs which uses Chernoff Distance (CD) as novel loss function.
- Using **COFFE** with fusion of LB and IB we achieved the topmost performance in comparison to individual FMs and baseline fusion techniques.
- We establish the first benchmark for SVDSA.

The code and models proposed in this study are available at:

* Contributed equally as a first authors.

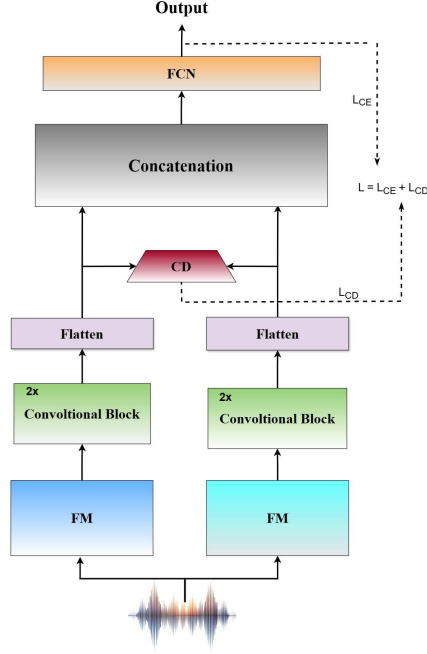


Figure 1: *Proposed Modeling Framework for fusion of FMs: COFFEE*; *CD* represents the Chernoff Distance; L_{CD} , L_{CE} , L stands for Chernoff Distance Loss, Cross-entropy loss, Total loss respectively

https://github.com/Helix-IIIT-Delhi/COFFEE-Singing_Voice_Deepfake

2. Foundation Models

In this section, we first discuss the SFMs followed by MFMs and MMFMs considered in our study.

Speech Foundation Models: We consider WavLM¹ [14] and Unispeech-SAT² [15] which are SOTA SFMs in SUPERB. Unispeech-SAT incorporates contrastive utterance-wise loss and speaker-aware training while WavLM does its pre-training through masked speech modeling and denoising. We consider the base versions of both WavLM and Unispeech-SAT of 94.70M and 94.68M parameters and pre-trained on 960 hours of english librispeech. We also consider Wav2vec2³ [14]. Wav2vec2 is a SFM that is trained in self-supervised manner and applies contrastive learning to masked speech inputs. We use the base version of Wav2vec2 with 95.04M parameters with pre-training done on english librispeech 960 hours. Further, we included SOTA multilingual SFMs such as XLS-R⁴ [16], Whisper⁵ [17], and MMS⁶ [18] in our study. XLS-R and MMS are built on top of Wav2vec2 architecture, while XLS-R being trained on 128 languages, MMS extends pre-training to 1406 languages. Whisper is vanilla transformer encoder-decoder architecture and pre-trained in a multi-task learning weakly supervised manner. We consider XLS-R, Whisper, MMS of 300M, 74M and 1B

¹<https://huggingface.co/facebook/wav2vec2-base>

²<https://huggingface.co/microsoft/unispeech-sat-base>

³<https://huggingface.co/facebook/wav2vec2-base>

⁴<https://huggingface.co/facebook/wav2vec2-xls-r-300m>

⁵<https://huggingface.co/openai/whisper-base>

⁶<https://huggingface.co/facebook/mms-1b>

parameters variants. Additionally, we included x-vector⁷ [19], a time-delay neural network trained for speaker recognition of 4.2M parameters. We consider it as it has shown effective performance in synthetic speech detection [10] and we thought it might be helpful for SVDSA.

Music Foundation Models: MERT series [20] represents a SOTA MFMs, specifically designed for providing intricate features from musical audio. These MFMs demonstrate exceptional performance across various music-related tasks, including instrument classification, singer identification, emotion score prediction and so on due to their extensive pretraining on diverse music datasets. We incorporate multiple MERT variants, namely *MERT-v1-330M*⁸, *MERT-v1-95M*⁹, *MERT-v0-public*¹⁰, and *MERT-v0*¹¹. Additionally, we include music2vec-v1 [21]¹² which follows a self-supervised training approach, allowing it to effectively capture the nuanced characteristics of music and provide generalized representations for diverse music information retrieval tasks. All MFMs excluding MERT-v1-330M are of 95M parameters each and MERT-v1-330M is of 330M parameters.

Multimodal Foundation Models: IB¹³ [22] is a MMFM that aligns diverse inputs such as images, audio, text, IMU, depth, and thermal data to a shared image representation space during its pre-training stage by employing an InfoNCE-based optimization strategy. Without requiring explicitly paired training data, IB shows strong generalization across modalities. Similarly, LanguageBind¹⁴ [23] aligns multiple data types—including video, depth, audio, and infrared—by anchoring them to a fixed language encoder through contrastive learning.

All FMs requires the input audio to be resampled at different rates: MERT-v1-330M and MERT-v1-95M operate at 24 kHz, whereas MERT-v0-public, MERT-v0, and music2vec-v1 process audio at 16 kHz. Both SFMs and MMFMs necessitate resampling audio to 16 kHz. For feature extraction, average pooling is applied to the final hidden layer of each frozen FM. music2vec-v1 and MERT variants yield 768-dimensional representations, except for MERT-v1-330M, which generates 1024-dimensional representations. Additionally, x-vector, Whisper produces 512-dimensional representations, while IB and LB output 1024 and 768 dimensions, respectively. For whisper, only the encoder is used. XLS-R, MMS produces 1280-dimension representations while Unispeech-SAT, WavLM, Wav2vec2 generates 768-dimensional representations.

3. Modeling

In this section, we first discuss the downstream network with individual FMs followed by our proposed framework for fusion of FMs, COFFEE. We implemented two distinct downstream for individual FMs—Fully Connected Network (FCN) and CNN. The CNN consists of two 1D convolutional layers with 64 and 128 filters (kernel size = 3), each followed by max pooling (pool size = 2). The features are then flattened and passed through a FCN block with a dense layer of 128. The output layer comprised

⁷<https://huggingface.co/speechbrain/spkrec-xvect-voxceleb>

⁸<https://huggingface.co/m-a-p/MERT-v1-330M>

⁹<https://huggingface.co/m-a-p/MERT-v1-95M>

¹⁰<https://huggingface.co/m-a-p/MERT-v0-public>

¹¹<https://huggingface.co/m-a-p/MERT-v0>

¹²<https://huggingface.co/m-a-p/music2vec-v1>

¹³<https://github.com/facebookresearch/ImageBind/tree/main>

¹⁴<https://github.com/PKU-YuanGroup/LanguageBind>

8 neurons with softmax as activation function that provides probabilities of SVD sources. The FCN model consisted the same modeling details as the FCN block in CNN.

3.1. COFFE

We propose a novel framework, **COFFE** for the fusion of FMs. The modeling architecture is given in Figure 1. First, the representations from the FMs are passed through two convolutional blocks consisting of 1D convolutional layers with maxpooling with the same modeling as done with individual FMs above. The features are then flattened. We use chernoff distance as novel loss function for aligning the representation space of two FMs. CD is beneficial as it effectively minimizes the separability between feature distributions of the FMs. The CD between two FMs feature space p and q , defined as:

$$L_{CD} = -\log \left(\sum_i p_i^s \cdot q_i^{1-s} \right) \quad (1)$$

where s balances their contributions. A higher CD value indicates greater separability and we aim for optimizing it to minimum for aligning the FMs. The features are then concatenated and finally passed through a FCN block with 128 neurons followed by an output layer with 8 neurons as the singfake source classes. Finally, we obtain the total loss function L that combines the L_{CD} and L_{CE} for joint optimization. L is given as: $L = L_{CE} + \lambda \cdot L_{CD}$, where L_{CE} , L_{CD} are cross-entropy and CD loss. λ is a hyperparameter. The number of trainable parameters varies between 3M and 8M, depending on the size of the FMs representations.

4. Experiment

4.1. Dataset

We utilized the CtrSVDD [24], a benchmark dataset specifically designed for SVDD and the audio samples are in Chinese and Japanese. We only consider the synthetic samples for our experiments and it includes 188,486 clips, totaling 260.34 hours. The synthetic samples are generated through 14 distinct synthesis methods (A01–A14). The dataset contains its own official split into train, dev and eval. As train and dev contains the samples generated by the same systems ranging from A01–A08, we used them as training and testing for our source classification models. **Training and Hyperparameter Details:** We keep the training epochs as 50 and use Adam as optimizer with cross-entropy as loss function. We keep the learning rate as 1e-3 and use dropout and early dropping for preventing overfitting. For the experiments with **COFFE**, we keep the value of s and λ as 0.3 and 0.1 after some initial experimentation.

4.2. Experimental Results

Although the dataset considered in our study lacks musical content, we include MFMs in our experiments, believing that their pre-training on musical data might provide them strength to implicitly capture rhythmic patterns in singing voices, and thus might benefit SVDSA. Table 1 presents the evaluation results of downstream models trained with various individual FMs for SVDSA. We use accuracy, F1 score, and EER (equal error rate) as the evaluation metrics for the evaluation of the models. Accuracy is a preferred metric for source attribution as shown by previous research for speech deepfake source attribution [8]. We have also used EER as it is the preferred metric for SVDD as well as various types of audio deepfake detection [1, 10]. For EER,

FM	FCN			CNN		
	Acc ↑	F1 ↑	EER ↓	Acc ↑	F1 ↑	EER ↓
UNI	45.56	42.22	22.25	48.76	46.56	18.95
W2V2	59.17	52.12	17.12	64.03	57.90	13.66
WM	35.22	31.26	27.87	38.96	36.91	20.18
XL	74.90	64.03	10.85	77.51	69.73	8.73
WS	67.92	64.69	12.32	72.65	65.11	10.70
MMS	76.36	67.56	8.97	80.41	76.83	7.31
XC	63.48	61.17	13.04	66.52	63.95	12.32
LB	79.69	77.26	6.98	82.37	79.80	5.35
IB	78.50	74.41	7.54	81.92	77.90	6.19
Mv1	40.87	37.02	25.72	47.13	42.98	19.79
M95	47.95	45.12	21.44	50.48	47.85	18.71
Mpub	53.03	47.19	19.52	55.66	52.49	17.77
M330	58.64	54.93	15.95	67.96	63.88	12.19
Mv0	45.47	42.81	24.58	48.55	45.53	19.65

Table 1: Evaluation Scores for different FMs; Abbreviations used: **UNI**: Unispeech-SAT, **W2V2**: Wav2vec2, **WM**: WavLM, **WS**: Whisper, **XL**: XLS-R, **XC**: x-vector, **MMS**: MMS, **HUB**: HuBERT, **LB**: LanguageBind, **IB**: ImageBind, **Mv1**: music2vec-v1, **M95**: MERT-v1-95M, **Mpub**: MERT-v0-public, **M330**: MERT-v1-330M, **Mv0**: MERT-v0; Scores are expressed as percentages (%); The abbreviations and information in this Table 1 are kept same for Table 2

we report the average one-vs-all scores. The results demonstrate that MMFMs consistently outperformed both SFMs and MFMs, thus proving our hypothesis that MMFMs will be the most effective for SVDSA due to their cross-modality pre-training them enables them to capture source-specific traits—such as timbre, pitch variations, and synthesis artifacts—by learning rich, complementary representations from diverse acoustic and contextual cues. Among the MMFMs reported the topmost performance with both FCN and CNN downstreams. Overall, the CNN models shows better performance than FCN models with most of the FMs. After MMFMs, the multilingual SFMs shows the second top performance and their results can be result of their exposure to multilingual data during their pre-training as the dataset used in our study consists of singing voices in Chinese and Japanese. Further, we see that monolingual SFMs (WavLM, Unispeech-SAT, Wav2vec2) shows lower performance than its multilingual counterparts, this can be attributed to the linguistic difference between the pre-training of the monolingual SFMs and the downstream data distribution. MFMs reported the lowest performance amongst different FMs. This shows their ineffectiveness in capturing source specific characteristics. Further, a surprising observation is the performance of x-vector, a SFM trained for speaker recognition as it is a much smaller SFM compared to other SFMs. It shows better performance than both monolingual SFMs and MFMs and this performance can be due to its speaker recognition pre-training which equips it with a stronger ability to distinguish source-specific traits. We also plot the t-SNE visualization of FMs raw representations in Figure 2. Through these plots, we observe better clustering across the source classes in MMFMs and providing support to our hypothesis and obtained experimental results.

Table 2 presents the evaluation scores for modeling with various combinations of SFMs. We use concatenation-based fusion as baseline technique. We keep the modeling the same as **COFFE** except the CD loss. We also keep the training details same as **COFFE** for fair comparison. We observe that fusion of FMs through **COFFE** generally shows better performance than baseline concatenation-based fusion. With **COFFE**, through the fusion of MMFMs LB and IB, we report the topmost performance across

Combinations	Concatenation			COFFE		
	ACC \uparrow	F1 \uparrow	EER \downarrow	ACC \uparrow	F1 \uparrow	EER \downarrow
UNI + W2V2	61.78	54.02	17.13	62.36	61.96	15.20
UNI + WM	52.33	44.28	23.11	54.96	53.61	19.63
UNI + XL	74.60	61.24	11.97	72.31	78.50	7.39
UNI + WS	68.49	62.67	13.04	73.18	71.37	9.62
UNI + MMS	73.23	68.49	9.16	73.31	78.22	7.94
UNI + XC	65.89	58.65	11.80	67.34	66.34	8.61
UNI + LB	78.86	74.00	8.22	80.91	79.36	5.34
UNI + IB	72.10	57.77	12.65	76.36	74.36	10.37
UNI + Mv1	59.28	52.29	19.09	47.13	42.98	16.43
UNI + M95	63.64	57.26	16.69	68.36	67.63	11.91
UNI + Mpub	62.83	55.81	14.97	66.95	65.09	9.33
UNI + M330	66.74	60.54	13.63	71.69	70.34	11.53
UNI + Mv0	60.91	53.86	18.35	63.85	62.31	13.31
W2V2 + WM	60.06	52.18	16.35	65.85	63.29	11.61
W2V2 + XL	73.05	65.89	10.13	74.31	73.35	6.39
W2V2 + WS	71.24	65.52	11.75	73.36	72.39	9.32
W2V2 + MMS	71.78	62.73	10.58	74.62	71.59	8.62
W2V2 + XC	68.15	57.41	12.65	75.35	72.64	10.37
W2V2 + LB	77.02	72.36	9.54	81.36	80.91	6.32
W2V2 + IB	71.48	58.20	12.14	76.61	75.30	9.33
W2V2 + Mv1	60.23	50.59	17.23	62.93	61.39	13.96
W2V2 + M95	63.46	54.13	16.79	65.23	64.94	10.63
W2V2 + Mpub	62.51	52.15	16.52	64.36	63.96	11.91
W2V2 + M330	64.98	55.19	15.01	67.96	66.11	9.61
W2V2 + Mv0	61.38	51.19	17.13	63.31	62.96	12.33
WM + XL	73.00	59.27	11.06	75.66	74.31	6.97
WM + WS	67.81	60.31	12.87	69.34	68.36	8.63
WM + MMS	75.87	66.99	10.34	76.93	77.23	10.01
WM + XC	65.41	58.69	12.16	66.31	64.96	8.19
WM + LB	78.48	73.50	8.58	81.63	80.14	8.13
WM + IB	69.41	59.61	9.98	78.54	70.13	8.64
WM + Mv1	54.19	47.05	22.41	56.96	55.14	19.17
WM + M95	58.18	50.63	18.84	61.08	60.31	16.42
WM + Mpub	58.34	49.78	16.92	60.27	52.67	16.17
WM + M330	63.42	57.45	14.54	65.63	64.49	11.16
WM + Mv0	56.46	49.49	20.31	59.61	58.31	16.37
XL + WS	79.41	68.59	7.80	82.64	81.37	6.09
XL + MMS	76.56	67.21	8.42	78.34	77.31	8.09
XL + XC	78.56	73.96	10.36	81.61	80.67	8.63
XL + LB	77.80	72.40	8.08	79.38	77.31	8.31
XL + IB	78.94	73.50	10.06	80.37	79.09	10.03
XL + Mv1	74.07	66.53	11.16	74.93	73.52	8.34
XL + M95	75.04	60.03	10.58	77.31	75.99	9.06
XL + Mpub	76.11	66.09	9.61	78.05	74.70	8.20
XL + M330	75.36	65.91	12.29	76.89	75.47	11.37
XL + Mv0	71.36	61.11	11.04	72.89	71.67	9.53
WS + MMS	77.62	70.21	8.24	81.64	80.34	6.32
WS + XC	73.56	71.69	10.96	75.28	74.31	8.94
WS + LB	78.64	74.07	8.36	81.39	80.31	6.17
WS + IB	72.38	68.88	10.87	76.64	75.13	9.84
WS + Mv1	69.78	64.32	12.75	71.18	70.61	9.49
WS + M95	69.27	64.83	11.57	72.58	71.23	9.82
WS + Mpub	70.48	65.90	12.08	72.71	69.57	11.80
WS + M330	71.77	67.33	11.40	74.56	73.64	8.26
WS + Mv0	69.71	64.42	12.72	73.64	72.54	10.67
MMS + XC	76.54	71.35	10.31	77.96	76.68	8.22
MMS + LB	79.05	75.43	9.75	83.68	82.27	9.35
MMS + IB	76.59	71.38	10.06	78.34	76.13	9.08
MMS + Mv1	74.30	70.67	9.87	76.34	74.46	9.03
MMS + M95	72.98	63.75	9.57	75.38	74.49	9.64
MMS + Mpub	76.13	67.10	8.82	78.93	77.08	7.51
MMS + M330	75.56	70.77	8.98	77.34	76.31	7.34
MMS + Mv0	74.25	68.40	11.95	76.68	75.39	10.62
XC + LB	82.02	80.15	6.54	83.64	82.66	4.67
XC + IB	75.89	68.45	11.65	77.63	75.58	9.85
XC + Mv1	65.23	63.56	14.63	67.73	66.69	11.63
XC + M95	64.23	57.88	12.36	66.38	65.59	10.16
XC + Mpub	61.35	57.46	11.36	63.28	61.03	9.34
XC + M330	64.65	58.98	13.65	66.37	64.49	11.44
XC + Mv0	59.63	55.41	11.65	62.38	60.34	10.74
LB + IB	89.62	83.88	3.75	91.16	90.03	3.63
LB + Mv1	77.60	72.30	8.91	78.63	77.62	8.03
LB + M95	77.70	72.20	8.65	80.37	78.86	7.39
LB + Mpub	77.63	72.88	9.20	79.78	76.62	8.33
LB + M330	80.40	77.59	7.63	82.29	81.17	7.39
LB + Mv0	79.06	75.84	8.04	82.23	81.18	7.93
IB + Mv1	68.06	55.21	13.22	71.94	68.83	11.38
IB + M95	72.49	71.56	14.38	73.94	71.28	11.07
IB + Mpub	71.47	57.79	16.02	72.29	71.13	13.03
IB + M330	69.95	57.57	15.48	74.24	71.16	12.92
IB + Mv0	68.53	58.68	14.89	70.38	70.08	9.36
Mv1 + M95	57.94	51.52	19.77	61.93	58.39	12.83
Mv1 + Mpub	57.57	49.67	18.17	59.93	58.81	13.93
Mv1 + M330	62.89	57.46	13.25	63.49	63.02	11.52
Mv1 + Mv0	53.54	46.78	22.63	56.67	55.37	10.34
M95 + Mpub	59.77	53.01	19.51	62.64	59.89	16.51
M95 + M330	60.17	59.09	15.12	63.38	62.97	11.09
M95 + Mv0	57.39	51.06	19.84	59.96	58.64	13.81
Mpub + M330	63.20	56.66	14.60	66.33	64.85	9.57
Mpub + Mv0	57.65	49.43	18.30	59.29	56.73	13.38
M330 + Mv0	61.06	54.20	16.02	64.34	63.08	9.38

Table 2: Evaluation Scores for various FM combinations

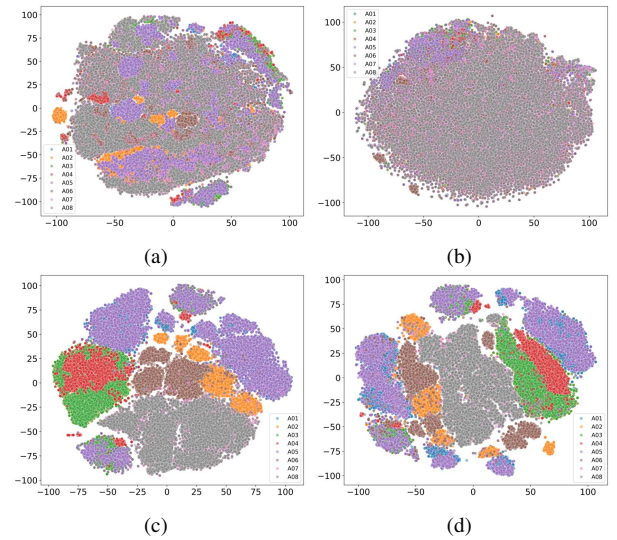


Figure 2: t-SNE Plots- (a) XLS-R (b) MERT-v1-330M (c) LanguageBind (d) ImageBind

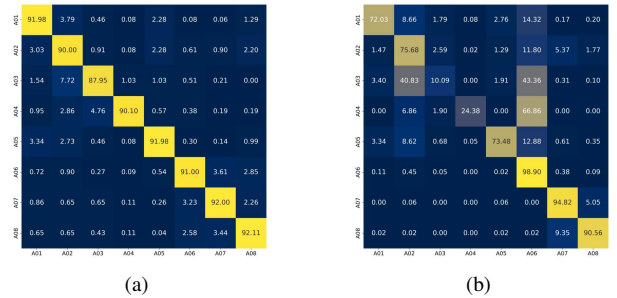


Figure 3: Confusion matrices: (a) COFFE(LB + IB) (b) CNN(LB); The x-axis represents the predicted values, while the y-axis denotes the true values

all the combinations of FMs as well as individual FMs. This shows that combination of MMFMs further brings out the complementary behavior among them. These results show the efficacy of **COFFE** for effective fusion of FMs for improved SVDSA. Additionally, the confusion matrices in Figure 3 further illustrate the improvement in classification accuracy when using **COFFE** with the fusion of LB and IB compared to individual LB with CNN. The results presented in our paper will act as a benchmark for future studies exploring SVDSA.

5. Conclusion

In this study, we introduce the task of SVDSA and demonstrate that MMFMs are the most effective for SVDSA. MMFMs such as IB and LB, excel in capturing source-specific traits like timbre, pitch manipulation, and synthesis artifacts due to their cross-modal pretraining. Building on these insights, we propose **COFFE**, a novel fusion framework that leverages Chernoff Distance as a loss function to enhance the integration of FMs. **COFFE** through the fusion of MMFMs achieves superior performance compared to individual FMs and baseline fusion approaches, establishing a strong baseline for future research in SVDSA. Our work also calls upon researchers to work upon our established benchmarks for further performance improvement of SVDSA.

6. References

- [1] Y. Zang, Y. Zhang, M. Heydari, and Z. Duan, "Singfake: Singing voice deepfake detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 156–12 160.
- [2] Y. Zang, J. Shi, Y. Zhang, R. Yamamoto, J. Han, Y. Tang, S. Xu, W. Zhao, J. Guo, T. Toda, and Z. Duan, "Ctrsvdd: A benchmark dataset and baseline analysis for controlled singing voice deepfake detection," in *Interspeech 2024*, 2024, pp. 4783–4787.
- [3] Y. Zhang, Y. Zang, J. Shi, R. Yamamoto, T. Toda, and Z. Duan, "Svdd 2024: The inaugural singing voice deepfake detection challenge," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 782–787.
- [4] T. Zhu, X. Wang, X. Qin, and M. Li, "Source tracing: Detecting voice spoofing," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022.
- [5] X. Yan, J. Yi, J. Tao, C. Wang, H. Ma, T. Wang, S. Wang, and R. Fu, "An initial investigation for detecting vocoder fingerprints of fake audio," in *Proc. of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, 2022.
- [6] C. Y. Zhang, J. Yi, J. Tao, C. Wang, and X. Yan, "Distinguishing neural speech synthesis models through fingerprints in speech waveforms," in *China National Conference on Chinese Computational Linguistics*. Springer, 2024, pp. 259–273.
- [7] N. Müller, F. Diekmann, and J. Williams, "Attacker attribution of audio deepfakes," in *Interspeech 2022*, 2022, pp. 2788–2792.
- [8] N. Klein, T. Chen, H. Tak, R. Casal, and E. Khoury, "Source tracing of audio deepfake systems," in *Interspeech 2024*, 2024, pp. 1100–1104.
- [9] K. Bhagtani, A. K. S. Yadav, P. Bestagini, and E. J. Delp, "Attribution of diffusion based deepfake speech generators," in *2024 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2024, pp. 1–6.
- [10] 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," in *NAACL-HLT*, 2024.
- [11] X.-B. Chen, H. Wu, J.-S. R. Jang, and H. yi Lee, "Singing voice graph modeling for singfake detection," *Interspeech 2024*, 2024.
- [12] A. Arunkumar, V. N. Sukhadia, and S. Umesh, "Investigation of ensemble features of self-supervised pretrained models for automatic speech recognition," *ArXiv*, vol. abs/2206.05518, 2022.
- [13] A. Guragain, T. Liu, Z. Pan, H. B. Sailor, and Q. Wang, "Speech foundation model ensembles for the controlled singing voice deepfake detection (ctrsvdd) challenge 2024," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 774–781.
- [14] 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.
- [15] S. Chen, Y. Wu, C. Wang, Z. Chen, Z. Chen, S. Liu, J. Wu, Y. Qian, F. Wei, J. Li, and X. Yu, "Unispeech-sat: Universal speech representation learning with speaker aware pre-training," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6152–6156, 2021.
- [16] A. Babu, C. Wang, A. Tjandra, K. Lakhota, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," in *Interspeech 2022*, 2022, pp. 2278–2282.
- [17] 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.
- [18] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi *et al.*, "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.
- [19] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, 2018.
- [20] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Xiao, C. Lin, A. Ragni, E. Benetos *et al.*, "Mert: Acoustic music understanding model with large-scale self-supervised training," *arXiv preprint arXiv:2306.00107*, 2023.
- [21] Y. Li, R. Yuan, G. Zhang, Y. Ma, C. Lin, X. Chen, A. Ragni, H. Yin, Z. Hu, H. He, E. Benetos, N. Gyenge, R. Liu, and J. Fu, "Map-music2vec: A simple and effective baseline for self-supervised music audio representation learning," *ArXiv*, vol. abs/2212.02508, 2022.
- [22] 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 *CVPR*, 2023.
- [23] 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," 2023.
- [24] Y. Zhang, Y. Zang, J. Shi, R. Yamamoto, J. Han, Y. Tang, T. Toda, and Z. Duan, "Svdd challenge 2024: A singing voice deepfake detection challenge (ctrsvdd track, training/development set)," 2024. [Online]. Available: <https://doi.org/10.5281/zenodo.10467648>