SOURCE TRACING OF SYNTHETIC SPEECH SYSTEMS THROUGH PARALINGUISTIC PRE-TRAINED REPRESENTATIONS

Girish*‡†, Mohd Mujtaba Akhtar††, Orchid Chetia Phukan‡†, Drishti Singh‡†
Swarup Ranjan Behera§, Pailla Balakrishna Reddy¶, Arun Balaji Buduru*, Rajesh Sharma‖**
*UPES, India, †V.B.S.P.U, India, ‡IIIT-Delhi, India, §Independent Researcher, India, ¶Reliance AI, India,
‖University of Tartu, Estonia, **Plaksha university,India
Correspondence:mmakhtar.research@gmail.com,orchidp@iiitd.ac.in

Abstract—In this work, we focus on source tracing of synthetic speech generation systems (STSGS). Each source embeds distinctive paralinguistic features—such as pitch, tone, rhythm, and intonation—into their synthesized speech, reflecting the underlying design of the generation model. While previous research has explored representations from speech pre-trained models (SPTMs), the use of representations from SPTM pre-trained for paralinguistic speech processing, which excel in paralinguistic tasks like synthetic speech detection, speech emotion recognition has not been investigated for STSGS. We hypothesize that representations from paralinguistic SPTM will be more effective due to its ability to capture source-specific paralinguistic cues attributing to its paralinguistic pre-training. Our comparative study of representations from various SOTA SPTMs, including paralinguistic, monolingual, multilingual, and speaker recognition, validates this hypothesis. Furthermore, we explore fusion of representations and propose `TRIO`, a novel framework that fuses SPTMs using a gated mechanism for adaptive weighting, followed by canonical correlation loss for inter-representation alignment and self-attention for feature refinement. By fusing TRILLsson (Paralinguistic SPTM) and x-vector (Speaker recognition SPTM), `TRIO` outperforms individual SPTMs, baseline fusion methods, and sets new SOTA for STSGS in comparison to previous works.

Index Terms—Source Tracing, Paralinguistic Pre-Trained Models, Synthetic Speech Generators

I. INTRODUCTION

Advancements in audio manipulation technology have blurred the line between real and synthetic speech. Modern text-to-speech (TTS) and voice conversion (VC) systems can produce highly realistic voices, enabling malicious actors to manipulate speech with remarkable accuracy. This growing challenge highlights the urgent need for reliable methods to detect and attribute synthetic speech. This advancement poses significant risks, as malicious entities can exploit synthetic speech for impersonation, fraud, and misinformation. As such the urgent need for robust synthetic speech detection (SSD) solutions becomes undeniable to safeguard trust in digital communication. As a remedy there has been sufficient research into SDD [1], [2], [3]. Also, the use of representations from speech pre-trained models (SPTMs) such as Wav2vec2,

WavLM, Whisper have captured recent attention within the community as these SPTMs provide performance benefit [4], [5], [6]. These SPTMs are either fine-tuned or used as feature extractors for extracting representations. Despite much advancement in synthetic speech detection, most of the previous research has mostly focused on distinguishing real and synthetically generated speech i.e. binary classification, but it is not sufficient to predict and mitigate misuse and improve forensic analysis.

To further enhance synthetic speech detection from forensic analysis, it is important to understand the exact tool used to generate the speech and this task is known as Source Tracing of Synthetic Speech Generation Systems (STSGS). It has recently captured attention within the community and plays a crucial role in improving the explainability of detection systems, enforcing accountability, and developing targeted countermeasures against malicious deepfake applications [7], [8], [9]. Each source (TTS, VC) imprint distinctive paralinguistic features—including pitch, tone, rhythm, and intonation—onto their synthesized speech, mirroring the underlying design principles and processing mechanisms of the respective generation models.

As such previous research on STSGS have investigated representations from various state-of-the-art (SOTA) SPTMs [10], [11], [12], [13] for understanding their capability for capturing such source-specific paralinguistic cues. However, they haven't investigated the usage of representations from SPTM pre-trained for paralinguistic speech processing such as TRILLsson [14] which have shown SOTA behavior for different paralinguistic tasks including tasks such as synthetic speech detection and speech emotion recognition. In this work, we solve this research gap and explore representations from paralinguistic SPTM for STSGS. *We hypothesize that paralinguistic SPTM representations will be the most effective for STSGS, as their specialized paralinguistic pre-training enables them to capture paralinguistic cues unique to each source more effectively than other SPTM representations.* To test this hypothesis, we conduct a comprehensive comparative study of various SOTA SPTMs, including paralinguistic, monolingual, multilingual, and speaker recogni-

---

† Contributed equally as first authors

tion. Our findings validate our hypothesis.

Additionally, inspired by prior research demonstrating performance gains through SPTMs representations fusion in related areas such as synthetic speech detection [15] and speech emotion recognition [16], we also explore this direction for STSGS. Phukan et al. [13] have made the initial exploration for fusion of SPTMs representation for STSGS, however, they have considered only a handful of SPTMs representations, here, in our study, we consider a wide range of SOTA SPTMs representations and also the inclusion of paralinguistic SPTM representations that has been missing and a major drawback in their study. To this end, we introduce **TRIO** (GaTed Canonical CorRelatIOn Attention Network), a novel framework for fusing SPTMs. **TRIO** employs a gated mechanism for adaptive weighting of representations, incorporates canonical correlation loss for better alignment between the representations, and utilizes self-attention for enhanced feature refinement. By fusing TRILLsson (a paralinguistic SPTM) with x-vector (a speaker recognition SPTM), **TRIO** achieves superior performance, outperforming individual SPTMs, baseline fusion techniques, and setting a new SOTA benchmark for STSGS in comparison to previous works.

In summary, the key contributions of this work are as follows:

- We carry out a comprehensive comparative analysis of various SOTA SPTMs representations to understand the capability of paralinguistic SPTM representations for STSGS. We show that representations from TRILLsson achieves the topmost performance amongst all other SPTMs representations.
- We introduce a novel framework, **TRIO** for effective fusion of SPTMs representations. **TRIO** uses a gated mechanism for adaptive representation weighting, applies canonical correlation loss for improved alignment, and employs self-attention for refined feature enhancement. By fusing TRILLsson and x-vector, **TRIO** surpasses individual SPTMs and baseline fusion methods, setting a new SOTA benchmark for STSGS compared to prior works.

To make our work more accessible and reproducible, we've shared the code and models at [1].

II. Speech Pre-trained Representations

In this section, we present a brief overview of the SPTMs used in our study. Wav2Vec2[2] [17], WavLM[3] [18], and Unispeech-SAT[4] [19] are monolingual SPTMs and we consider their base versions pre-trained on LibriSpeech (960 hours of English). Wav2Vec2 was trained to solve a contrastive learning objective, WavLM was pre-trained for solving masked speech modeling and speech denoising simultaneously while Unispeech-SAT was trained in a multi-task speaker-aware format. Both WavLM and Unispeech-SAT

have reported SOTA performance in SUPERB. We consider XLS-R[5] [20], Whisper[6] [21], and MMS[7] [22] for multilingual SPTMs. We consider their 300M, 74M, and 1B parameters version for XLS-R, Whisper, and MMS respectively. XLS-R, Whisper, MMS were pre-trained on 128, 96, and over 1400 languages respectively. XLS-R and MMS follows Wav2vec2 architecture and pre-trained in a contrastive learning approach while Whisper is a vanilla transformer encoder-decoder architecture and trained in a multi-task manner. We also consider speaker recognition SPTMs such as x-vector[8] [23] and ECAPA[9] [24] as they have shown its effectiveness for synthetic speech detection [15] as well as STSGS [13]. However, Phukan et al. [13] only considered x-vector in their study and here, in our study, we included, ECAPA, which shows further improvement over x-vector in speaker recognition tasks. Both x-vector and ECAPA are trained on Voxceleb1 + Voxceleb2. As paralinguistic SPTM, we consider TRILLsson[10] [14]. It is a distilled model from the SOTA universal paralinguistic conformer (CAP12). TRILLsson representations shows SOTA performance across various paralinguistic tasks such as speech emotion recognition, synthetic speech detection, speaker recognition, and we use the version with 63M parameters. Additionally, we also add Wav2Vec2-emo[11], a SPTM fine-tuned for SER because SER is inherently a paralinguistic application. Before passing the speech samples to SPTMs, we resample them to 16KHz and extract representations from the last hidden state of the frozen SPTMs by mean pooling. We extract representations of 192 for ECAPA; 512 for x-vector, Whisper (We use its encoder); 768 for Wav2vec2, WavLM, Unispeech-SAT, Wav2vec2-emo; 1024 for TRILLsson; 1280 for XLS-R, MMS.

III. Modeling

In this section, we discuss the downstream models used with individual representations followed by the proposed framework, **TRIO** for fusion of SPTMs representations. We use fully connected network (FCN) and CNN as downstream models as they have preferred by previous research as effective downstream networks [15], [13]. The CNN model consists of two convolutional blocks that receives SPTMs representations as input with 1D-CNN layers of 128 and 64 filters of kernel size 3 with each 1D-CNN layer followed by maxpooling. Then we flatten the outputs and use a FCN block that consists of two dense layers with 90 and 45 neurons each followed by the final output layer that uses softmax as activation function and outputs probabilities of the source classes. The FCN model follows the same modeling paradigm as used for the FCN block in the CNN model. The number of trainable parameters in FCN models ranges 0.6 to 0.8M while for CNN models, it varies between 0.8 to 1.2M, depending on the input representations dimensionality.

---

[1]https://github.com/Helix-IIIT-Delhi/TRIO-Source_Tracing

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

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

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

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

[6]https://huggingface.co/openai/whisper-base

[7]https://huggingface.co/facebook/mms-1b

[8]https://huggingface.co/speechbrain/spkrec-xvect-voxceleb

[9]https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

[10]https://www.kaggle.com/models/google/trillsson

[11]https://huggingface.co/speechbrain/emotion-recognition-wav2vec2-IEMOCAP
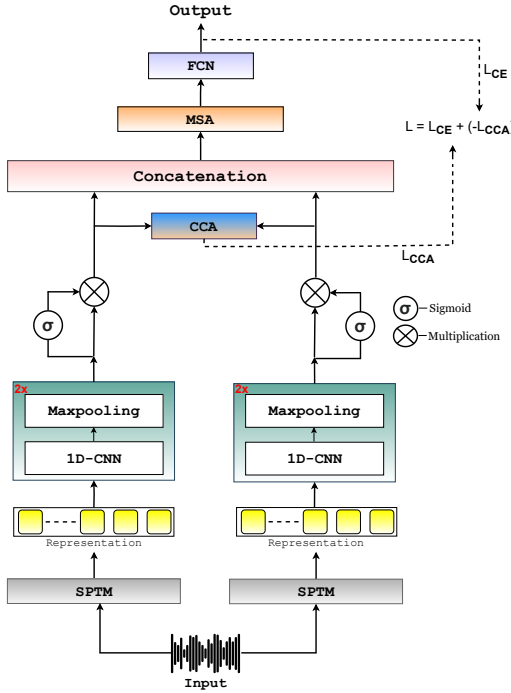
Fig. 1: Proposed Framework: **TRIO**

A. *TRIO*

The architecture diagram of the proposed framework, **TRIO** for fusion of SPTMs representations is shown in Figure 1. **TRIO** leverages a gated mechanism to adaptively weight representations, integrates canonical correlation loss to improve alignment between them, and applies self-attention for more effective feature refinement. First, the SPTMs representations are passed through to two convolutional blocks that uses same modeling as used for individual representational modeling above. Suppose, X and Y are features from two SPTMs branches after the flattening them. Then the flattened features are passed through gated mechanism that consists of a sigmoid gate and the outputs are G_X and G_Y. After that, we perform element wise-multiplication with the original features, $\hat{X} = G_X \odot X, \quad \hat{Y} = G_Y \odot Y$ to extract the most relevant features. Next, the refined features are aligned using canonical correlation analysis (CCA) as a novel loss function, which maximizes the correlation between $\hat{X}$ and $\hat{Y}$. Higher CCA means better alignment. The CCA loss is formulated as:

$$\mathcal{L}_{CCA} = \text{tr}\left((\Sigma_{\hat{X}\hat{X}})^{-1/2} \Sigma_{\hat{X}\hat{Y}} (\Sigma_{\hat{Y}\hat{Y}})^{-1/2} \right)$$

where $\Sigma_{\hat{X}\hat{X}}$ and $\Sigma_{\hat{Y}\hat{Y}}$ are the covariance matrices of $\hat{X}$ and $\hat{Y}$, $\Sigma_{\hat{X}\hat{Y}}$ is the cross-covariance matrix between $\hat{X}$ and $\hat{Y}$. $\text{tr}(\cdot)$ denotes the trace operation. $\mathcal{L}_{CCA}$ ensures that the representations $\hat{X}$ and $\hat{Y}$ are maximally correlated, thereby improving their alignment. After aligning the features to a joint representational space, we concatenate the features from the two SPTMs representation networks. Following this, we use a self-attention mechanism, which computes the queries Q, keys K, and values V as: $Q = X_{\text{concat}}W_Q, \quad K = $

$X_{\text{concat}}W_K, \quad V = X_{\text{concat}}W_V$ where X_{concat} represents the concatenated features from SPTMs representations branches. The attention scores are then computed using the scaled dot-product attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} \right) V$$

Then the features are passed through a FCN block of two dense layers with 90 and 45 neurons followed by a output layer with softmax activation that outputs probabilities. We perform joint optimization with the cross entropy loss $\mathcal{L}_{CCA}$. Finally, the total loss $\mathcal{L}$ is given as: $\mathcal{L} = \mathcal{L}_{CE} + (-\lambda \cdot \mathcal{L}_{CCA})$ where λ is a hyperparameter controlling the importance of $\mathcal{L}_{CCA}$. The negative sign before $\mathcal{L}_{CCA}$ is used because $\mathcal{L}_{CCA}$ is formulated to maximize correlation, while loss functions are typically minimized in optimization. By using a negative sign, we effectively encourage the model to maximize correlation while jointly minimizing $\mathcal{L}_{CE}$. The trainable parameters range from 1.3 to 1.5M.

IV. EXPERIMENTS

A. *Dataset*

We use two benchmark synthetic speech detection databases: ASVSpoof 2019 (ASV) [25] and FAD Chinese Dataset (CFAD) [26]. ASV contains both real and synthetic speech samples from 19 generative systems, recorded at 16 kHz. Real recordings feature diverse speakers with varying accents and speaking styles, while synthetic samples were generated using SOTA VC and TTS methods. We merged the train, validation, and test splits for ASV, resulting in 19 synthetic speech source classes (A01 to A19). We followed 5-fold cross-validation for ASV, with 4 folds used for training and one fold for testing. CFAD is a chinese dataset and features real and synthetic samples from 12 speech synthesis techniques including SOTA TTS and VC systems. We use the official dataset split for training, validating and evaluation of the models.

Training and Hyperparameter Details: The models are trained for 50 epochs with a batch size of 32, utilizing Adam as optimizer and cross-entropy as the loss function. For experiments with **TRIO**, we kept the value of λ fixed as 0.3 throughout the experiments as preliminary exploration yielded optimal results. Dropout and Early stopping are used for mitigating overfitting.

B. *Experimental Results*

Table I presents the evaluation scores of downstream networks trained on different SPTMs representations. We use accuracy and equal error rate (EER) as the evaluation metrics following previous research on STSGS [13]. We report EER by computing the average scores using a one-vs-all approach. For ASV, we report the average of five folds scores and for CFAD, we report the scores obtained in the official evaluation set. Our findings indicate that representations from TRILLsson (paralinguistic SPTM) consistently achieve the highest attribution accuracy and the lowest EER, significantly outperforming other representations. This reinforces their

498

PTMs	ASV				CFAD			
	FCN		CNN		FCN		CNN	
	A ↑	EER ↓	A ↑	EER ↓	A ↑	EER ↓	A ↑	EER ↓
W2V	46.24	15.44	63.76	6.78	51.37	24.20	76.59	9.19
WV	35.42	14.22	47.49	11.53	34.29	25.63	37.83	21.20
US	45.68	23.19	56.48	10.04	45.93	34.91	73.65	17.64
XR	65.67	10.53	80.24	5.03	52.43	17.32	77.98	8.54
WP	76.92	8.88	88.43	4.61	72.17	13.85	86.41	7.91
MMS	82.21	7.90	89.29	4.17	73.54	13.51	89.78	6.95
XV	89.46	5.43	96.63	2.13	76.85	11.63	92.33	4.47
EP	85.17	4.87	93.79	4.04	74.29	10.45	88.61	4.93
W2V-emo	72.59	9.65	82.29	6.32	70.53	12.96	85.39	12.51
T	92.53	4.79	97.16	1.69	78.91	8.63	92.81	3.37

TABLE I: Accuracy and EER in %; Abbreviations used: Wav2vec2 (W2V), WavLM (WV), Unispeech (US), XLS-R (XR), Whisper (WP), MMS (MMS), x-vector (XV), ECAPA (EP), Wav2vec2-emo (W2V-emo), TRILLsson (T); The abreviations used here are kept same for Table II

Pairs	ASV				CFAD			
	Concat		TRIO		Concat		TRIO	
	A↑	EER↓	A↑	EER↓	A↑	EER↓	A↑	EER↓
W2V + WV	94.31	7.69	95.97	7.62	88.79	4.89	94.28	4.61
W2V + US	92.55	8.34	94.85	7.58	85.25	9.28	91.39	8.85
W2V + XR	95.64	8.73	97.14	7.69	93.86	7.95	94.13	8.63
W2V + WP	96.79	7.56	95.96	7.59	94.01	8.94	93.59	7.94
W2V + MMS	94.60	7.62	96.28	6.28	89.97	8.55	93.54	8.51
W2V + XV	96.21	7.36	95.17	7.39	86.54	7.49	89.73	7.37
W2V + EP	93.08	6.59	95.25	7.14	86.21	9.54	92.62	7.56
W2V + W2V-emo	92.85	6.48	96.64	6.59	88.67	8.58	92.47	7.58
W2V + T	97.50	5.86	98.21	5.08	95.85	6.08	96.23	5.89
WV + US	86.79	6.22	88.57	4.93	78.61	9.05	89.28	8.19
WV + XR	85.91	5.30	87.32	4.78	91.59	8.79	91.68	7.54
WV + WP	93.46	6.76	95.35	5.04	93.66	8.89	95.73	7.71
WV + MMS	90.31	6.40	92.39	4.97	90.55	8.01	93.23	8.59
WV + XV	94.89	5.49	94.72	4.30	90.27	8.81	93.79	8.69
WV + EP	93.84	5.06	94.29	5.87	93.21	9.81	93.85	7.29
WV + W2V-emo	88.69	4.99	93.51	4.29	94.67	8.63	94.89	7.63
WV + T	95.81	4.55	95.16	4.33	95.29	7.86	95.21	7.21
US + XR	89.28	5.36	84.61	5.23	79.20	8.11	84.62	7.06
US + WP	91.59	6.04	93.82	5.27	81.82	9.24	85.06	7.29
US + MMS	90.55	5.22	92.38	4.51	89.63	7.99	91.50	5.72
US + XV	92.29	5.54	97.63	4.69	88.26	8.14	92.72	5.85
US + EP	91.97	5.66	94.27	4.76	87.72	8.39	93.50	6.53
US + W2V-emo	92.32	5.49	94.93	4.47	90.28	8.06	92.85	5.69
US + T	93.52	4.92	95.25	4.22	91.63	7.02	94.23	4.86
XR + WP	94.81	5.06	95.53	4.11	90.62	5.49	95.36	5.31
XR + MMS	94.59	5.72	95.83	5.14	92.36	6.52	94.82	5.17
XR + XV	94.27	4.95	95.37	5.35	90.89	5.30	93.81	5.52
XR + EP	93.51	4.92	94.43	4.26	91.76	5.87	92.29	4.49
XR + W2V-emo	93.84	5.29	94.11	4.64	93.83	5.19	94.08	3.85
XR + T	94.62	4.38	96.89	4.05	94.05	4.84	95.13	3.91
WP + MMS	93.59	4.93	94.44	4.48	92.52	6.29	92.89	4.96
WP + XV	95.13	5.31	96.01	4.29	95.11	5.19	97.16	4.14
WP + EP	93.81	4.89	94.06	4.21	93.66	4.81	92.28	4.23
WP + W2V-emo	94.26	4.73	95.24	4.02	92.98	4.51	94.08	3.85
WP + T	94.89	3.95	95.31	3.24	95.21	4.09	97.52	3.09
MMS + XV	96.89	3.84	97.51	3.53	93.53	5.27	94.48	4.61
MMS + EP	95.67	3.18	96.11	2.96	91.13	3.59	92.34	3.38
MMS + W2V-emo	96.91	3.08	97.86	2.93	92.89	4.19	93.82	3.91
MMS + T	97.17	2.84	98.21	2.72	93.86	3.53	94.28	3.01
XV + EP	97.16	4.24	98.04	4.15	94.22	4.21	95.68	4.03
XV + W2V-emo	97.25	4.31	98.14	4.01	95.14	4.63	96.55	4.14
XV + T	98.38	0.36	99.56	0.19	97.28	1.29	99.04	0.95
EP + W2V-emo	87.61	8.53	89.93	7.21	76.28	10.64	79.94	8.28
EP + T	92.28	2.10	94.81	1.83	79.24	7.61	82.38	5.53
W2V-emo + T	97.39	0.45	97.56	0.39	96.38	1.49	97.16	0.99

TABLE II: Accuracy and EER in %
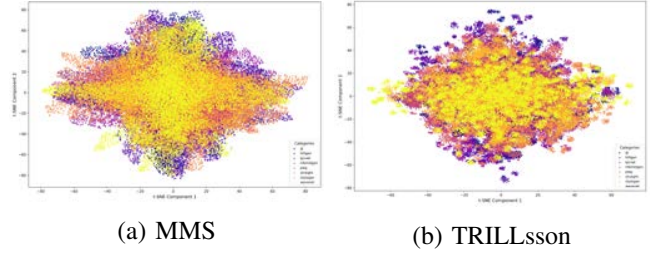

(a) MMS
(b) TRILLsson

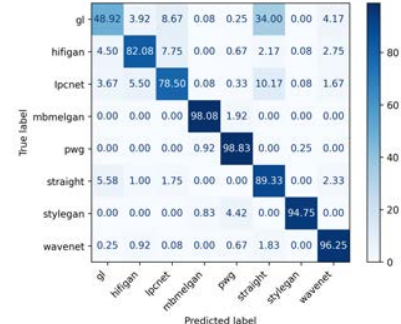Fig. 2: t-SNE Plots for CFAD

Fig. 3: Confusion Matrix for CFAD using **TRIO (x-vector + TRILLsson)**

ability to capture source-specific paralinguistic cues, which are crucial for distinguishing synthetic speech sources. This validates our hypothesis that paralinguistic SPTM representations will be the most effective for STSGS attributing to their paralinguistic pre-training. Among all the other SPTMs, speaker recognition SPTMs (x-vector and ECAPA) showed comparatively good performance. This suggests that their pre-training for speaker recognition tasks enhances their ability to capture source-specific cues, contributing to improved performance in STSGS. Additionally, we observe that monolingual SPTMs reports the lowest performance for both the datasets showing its inability to capture source specific cues. Overall, the CNN models showed better performance than its FCN counterparts. We also plot the t-SNE plots of raw representations of MMS and TRILLsson in Figure 2. We observe better cluster across the source classes for TRILLsson and this supports our obtained results and further amplify the credibility of the proposed hypothesis.

Table II presents the results of fusion of various SPTMs representations. We use concatention-based fusion as the baseline fusion technique. We keep the same network for concatenation-based fusion technique as the proposed framework, **TRIO**. However, we remove the gated mechansim, CCA loss and self-attention refinement block. We keep the training details as same as for experiments with **TRIO**. Our results indicate that fusion of representations through **TRIO** outperforms the baseline fusion technique, demonstrating its effectiveness in integrating diverse SPTM representations. Notably, the best performance across both the datasets was achieved by fusing x-vector and TRILLsson representations

using **TRIO**, highlighting the complementary nature of these representations. Further, we observe that fusion of TRILLsson with speaker recognition and multilingual SPTMs shows comparatively good performance than fusion of monolingual SPTMs with each other. Overall, fusion of SPTMs representations improved performance than the performance with individual representations. We also plot the confusion matrices of CNN trained with **TRIO** with fusion of x-vector and TRILLsson in Figure 3.

Comparison with SOTA: We compare our best performing model **TRIO** with x-vector and TRILLsson with previous SOTA work [13]. They reported accuracy and EER of 98.91% and 0.26% on ASV, while for CFAD, they reported 99.01% and 1.07%. While we report accuracy and EER: 99.56% and 0.19% on ASV, 99.04% and 0.95% on CFAD. This top performance shows that our work sets the new SOTA for STSGS.

V. CONCLUSION

In our study, we show the effectiveness of utilizing paralinguistic SPTMs representations for STSGS. By capturing source-specific paralinguistic cues, these representations outperform representations from various other SOTA SPTMs. Further, we propose **TRIO**, a novel framework for fusion of representations. By integrating TRILLsson and x-vector representations through **TRIO**, we show topmost performance surpassing individual SPTMs representations and baseline fusion methods as well as report SOTA results in STSGS compared to previous SOTA work. Our findings serve as a valuable reference for future studies in selecting appropriate SPTMs representations for STSGS and highlight the potential of combining SPTMs representations for further enhancing STSGS.

REFERENCES

[1] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi *et al.*, "ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," in *Proc. of INTERSPEECH*, 2015.

[2] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, and K. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. of INTERSPEECH*, 2019.

[3] 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.

[4] 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.

[5] P. Kawa, M. Plata, M. Czuba, P. Szymański, and P. Syga, "Improved deepfake detection using whisper features," in *Interspeech 2023*, 2023, pp. 4009–4013.

[6] 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.

[7] 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.

[8] C. Y. Zhang, J. Yi, J. Tao, C. Wang, and X. Yan, "Distinguishing neural speech synthesis models through fingerprints in speech waveforms," *ArXiv*, vol. abs/2309.06780, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:261705832

[9] 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.

[10] N. Klein, T. Chen, H. Tak, R. Casal, and E. Khoury, "Source tracing of audio deepfake systems," in *Interspeech 2024*, 2024, pp. 1100–1104.

[11] X. Yan, J. Yi, J. Tao, and J. Chen, "Audio deepfake attribution: An initial dataset and investigation," 2024. [Online]. Available: https://arxiv.org/abs/2208.10489

[12] 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)*, 2024, pp. 1–6.

[13] O. C. Phukan, D. Singh, S. R. Behera, A. B. Buduru, and R. Sharma, "Investigating prosodic signatures via speech pre-trained models for audio deepfake source attribution," *arXiv preprint arXiv:2412.17796*, 2024.

[14] J. Shor and S. Venugopalan, "Trillsson: Distilled universal paralinguistic speech representations," in *Interspeech 2022*, 2022, pp. 356–360.

[15] O. Chetia Phukan, G. Kashyap, A. B. Buduru, and R. Sharma, "Heterogeneity over homogeneity: Investigating multilingual speech pre-trained models for detecting audio deepfake," in *Findings of the Association for Computational Linguistics: NAACL 2024*, K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 2496–2506. [Online]. Available: https://aclanthology.org/2024.findings-naacl.160

[16] Y. Wu, P. Yue, C. Cheng, and T. Li, "Investigation of ensemble of self-supervised models for speech emotion recognition," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2023, pp. 988–995.

[17] 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.

[18] 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.

[19] S. Chen, Y. Wu, C. Wang, Z. Chen, Z. Chen, S. Liu, J. Wu, Y. Qian, F. Wei, J. Li *et al.*, "Unispeech-sat: Universal speech representation learning with speaker aware pre-training," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6152–6156.

[20] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, 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.

[21] 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.

[22] 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.

[23] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[24] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," 2020.

[25] 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.

[26] H. Ma, J. Yi, C. Wang, X. Yan, J. Tao, T. Wang, S. Wang, and R. Fu, "Cfad: A chinese dataset for fake audio detection," *Speech Communication*, vol. 164, p. 103122, 2024.