Multilingual Source Tracing of Speech Deepfakes: A First Benchmark

Xi Xuan[1,2,*], *Yang Xiao*[3,4], *Rohan Kumar Das*[4], *Tomi Kinnunen*[1]

[1]School of Computing, University of Eastern Finland, Finland
[2]Department of Linguistics and Translation, City University of Hong Kong, Hong Kong SAR
[3]School of Computing and Information Systems, The University of Melbourne, Australia
[4]Fortemedia Singapore, Singapore

*Corresponding author: xi.xuan@uef.fi

Abstract

Recent progress in generative AI has made it increasingly easy to create natural-sounding deepfake speech from just a few seconds of audio. While these tools support helpful applications, they also raise serious concerns by making it possible to generate convincing fake speech in many languages. Current research has largely focused on detecting fake speech, but little attention has been given to tracing the source models used to generate it. This paper introduces the first benchmark for multilingual speech deepfake source tracing, covering both mono- and cross-lingual scenarios. We comparatively investigate DSP- and SSL-based modeling; examine how SSL representations fine-tuned on different languages impact cross-lingual generalization performance; and evaluate generalization to unseen languages and speakers. Our findings offer the first comprehensive insights into the challenges of identifying speech generation models when training and inference languages differ. The dataset, protocol and code are available at https://github.com/xuanxixi/Multilingual-Source-Tracing.

Index Terms: Source Tracing, Speech Deepfakes, Cross-Lingual generalization, Linguistic Diversity, Unseen Speaker

1. Introduction

Recent advances in Generative AI (GenAI) have resulted in an unprecedented surge in synthetic data generation. The National Security Agency (NSA), Federal Bureau of Investigation (FBI), and Department of Homeland Security (DHS) recently released a joint report[1]. It warns that synthetic media, especially deepfake content, is now spreading quickly across many languages worldwide. This warning comes at a time when generative AI has made strong progress. Voice synthesis tools, including text-to-speech (TTS) [1] and voice conversion (VC) [2], can now create very natural speech from just a few seconds of audio [3]. These tools help with positive applications like virtual assistants. However, they can also be used to create harmful deepfake speech in several languages. These harmful use cases include phone scams, disinformation and defaming campaigns, and spoofing voice biometric systems to mention a few [4–8].

To address the growing threats of deepfake speech, international challenges like ASVspoof [9, 10] and audio deepfake detection (ADD) [11, 12] promote development of new defense methods to detect deepfake speech [13–15]. Some of these detectors are already very accurate, reaching equal error rates (EERs) below 0.5% on ASVspoof19 [16]. But checking whether the audio is real or fake is insufficient; it is imperative to trace the source of a deepfake speech sample—who created it (or, more practically, *which generative architecture is a likely origin of the deepfake speech sample*). In forensic science, source tracing (ST) of evidentiary materials, including telephone recordings [17], digital images [18], and text documents [19], has been extensively studied for law enforcement applications. Prior research from audio forensics has established methodologies for tracing microphones [20, 21], cell phones [22,23], and caller networks [24,25]. While audio forensic source analysis is a well-established field, ST for speech deepfakes has emerged as a new and pressing research challenge, receiving thus far relatively little attention.

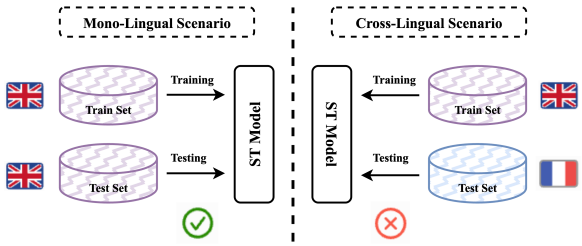

Figure 1: *Illustration of mono-lingual and cross-lingual scenarios on source tracing systems. The English data-trained system works well with English data (left) but fails with other languages (right).*

Some early studies on ST give promising results. For example, [26] compared different ST system designs (end-to-end or two-stages), and [27] proposed a general ST framework which was later extended to a benchmark for neural-codec models [28]. However, these studies are all limited to a single language (specifically, English) for training and testing. They do not study what happens when the model is trained in one language but tested in another. As illustrated in Figure 1, the language mismatch effects refer to the decrease in performance resulting from discrepancies between the languages used during training and inference. These effects have been extensively observed since different languages (and language families) differ at lexical, prosodic, and phonotactic levels – any data-driven model trained on one language only is expected to be overfit to the training language, hindering generalization to new languages. This general problem, called *language mismatch*, is well-known in many tasks including translation [29–31], speaker anonymization [32], and speaker verification [33–38]. Its impact on ST, however, remains thus far unknown. At the same time, audio language models are growing fast. They now support thousands of languages [39]. With only a short voice sample, they can create high-quality speech in any target language [40]. This makes it easier to create fake speech that can

[1]https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf

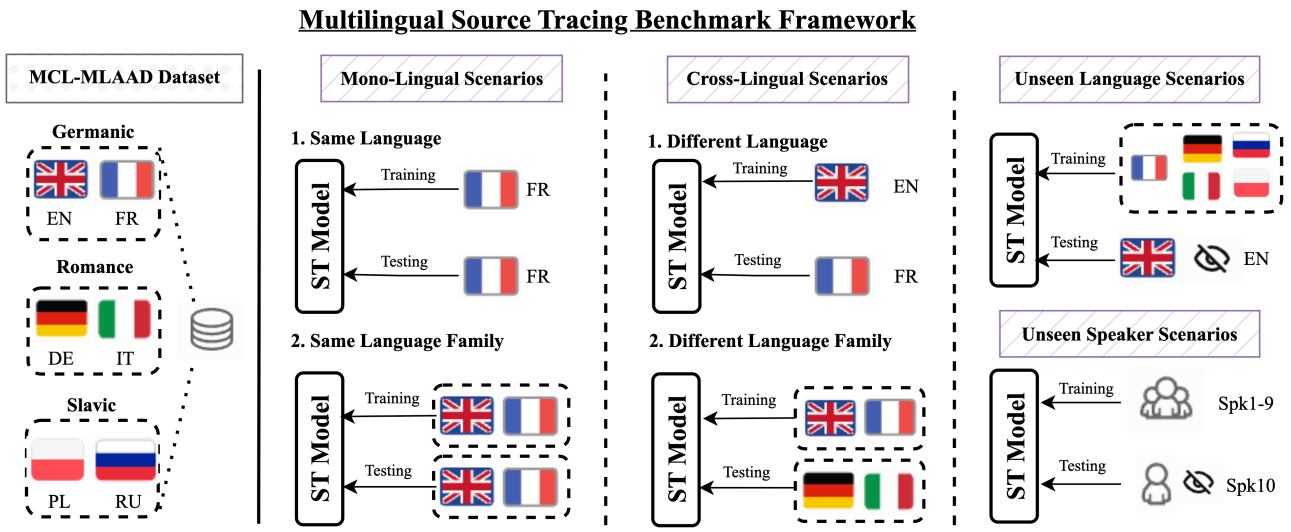

Figure 2: *Overview of the multilingual source tracing (ST) benchmark framework. The framework evaluates model generalization across languages and speakers through four experimental scenarios: (1) Monolingual (same-language/family training and testing); (2) Crosslingual (train on one language/family, test on another language/family); (3) Unseen languages generalization (train on multiple languages, test on unseen language); (4) Unseen speaker generalization (train on multiple languages and speakers, test on unseen speakers across multiple languages).*

escape current detection tools. While some new datasets include many languages [41–43], they all focus on the detection task, not tracing which model made them. Our study seeks to fill this knowledge gap.

Recent related studies have advanced multilingual deepfake *detection*. For instance, [44] studied how multilingual pretrained models can detect speech deepfakes across accents and tones, whereas [45] transferred detection knowledge across languages using domain adaptation methods. The study in [46], in turn, proposed accent-based data expansion to reduce the effects of language mismatch [46]. Finally, [47] tested multilingual ability of detection models, demonstrating that training only on English limits their effectiveness in other languages. Despite progressing understanding of the impact of language in the context of deepfakes, all the prior studies share a key limitation: they focus on detection, rather than ST.

In contrast, our work shifts the focus from binary detection to ST. This task brings new challenges, especially in cross-lingual conditions where the model must trace the origin of fake speech even when the training and testing languages differ. To address this gap, we propose to first explore a benchmark for ST in both mono- and cross-lingual scenarios, and evaluate it across multiple feature types, language settings, and low-resource conditions. **We introduce the first multilingual benchmark for ST of speech deepfakes, covering both mono-lingual and cross-lingual scenarios. Furthermore, this work is the first to explore the effects of unseen languages and speakers in ST tasks**. The speech data itself is drawn from Multi-Language Audio Anti-Spoofing (MLAAD) dataset [43] used in multiple recent studies, including studies under Interspeech 2025 special session *Source tracing: The origins of synthetic or manipulated speech*. Despite consisting of an extensive collection of deepfake samples in multiple languages, MLAAD does not contain an evaluation protocol to address ST in both mono-lingual and cross-lingual scenarios. Our new benchmark is intended as a reproducible starting point to foster further work in performance assessment and improving multilingual ST models.

Our immediate scientific aim is to better understand how language differences and resource limitations influence ST performance. We address seven key research questions (RQs):

RQ1 What is the performance of ST models in monolingual scenarios?

RQ2 How effectively do ST models generalize to cross-lingual scenarios?

RQ3 Does training and testing within the same language family improve cross-lingual performance?

RQ4 How do digital signal processing (DSP) and self-supervised learning (SSL)-based ST models compare in cross-lingual generalization?

RQ5 How do training strategies and linguistic diversity of pre-training corpora affect the cross-lingual generalization performance of SSL-based ST models?

RQ6 How do unseen languages impact ST models generalization?

RQ7 How do unseen speakers impact ST models robustness against shortcut learning?

By answering these questions, our work offers the first systematic evaluation of multilingual ST and lays a foundation for future research in this area.

2. Multilingual Source Tracing Benchmark

As shown in Figure 2, to comprehensively and fair evaluate Multilingual ST model performance in both **M**ono- and **C**ross-**L**ingual scenarios, we propose the linguistically-balanced dataset **MCL**-MLAAD and establish novel protocols for thorough assessment of generalization capabilities.

2.1. Dataset

Accurate evaluation of Multil-lingual ST models needs a dataset with balanced language and TTS synthesizer distribution. For this, we present the Lingual-Balanced MLAAD dataset. It is an improved version of the original MLAAD corpus [43][2], which contains 420.7 hours of synthetic speech in 38 languages, pro-

[2] https://deepfake-total.com/sourcetracing

duced by 91 TTS models (including 32 types of architectures). This data was created using the M-AILABS multilingual audio source[3]. However, the original corpus has two limitations. First, no single TTS model covers all 38 languages. Second, each language does not include all the 91 TTS models. These limitations hinder multilingual generalization studies.

To address this, we built a refined version of MLAAD v5, referred to as MCL-MLAAD. We select six languages from three language families: Germanic (English, German), Romance (French, Italian), and Slavic (Polish, Russian). We also include four popular TTS architectures: Griffin-Lim, Bark, XTTS v1.1, and XTTS v2. Our dataset design aligns with the partitioning methods in [47]. To simulate diverse acoustic environments, the dataset includes four types of noise perturbations—namely, noise, music, babble, and reverberation from MUSAN [48]—systematically applied to each clean utterance. Thus, five acoustic variants (the clean original and its four perturbed counterparts) are generated per utterance, enhancing real-world robustness. Due to uneven spoofing attack distribution, we partitioned the data into train, dev, and test sets with a 60:20:20 ratio for each language as detailed in Table 1.

Table 1: *Dataset Statistics for Different Languages*

Language Family	Language	Code	Samples (×5)	Dur (×5)
Germanic	English	en	2,100	4h 37m 59.59s
	German	de	2,100	4h 35m 7.89s
Romance	French	fr	2,100	5h 6m 38.35s
	Italian	it	2,100	4h 14m 4.82s
Slavic	Polish	pl	2,100	5h 12m 4.68s
	Russian	ru	1,200	3h 11m 16.03s
Total	-	-	**11,700**	**26h 57m 11s**

2.2. Protocols

Our benchmark evaluates source tracing capabilities through four experimental protocols designed to isolate critical dimensions of multilingual generalization:

2.2.1. Mono- & Cross-Lingual Protocol

We trained ST models using only one language each: English, German, French, Italian, Polish, or Russian. These models help us study how the model performs within a single language. Later, we also use them for cross-lingual experiments by testing them on different languages.

2.2.2. Mono- & Cross Language Family Protocol

To explore how language families affect performance, we grouped the six languages into three families: Germanic, Romance, and Slavic. We then trained one model per family. These models help examine whether training on one language group improves performance on related languages.

2.2.3. Seen & Unseen Languages Protocol

To investigate the generalization capability for unseen languages, we employ a leave-one-language-out experimental protocol. Our work explicitly contrasts "seen" and "unseen" language conditions and rigorously analyzes how pre-training data affects cross-lingual robustness.

[3] https://huggingface.co/datasets/mueller91/MLAAD

2.2.4. Seen & Unseen Speakers Protocol

Besides content-related variability (which manifests as phonemic differences between languages), speaker-related factors can be expected to impact ST performance. For instance, if training data for TTS method A represents only one voice identity 1 while training data for TTS method B represents only voice identity 2, a model may learn to differentiate the two speakers, as opposed to the two sources—an instance of *shortcut learning* [49]. To analyze the impact of seen vs. unseen speakers, we use a leave-one-language-out protocol: the model is trained on all but one language, which serves as an unseen test language.

This analysis brings up some new challenges. As opposed to language labels, the MLAAD metadata does not contain speaker labels. Moreover, arguably synthetic speech does not even *have* crisply-defined speaker identity (only targeted speaker identity). These necessitate approximate, *pseudo-speaker* labels that we obtain through an approach similar to [26, Section 3.3]. We use an off-the-shelf[4] ECAPA2 [50] model to extract speaker embeddings from each synthetic utterance, followed by clustering the resulting $11,700$ embeddings using spherical k-means [51], a method suitable for clustering length-normalized $d = 192$ dimensional embeddings. We first run 10 repeats (each with different random initialization) of spherical k-means for each cluster count in $k \in [1, 100]$ and use an 'elbow' criterion [52] to set the number of clusters. We then repeat clustering with the selected cluster count ($k = 18$) using 100 restarts. The cluster assignments of speaker embeddings give us unique pseudo-speaker label per utterance.

Table 2: *Protocol statistics for speaker effect analysis.*

Unseen Language	Utterance Count		Threshold
	Seen spk.	Unseen spk.	
English	938	982	0.085
German	1,029	891	0.075
French	908	1,012	0.079
Italian	919	1,001	0.077
Polish	1,026	894	0.077
Russian	1,064	1,036	0.074

A pseudo-speaker i is defined as *seen* if the empirical speaker prior in the combined training and development data $P_i \equiv N_i^{\text{tra+val}}/N_{\text{total}}^{\text{tra+val}}$ exceeds threshold θ, *and* speaker i is present in the test data. Here, $N_i^{\text{tra+val}}$ and $N_{\text{total}}^{\text{tra+val}}$ denote speaker-specific and total training-development samples respectively. Here, θ mitigates the inherent trade-off between strict speaker exclusion (which causes severe data scarcity) and complete inclusion (which induces class imbalance), ensuring statistically viable group comparisons. We balanced the number of test samples between the seen and unseen groups. The per-language seen/unseen utterance counts and thresholds summarized in Table 2.

3. Multilingual Source Tracing Methods

3.1. ST Models

This section introduces the models used for ST tasks. We first study how different input features influence performance

[4] https://huggingface.co/Jenthe/ECAPA2

in both mono-lingual and cross-lingual settings. Based on the type of front-end feature, we group the models into two categories: DSP-based and SSL-based. The architectural diagrams of both model types are presented in Figure 3.

3.1.1. Models with DSP front-end

We developed three models using classic digital signal processing features: LFCC-ResNet18, LFCC-AASIST and LFCC-ECAPA-TDNN. LFCCs (linear frequency cepstral coefficients) were used in early synthetic speech detectors [53]. We consider three different backends: (1) AASIST [54] is an audio anti-spoofing method using integrated spectro-temporal graph attention networks; (2) ResNet18 [55] is designed to learn local spectral patterns using residual blocks; (3) ECAPA-TDNN [56] focuses on capturing global information using attention mechanisms and multi-scale features.

3.1.2. Models with SSL front-end

We also develop eight models that use self-supervised learning (SSL) front-ends with a shared back-end, AASIST [54]. These front-ends include: (1) two foundation models—XLS-R-300M [57], trained on multilingual data [58–62], and wav2vec2.0 Large LV-60 [63], trained on English [64]; (2) six versions of XLS-R fine-tuned on specific languages. The XLS-R builds on wav2vec2.0 by enabling cross-lingual learning. It does this through shared quantization over encoded features, allowing different languages to share acoustic representations [65]. The front-end details are summarized in Table 3.

Table 3: *Evaluation of self-supervised representations for ST. All SSL models have 300M parameters. Base models include wav2vec2.0 Large LV-60 and XLS-R-300M. Language-specific fine-tuned variants are based on large-xlsr-53, trained on six languages (en, de, fr, it, pl, ru). The datasets abbreviations are: Librispeech (LL) [64], CommonVoice (CV) [60], BABEL (BBL) [62], multilingual Librispeech (MLS) [59], VoxPopuli (VP) [58], and VoxLingua107 (VL) [61].*

Name		Pretraining		Fine-tuning	Datasets
Model		Dur. (h)	Langs.	Lang.	
wav2vec2					
1	large-lv60	53k	en	–	LL
2	xls-r-300m	436k	many	–	CV, BBL, MLS, VP, VL
Fine-tuned variants					
3	large-xlsr-53-en	56k	many	en	CV-en
4	large-xlsr-53-de	56k	many	de	CV-de
5	large-xlsr-53-fr	56k	many	fr	CV-fr
6	large-xlsr-53-it	56k	many	it	CV-it
7	large-xlsr-53-pl	56k	many	pl	CV-pl
8	large-xlsr-53-ru	56k	many	ru	CV-ru

3.2. Implementation details

All audio samples were downsampled to 16 kHz and trimmed or padded to 4 seconds (64,000 samples). Multiclass cross-entropy loss was used to train all the models with a batch size of 16, and 50 epochs. We used an initial learning rate of 5×10^{-4}, and selected the final model based on the lowest development loss.

3.2.1. Models with DSP front-end

LFCCs are extracted using 20ms window and 10ms shift, producing feature matrices of shape (80, 399), where 80 is the number of LFCCs and 399 is the number of frames. We implemented two classifiers: ResNet18[5] and ECAPA-TDNN[6].

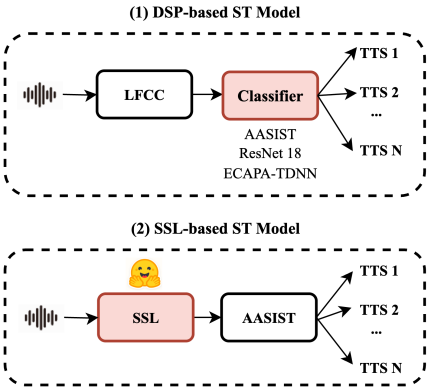

Figure 3: *Architecture Comparison of DSP-based and SSL-based Source Tracing (ST) Models. Red components represent variable components that can be replaced or adjusted in different scenarios, while white components represent fixed components that remain consistent across the models.*

3.2.2. Models with SSL front-end

Each SSL front-end converts raw audio into a matrix of size 199×1024, where 199 is the time frame and 1024 is the feature dimension. These were linearly projected to 128 dimensions before being passed to the AASIST classifier[7]. During training, we used Mixup [66, 67], a method that blends two samples and their labels. The mixing ratio λ is drawn from a Beta distribution with parameters $\alpha = 0.5$. This process keeps the input size unchanged and adds a label smoothing effect, which helps improve model generalization.

3.2.3. Metrics

Experimental results were quantified using (Macro-averaged) Macro-F1 metric: Macro-F1 is defined as

$$\text{Macro-F1} = \frac{2 \cdot \overline{P} \cdot \overline{R}}{\overline{P} + \overline{R}},$$

where $\overline{P}$ and $\overline{R}$ denote the average precision and recall across all classes (synthesizer). Precision and recall are computed per-class as

$$P = \frac{TP}{TP + FP} \quad \text{and} \quad R = \frac{TP}{TP + FN},$$

respectively, with TP, FP, and FN representing true positives, false positives, and false negatives. This aligns with evaluation metrics from recent studies [26].

[5] https://github.com/hubert10/ResNet18_from_Scratch_using_PyTorch/blob/main/resnet18.py

[6] https://github.com/TaoRuijie/ECAPA-TDNN/blob/main/model.py

[7] https://github.com/clovaai/aasist/blob/main/models/AASIST.py

4. Results and Discussion

4.1. Mono-Lingual Performance (RQ1)

As shown in the diagonal entries of Table 4, monolingual performance demonstrates that W2V2(*xx*)-AASIST achieves highest macro-F1 score of 97.91%, indicating that language-specific fine-tuning enhance phonetic differentiation. Notably, LFCC-ECAPA-TDNN attains 97.78% (18.98% higher than LFCC-AASIST), indicating its back-end could more effectively capture subtle artifacts.

4.2. Cross-Lingual Transfer Performance (RQ2)

As shown in the off-diagonal entries of Table 4, LFCC-ECAPA-TDNN achieved optimal cross-lingual performance (88.40%), exceeding LFCC-AASIST by 33.94%, suggesting its back-end better handle phonological variations. W2V2EN-AASIST showed strong English→other transfers (average 95.76%) but lower performance for non-English pairs, reflecting English-only pretrained SSL bias. Low-resource language transfers exhibit significant performance degradation compared to high-resource pairs, highlighting persistent challenges in modeling typologically distant languages.

4.3. Language Family Effects (RQ3)

As shown in Figure 4, cross-family performance analysis reveals consistent advantages for monofamily transfers compared to cross-family settings across all architectures. DSP models demonstrate superior robustness, maintaining minimal performance variance between language pairs while achieving strong cross-family generalization. In contrast, XLSR-AASIST exhibits significant performance degradation under cross-family conditions despite comparable effectiveness in monofamily scenarios, highlighting its heightened sensitivity to linguistic distance. Notably, DSP models preserve near-optimal performance for typologically distant language groups, whereas XLSR-AASIST shows pronounced disparities, suggesting fundamental differences in cross-linguistic divergence modeling.

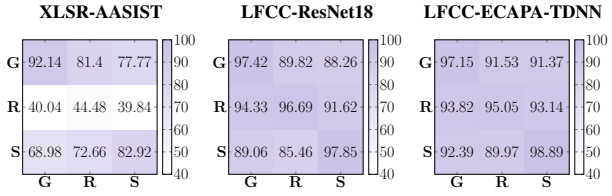

Figure 4: *Heatmaps illustrating the Macro F1-score (%) comparison across XLSR-AASIST, LFCC-ResNet18, and LFCC-ECAPA-TDNN models. Rows indicate source language families used for training (**G: Germanic**, **R: Romance**, **S: Slavic**). Columns indicate target language families used for evaluation.*

4.4. Comparison between DSP and SSL Models (RQ4)

As shown in the off-diagonal entries of Table 4, the SSL-based models (top three) and DSP-based models (bottom three) are displayed their cross-lingual performance. When backend architectures are comparable (top four subtables), SSL models English-only pretrained and language-specific fine-tuning SSL front-end demonstrate competitive cross-lingual performance, suggesting that domain adaptation strategies can mitigate inherent language biases. However, DSP-based methods with robust backend ECAPA-TDNN exhibit superior cross-lingual stability, particularly under low-resource conditions. These findings

Table 4: *Performance comparison (macro F1-scores, %) of SSL-based (top three) and DSP-based (bottom three) models across six languages, with the highest macro F1-scores for monolingual (mono) and cross-lingual (cross) settings underlined. SSL-based models: XLSR denotes XLS-R-300M (multilingual pretrained), W2V2EN denotes wav2vec2.0 Large LV-60 (English-only pretrained), and W2V2 (xx) denotes XLS-R (language-specific fine-tuned, language code: xx). Cells show source **(S)**→target **(T)** language performance, with diagonals indicating monolingual results and off-diagonals cross-lingual transfer.*

S\T	en	de	fr	it	pl	ru
XLSR-AASIST (mono:80.05; cross:59.75)						
en	**93.55**	91.89	78.09	73.21	81.11	64.95
de	80.46	**88.53**	74.78	75.17	83.28	50.11
fr	33.94	35.78	**41.43**	36.95	41.19	25.42
it	76.00	63.55	75.17	**82.08**	70.05	59.25
pl	51.36	53.12	68.31	52.23	**92.64**	39.45
ru	54.10	46.26	47.42	53.98	55.82	**82.07**
W2V2EN-AASIST (mono:92.84; cross:78.18)						
en	**99.51**	98.64	95.06	93.09	96.61	91.63
de	90.95	**66.21**	89.76	64.75	67.03	71.47
fr	93.68	92.74	**98.62**	90.40	98.42	90.17
it	96.64	96.13	93.75	**97.35**	97.17	94.09
pl	50.08	60.70	76.30	57.39	**98.93**	57.62
ru	43.59	36.72	50.28	61.04	49.45	**96.40**
W2V2 (*xx*)-AASIST (mono:97.91; cross:73.45)						
en	**99.12**	97.85	93.66	95.22	95.76	91.06
de	92.63	**96.37**	90.93	83.58	95.03	70.95
fr	88.29	91.91	**97.36**	87.57	97.03	88.12
it	94.45	88.27	92.53	**97.53**	94.85	89.33
pl	43.90	54.43	66.69	44.78	**99.16**	35.31
ru	47.35	37.63	59.06	60.85	62.40	**97.91**
LFCC-AASIST (mono:78.80; cross:54.46)						
en	**82.58**	62.71	57.85	59.53	64.49	54.28
de	56.51	**74.33**	60.50	57.15	72.57	49.25
fr	58.35	62.12	**78.40**	62.28	73.17	54.79
it	65.48	59.42	55.90	**73.56**	65.58	58.48
pl	42.53	54.35	55.63	49.01	**87.18**	43.84
ru	32.30	30.26	32.48	37.87	45.25	**76.76**
LFCC-ResNet18 (mono:95.76; cross:79.23)						
en	**97.32**	87.00	79.49	81.13	88.92	64.04
de	93.60	**96.44**	89.99	85.52	93.78	71.86
fr	90.32	87.43	**94.86**	84.22	90.34	72.08
it	91.98	87.36	86.29	**92.74**	91.12	82.79
pl	80.53	84.56	79.76	70.83	**97.37**	55.33
ru	65.91	56.71	56.73	62.84	64.34	**95.82**
LFCC-ECAPA-TDNN (mono:97.78; cross:88.40)						
en	**98.42**	94.67	92.44	89.82	94.87	78.13
de	95.95	**98.03**	95.73	93.53	96.32	89.26
fr	88.86	94.65	**96.59**	89.78	94.06	92.87
it	96.45	96.91	95.63	**97.19**	97.19	84.69
pl	79.26	87.46	81.04	72.73	**98.76**	60.01
ru	83.73	84.66	84.45	81.43	85.48	**97.98**

Table 5: *Macro-Averaged F1 Scores (↑) under Leave-One-Language-Out Setting for Seen and Unseen Languages; Method A: XLSR-AASIST; Method B: LFCC-ResNet18; Method C: LFCC-ECAPA-TDNN.*

Method	Seen Languages							Unseen Languages						
	-en	-de	-fr	-it	-pl	-ru	Avg	en	de	fr	it	pl	ru	Avg
A	54.50	72.09	95.73	50.18	82.52	48.83	67.31	49.61	72.46	94.39	54.12	85.40	25.35	63.56
B	**97.75**	**97.49**	97.68	**98.21**	**97.75**	95.75	**97.44**	96.80	**97.06**	95.55	93.00	97.88	**99.20**	**96.58**
C	97.62	94.99	**97.96**	97.53	97.45	**98.26**	97.30	**96.82**	96.75	**97.16**	**94.92**	**98.62**	91.91	96.03

Table 6: *Macro-averaged F1 scores (↑) for three methods on Unseen Languages, evaluated under Seen and Unseen Speaker conditions. Method A: XLSR-AASIST, Method B: LFCC-ResNet18, Method C: LFCC-ECAPA-TDNN.*

Method	Seen Speakers							Unseen Speakers						
	en	de	fr	it	pl	ru	Avg	en	de	fr	it	pl	ru	Avg
A	54.66	52.69	96.19	40.62	84.70	48.11	62.83	54.63	75.78	93.97	49.88	78.91	49.13	67.05
B	98.04	**86.76**	97.64	**93.39**	97.59	98.16	**95.26**	**96.84**	**94.69**	97.65	**97.45**	**97.01**	**98.35**	**97.00**
C	**98.67**	80.63	**98.33**	90.53	**98.23**	**98.42**	94.14	96.82	91.73	**98.37**	96.12	96.61	97.71	96.23

indicate that while SSL models benefit from language-specific fine-tuning to bridge linguistic gaps, DSP architectures inherently prioritize language-agnostic patterns through their signal processing pipelines, offering a more resilient framework for cross-lingual source tracing when paired with advanced back-end designs.

4.5. Training Strategy Effects (RQ5)

As shown in Table 4, this section compares the SSL-based models (top three) and LFCC-based models (bottom three) under varying training strategies to assess their cross-lingual generalization capabilities. Among SSL models, multilingual SSL pretraining (XLSR-AASIST) exhibits lower monolingual performance and weaker cross-lingual transfer capabilities, particularly in low-resource settings. In contrast, English-only SSL pretraining (W2V2EN-AASIST) achieves stronger monolingual performance and moderate cross-lingual effectiveness but retains asymmetric transfer biases between language pairs. Language-specific fine-tuning (W2V2(xx)-AASIST) further enhances monolingual results while showing limited improvement in cross-lingual scenarios. Conversely, LFCC-ECAPA-TDNN demonstrate stronger cross-lingual robustness with comparable monolingual performance, maintaining stable accuracy across both high- and low-resource language groups. These findings indicate that (1) SSL models heavily depend on pretraining language coverage and fine-tuning strategies, whereas (2) LFCC-based approaches inherently prioritize language-agnostic acoustic features, enabling more reliable generalization across linguistic and resource-diverse conditions.

4.6. Unseen Languages Generalization Experiment (RQ6)

As shown in Table 5, the leave-one-language-out evaluation demonstrates generalization capabilities to unseen languages. The results reveal that XLSR-AASIST exhibits notable performance degradation on unseen languages, highlighting limitations due to its reliance on language-specific pretraining data. Furthermore, a trade-off between local and global modeling is observed: LFCC-ResNet18 effectively preserves local phoneme boundaries but struggles with cross-lingual prosody modeling, whereas LFCC-ECAPA-TDNN aggregates multi-scale temporal features, capturing both local articulatory details and global prosodic features through hierarchical modeling.

4.7. Unseen Speakers Generalization Experiment (RQ7)

Finally, Table 6 reports the seen/unseen speaker analysis following the leave-one-language out setup with pseudo-speaker labels, broken down according to the held-out language and the three models. The findings related to the six languages and the three models are in line with the previous analyses. As for the relative performance for seen/unseen speakers, unlike was hypothesized, no apparent trends are visible—the results are dependent both on the held-out language and the model. While this may suggest that the investigated models can be robust to speaker factors, the number of speakers is low and the speaker labels were derived through clustering process. Another future study with larger number speakers and known target speaker labels is needed to validate these preliminary findings.

5. Conclusion

In this work, we establish the first multilingual benchmark for speech deepfake source tracing, covering both monolingual and cross-lingual scenarios across six languages and two model categories (DSP- and SSL-based models). Furthermore, we first explore the effects of unseen languages and speakers in ST tasks. Our findings reveal three key insights: First, in monolingual scenarios, SSL front-ends fine-tuned on language-specific data outperform both multilingual/English-only pretrained SSL front-ends and LFCC front-ends. Second, LFCC features combined with ResNet or ECAPA-TDNN backends demonstrate superior cross-lingual generalization. Third, while cross-lingual generalization is stronger within the same language family, significant performance variations persist across language pairs.

We also explored the impact of speaker variability, finding no consistent performance gap between seen and unseen pseudo-speakers, though results fluctuated across held-out languages and model types. This indicates a potential robustness to speaker variation, yet also highlights current limitations, including reliance on unsupervised speaker clustering and the lack of ground-truth speaker labels. As a future direction, we encourage further validation using larger, labeled datasets to better understand the interplay between language, speaker, and model-specific factors in deepfake attribution tasks. Our benchmark aims to establish a foundation for this emerging field and promote further research into multilingual, speaker-aware, and model-specific audio deepfake forensics.

6. References

[1] W. Liu, J. Bai, X. Cheng, J. Zuo, Z. Jiang, S. Ji, M. Fang, X. Yang, Q. Yang, and Z. Zhao, "Voxpopulitts: a large-scale multilingual tts corpus for zero-shot speech generation," in *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, pp. 10 293–10 297.

[2] J. Yao, Y. Yang, Y. Lei *et al.*, "Promptvc: Flexible stylistic voice conversion in latent space driven by natural language prompts," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 571–10 575.

[3] M. Li, Y. Ahmadiadli, and X. P. Zhang, "A survey on speech deepfake detection," *ACM Computing Surveys*, vol. 58, no. 3, pp. 1–42, 2025, just Accepted.

[4] S. V. J. Kolupuri, A. Paul, R. S. Bhowmick, and et al., "Scams and frauds in the digital age: Ml-based detection and prevention strategies," in *Proceedings of the 26th International Conference on Distributed Computing and Networking*, 2025, pp. 340–345.

[5] W. Zhang and C. Luo, "Ge-gnn: Gated edge-augmented graph neural network for fraud detection," *IEEE Transactions on Big Data*, vol. 11, no. 4, pp. 1664–1676, 2025.

[6] B. Ding, R. Han, Z. Ma, and X. Xuan, "Crowd density estimation based on multi-level attention maps," in *2021 IEEE 5th Information Technology,Networking,Electronic and Automation Control Conference (ITNEC)*, vol. 5, 2021, pp. 1759–1765.

[7] W. Zhang, D. Xu, G. Yao, X. Lin, R. Guan, C. Du, R. Han, X. Xuan, and C. Luo, "Frect: Frequency-augmented convolutional transformer for robust time series anomaly detection," in *Advanced Intelligent Computing Technology and Applications*, D.-S. Huang, W. Chen, Y. Pan, and H. Chen, Eds. Singapore: Springer Nature Singapore, 2025, pp. 15–26.

[8] W. Zhang and C. Luo, "Decomposition-based multi-scale transformer framework for time series anomaly detection," *Neural Networks*, vol. 187, p. 107399, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608025002783

[9] X. Liu, X. Wang, M. Sahidullah *et al.*, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2023.

[10] X. Wang, H. Delgado, H. Tak *et al.*, "Asvspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale," in *Proceedings of The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*. ISCA, 2024, pp. 1–8.

[11] J. Yi, R. Fu, J. Tao *et al.*, "Add 2022: The first audio deep synthesis detection challenge," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9216–9220.

[12] J. Yi, J. Tao, R. Fu *et al.*, "Add 2023: The second audio deepfake detection challenge," in *Proceedings of CEUR Workshop*, vol. 3597, 2023, pp. 125–130, iSSN: 1613-0073.

[13] T. P. Doan, H. Dinh-Xuan, T. Ryu *et al.*, "Trident of poseidon: A generalized approach for detecting deepfake voices," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '24. ACM, 2024, pp. 2222–2235.

[14] Y. Xiao and R. K. Das, "Xlsr-mamba: A dual-column bidirectional state space model for spoofing attack detection," *IEEE Signal Processing Letters*, vol. 32, pp. 1276–1280, 2025.

[15] T. Liu, D.-T. Truong, R. K. Das, K. A. Lee, and H. Li, "Nes2net: A lightweight nested architecture for foundation model driven speech anti-spoofing," 2025. [Online]. Available: https://arxiv.org/abs/2504.05657

[16] M. Li and X. P. Zhang, "Interpretable temporal class activation representation for audio spoofing detection," in *Proceedings of Interspeech*, 2024, pp. 1120–1124.

[17] M. Faundez-Zanuy, J. J. Lucena-Molina, and M. Hagmüller, "Speech watermarking: An approach for the forensic analysis of digital telephonic recordings," *Journal of Forensic Sciences*, vol. 55, no. 4, pp. 1080–1087, 2018, impact Factor: 1.5 (2023).

[18] A. Swaminathan, M. Wu, and K. J. R. Liu, "Digital image forensics via intrinsic fingerprints," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 1, pp. 101–117, March 2008.

[19] T. Richter, S. Escher, D. Schönfeld *et al.*, "Forensic analysis and anonymization of printed documents," in *Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security*. ACM, 2018, pp. 127–138.

[20] M. Qamhan, Y. A. Alotaibi, and S. A. Selouani, "Transformer for authenticating the source microphone in digital audio forensics," *Forensic Science International: Digital Investigation*, vol. 45, p. 301539, 2023.

[21] Z. Lin, J. Wang, R. Li, F. Shen, and X. Xuan, "Primek-net: Multi-scale spectral learning via group prime-kernel convolutional neural networks for single channel speech enhancement," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[22] W. Jansen and R. Ayers, "Guidelines on cell phone forensics," National Institute of Standards and Technology (NIST), NIST Special Publication 800-101, 2007, nIST SP 800-101 Rev. 1. [Online]. Available: https://csrc.nist.gov/publications/detail/sp/800-101/rev-1/final

[23] C. Hanilçi and T. Kinnunen, "Source cell-phone recognition from recorded speech using non-speech segments," *Digital Signal Processing*, vol. 35, pp. 75–85, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1051200414002565

[24] S. Catanese, E. Ferrara, and G. Fiumara, "Forensic analysis of phone call networks," *Social Network Analysis and Mining*, vol. 3, no. 1, pp. 15–33, March 2013.

[25] W. Zhang, D. Xu, X. Xuan, L. Jiang, G. Yao, R. Han, X. Lang, and C. Luo, "Addressing noise and stochasticity in fraud detection for service networks," 2025. [Online]. Available: https://arxiv.org/abs/2505.00946

[26] N. Klein, T. Chen, H. Tak *et al.*, "Source tracing of audio deepfake systems," in *Proceedings of Interspeech*. ISCA, 2024, pp. 1100–1104.

[27] Y. Xie, R. Fu, Z. Wen *et al.*, "Generalized source tracing: Detecting novel audio deepfake algorithm with real emphasis and fake dispersion strategy," in *Proceedings of Interspeech*. ISCA, 2024, pp. 4833–4837.

[28] Y. Xie, X. Wang, Z. Wang *et al.*, "Neural codec source tracing: Toward comprehensive attribution in open-set condition," *arXiv preprint arXiv:2501.06514*, 2025, preprint.

[29] X. Li, C. Wang, Y. Tang *et al.*, "Multilingual speech translation from efficient finetuning of pretrained models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2021, pp. 827–838.

[30] K. kui Sin, X. Xuan, C. Kit, C. H. yan Chan, and H. H. kin Ip, "Solving the unsolvable: Translating case law in hong kong," 2025. [Online]. Available: https://arxiv.org/abs/2501.09444

[31] X. Xuan, K. kui Sin, Y. Zhou, and C. Kit, "Translaw: Benchmarking large language models in multi-agent simulation of the collaborative translation," 2025. [Online]. Available: https://arxiv.org/abs/2507.00875

[32] S. Meyer, F. Lux, and N. T. Vu, "Probing the feasibility of multilingual speaker anonymization," in *Proceedings of Interspeech*, 2024, pp. 4448–4452.

[33] Z. Song, L. He, P. Wang *et al.*, "Introducing multilingual phonetic information to speaker embedding for speaker verification," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 091–10 095.

[34] X. Xuan, J. Dong, and T. Xuan, "Research on front-end of asv system based on mel spectrum in noise scenario," in *2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, vol. 10, 2022, pp. 2638–2642.

[35] X. Xuan and R. Han, "Research on acoustic feature extractor for automatic speaker verification systerm," in *2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, vol. 10, 2022, pp. 2628–2633.

[36] X. Xuan, R. Jin, T. Xuan, G. Du, and K. Xuan, "Multi-scene robust speaker verification system built on improved ecapa-tdnn," in *2022 IEEE 6th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 2022, pp. 1689–1693.

[37] X. Xuan, R. Han, and B. Ding, "Research on speaker identification models based on cnn and additive angular margin loss," in *2021 2nd International Conference on Electronics, Communications and Information Technology (CECIT)*, 2021, pp. 1046–1050.

[38] X. Xuan, R. Han, and J. Gao, "Conformer-based speaker recognition model for real-time multi-scenarios," *Computer Engineering and Applications*, vol. 60, no. 7, pp. 147–156, 2024.

[39] V. Pratap, A. Tjandra, B. Shi *et al.*, "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.

[40] S. Chen, C. Wang, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[41] Y. Hou, H. Fu, C. Chen *et al.*, "Polyglotfake: A novel multilingual and multimodal deepfake dataset," in *International Conference on Pattern Recognition*. Cham: Springer Nature Switzerland, 2024, pp. 180–193.

[42] X. Qi, H. Gu, J. Yi *et al.*, "Madd: A multi-lingual multi-speaker audio deepfake detection dataset," in *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2024, pp. 466–470.

[43] N. M. Müller, P. Kawa, W. H. Choong *et al.*, "Mlaad: The multi-language audio anti-spoofing dataset," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–7.

[44] O. C. Phukan, G. Kashyap, A. B. Buduru *et al.*, "Heterogeneity over homogeneity: Investigating multilingual speech pre-trained models for detecting audio deepfake," in *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024, pp. 2496–2506.

[45] Z. Ba, Q. Wen, P. Cheng, Y. Wang, F. Lin, L. Lu, and Z. Liu, "Transferring audio deepfake detection capability across languages," in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 2033–2044.

[46] T. Liu, I. Kukanov, Z. Pan, Q. Wang, H. B. Sailor, and K. A. Lee, "Towards quantifying and reducing language mismatch effects in cross-lingual speech anti-spoofing," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 1185–1192.

[47] B. Marek, P. Kawa, and P. Syga, "Are audio deepfake detection models polyglots?" *arXiv preprint arXiv:2412.17924*, 2024.

[48] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[49] R. Geirhos *et al.*, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020.

[50] J. Thienpondt and K. Demuynck, "Ecapa2: A hybrid neural network architecture and training strategy for robust speaker embeddings," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[51] I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Machine Learning*, vol. 42, no. 1, pp. 143–175, 2001. [Online]. Available: https://doi.org/10.1023/A:1007612920971

[52] T. M. Kodinariya and P. R. Makwana, "Review on determining number of cluster in k-means clustering," *International Journal*, vol. 1, no. 6, pp. 90–95, 2013.

[53] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proceedings of Interspeech*, 2015, pp. 2087–2091.

[54] J. Jung, H. S. Heo, H. Tak *et al.*, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6367–6371.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[56] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proceedings of Interspeech*, 2020, pp. 3830–3834, iSSN 2308-457X.

[57] A. Conneau, A. Baevski, R. Collobert, and A. Mohamed, "Unsupervised cross-lingual representation learning for speech recognition," in *Proceedings of Interspeech*, 2021, pp. 2426–2430.

[58] C. Wang, M. Riviere, A. Lee *et al.*, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 993–1003.

[59] V. Pratap, Q. Xu, A. Sriram *et al.*, "MLS: A large-scale multilingual dataset for speech research," in *Proceedings of Interspeech*, 2020, pp. 2757–2761.

[60] R. Ardila, M. Branson, K. Davis *et al.*, "Common Voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.

[61] J. Valk and T. Alumäe, "Voxlingua107: A dataset for spoken language recognition," in *Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 652–658.

[62] M. J. F. Gales, K. M. Knill, A. Ragni *et al.*, "Speech recognition and keyword spotting for low-resource languages: Babel project research at cued," in *Proceedings of the Fourth International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU-2014)*. International Speech Communication Association (ISCA), 2014, pp. 16–23.

[63] A. Baevski, Y. Zhou, A. Mohamed, and et al., "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.

[64] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[65] O. Pascu, A. Stan, D. Oneata *et al.*, "Towards generalisable and calibrated audio deepfake detection with self-supervised representations," in *Proceedings of Interspeech*, 2024, pp. 4828–4832.

[66] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: https://arxiv.org/abs/1710.09412

[67] A. Cui, C. Zhao, X. Deng, G. Jiang, Y. Yang, G. Yao, R. Han, W. Zhang, and X. Xuan, "Unlocking the full potential of separable convolutions on tensor cores," in *International Conference on Intelligent Computing*. Springer, 2025, pp. 39–50.