



# Language-Conditioned Audio Adaptation for Multilingual Audio-Visual Deepfake Detection

Gurleen Kaur Sihra<sup>1</sup>, Dr. Mohammad Abuzar Ansari<sup>2</sup>

<sup>2</sup>Head of Department

<sup>1,2</sup>Department of Data Science, SIES College of Arts, Science and Commerce (Empowered Autonomous)  
Sion (West), Mumbai, India

<sup>1</sup>gurleenkaurfmds2526134@ascs.sies.edu.in

<sup>2</sup>abuzara@sies.edu.in

## How to Cite this Article:

Sihra, G. K. (2026). Language-Conditioned Audio Adaptation for Multilingual Audio-Visual Deepfake Detection. International Journal of Creative and Open Research in Engineering and Management, 2(10), 1-9.  
<https://doi.org/10.55041/ijcope.v2i10.116>

## License:

This article is published under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

© The Author(s). Published by International Journal of Creative and Open Research in Engineering and Management.



<https://doi.org/10.55041/ijcope.v2i10.116>

**Abstract**—Audio-visual deepfake detectors are increasingly deployed in multilingual settings, yet language-dependent speech variability can alter the behavior of audio branches trained without explicit language conditioning. This paper evaluates a MAVDD-style audio-visual detector on the French, Spanish, and Russian subsets of PolyGlottFake and proposes a lightweight language-conditioned MFCC adapter routed by predictions from a frozen MMS-LID-126 model. The Xception visual branch, the MFCC-CNN audio classifier, and the language identifier remain frozen; only 104 scalar adapter parameters are trained. To prevent source leakage, all manipulated variants derived from the same raw source are assigned to a single split. On the untouched 1,311-video test set, baseline hard-OR fusion obtains 91.15% accuracy, 62.58% balanced accuracy, 94.90% fake recall, and 95.28% fake F1. The language-aware system reaches 94.28% accuracy, 59.30% balanced accuracy, 98.87% fake recall, and 97.02% fake F1. The audio branch alone improves from 73.53% to 90.08% accuracy and from 59.86% to 66.31% ROC-AUC. Of 57 changed final decisions, 49 are baseline errors corrected and 8 are newly introduced errors; real recall falls from 30.26% to 19.74%. An exact paired McNemar test gives  $p \approx 2.72 \times 10^{-8}$ . We also report that a trivial all-fake predictor attains 94.20% accuracy and 97.01% fake F1 on this test set, so the full-system accuracy gain should be read as increased fake sensitivity rather than improved class separation. Post-hoc Integrated Gradients, adapter-parameter visualization, and Grad-CAM provide qualitative evidence consistent with this trade-off.

**Index Terms**—audio-visual deepfake detection, multilingual deepfake detection, language identification, MFCC, Xception, multimodal learning, language-conditioned adaptation, explainable AI.

## I. INTRODUCTION

Deepfake generation has progressed from isolated face manipulation to multimodal media in which speech, facial motion, or both are synthesized. This motivates detectors that jointly analyze visual and acoustic evidence rather than relying on a single modality. Audio-visual datasets such as FakeAVCeleb [3] and multimodal detectors such as AVoid-DF [4], MultimodalTrace [5], and MIS-AVoiDD [6] demonstrate the value of cross-modal information, while MAVDD provides a practical decision-level architecture combining an Xception visual branch with an MFCC-based audio branch [2].

A second challenge is linguistic variation. Spectral speech representations can encode phonetic, prosodic, and temporal patterns that vary across languages, while synthesis systems may introduce language-dependent artifacts. PolyGlottFake was created specifically to expose multilingual and multimodal deepfake detection challenges, covering seven languages and multiple audio and lip-synchronization techniques [1]. However, a detector that performs well in aggregate may still behave differently across languages or may become oversensitive to the majority fake class.

This work investigates a deliberately narrow question: can explicit language-conditioned processing improve the audio component of an existing audio-visual detector without retraining its visual or audio classifiers? We retain a controlled scope: PolyGlottFake only; French, Spanish, and Russian only;

a MAVDD-style Xception+MFCC detector; and hard-OR decision fusion. The proposed modification is a small affine adapter applied to MFCC coefficients and selected by a frozen language-identification (LID) model.

The contributions are threefold. First, we establish a source-disjoint multilingual baseline across French, Spanish, and Russian. Second, we develop an LID-routed language-conditioned MFCC adapter with only 104 trainable scalars while keeping Xception, the MFCC-CNN, and LID frozen. Third, we compare baseline and proposed systems with class-sensitive metrics, paired significance testing, source-group uncertainty analysis, language- and route-level breakdowns, a trivial all-fake reference, and post-hoc explainability. The central result is deliberately nuanced: the proposed method improves accuracy, fake recall, and fake F1, but decreases balanced accuracy because hard-OR fusion converts additional real videos into false positives.

## II. RELATED WORK

### A. Audio-Visual Deepfake Detection

Multimodal deepfake detection combines evidence from speech and facial imagery. FakeAVCeleb introduced an audio-video benchmark containing authentic and manipulated modality combinations [3]. AVoid-DF learns audio-visual inconsistencies with temporal-spatial encoding and cross-modal classification [4]. MultimodalTrace learns joint audiovisual representations rather than combining only final decisions [5], while MIS-AVoiDD explicitly separates modality-invariant and modality-specific representations to improve fusion [6]. In



contrast, MAVDD uses separate visual and audio classifiers and labels a video real only when both branches predict real [2]. This decision-level design directly supports the present controlled study because the audio branch can be modified without retraining the visual branch.

### B. Multilingual Deepfake Detection

PolyGlottFake was designed to address the limited language coverage of earlier multimodal deepfake datasets [1]. It contains 15,238 videos (766 real and 14,472 fake) across English, French, Spanish, Russian, Chinese, Arabic, and Japanese, with multiple text-to-speech/voice-conversion and lip-synchronization techniques. The present work uses only French, Spanish, and Russian so that language-dependent behavior can be studied while keeping the experimental scope controlled.

### C. Language Identification and Explainability

The language router uses the public MMS-LID-126 checkpoint associated with Meta's Massively Multilingual Speech (MMS) project [8]. MMS extends wav2vec 2.0-style speech modeling [9] to large multilingual coverage, making it suitable as a frozen routing model rather than a trainable component of the detector. For interpretation, Grad-CAM [10] is used for the Xception branch [7] and Integrated Gradients [11] is used for the adapted MFCC branch. These explanations are strictly post-hoc and are not used for training, hyperparameter selection, or test-set model selection.

## III. METHODOLOGY

### A. Dataset and Three-Language Scope

Only PolyGlottFake [1] is used. Real videos are drawn from the fr, es, and ru folders; fake videos are drawn from to\_fr, to\_es, and to\_ru. Initial metadata contained 6,620 selected videos. File-integrity checks found 68 missing fake files (50 French, 7 Spanish, and 11 Russian), leaving 6,552 usable videos: 6,174 fake and 378 real. No missing file was recreated. The final clean language counts are French: 2,195 fake/120 real; Spanish: 2,231 fake/122 real; and Russian: 1,748 fake/136 real. The untouched test set contains 1,311 videos: French 438 fake/24 real, Spanish 447 fake/25 real, and Russian 350 fake/27 real.

TABLE I  
SELECTED POLYGLOTFAKE SUBSET AND SOURCE-DISJOINT SPLIT

| Split      | Videos | Fake  | Real |
|------------|--------|-------|------|
| Train      | 3,931  | 3,704 | 227  |
| Validation | 1,310  | 1,235 | 75   |
| Test       | 1,311  | 1,235 | 76   |
| Total      | 6,552  | 6,174 | 378  |

### B. Source-Disjoint Data Partitioning

PolyGlottFake contains multiple manipulated variants derived from the same underlying raw source video. A naive video-level random split can therefore leak source identity and content across training and test sets. We group all fake variants sharing the same raw source and force each group into one split. The resulting source-disjoint partition (Table I) contains 3,931 training, 1,310 validation, and 1,311 test videos; the test split comprises 229 source groups. No source group occurs in more than one split, and all reported final test results are obtained without test-time training or parameter updates.

### C. Baseline MAVDD-Style Detector

The baseline follows the released MAVDD-style Xception+MFCC architecture [2] but is reconstructed to match the released H5 weights used in this study. The Xception backbone [7] is instantiated with weights=None,

include\_top=False, and input size  $128 \times 128 \times 3$ ; no ImageNet initialization is loaded during reconstruction or evaluation. The first 50 backbone layers are marked non-trainable before the released weights are loaded. The classifier head is global average pooling, dropout 0.7, Dense(128, ReLU), dropout 0.7, and Dense(1, sigmoid). Output values are interpreted as P(real), with values above 0.5 classified as real.

Visual preprocessing replicates the evaluation pipeline used for the baseline. OpenCV frames are converted from BGR to RGB and passed to the FER facial-expression library [21]. If multiple faces are detected, the face whose strongest emotion probability is highest is selected. The first detected face is saved, and subsequent crops are saved when the dominant emotion label changes. Crops are resized to  $128 \times 128$  and, at classification time, converted to RGB float arrays and divided by 255.0. The video-level visual decision is obtained by majority voting over crop predictions; ties are classified as fake. Six test videos yield no valid visual prediction.

The audio branch uses the first 5 s of audio resampled to 22,050 Hz. Thirteen mel-frequency cepstral coefficients (MFCCs) [15], extracted with librosa [17], are zero-padded or truncated to 109 time steps, producing a  $13 \times 109 \times 1$  input. The frozen MFCC-CNN consists of Conv2D(64,  $3 \times 3$ , ReLU), max pooling, Conv2D(32,  $3 \times 3$ , ReLU), max pooling, flattening, Dense(128, ReLU), dropout 0.5, Dense(64, ReLU), and Dense(1, sigmoid). Its output is also P(real) with a 0.5 threshold.

MAVDD-style decision fusion is a hard-OR rule [2]. Let  $C_v$  and  $C_a$  denote the visual and audio decisions. A video is real only when both branches are real; otherwise it is fake:

$$C_{\text{final}} = \begin{cases} \text{REAL}, & C_v = \text{REAL and } C_a = \text{REAL}; \\ \text{FAKE}, & \text{otherwise.} \end{cases} \quad (1)$$

For the six test videos with no valid visual prediction, fusion falls back to the available audio decision in both baseline and proposed evaluations. Because the proposed method does not modify face extraction, Xception weights, crop preprocessing, or visual voting, the exact baseline visual predictions are reused in the final comparison.

### D. Frozen Language Identification

For language routing, the same video audio is independently resampled to 16 kHz and passed to the frozen facebook/mms-lid-126 checkpoint, a wav2vec 2.0 sequence-classification model from the MMS family [8], [9]. Predictions fra, spa, and rus route to the French, Spanish, and Russian adapters. Every other predicted language is assigned to an explicit other route. Predictions outside the three target languages are not force-mapped to a target language. The same predicted-routing procedure is used during training, validation, and testing.

### E. Language-Conditioned MFCC Adapter

The proposed component is a per-route affine transformation applied coefficient-wise to the MFCC matrix before the frozen MFCC-CNN. For route  $l$  and MFCC matrix  $X$ , the adapted representation is

$$X' = \gamma_l \odot X + \beta_l, \quad (2)$$

where  $\gamma_l$  and  $\beta_l$  are 13-dimensional learnable scale and shift vectors broadcast over time, and  $\odot$  denotes element-wise multiplication. Four routes (fra, spa, rus, other) produce  $4 \times 13$  scale and  $4 \times 13$  shift parameters, for only 104 trainable scalars. Identity initialization uses  $\gamma = 1$  and  $\beta = 0$ , so the proposed system exactly reproduces the original MFCC input before training. A numerical identity check gives a maximum initial difference of  $8.94 \times 10^{-8}$ , and the frozen MFCC-CNN weights remain unchanged. Fig. 1 summarizes the pipeline.

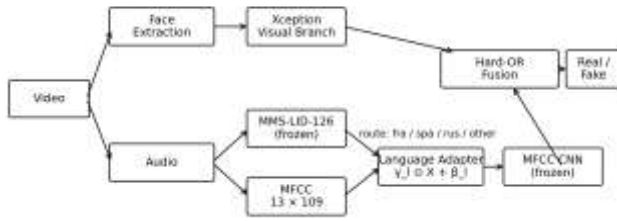


Fig. 1. Proposed language-conditioned MAVDD-style pipeline. Xception, LID, and MFCC-CNN are frozen; only the language adapter is trained.

#### F. Adapter Training

The Xception visual branch, MFCC-CNN, and LID model are frozen throughout adapter training. The 104 adapter parameters are optimized with Adam [16] (learning rate 0.001) for 20 epochs with batch size 16 using the 3,931-video training split and 1,310-video validation split. Validation balanced accuracy is the predefined checkpoint criterion because of severe class imbalance. The best checkpoint occurs at epoch 19 with validation balanced accuracy 0.5998, validation accuracy 0.9183, fake recall 0.9595, fake F1 0.9568, and confusion matrix  $\begin{bmatrix} 18 & 57 \\ 50 & 1185 \end{bmatrix}$  using rows Actual Real/Fake and columns Predicted Real/Fake.

#### G. Evaluation and Statistical Analysis

Primary metrics are accuracy, balanced accuracy [14], fake-class precision, fake-class recall, and fake-class F1; ROC-AUC is additionally reported for continuous branch scores. Fake is treated as the positive class, while balanced accuracy gives equal weight to real and fake recall. The final comparison is paired because baseline and proposed systems are evaluated on the same 1,311 videos. Statistical significance of changed decisions is assessed with the exact McNemar test [12]. To respect source dependence, we also aggregate by the 229 test source groups and perform a source-group sign-flip/randomization test, together with source-group bootstrap [13] confidence intervals for metric differences. Confusion matrices use rows for actual Real/Fake and columns for predicted Real/Fake.

#### H. Post-hoc Explainability

Explainability is used only after model selection and is not used to train, tune, or select the detector. Integrated Gradients [11] is applied to the adapted MFCC representation using a zero-MFCC reference; absolute attribution magnitudes are normalized for visualization and interpreted as local sensitivity rather than causal evidence. Grad-CAM [10] is applied to the exact cached baseline face crops. To avoid saturated sigmoid gradients, the Grad-CAM target is the pre-sigmoid real logit for a real explanation and its negative for a fake explanation. The underlying Xception activation map is  $4 \times 4$  and is therefore spatially coarse. Grad-CAM explains individual face-crop predictions, not the entire video-level majority-vote decision.

### IV. RESULTS

#### A. Language Identification

The frozen LID model achieves 77.23% accuracy on the training split, 77.71% on validation, and 77.80% on test (Table II). Test accuracy is highest for French (88.96%), followed by Russian (72.94%) and Spanish (70.76%). Thus, 1,020/1,311 test videos are routed to the correct target language and 291/1,311 (22.20%) are LID errors. On this test set, every LID error was routed to the explicit other route; no video was confused between French, Spanish, and Russian. The resulting routing distribution (411 fra, 334 spa, 275 rus, and 291 other) therefore equals the per-language correct counts plus the 291 other

predictions. Oracle language labels are never used for routing at test time.

TABLE II  
LID ACCURACY (%)

| Split      | Overall | French | Spanish | Russian |
|------------|---------|--------|---------|---------|
| Train      | 77.23   | 88.34  | 72.17   | 69.91   |
| Validation | 77.71   | 88.36  | 74.20   | 68.97   |
| Test       | 77.80   | 88.96  | 70.76   | 72.94   |

#### B. Baseline Performance

The baseline full detector achieves 91.15% accuracy and 95.28% fake-class F1, but balanced accuracy is only 62.58% because the test set is strongly imbalanced and the hard-OR rule generates many real false positives. The final confusion matrix is  $\begin{bmatrix} 23 & 53 \\ 63 & 1172 \end{bmatrix}$ , corresponding to real recall 30.26% and fake recall 94.90%. Visual-only performance on the 1,305 videos with valid visual predictions is stronger in balanced accuracy and ROC-AUC than audio-only performance, whereas hard-OR fusion raises overall fake sensitivity (Table III).

TABLE III  
BASELINE BRANCH AND FUSION TEST PERFORMANCE (%)

| System         | Acc.  | Bal. Acc. | Fake F1 | ROC-AUC |
|----------------|-------|-----------|---------|---------|
| Visual only*   | 79.46 | 76.75     | 87.98   | 84.47   |
| Audio only     | 73.53 | 55.08     | 84.39   | 59.86   |
| Hard-OR fusion | 91.15 | 62.58     | 95.28   | —       |

\*Visual metrics use 1,305 videos with valid visual predictions.

#### C. Effect of Language-Conditioned Audio Adaptation

The adapted audio branch shows the clearest effect of the adapter. Accuracy rises from 73.53% to 90.08%, fake recall from 75.95% to 94.17%, fake F1 from 84.39% to 94.71%, balanced accuracy from 55.08% to 58.93%, and ROC-AUC from 59.86% to 66.31% (Table IV). Fake precision remains high (94.94% original; 95.25% adapted). Because the underlying MFCC-CNN is frozen, these changes are attributable to the learned route-specific affine transformation rather than to retraining the classifier. We note that the adapted audio ROC-AUC of 66.31% remains modest in absolute terms.

TABLE IV  
FROZEN AUDIO BRANCH: ORIGINAL VS. LANGUAGE-ADAPTED (%)

| Audio    | Acc.   | Bal. Acc. | Fake Recall | Fake F1 | AUC   |
|----------|--------|-----------|-------------|---------|-------|
| Original | 73.53  | 55.08     | 75.95       | 84.39   | 59.86 |
| Adapted  | 90.08  | 58.93     | 94.17       | 94.71   | 66.31 |
| $\Delta$ | +16.55 | +3.85     | +18.22      | +10.32  | +6.45 |

#### D. Full Audio-Visual Detector

At the full-detector level, the proposed method improves accuracy from 91.15% to 94.28%, fake recall from 94.90% to 98.87%, and fake F1 from 95.28% to 97.02% (Table V, Fig. 2). Fake false negatives fall from 63 to 14. However, balanced accuracy decreases from 62.58% to 59.30% and real recall falls from 30.26% to 19.74% because real-to-fake errors increase from 53 to 61. Fake precision decreases slightly from 95.67% to 95.24%. The proposed confusion matrix is  $\begin{bmatrix} 15 & 61 \\ 14 & 1221 \end{bmatrix}$ .

TABLE V  
FULL DETECTOR: BASELINE VS. LANGUAGE-AWARE (%)

| Metric            | Baseline | Proposed | $\Delta$ (pp) |
|-------------------|----------|----------|---------------|
| Accuracy          | 91.15    | 94.28    | +3.13         |
| Balanced accuracy | 62.58    | 59.30    | -3.28         |
| Fake precision    | 95.67    | 95.24    | -0.43         |
| Fake recall       | 94.90    | 98.87    | +3.97         |



|             |       |       |        |
|-------------|-------|-------|--------|
| Fake F1     | 95.28 | 97.02 | +1.74  |
| Real recall | 30.26 | 19.74 | -10.52 |

Exactly 57 final decisions change: 49 baseline errors become correct and 8 previously correct predictions become wrong, yielding a net gain of 41 correct videos. The video-level exact McNemar test on 49 versus 8 discordant outcomes gives  $p \approx 2.72 \times 10^{-8}$ , indicating that the paired accuracy change is unlikely to be explained by symmetric decision noise. A source-group sign-flip/randomization test gives  $p \approx 6.5 \times 10^{-5}$ , supporting the same conclusion while respecting grouping by raw source.

Source-group bootstrap intervals further characterize the trade-off. Approximate 95% intervals for the proposed-minus-baseline differences are: accuracy +3.13 percentage points (CI  $\approx +1.68$  to  $+4.74$ ), balanced accuracy -3.28 points ( $\approx -7.06$  to  $+0.08$ ), fake precision -0.43 points ( $\approx -0.90$  to  $-0.04$ ), fake recall +3.97 points ( $\approx +2.52$  to  $+5.62$ ), and fake F1 +1.74 points ( $\approx +0.94$  to  $+2.64$ ). Thus, the improvements in accuracy, fake recall, and fake F1 are consistent across source resamples, whereas the balanced-accuracy decline is less precisely estimated.

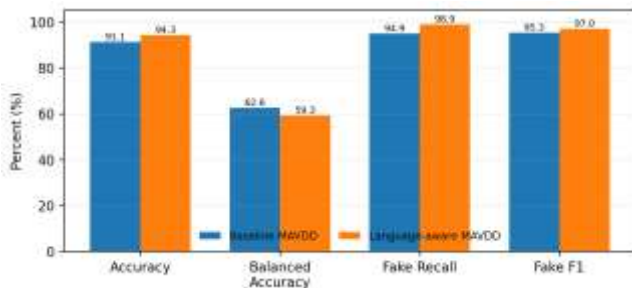


Fig. 2. Overall baseline and language-aware full-detector performance on the untouched 1,311-video test set.

#### E. Comparison with a Trivial All-Fake Reference

Because the test set contains 1,235 fake and 76 real videos, a predictor that labels every video as fake attains  $1235/1311 = 94.20\%$  accuracy, 100% fake recall, 97.01% fake F1, and 50.00% balanced accuracy. This reference is computed directly from the test-set class counts in Table I and involves no model. The proposed system (94.28% accuracy, 97.02% fake F1) is therefore close to this reference on accuracy and fake F1, while retaining a non-zero real recall (19.74%) and a higher balanced accuracy (59.30%). The baseline system (91.15% accuracy) is below the all-fake reference in accuracy. Overall accuracy and fake F1 should consequently be interpreted together with balanced accuracy and real recall.

#### F. Per-Language Performance

Accuracy and fake F1 improve in all three languages, whereas balanced accuracy decreases in each language (Table VI). French accuracy increases from 91.77% to 95.24%, Spanish from 90.68% to 93.86%, and Russian from 90.98% to 93.63%. Fake recall rises to 99.54%, 98.88%, and 98.00%, respectively. Russian retains the highest balanced accuracy after adaptation (67.52%), reflecting better real-class recognition than French or Spanish under the same hard-OR fusion rule. The per-language all-fake reference accuracies computed from the test counts are 94.81% (French, 438/462), 94.70% (Spanish, 447/472), and 92.84% (Russian, 350/377); the proposed Spanish accuracy (93.86%) is below its all-fake reference. Each language contains only 24–27 real test videos, so per-language real-class estimates are imprecise.

TABLE VI  
PER-LANGUAGE FULL-DETECTOR RESULTS (%). B = BASELINE, P = PROPOSED; AF = ALL-FAKE REFERENCE ACCURACY

| Lang.   | Acc. B | Acc. P | Acc. AF | Bal. B | Bal. P | F1 B  | F1 P  |
|---------|--------|--------|---------|--------|--------|-------|-------|
| French  | 91.77  | 95.24  | 94.81   | 62.19  | 58.11  | 95.64 | 97.54 |
| Spanish | 90.68  | 93.86  | 94.70   | 53.54  | 51.44  | 95.08 | 96.82 |
| Russian | 90.98  | 93.63  | 92.84   | 71.22  | 67.52  | 95.10 | 96.62 |

#### G. Manipulation and Routing Observations

Technique-level analysis is descriptive because manipulation technique, source identity, and language are not independently randomized. Among fake test videos, recall improves for Bark (91.42%→98.52%), MicroTts (92.59%→98.15%), and XTTS (98.27%→99.42%), while Tacotron remains 100%. Recall also improves for both synchronization techniques: VideoRetalking 94.00%→98.80% and Wav2Lip 95.71%→98.93%. These observations describe where corrections occur but do not establish causal superiority for any manipulation type.

The test routing distribution is 411 fra, 334 spa, 275 rus, and 291 other; all 291 LID errors fall in the other route. The other route shows the largest balanced-accuracy decline (58.01%→49.64%) despite an accuracy increase (91.07%→94.50% for the LID-incorrect subset overall). Correct-LID samples likewise improve in accuracy (91.18%→94.22%). Thus, gains are not limited to correctly routed samples, but routing errors are especially risky for the scarce real class: among 14 real videos with incorrect LID, the proposed system classifies all 14 as fake, whereas the baseline classifies three correctly.

#### H. Explainability Results

The learned adapter parameters are route dependent (Fig. 3). Most non-identity scaling and shifting occurs in lower and middle MFCC coefficients, while coefficients 11–13 retain exactly  $\gamma = 1$  and  $\beta = 0$  for all four routes. This does not imply that the final coefficients are unimportant; it only means that training did not modify them through the adapter. The distinct fra, spa, rus, and other patterns show that the proposed component learned different coefficient-wise transformations for different predicted-language routes.

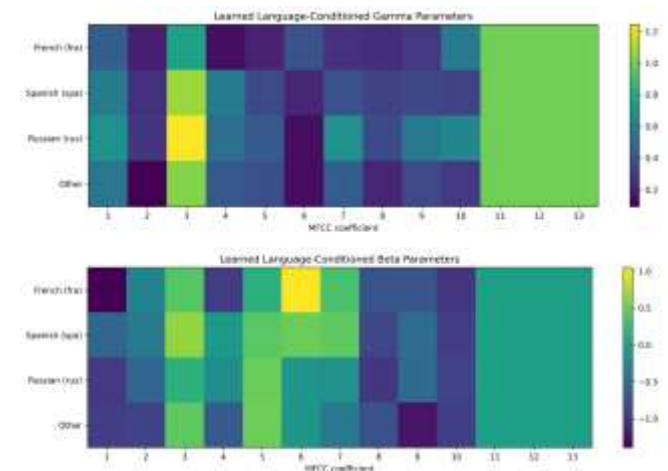


Fig. 3. Learned language-conditioned adapter parameters. Identity corresponds to  $\gamma = 1$  and  $\beta = 0$ .

Integrated Gradients provides a complementary instance-level view (Fig. 4). For the Spanish fake en\_17\_to\_es\_Bark.mp4, the original audio classifier gives  $P(\text{real}) = 0.99998$ , while the spa adapter reduces  $P(\text{real})$  to 0.01025 and correctly predicts fake; attribution is concentrated primarily in lower-order MFCC coefficients at selected time regions. Similar corrections occur for es\_61\_to\_fr\_Bark.mp4

(0.96412  $\rightarrow$  0.10680) and es\_35\_to\_ru\_Xtts.mp4 (0.87558  $\rightarrow$  0.01235). The real French failure fr\_7.mp4 moves in the same direction (0.95705  $\rightarrow$  0.37968), causing the adapted audio branch to predict fake and, through hard-OR fusion, producing a false positive.

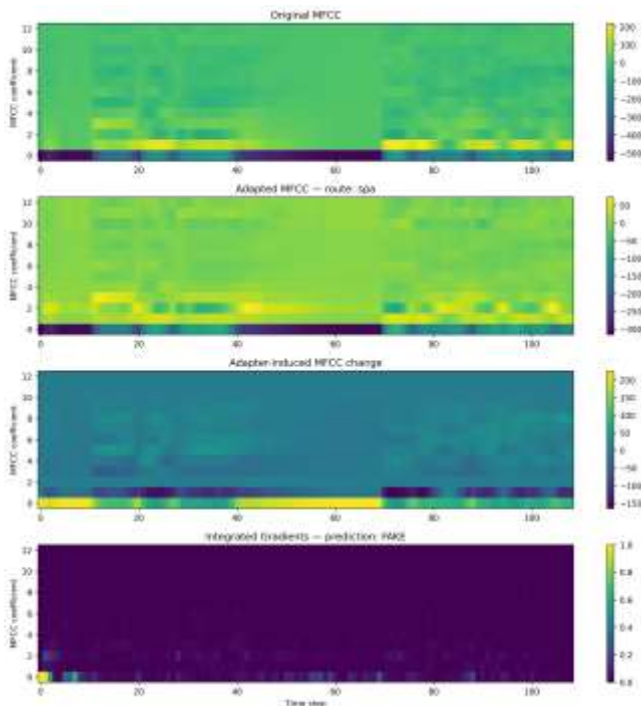


Fig. 4. Audio explainability for Spanish fake en\_17\_to\_es\_Bark.mp4: original MFCC, adapted MFCC, adapter-induced change, and absolute normalized Integrated Gradients attribution. MFCC panels use their displayed value scales.

Grad-CAM analysis of representative baseline face crops complements the audio explanation (Fig. 5). The selected face crop from each of the three corrected fake videos is strongly classified as real by Xception ( $P(\text{real}) \approx 1.000$ ), while the real failure fr\_7.mp4 is also visually real ( $P(\text{real}) = 0.992$ ). Across these representative examples, Grad-CAM frequently emphasizes upper facial or head-boundary regions rather than the mouth. Because the maps arise from a  $4 \times 4$  activation grid and only a small set of representative crops is examined, this observation is qualitative and should not be generalized to the entire dataset.



Fig. 5. Logit-based Grad-CAM on representative baseline face crops (each panel group shows face crop, Grad-CAM map, and overlay). (a)–(c): fake videos corrected after language-aware audio adaptation (French fake es\_61\_to\_fr\_Bark, Spanish fake en\_17\_to\_es\_Bark, Russian fake es\_35\_to\_ru\_Xtts); (d): French real video fr\_7 changed from correct to incorrect by hard-OR fusion. Grad-CAM is crop-level and spatially coarse.

## V. DISCUSSION

### A. What Language Conditioning Improves

The strongest evidence for the proposed mechanism is the change in the frozen audio branch. With only 104 trainable scalars, audio accuracy increases by 16.55 points, fake recall by 18.22 points, fake F1 by 10.32 points, and ROC-AUC by 6.45 points. The MFCC-CNN weights are unchanged, so the adapter is not a fine-tuning of the classifier; it is a language-conditioned

input transformation. This is consistent with the hypothesis that a fixed audio detector can benefit from route-specific normalization or reweighting of cepstral structure. We stress, however, that this study does not include ablations (e.g., a single shared adapter or oracle routes) that would isolate the contribution of language routing from that of in-domain recalibration, and the gains on LID-correct and LID-incorrect subsets are of similar size. The language-specific interpretation is therefore a hypothesis rather than an established result.

### B. Why Balanced Accuracy Declines After Fusion

The full-system result exposes an important limitation of hard-OR fusion. Increasing the audio branch's tendency to flag fake content is highly effective for the majority fake class: 49 baseline mistakes are corrected and fake recall approaches 99%. However, hard-OR fusion requires both modalities to say real, so any extra audio false positive directly becomes a final false positive even when the visual branch is correct. The eight newly wrong decisions and the drop in real recall from 30.26% to 19.74% are therefore not contradictory to the audio-branch improvement; they are a predictable consequence of the fusion rule operating on an imbalanced test distribution.

### C. Role of Imperfect LID

LID accuracy of 77.80% means routing is imperfect; on the test set the errors manifest as predictions outside the three target languages rather than as confusions among them. The explicit other route and the use of predicted, rather than oracle, routes during adapter training reduce train-test mismatch. Accuracy gains occur on both LID-correct and LID-incorrect subsets, indicating that the adapter does not depend on perfect routing. Nevertheless, the real-video analysis shows that incorrect routing can be costly under hard-OR fusion, especially because only 76 real videos are present in the test set. Future routing schemes should therefore consider uncertainty rather than treating a single LID label as fully reliable.

### D. Interpretation of XAI

The explainability results are consistent with the quantitative findings but are not causal evidence. Adapter heatmaps show where route-specific coefficient transforms were learned; Integrated Gradients shows local sensitivity of the adapted audio decision; and Grad-CAM shows spatial evidence for individual Xception crop predictions. Together, the selected cases illustrate the observed mechanism: a fake video can be visually judged real and originally judged real by audio, then be corrected when the adapted audio branch flips to fake; conversely, a real video can be visually correct but become a final false positive when adapted audio flips to fake.

## VI. LIMITATIONS AND FUTURE WORK

The study has several limitations. First, only French, Spanish, and Russian are evaluated, and all three languages appear in training, validation, and test splits; the findings therefore describe multilingual evaluation and do not establish universal multilingual or zero-shot cross-lingual generalization. Second, the selected subset is highly imbalanced, with only 76 real videos in the test split; accuracy and fake F1 therefore favor the majority class, and a trivial all-fake predictor achieves comparable accuracy and fake F1, so results must be interpreted together with balanced accuracy and real recall. Third, the experiment uses released MAVDD-style weights rather than retraining the original detector on PolyGlottFake, making this an adaptation/evaluation study rather than a new end-to-end detector. Fourth, the LID model is external to the deepfake task and is imperfect. Fifth, hard-OR decision fusion is intentionally preserved for comparability but is poorly calibrated for a more



sensitive audio branch. Sixth, no ablation isolates language routing from generic in-domain recalibration, and results come from a single training run. Finally, XAI is qualitative and based on representative cases rather than a dataset-wide explanation benchmark.

Future work should prioritize fusion rather than adding complexity to the adapter. Calibrated score fusion, learned decision fusion, or real-class-aware thresholding could retain the fake-recall gains while reducing real false positives. LID confidence could also be used for soft routing or mixtures of adapters instead of a single hard route. Additional languages, ablations (shared, oracle, and random routing), repeated runs, and more balanced real/fake evaluation would be required before claiming broad multilingual robustness.

## VII. REPRODUCIBILITY AND ETHICAL USE

All experiments use fixed source-disjoint train/validation/test CSVs and an untouched final test split. The 68 missing files identified during integrity checking are excluded rather than reconstructed. Adapter checkpointing uses validation balanced accuracy only; no parameter is selected using test labels. The implementation environment used Python 3.11.9, TensorFlow 2.20.0 [18], Keras 3.13.0, librosa 0.11.0 [17], OpenCV 4.11.0, NumPy 1.26.4, Pillow 10.2.0, FER 22.5.1 [21], PyTorch 2.2.2 [19], torchvision 0.17.2, and Transformers 4.49.0 [20]. PolyGlottFake is used under its academic/research access terms [1]; raw dataset videos are not redistributed. Code, split definitions, derived metadata, and result tables can be shared independently of the restricted raw media.

## VIII. CONCLUSION

This paper establishes a controlled multilingual baseline for a MAVDD-style detector on French, Spanish, and Russian PolyGlottFake data and introduces a lightweight LID-based MFCC adapter. The adapter substantially strengthens the frozen audio branch and improves full-system accuracy, fake recall, and fake F1 on the untouched test set. However, the same change increases false positives on real videos under hard-OR fusion, lowering balanced accuracy and real recall, and the resulting accuracy is close to that of a trivial all-fake predictor. The main conclusion is therefore not that language conditioning universally improves the detector, but that it improves fake sensitivity while exposing a fusion-calibration problem. Language-aware audio adaptation is promising, yet it should be paired with fusion mechanisms that explicitly protect minority-class real performance and validated with ablations that isolate the effect of language routing.

## REFERENCES

- [1] Y. Hou, H. Fu, C. Chen, Z. Li, H. Zhang, and J. Zhao, "PolyGlottFake: A novel multilingual and multimodal DeepFake dataset," in *Pattern Recognition—27th Int. Conf., ICPR 2024, Lecture Notes in Computer Science*, vol. 15314, 2025, pp. 180–193, doi: 10.1007/978-3-031-783418\_12.
- [2] A. A. Bekheet, G. Khoriba, and A. S. Ghoneim, "Development of a multimodal framework for deepfake detection: Combining visual and audio analysis," in *Proc. 10th World Congr. Electrical Eng. and Computer Systems and Sciences (EECSS'24)*, Barcelona, Spain, 2024, Paper MVML 115, doi: 10.11159/mvml24.115.
- [3] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A novel audio-video multimodal deepfake dataset," in *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2021.
- [4] W. Yang, X. Zhou, Z. Chen, B. Guo, Z. Ba, Z. Xia, X. Cao, and K. Ren, "AVoid-DF: Audio-visual joint learning for detecting deepfake," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 2015–2029, 2023, doi: 10.1109/TIFS.2023.3262148.
- [5] M. A. Raza and K. M. Malik, "MultimodalTrace: Deepfake detection using audiovisual representation learning," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 993–1000, doi: 10.1109/CVPRW59228.2023.00106.
- [6] V. S. Katamneni and A. Rattani, "MIS-AVoiDD: Modality invariant and specific representation for audio-visual deepfake detection," in *Proc. 22nd IEEE*

- Int. Conf. Machine Learning and Applications (ICMLA)*, 2023, pp. 1371–1378, doi: 10.1109/ICMLA58977.2023.00207.
- [7] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1251–1258, doi: 10.1109/CVPR.2017.195.
- [8] V. Pratap et al., "Scaling speech technology to 1,000+ languages," *J. Mach. Learn. Res.*, vol. 25, no. 97, pp. 1–52, 2024.
- [9] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12449–12460.
- [10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.
- [11] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Machine Learning (ICML)*, vol. 70, 2017, pp. 3319–3328.
- [12] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947, doi: 10.1007/BF02295996.
- [13] B. Efron, "Bootstrap methods: Another look at the jackknife," *Ann. Statist.*, vol. 7, no. 1, pp. 1–26, 1979.
- [14] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *Proc. 20th Int. Conf. Pattern Recognition (ICPR)*, 2010, pp. 3121–3124.
- [15] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980.
- [16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2015.
- [17] B. McFee et al., "librosa: Audio and music signal analysis in Python," in *Proc. 14th Python in Science Conf. (SciPy)*, 2015, pp. 18–25.
- [18] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Systems Design and Implementation (OSDI)*, 2016, pp. 265–283.
- [19] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 8024–8035.
- [20] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proc. EMNLP 2020: System Demonstrations*, 2020, pp. 38–45.
- [21] J. Shenk, "FER: Facial expression recognition (Python package)," 2021. [Online]. Available: <https://github.com/justinshenk/fer>