WHAT SPATIAL CUES DO MULTICHANNEL REPLAY DETECTORS RELY ON? A PHYSICALLY INTERPRETABLE ANALYSIS VIA RELATIVE TRANSFER FUNCTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Replay attacks have become a major security threat as intelligent voice interaction systems are widely deployed. Existing multichannel detectors typically use spatial information between microphones, but their dependence on specific spatial features remains unclear. We use physically interpretable relative transfer functions (RTFs) as control variables to adjust inter-channel magnitude ratios and phase differences. Fixed templates are selected using only detection scores from a small set of recordings, without access to model internals. Experiments show that the selected templates can cause replay speech to be misclassified as genuine on recordings and some detectors not used in selection. In speech-preservation evaluation, transcripts and speaker representations remain highly consistent before and after processing. RTF control thus connects changes in model decisions to explicit inter-channel acoustic quantities and enables comparisons of detector sensitivity to spatial changes.

1 INTRODUCTION

Smart speakers and voice assistants are widely used for device control, authentication, and personalized services. Secure interaction requires these systems to recognize speech content and speaker identity while also verifying the authenticity of the sound source (Meng et al., 2022; Liu et al., 2023). Replay attacks reproduce speech content and speaker characteristics through ordinary recording and playback devices to impersonate live speech, posing a major threat to voice interaction systems (Kinnunen et al., 2017). Propagation distance, environmental noise, and device differences further complicate far-field detection (Gong et al., 2019). Microphone arrays capture spatial relationships between acoustic observations at different positions, providing cues for distinguishing genuine speech from loudspeaker playback.

Existing replay detectors primarily extract cues from acoustic artifacts introduced by the recording and playback chain. Single-channel methods emphasize device features (He et al., 2024) and channel magnitude responses (Bekiryazıcı et al., 2026). In broader audio deepfake detection research, self-supervised representations improve generalization (Pascu et al., 2024), and separating semantic and acoustic representations protects content privacy (Li et al., 2024a); ASVspoof 5 extends evaluation to more complex data and adversarial conditions (Wang et al., 2026). Multichannel methods represent spatial channel relationships through magnitude and phase spectra (Li et al., 2021), cross-microphone energy dynamics (Yang et al., 2023), learnable beamforming (Neri & Virtanen, 2025b), and acoustic maps (Neri & Virtanen, 2026). The specific inter-channel structures on which these models rely remain to be established, particularly the respective roles of magnitude ratios and phase differences in their decisions.

Audio attacks actively modify inputs and observe decision responses, providing interventions for studying detector robustness and decision mechanisms. Gradient perturbations (Kawa et al., 2023) and post-processing such as noise addition (Schäfer et al., 2025) can induce misclassification. Optimized convolutional filters can be reused across utterances (Panariello et al., 2023), and some attacks transfer to other models (Farooq et al., 2025). These modifications may also degrade audio quality

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(Todisco et al., 2024) or disrupt speech and speaker recognition while preserving meaning (Yu et al., 2023). Connecting decision changes to specific spatial cues therefore requires interpretable control of inter-channel structure, together with evaluation of speech-content and speaker-representation preservation.

Relative transfer functions (RTFs) describe magnitude ratios and phase differences relative to a reference channel (Cohen, 2004). We use RTFs as physically interpretable control variables to analyze detector responses by adjusting inter-channel magnitude–phase relationships. As Figure 1 shows, empirical templates are estimated from genuine recordings, whereas parameterized templates are selected from predefined candidates using detection scores from a small set of recordings. Selected templates remain fixed when applied to recordings and models not used in selection. No per-recording optimization or access to model architectures, parameters, or gradients is required.

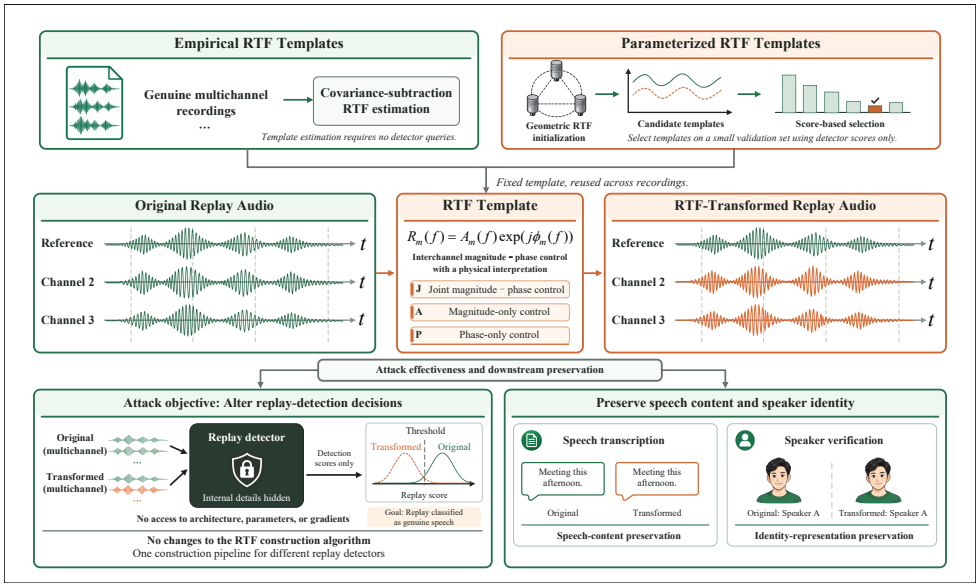


Figure 1: Physically interpretable analysis through RTFs. Templates are estimated from genuine recordings or selected from parameterized candidates using detection scores. Inter-channel magnitude–phase control is evaluated for attacks across recordings and models, and for speech preservation. J, A, and P denote joint, magnitude-ratio-only, and phase-difference-only control.

We evaluate both detector decisions and speech preservation. Our main contributions are:

- **Physically interpretable control of inter-channel structure.** We use RTFs to adjust magnitude ratios and phase differences under the same reference speech, and analyze detector dependence on inter-channel structure through decision responses.
- **Vulnerabilities in multichannel replay detectors.** Fixed RTF templates can cause replay speech to be misclassified as genuine on recordings and some detectors not used in selection, indicating sensitivity to particular changes in inter-channel structure.
- **Speech preservation under inter-channel control.** Complete utterances retain high transcript agreement and speaker-representation similarity after magnitude-ratio and phase-difference control.

2 Background and Related Work

2.1 Multichannel Replay Attack Detection

ReMASC supports multichannel replay-detection evaluation (Gong et al., 2019). Spatial representations use channel magnitude–phase spectra (Li et al., 2021), cross-microphone sound-field dynamics in VoShield (Yang et al., 2023), jointly learned beamforming and classification (Neri & Virtanen, 2025b), or directional acoustic maps (Neri & Virtanen, 2026). Their different treatment of channel relationships motivates controlled transformations for comparing detector responses.

Multichannel replay detection is sensitive to microphone-array mismatches (Neri & Virtanen, 2025a). Audio-forgery detectors also respond to changes in data distributions (Müller et al., 2024), synthesis systems and codecs (Li et al., 2024b), neural-codec generation (Wu et al., 2024), playback and re-recording (Zhang et al., 2026), and wording even with largely preserved meaning (Nguyen et al., 2025). Beyond evaluating stability under changing conditions, we control inter-channel magnitude ratios and phase differences to examine responses to specific spatial relationships.

2.2 Audio Adversarial Attacks and Black-Box Transfer

Structured audio attacks reuse transformations across utterances. Malafide optimizes linear filters (Panariello et al., 2023); Farooq et al. (2025) use surrogate-model ensembles for transferable attacks. Our RTF templates provide explicit magnitude–phase interpretations. Selected by detection scores and fixed across recordings and models, they connect reusable attacks to spatial-structure analysis.

Attack effectiveness complements speech-preservation evaluation. Malacopula improves speaker-verification attacks while reducing audio quality (Todisco et al., 2024); SMACK can change speech- and speaker-recognition outputs while preserving meaning (Yu et al., 2023). We evaluate detector decisions, transcripts, and speaker representations together to assess RTF transformations across speech tasks.

3 Method

Using the same reference speech, we control inter-channel magnitude ratios, phase differences, or both. Empirical templates come from genuine recordings. Parameterized templates combine array geometry with magnitude–phase parameters, fixed after score-based selection.

3.1 RTF Representation and Magnitude–Phase Control

Let f denote frequency, $H_m(f)$ the effective source-to-channel transfer function for channel m, and r the reference channel. Where $H_r(f) \neq 0$, the RTF is (Cohen, 2004)

$$R_m(f) = \frac{H_m(f)}{H_r(f)} = A_m(f) e^{\mathrm{j}\phi_m(f)}, \qquad R_r(f) = 1. \tag{1}$$

The RTF magnitude $A_m(f) = |R_m(f)|$ and argument $\phi_m(f) = \arg R_m(f)$ represent the target-to-reference magnitude ratio and phase difference; j is the imaginary unit.

We compare joint (J), magnitude-ratio-only (A), and phase-difference-only (P) control using the same template:

$$R_m^{\mathrm{J}}(f) = A_m(f) e^{\mathrm{j}\phi_m(f)}, \qquad R_m^{\mathrm{A}}(f) = A_m(f), \qquad R_m^{\mathrm{P}}(f) = e^{\mathrm{j}\phi_m(f)}. \tag{2}$$

J retains the complete template magnitude–phase relationship. A retains the magnitude ratio and sets the phase difference to 0; P retains the phase difference and sets the magnitude ratio to 1. We also define a unit RTF, $R_m^{\mathrm{unit}}(f) = 1$, with magnitude ratio 1 and phase difference 0. It provides a reconstruction reference without inter-channel magnitude or phase differences for empirical-template preservation evaluation, as specified in Appendix D.2. All conditions share the reference speech and reconstruction procedure.

For recording i, the reference waveform $x_{r,i}$ has short-time Fourier transform $X_{r,i}(f,k) = \mathrm{STFT}\{x_{r,i}\}(f,k)$, where k indexes frames. Let q denote J, A, P, or unit RTF, and $R_{m,i}^q$ the corresponding template response. Multiplying it by the reference spectrum yields the constructed spectrum $\widetilde{Z}_{m,i}^q$; inverse STFT, iSTFT, produces waveform $\widetilde{z}_{m,i}^q$:

$$\begin{aligned} \widetilde{Z}_{m,i}^q(f,k) &= R_{m,i}^q(f) X_{r,i}(f,k), \qquad m \neq r, \\ \widetilde{z}_{m,i}^q(n) &= \mathrm{iSTFT}\{\widetilde{Z}_{m,i}^q\}(n), \qquad \widetilde{z}_{r,i}^q(n) = x_{r,i}(n). \end{aligned} \tag{3}$$

Here n indexes samples, and the template remains constant within each recording. The channel waveforms form the multichannel signal $\widetilde{\mathbf{z}}_i^q$. A common positive gain g_i produces detector input $\mathbf{y}_i^q = g_i \widetilde{\mathbf{z}}_i^q$. Paired conditions within a configuration share this gain, preserving their inter-channel magnitude–phase relationships.

3.2 Empirical RTF Estimation

Covariance subtraction (CS) estimates empirical RTFs from genuine recordings (Markovich-Golan et al., 2018). With a single dominant source, local stationarity, and a narrowband approximation,

$$X_m(f,k) \approx H_m(f)S(f,k) + V_m(f,k), \tag{4}$$

Here X_m, S, and V_m are the STFTs of the observation, source, and noise plus modeling residual. H_m is approximately stable during estimation. For zero-mean, second-order uncorrelated S and V_m, CS subtracts noise-dominated-frame statistics from speech-present-frame statistics to reduce noise effects.

Let $\mathcal{T}^{\mathrm{sp}}$ and $\mathcal{T}^{\mathrm{ns}}$ be speech-present and noise-dominated frame sets. A statistic $h(f,k)$ has mean $\langle h \rangle_\ell = |\mathcal{T}^\ell|^{-1} \sum_{k \in \mathcal{T}^\ell} h(f,k)$, with class $\ell \in \{\mathrm{sp}, \mathrm{ns}\}$ and frame count $|\mathcal{T}^\ell|$. Cross-power and reference auto-power are $C_{mr}^\ell = \langle X_m X_r^* \rangle_\ell$ and $P_r^\ell = \langle |X_r|^2 \rangle_\ell$; X_r^* denotes complex conjugation. Neglecting residual finite-frame cross-terms gives

$$\begin{aligned} C_{mr}^\ell &\approx H_m H_r^* P_s^\ell + C_{V_m V_r}^\ell, \\ P_r^\ell &\approx |H_r|^2 P_s^\ell + P_{V_r}^\ell, \end{aligned} \tag{5}$$

Here $P_s^\ell = \langle |S|^2 \rangle_\ell$ is source power, while $C_{V_m V_r}^\ell = \langle V_m V_r^* \rangle_\ell$ and $P_{V_r}^\ell = \langle |V_r|^2 \rangle_\ell$ are noise cross-power and reference-channel noise power; frequency arguments are omitted. If noise second-order statistics are approximately stable across the two frame classes, subtraction approximately cancels the noise terms. Denoting differences by Δ yields

$$\begin{aligned} \Delta C_{mr} &:= C_{mr}^{\mathrm{sp}} - C_{mr}^{\mathrm{ns}} \approx H_m H_r^* \Delta P_s, \\ \Delta P_r &:= P_r^{\mathrm{sp}} - P_r^{\mathrm{ns}} \approx |H_r|^2 \Delta P_s, \end{aligned} \qquad \Delta P_s = P_s^{\mathrm{sp}} - P_s^{\mathrm{ns}}. \tag{6}$$

The difference ratio cancels common speech power. Denominator regularizer $\varepsilon > 0$ stabilizes the RTF estimate $\widehat{R}_m$:

$$\begin{aligned} \widehat{R}_m(f) &= \frac{\Delta C_{mr}(f)}{\Delta P_r(f) + \varepsilon} \\ &\approx \frac{H_m(f) H_r^*(f) \Delta P_s(f)}{|H_r(f)|^2 \Delta P_s(f) + \varepsilon} \approx R_m(f), \qquad \varepsilon > 0. \end{aligned} \tag{7}$$

This requires $\Delta P_s > 0$ and noise-subtracted reference speech power much larger than ε. Changing noise statistics and propagation paths also affect estimation bias. Appendix E.1 gives the derivation.

We stabilize the covariance-difference matrix before extracting cross-power and reference power and taking their ratio. Appendix F.1 specifies stabilization and frequency-bin constraints.

To reduce variability, we pool cross-power and reference auto-power over recordings $\mathcal{U}_c$ from scene c. For statistic $h_u(f,k)$ from recording u with valid frames $\mathcal{T}_u^\ell$ of class ℓ, the pooled mean is

$$\langle h \rangle_{\ell,c} = \frac{\displaystyle\sum_{u \in \mathcal{U}_c} \sum_{k \in \mathcal{T}_u^\ell} h_u(f,k)}{\displaystyle\sum_{u \in \mathcal{U}_c} |\mathcal{T}_u^\ell|}, \qquad \ell \in \{\mathrm{sp}, \mathrm{ns}\}. \tag{8}$$

Speech-present ($\ell = \mathrm{sp}$) and noise-dominated ($\ell = \mathrm{ns}$) frames are pooled separately, weighting recordings by valid-frame counts. Equations (6) and (7) then give the scene RTF. Pooled recordings share the environment, source position, devices, array geometry, and reference channel, with approximately stable propagation and noise statistics.

Predefined, environment-prioritized scene rules assign templates to replays; estimation and assignment require no detector queries. Test/LOSO excludes the target speaker from the reference pool; Train uses independent training recordings. Equations (2) and (3) then apply magnitude–phase control and reconstruction. Appendix B.2 gives the assignment rules.

3.3 Parameterized RTF Construction

Parameterized templates adjust magnitude–phase relationships relative to virtual source–array geometry. Microphone m at $\mathbf{p}_m$ and virtual source $\mathbf{p}_{\mathrm{src}}$ have distance $\delta_m = \|\mathbf{p}_m - \mathbf{p}_{\mathrm{src}}\|_2 > 0$, where $\|\cdot\|_2$ is the Euclidean norm. For sound speed c_0, the nominal free-field point-source response $G_m(f; \mathbf{p}_{\mathrm{src}})$ is

$$G_m(f; \mathbf{p}_{\mathrm{src}}) = \frac{e^{-\mathrm{j}2\pi f \delta_m / c_0}}{4\pi \delta_m}, \tag{9}$$

The target-to-reference ratio gives geometric RTF R_m^{geo}, with source-to-reference distance δ_r:

$$R_m^{\mathrm{geo}}(f) = \frac{G_m(f; \mathbf{p}_{\mathrm{src}})}{G_r(f; \mathbf{p}_{\mathrm{src}})} = \frac{\delta_r}{\delta_m} \exp\left[-\mathrm{j}2\pi f \frac{\delta_m - \delta_r}{c_0}\right]. \tag{10}$$

Distance ratios and path differences set relative magnitude and propagation delay. This geometric baseline supports magnitude–phase control; derived candidates may depart from responses at actual source positions.

For three adjacent nominal bands, band v has magnitude and phase parameters $\alpha_{A,v}$ and $\alpha_{\phi,v}$, $v = 1, 2, 3$. Tukey weights $\omega_v(f)$ smoothly combine them into control functions $\alpha_A(f)$ and $\alpha_\phi(f)$:

$$\alpha_A(f) = \sum_{v=1}^{3} \omega_v(f) \alpha_{A,v}, \qquad \alpha_\phi(f) = \sum_{v=1}^{3} \omega_v(f) \alpha_{\phi,v}. \tag{11}$$

Overlapping weights are normalized by their sum; single-window weights are unchanged. Outermost weights taper to zero beyond nominal boundaries, smoothly returning the RTF to a unit response. Appendix F.3 defines the windows and boundaries.

Collecting the three bands' magnitude and phase parameters into $\boldsymbol{\alpha}$ gives the parameterized RTF

$$R_m(f; \boldsymbol{\alpha}) = \left(\frac{\delta_r}{\delta_m}\right)^{\alpha_A(f)} \exp\left[-\mathrm{j}2\pi f \frac{\delta_m - \delta_r}{c_0} \alpha_\phi(f)\right]. \tag{12}$$

Magnitude parameters control relative gain; phase parameters scale geometric phase differences. All parameters set to 1 recover the baseline in flat and overlapping regions; 0 gives a unit RTF. A phase parameter of 1 retains geometric phase, whereas condition A in Equation (2) sets phase difference to zero.

Parameters $\boldsymbol{\alpha}$ are generated within predefined ranges, independently of detector feedback. Candidates satisfying numerical-validity, relative-gain, and equivalent-delay constraints form finite set Θ; Appendix C.1 gives the settings.

3.4 Template Selection and Reuse

Selection uses $\mathcal{D}_{\mathrm{sel}}$, comprising $|\mathcal{D}_{\mathrm{sel}}|$ recordings independent of evaluation. Source detector d outputs replay score $s_d(\cdot)$, with lower scores favoring genuine speech. For candidate $\boldsymbol{\alpha}$, Equation (3)

and common-gain scaling yield J input $\mathbf{y}_i^{\mathrm{J}}(\boldsymbol{\alpha})$. Native preprocessing and scoring select parameters $\boldsymbol{\alpha}_d^\star$ minimizing the mean score:

$$\boldsymbol{\alpha}_d^\star \in \arg\min_{\boldsymbol{\alpha}\in\Theta} \frac{1}{|\mathcal{D}_{\mathrm{sel}}|} \sum_{i\in\mathcal{D}_{\mathrm{sel}}} s_d\big(\mathbf{y}_i^{\mathrm{J}}(\boldsymbol{\alpha})\big)\,. \tag{13}$$

Selection uses scores alone, with no access to model architecture, parameters, or gradients. A predefined rule breaks ties. Equation (2) derives A and P from the selected J template.

Template parameters and thresholds remain fixed for evaluation on held-out recordings of source detector d and detectors not used in selection. Appendix F gives the procedure; Appendix C.3 specifies cross-model evaluation.

4 Experiments

We compare genuine and replay RTFs, examine detector responses to inter-channel magnitude–phase control, and test fixed-template reuse across recordings and models. Complete utterances assess transcript and speaker-representation preservation.

4.1 Experimental Setup

ReMASC has speaker-disjoint training (Train), development (Dev), and test (Test) sets. D1–D4 and seven detectors, including Acoustic Maps (AM), give 28 model–array combinations. We fix model weights, native preprocessing, and Dev-selected thresholds. Appendix B.1 reports settings and original detection baselines.

We compare transcripts before and after processing with three automatic speech recognition (ASR) systems: Whisper small.en, Whisper large-v3-turbo (Radford et al., 2023), and Parakeet (NVIDIA, 2025). Cosine similarity of ECAPA-TDNN (Desplanques et al., 2020), WeSpeaker (Wang et al., 2023), and X-vector (Snyder et al., 2018) embeddings measures speaker-representation consistency. Appendix D specifies model configurations.

Conversion rate $\mathrm{CR}[\%] = 100N_{\mathrm{converted}}/N_{\mathrm{eligible}}$ measures replay-to-genuine misclassification. N_{eligible} counts originally correctly detected replays; $N_{\mathrm{converted}}$ counts those classified as genuine after processing. Construction or scoring failures stay in the denominator.

4.2 RTF Differences Between Genuine and Replay Speech

We compare genuine and replay channel relationships using indoor D3 recordings. Each Figure 2 pair shares the speaker and utterance, controlling identity and content differences.

Figure 2 shows each channel's magnitude ratio and phase difference relative to the reference. Within the common valid frequency ranges of these pairs, genuine and replay speech differ clearly. Within each class, magnitude–phase relationships remain relatively similar across utterances, showing some cross-utterance stability.

We compute separate magnitude-ratio and phase-difference distances to quantify between-class differences and within-class variation, under speaker matching and joint speaker–position matching. In Figure 3(A), genuine–replay distances generally exceed within-class distances under both conditions. In Figure 3(B), between-class mean distances exceed within-class means for every evaluated speaker. Distance definitions and statistics appear in Appendix A.

4.3 Empirical Templates and Detector Decisions

To examine detector responses to these structural differences, we apply RTFs from genuine Test recordings to replays. J, A, and P distinguish magnitude-ratio and phase-difference effects.

In Table 1, joint, magnitude-only, and phase-only control induce replay misclassification under different configurations. Separate control sometimes matches or exceeds joint control. Effects

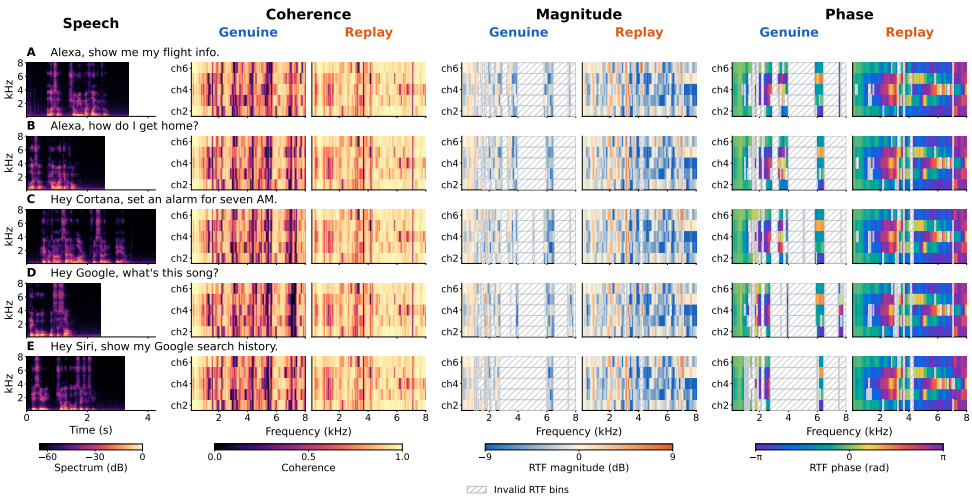
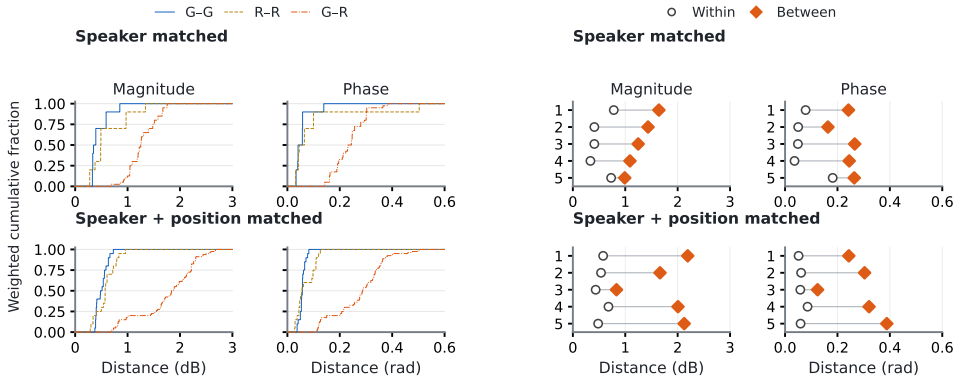


Figure 2: Genuine/replay channel relationships for matched speakers and utterances. Rows show utterances with reference spectrograms, coherence, magnitude ratios, and phase differences. Selection considers valid-frequency coverage, within-class similarity across utterances, and between-class differences. Rows share axes and color scales; hatching marks invalid RTF bins. Reference ch1 corresponds to experimental ch0.



(A) RTF distance distributions

(B) Per-speaker mean distances

Figure 3: Within- and between-class distances of inter-channel magnitude–phase relationships. (A) Weighted cumulative distributions; G and R denote genuine and replay. (B) Speaker means, with lines joining each speaker's within- and between-class means. Top: speaker matching; bottom: speaker–position matching. Common-valid-bin screening retains 6 of 56 and 16 of 44 quartets, respectively; each set covers 5 speakers. Magnitude-ratio and phase-difference distances are in dB and rad.

Table 1: Replay conversion rates (%) after empirical RTF template replacement. Templates are estimated from genuine Test recordings. Within each model–array combination, J, A, and P share the denominator of originally correctly detected replays.

Model	D1			D2			D3			D4		
	J	A	P	J	A	P	J	A	P	J	A	P
AM	7.30	0.21	6.33	86.58	78.25	79.55	86.42	35.58	84.75	72.56	58.39	76.59
M-ALRAD	0.25	0.76	2.69	73.48	66.86	85.77	41.34	82.76	57.92	27.52	12.26	71.75
MC-RAD	2.34	2.90	5.06	35.07	25.53	27.47	38.93	50.41	32.34	26.82	32.94	26.91
MIAD	0.00	0.00	0.00	54.31	0.00	63.11	14.58	4.83	15.50	40.41	43.41	29.97
VoShield	55.47	67.29	100.00	84.72	83.82	47.64	100.00	100.00	50.28	18.12	16.64	59.72
RawBMamba	6.62	6.17	4.99	4.92	4.92	4.35	3.63	4.24	4.07	6.18	6.36	4.98
mch-SSL	1.29	1.11	2.66	2.18	2.18	1.94	3.15	3.23	2.74	8.37	7.63	5.22

vary across models and arrays, revealing different sensitivities to inter-channel magnitude–phase relationships.

RawBMamba and mch-SSL retain low conversion across all four arrays. Our multichannel version of RawBMamba (Chen et al., 2024) uses a shared encoder to extract channel-wise features and averages the channel embeddings. mch-SSL (Yamaguchi et al., 2025) likewise extracts channel-wise self-supervised features before VGG fusion and classification. These processing strategies suggest that the models may rely more on within-channel acoustic cues.

We test this interpretation by copying one source-channel waveform to every input channel, preserving its speech features while removing inter-channel differences.

Table 2: Bidirectional decision-flip rates (%) after channel copying. The source is the first channel, ch0. Each direction uses the corresponding count of originally correctly classified genuine or replay recordings as its denominator. Pooled rates aggregate original counts across arrays. Other copying sources appear in Appendix B.3.

Array	RawBMamba		mch-SSL	
	G→R	R→G	G→R	R→G
D1	4.13	4.44	1.10	2.06
D2	6.13	4.51	2.66	1.94
D3	7.47	4.50	5.42	2.42
D4	5.27	5.44	4.58	2.49
Pooled	5.87	4.71	3.60	2.22

Both models remain largely stable under channel copying (Table 2). Across all copying sources, four arrays, and both flip directions, maxima are 7.47% for RawBMamba and 6.19% for mch-SSL. Together with low empirical-template conversion, this supports greater reliance on within-channel acoustic cues.

4.4 Searched-Template Reuse Across Recordings and Models

Without genuine references from the evaluation scene, we select parameterized RTF templates using scores from a small Dev replay set, then evaluate new recordings and other detectors.

We independently select one joint three-band, six-parameter band6 template per model–array combination. A and P derive from J; parameters stay fixed for full-Test evaluation.

Table 3: Full-Test conversion rates (%) of searched templates. Joint band6 templates are selected on Dev; A and P are derived from the selected template. Conditions within each model–array combination share the originally correctly detected Test replay denominator. Search settings and parameterization comparisons appear in Appendix C.

Model	D1			D2			D3			D4		
	J	A	P	J	A	P	J	A	P	J	A	P
AM	99.68	0.21	99.68	99.68	73.92	99.68	100.00	48.00	100.00	100.00	84.22	100.00
M-ALRAD	95.88	10.10	96.38	97.90	92.24	97.41	100.00	100.00	100.00	100.00	99.68	100.00
MC-RAD	32.30	5.06	32.30	67.84	10.95	66.70	60.34	42.19	59.35	44.91	37.71	43.92
MIAD	0.00	0.00	0.00	16.07	0.00	15.74	94.00	4.00	93.58	100.00	37.67	100.00
VoShield	100.00	100.00	100.00	100.00	100.00	100.00	98.72	0.00	1.12	0.08	0.00	0.00
RawBMamba	5.99	5.08	5.99	9.02	8.61	7.79	4.76	5.19	5.10	7.19	7.19	7.10
mch-SSL	3.17	3.08	3.08	4.43	2.33	4.04	3.07	2.26	3.47	6.22	3.57	5.98

Table 3 shows high joint conversion for AM and M-ALRAD across all four arrays. Phase-only results equal or approach joint results for AM and MC-RAD, so phase differences alone retain similar conversion effects. VoShield has markedly higher joint conversion than separate control in some configurations, with effects depending on joint magnitude–phase changes. RawBMamba and mch-SSL retain low conversion, consistent with their responses in the empirical-template experiment.

We compare Dev and full-Test conversion using the same template and threshold for each model–array combination.

AM and M-ALRAD reach full-Test conversion of 99.68%–100.00% and 95.88%–100.00%, respectively, across four arrays (Table 4). Thus, templates selected on a small Dev set retain high conversion on new recordings for these detectors.

Cross-model evaluation transfers fixed source templates to other detectors on full Test across four arrays; target feedback is used only for evaluation. Mean conversion over 168 pairs is 35.49%.

Table 4: Fixed-template conversion rates (%) on the Dev selection set and full Test. Each array's selection set consists of originally correctly classified replays from 64 preselected recordings. Each model–array combination uses the joint band6 template selected on this set, with a fixed detection threshold.

Model	D1		D2		D3		D4	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
AM	100.00	99.68	100.00	99.68	100.00	100.00	100.00	100.00
M-ALRAD	96.55	95.88	100.00	97.90	100.00	100.00	100.00	100.00
MC-RAD	55.17	32.30	74.14	67.84	41.51	60.34	48.08	44.91
MIAD	0.00	0.00	36.21	16.07	98.31	94.00	100.00	100.00
VoShield	100.00	100.00	100.00	100.00	98.31	98.72	0.00	0.08
RawBMamba	7.14	5.99	6.78	9.02	6.90	4.76	5.45	7.19
mch-SSL	1.56	3.17	6.35	4.43	3.33	3.07	7.81	6.22

Transfer is directional: on D3, VoShield's template gives 100.00% conversion on M-ALRAD, versus 0.00% in reverse. Both template structure and target responses affect transfer (Appendix C.3).

4.5 Speech-Content and Speaker-Representation Preservation

We assess transcript and speaker-representation preservation on complete utterances. Independently selected by pre-processing speech quality, 440 replay files from 8 speakers support before/after comparisons of each utterance.

Sentence agreement is the fraction of valid pairs with identical normalized transcripts before and after processing. Word-edit rate divides total word edits by the total word count of the original transcripts, which serve as references. Speaker evaluation reports mean and fifth-percentile (P05) embedding cosine similarity.

Table 5: Transcript agreement and speaker-representation similarity after RTF control. Both template families use J and reference the original target channel, pooling D1–D4 over their respective valid pairs. Empirical templates use Test/LOSO. Searched templates combine 56 shared2 and band6 configurations, yielding 53 templates after deduplication. Sentence agreement includes only pairs with valid, nonempty transcripts; ↑ and ↓ indicate higher and lower is better. Settings and complete results appear in Appendix D. Small.en and Turbo denote Whisper small.en and Whisper large-v3-turbo.

(a) Transcript consistency

ASR	Empirical J		Searched J	
	Agreement ↑	Word edit ↓	Agreement ↑	Word edit ↓
Small.en	93.14	1.62	94.49	1.23
Turbo	95.11	1.09	96.05	1.10
Parakeet	87.81	3.00	88.45	2.80

(b) Speaker-representation similarity

Encoder	Empirical J		Searched J	
	Mean ↑	P05 ↑	Mean ↑	P05 ↑
ECAPA-TDNN	0.9500	0.8889	0.9608	0.9042
WeSpeaker	0.9513	0.8944	0.9612	0.9030
X-vector	0.9938	0.9877	0.9958	0.9903

Both template families retain high sentence agreement across three ASR systems, with word-edit rates at most 3.00% (Table 5(a)). All three speaker encoders have mean cosine similarities of at least 0.9500 (Table 5(b)).

5 Discussion

Empirical and searched templates change detector decisions across multiple configurations. J/A/P comparisons reveal differing sensitivities to magnitude ratios and phase differences. Low responses to template replacement and channel copying suggest that RawBMamba and mch-SSL rely more on within-channel acoustic cues. These complementary interventions reveal detector use of inter-channel relationships.

Parameterized templates selected using scores from a small set of recordings can be reused on new recordings and some other detectors, reducing dependence on genuine reference recordings from the evaluation scene and per-recording search. In independent complete-utterance evaluation, transcripts and speaker representations retain high consistency. These results connect decision responses to reusable input transformations with explicit inter-channel acoustic quantities.

Future work will use physical playback and cross-scene experiments to examine RTF control in actual acoustic environments, and explore RTF-based data augmentation to improve detector robustness to spatial changes.

AI USAGE STATEMENT

The authors independently developed the core research ideas and theoretical contributions. We primarily used GPT-6 Astra for language polishing and grammar revision, and GPT-5.6 Sol to assist with coding. The authors take full responsibility for the final manuscript, code implementation, and experimental results.

REFERENCES

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. XLS-R: Self-Supervised Cross-Lingual Speech Representation Learning at Scale. In *Interspeech 2022*, pp. 2278–2282, 2022. doi: 10.21437/Interspeech.2022-143.

Şule Bekiryazıcı, Cemal Hanilçi, and Neyir Ozcan. Toward Robust Replay Attack Detection in Automatic Speaker Verification: A Study of Spectrum Estimation and Channel Magnitude Response Modeling. *Computer Speech & Language*, 98:101906, 2026. doi: 10.1016/j.csl.2025.101906.

Yujie Chen, Jiangyan Yi, Jun Xue, Chenglong Wang, Xiaohui Zhang, Shunbo Dong, Siding Zeng, Jianhua Tao, Zhao Lv, and Cunhang Fan. RawBMamba: End-to-End Bidirectional State Space Model for Audio Deepfake Detection. In *Interspeech 2024*, pp. 2720–2724, 2024. doi: 10.21437/Interspeech.2024-698.

Israel Cohen. Relative Transfer Function Identification Using Speech Signals. *IEEE Transactions on Speech and Audio Processing*, 12(5):451–459, 2004. doi: 10.1109/TSA.2004.832975.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Interspeech 2020*, pp. 3830–3834, 2020. doi: 10.21437/Interspeech.2020-2650.

Muhammad Umar Farooq, Awais Khan, Kutub Uddin, and Khalid Mahmood Malik. Transferable Adversarial Attacks on Audio Deepfake Detection. In *WACV Workshops 2025*, pp. 1640–1649, 2025. URL https://openaccess.thecvf.com/content/WACV2025W/MAPA/html/Farooq_Transferable_Adversarial_Attacks_on_Audio_Deepfake_Detection_WACVW_2025_paper.html.

Yuan Gong, Jian Yang, Jacob Huber, Mitchell MacKnight, and Christian Poellabauer. ReMASC: Realistic Replay Attack Corpus for Voice Controlled Systems. In *Interspeech 2019*, pp. 2355–2359, 2019. doi: 10.21437/Interspeech.2019-1541.

Yuan Gong, Jian Yang, and Christian Poellabauer. Detecting Replay Attacks Using Multi-Channel Audio: A Neural Network-Based Method. *IEEE Signal Processing Letters*, 27:920–924, 2020. doi: 10.1109/LSP.2020.2996908.

Daniel W. Griffin and Jae S. Lim. Signal Estimation from Modified Short-Time Fourier Transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984. doi: 10.1109/TASSP.1984.1164317.

Mingrui He, Longting Xu, Han Wang, Mingjun Zhang, and Rohan Kumar Das. Device Feature based on Graph Fourier Transformation with Logarithmic Processing For Detection of Replay Speech Attacks. In *Odyssey 2024*, pp. 137–144, 2024. doi: 10.21437/odyssey.2024-20.

Piotr Kawa, Marcin Plata, and Piotr Syga. Defense Against Adversarial Attacks on Audio DeepFake Detection. In *Interspeech 2023*, pp. 5276–5280, 2023. doi: 10.21437/Interspeech.2023-409.

Tomi Kinnunen, Md Sahidullah, Hector Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection. In *Interspeech 2017*, pp. 2–6, 2017. doi: 10.21437/Interspeech.2017-1111.

Xinfeng Li, Kai Li, Yifan Zheng, Chen Yan, Xiaoyu Ji, and Wenyuan Xu. SafeEar: Content Privacy-Preserving Audio Deepfake Detection. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3585–3599, 2024a. doi: 10.1145/3658644.3670285.

Yuang Li, Min Zhang, Mengxin Ren, Xiaosong Qiao, Miaomiao Ma, Daimeng Wei, and Hao Yang. Cross-Domain Audio Deepfake Detection: Dataset and Analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4977–4983, 2024b. doi: 10.18653/v1/2024.emnlp-main.286.

Zhuohang Li, Cong Shi, Tianfang Zhang, Yi Xie, Jian Liu, Bo Yuan, and Yingying Chen. Robust Detection of Machine-Induced Audio Attacks in Intelligent Audio Systems with Microphone Array. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1884–1899, 2021. doi: 10.1145/3460120.3484755.

Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and Kong Aik Lee. ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2507–2522, 2023. doi: 10.1109/TASLP.2023.3285283.

Shmulik Markovich-Golan, Sharon Gannot, and Walter Kellermann. Performance Analysis of the Covariance-Whitening and the Covariance-Subtraction Methods for Estimating the Relative Transfer Function. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pp. 2499–2503, 2018. doi: 10.23919/EUSIPCO.2018.8553007.

Yan Meng, Jiachun Li, Matthew Pillari, Arjun Deopujari, Liam Brennan, Hafsah Shamsie, Haojin Zhu, and Yuan Tian. Your Microphone Array Retains Your Identity: A Robust Voice Liveness Detection System for Smart Speakers. In *31st USENIX Security Symposium*, pp. 1077–1094, 2022. URL https://www.usenix.org/conference/usenixsecurity22/presentation/meng.

Nicolas M. Müller, Nicholas Evans, Hemlata Tak, Philip Sperl, and Konstantin Böttinger. Harder or Different? Understanding Generalization of Audio Deepfake Detection. In *Interspeech 2024*, pp. 2705–2709, 2024. doi: 10.21437/INTERSPEECH.2024-247.

Michael Neri and Tuomas Virtanen. Impact of Microphone Array Mismatches to Learning-Based Replay Speech Detection. In *Proceedings of the 33rd European Signal Processing Conference (EUSIPCO)*, pp. 1243–1247, 2025a. doi: 10.23919/EUSIPCO63237.2025.11226319.

Michael Neri and Tuomas Virtanen. Multi-Channel Replay Speech Detection Using an Adaptive Learnable Beamformer. *IEEE Open Journal of Signal Processing*, 6:530–535, 2025b. doi: 10.1109/OJSP.2025.3568758.

Michael Neri and Tuomas Virtanen. Multi-Channel Replay Speech Detection using Acoustic Maps. In *Proceedings of the 34th European Signal Processing Conference (EUSIPCO)*, pp. 1212–1216, 2026. URL https://eurasip.org/Proceedings/Eusipco/Eusipco2026/pdfs/0001212.pdf.

Binh Nguyen, Shuju Shi, Ryan Ofman, and Thai Le. What You Read Isn't What You Hear: Linguistic Sensitivity in Deepfake Speech Detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15741–15755, 2025. doi: 10.18653/v1/2025.emnlp-main.794.

NVIDIA. Parakeet TDT 0.6B V2. Model card, 2025. Released May 1, 2025. Accessed September 24, 2026. URL https://huggingface.co/nvidia/parakeet-tdt-0.6b-v2.

Michele Panariello, Wanying Ge, Hemlata Tak, Massimiliano Todisco, and Nicholas Evans. Malafide: A Novel Adversarial Convolutive Noise Attack Against Deepfake and Spoofing Detection Systems. In *Interspeech 2023*, pp. 2868–2872, 2023. doi: 10.21437/Interspeech.2023-703.

Octavian Pascu, Adriana Stan, Dan Oneata, Elisabeta Oneata, and Horia Cucu. Towards Generalisable and Calibrated Audio Deepfake Detection with Self-supervised Representations. In *Interspeech 2024*, pp. 4828–4832, 2024. doi: 10.21437/Interspeech.2024-1302.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Chandan K A Reddy, Vishak Gopal, and Ross Cutler. DNSMOS P.835: A Non-Intrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 886–890, 2022. doi: 10.1109/ICASSP43922.2022.9746108.

Karla Schäfer, Jeong-Eun Choi, and Martin Steinebach. Audio Deepfake Detection Under Post-Processing Attack. In *33rd European Signal Processing Conference (EUSIPCO 2025)*, pp. 586–590, 2025. doi: 10.23919/EUSIPCO63237.2025.11226066.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-Vectors: Robust DNN Embeddings for Speaker Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, 2018. doi: 10.1109/ICASSP.2018.8461375.

Massimiliano Todisco, Michele Panariello, Xin Wang, Héctor Delgado, Kong Aik Lee, and Nicholas Evans. Malacopula: adversarial automatic speaker verification attacks using a neural-based generalised Hammerstein model. In *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, pp. 94–100, 2024. doi: 10.21437/ASVspoof.2024-14.

Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian. Wespeaker: A Research and Production Oriented Speaker Embedding Learning Toolkit. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. URL https://arxiv.org/abs/2210.17016.

Xin Wang, Héctor Delgado, Hemlata Tak, Jee-weon Jung, Hye-jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, Nicholas Evans, Kong Aik Lee, Junichi Yamagishi, Myeonghun Jeong, Ge Zhu, Yongyi Zang, You Zhang, Soumi Maiti, Florian Lux, Nicolas Müller, Wangyou Zhang, Chengzhe Sun, Shuwei Hou, Siwei Lyu, Sébastien Le Maguer, Cheng Gong, Hanjie Guo, Liping Chen, and Vishwanath Singh. ASVspoof 5: Design, Collection and Validation of Resources for Spoofing, Deepfake, and Adversarial Attack Detection Using Crowdsourced Speech. *Computer Speech & Language*, 95:101825, 2026. doi: 10.1016/j.csl.2025.101825.

Haibin Wu, Yuan Tseng, and Hung-yi Lee. CodecFake: Enhancing Anti-Spoofing Models Against Deepfake Audios from Codec-Based Speech Synthesis Systems. In *Interspeech 2024*, pp. 1770–1774, 2024. doi: 10.21437/Interspeech.2024-2093.

Takuo Yamaguchi, Sayaka Shiota, and Naohiro Tawara. Investigating Self-Supervised Learning-Based Front-End for Multi-Channel Replay Attack Detection. In *APSIPA Annual Summit and Conference*, pp. 2098–2103, 2025. doi: 10.1109/APSIPAASC65261.2025.11249385.

Qiang Yang, Kaiyan Cui, and Yuanqing Zheng. VoShield: Voice Liveness Detection with Sound Field Dynamics. In *IEEE Conference on Computer Communications (INFOCOM)*, pp. 1–10, 2023. doi: 10.1109/INFOCOM53939.2023.10229038.

Zhiyuan Yu, Yuanhaur Chang, Ning Zhang, and Chaowei Xiao. SMACK: Semantically Meaningful Adversarial Audio Attack. In *32nd USENIX Security Symposium*, pp. 3799–3816, 2023. URL https://www.usenix.org/conference/usenixsecurity23/presentation/yu-zhiyuan-smack.

Tong Zhang, Yihuan Huang, and Yanzhen Ren. EchoFake: A Replay-Aware Dataset for Practical Speech Deepfake Detection. In *ICASSP 2026*, pp. 15502–15506, 2026. doi: 10.1109/ICASSP55912.2026.11464778.

This appendix provides experimental settings, complete results, and theoretical derivations. Appendix A describes RTF visualization and distance statistics; B and C report empirical-template and parameterized-search experiments; D evaluates speech-content and speaker-representation preservation; E and F provide theoretical derivations and algorithms, respectively.

J, A, and P denote joint, magnitude-ratio-only, and phase-difference-only control; AM abbreviates Acoustic Maps. Rates are percentages with two decimal places. AUC, cosine similarity, and its quantiles use four decimals. Magnitude-ratio and phase-difference distances are in dB and rad, respectively, with three decimals. NA denotes an inapplicable statistic or empty denominator; 0 denotes a valid denominator with no conversion.

A RTF Visualization and Distance Statistics

A.1 Paired Recordings and RTF Estimation

To examine RTF structure across utterances, we select five genuine–replay pairs from ReMASC D3 training recordings. Each pair shares speaker identity and utterance content, while the five pairs use different utterances, controlling identity and content differences within pairs. The array, environment, speaker, and position are fixed to ReSpeaker Core V2.0, Env2, Speaker 29, and Position 25. The replay-source recorder and playback device are also fixed. Parakeet and Whisper verify paired content. Example selection considers valid-frequency coverage, within-class similarity across utterances, and between-class differences.

RTF estimation. Complete recordings sampled at 44.1 kHz are analyzed by STFT with a 1024-point Hann window and a 256-sample hop. Visualization channels are numbered from 1: reference ch1 corresponds to ch0 in the other experiments, and ch2–ch6 correspond to the remaining five channels. Frames are ranked by reference-channel spectral energy within 80–16000 Hz. Approximately the highest 50.00% and lowest 20.00% estimate speech-present and noise statistics, respectively, followed by covariance subtraction to obtain the RTF.

Coherence and valid frequency bins. We compute observed coherence to measure the stability of channel relationships at each frequency:

$$\gamma^2_{m1}(f) = \frac{|\widehat{C}^{\mathrm{sp}}_{m1}(f)|^2}{\widehat{P}^{\mathrm{sp}}_{m}(f)\widehat{P}^{\mathrm{sp}}_{1}(f)}.$$

Here f is frequency and m indexes non-reference channels. $\widehat{C}^{\mathrm{sp}}_{m1}$ is the observed cross-power spectrum over speech-present frames, and $\widehat{P}^{\mathrm{sp}}_{m}$ and $\widehat{P}^{\mathrm{sp}}_{1}$ are the corresponding auto-power spectra. Superscript sp indicates speech-present frames, whose statistics contain both speech and noise. The RTF estimate is $\widehat{R}_m(f)$, with magnitude ratio and phase difference represented by $20\log_{10}|\widehat{R}_m(f)|$ and $\arg\widehat{R}_m(f)$.

Valid bins must satisfy coherence, noise-subtracted reference-power, and numerical-stability requirements, with coherence threshold 0.6. RTF comparisons use only valid bins; hatching marks invalid bins. Coherence panels display all observed values.

Display and local comparison. Feature maps cover 100–8000 Hz. Color scales are [0,1] for coherence, ± 9 dB for RTF magnitude ratios, and $[-\pi, \pi]$ for phase differences, using a cyclic phase scale. Magnitudes outside the display range receive endpoint colors; statistics retain the original estimates. Reference-channel spectrograms use a 512-point Hann window and a 128-sample hop, with shared peak normalization and a dB color scale.

Based on common valid coverage and local differences across utterances, we select ch5 at 6.0–6.5 kHz for phase comparison. All utterances use the same channel, band, and common valid bins. Circular angular distance accounts for phase periodicity.

Within each class of paired recordings, inter-channel magnitude–phase relationships are relatively similar across utterances, while genuine and replay speech exhibit different frequency structures.

A.2 Matched Samples and Distance Statistics

We further compare within- and between-class channel differences through distance statistics. Samples come from the same array's training split. Before reading audio or estimating RTFs, we sample by a fixed-seed hash order and then retain Env2 recordings. We independently draw 56 quartets under speaker matching and 44 under joint speaker–position matching. Common-valid-bin screening retains 6 and 16 quartets, respectively, each covering 5 speakers.

Each quartet contains two genuine and two replay recordings, yielding six recording pairs. RTF estimation follows Appendix A.1, using at most the first 8 s of each recording. All six pairs share the intersection of valid bins across the four recordings. This intersection must contain at least 100 channel–frequency cells, with at least two channels having 20 valid bins each. Over these cells, magnitude-ratio distance is the median absolute difference in log-magnitude ratios, and phase-difference distance is the median circular angular distance.

Statistical weights. Each quartet yields one genuine–genuine, one replay–replay, and four genuine–replay distances. In weighted empirical cumulative distributions, quartets from the same speaker receive equal weight, and speakers have equal total weight. Each between-class distance receives one quarter of its quartet's between-class weight. For speaker-level comparisons, the two within-class and four between-class distances are averaged separately within each quartet, followed by arithmetic averaging over that speaker's quartets. Connecting lines indicate the difference between between-class and within-class means.

Distance results. Under joint speaker–position matching, weighted mean magnitude-ratio distances for genuine–genuine, replay–replay, and genuine–replay pairs are 0.504, 0.581, and 1.764 dB; corresponding phase-difference distances are 0.057, 0.069, and 0.276 rad. Across speakers, between-class minus within-class mean differences range from 0.397 to 1.643 dB for magnitude ratios and from 0.066 to 0.329 rad for phase differences. All between-class means exceed their within-class counterparts.

Class differences in inter-channel magnitude ratios and phase differences remain after matching speaker and position. Introducing genuine RTFs into replay speech further tests their influence on detector decisions.

B Empirical Templates and Channel Copying

B.1 Data Splits and Detection Baselines

Data splits. Detection experiments cover ReMASC arrays D1–D4. Train, Dev, and Test contain 35, 5, and 10 speakers, respectively, with no overlap. We select genuine and human-speech replay recordings satisfying `speech_type`$\in \{2, 3\}$, `speaker`> 0, and `source_recorder`$\neq 3$; these fields denote speech type, speaker ID, and source-recorder ID. Env1–Env4 correspond to outdoor, indoor 1, indoor 2, and in-vehicle environments. Table B1 lists recording counts.

Table B1: Recording counts by array. The counting unit is the recording file.

Array	Channels	Sample rate (Hz)	Train	Dev	Test
D1	2	44100	4585	772	1633
D2	4	44100	5240	886	1966
D3	6	44100	5015	847	1905
D4	7	16000	5321	840	1920
Total	—	—	20161	3345	7424

Models and input processing. Each model–array combination uses a fixed checkpoint, native preprocessing, and a Dev-selected threshold. Acoustic Maps (AM), M-ALRAD, MIAD, VoShield, and mch-SSL take the first second at the native sampling rate, with right zero-padding when shorter. AM, M-ALRAD, MIAD, and mch-SSL preserve input amplitude scale during preprocessing.

AM uses its native acoustic-map extractor and sampling rate. M-ALRAD uses 32 ms windows, corresponding to 1411 and 512 samples at 44.1 and 16 kHz. MIAD extracts magnitude and phase features with a 512-point FFT; VoShield extracts sound-field dynamics (SFD) from all unordered channel pairs. mch-SSL (Yamaguchi et al., 2025) converts inputs to 16 kHz, extracts features using

an XLS-R (Babu et al., 2022) frontend with pretrained weights, and classifies them through a VGG backend.

MC-RAD (Gong et al., 2020) resamples constructed D4 waveforms to 44.1 kHz, takes the first second with zero-padding if needed, and applies a common peak-normalization factor across channels. Our multichannel RawBMamba (Chen et al., 2024) first takes the valid first-second segment at the native rate, resamples it to 16 kHz, and pads to 16000 samples. It then extracts and averages channel embeddings.

Replay scores. AM, M-ALRAD, MIAD, and mch-SSL use class-1 softmax probabilities; VoShield uses sigmoid output. MC-RAD uses the class-1 minus class-0 logit, where logits are classification scores before probability transformation. RawBMamba uses the class-0 minus class-1 output of its cosine classification head. Higher scores favor replay for every model; score magnitudes are compared within each model.

Baselines and metrics. Table B2 reports performance on original Test recordings for all 28 model–array combinations. ACC is the fraction correctly classified, and AUC is the area under the receiver operating characteristic curve. TP and TN count originally correctly classified replay and genuine recordings, respectively; replay-to-genuine and genuine-to-replay conversion rates use TP and TN as their denominators. Construction or scoring failures remain in the denominator; empty denominators are NA. Overall rates are computed after separately pooling conversion counts and denominators.

Table B2: Detector performance on original Test recordings. ACC is in %; AUC has four decimal places.

Model	D1		D2		D3		D4	
	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
AM	0.7963	73.79	0.8651	74.21	0.9612	90.03	0.8739	80.26
M-ALRAD	0.9488	91.06	0.9691	89.93	0.9702	90.03	0.9706	90.99
MC-RAD	0.9082	83.28	0.9373	84.74	0.8925	82.20	0.8858	80.16
MIAD	0.9308	85.43	0.8893	81.74	0.9028	84.51	0.9050	82.03
VoShield	0.9493	85.12	0.8399	71.87	0.9404	88.61	0.9608	88.02
RawBM.	0.9624	89.77	0.9545	88.56	0.9360	85.98	0.9248	81.20
mch-SSL	0.9841	93.75	0.9842	94.10	0.9661	92.18	0.9344	88.96

B.2 Empirical-Template Estimation and Source Comparison

Empirical templates are estimated from genuine recordings. We compare Test/LOSO and Train sources to determine whether templates still affect detector decisions when reference recordings change. Each source covers 28 model–array combinations.

Template estimation. We use the first channel, ch0, as reference and estimate templates from at most the first 8 s of source recordings. STFT uses a Hann window, a 1024-point FFT, and a 256-sample hop, over 80 Hz to $\min(16\,\mathrm{kHz}, \mathrm{Nyquist})$. The lowest-energy 20.00% of frames estimate noise; the highest-energy 50.00% estimate speech plus noise. These sets are disjoint. Estimation fails with fewer than 8 noise frames or 16 speech-present frames.

Complex RTFs are estimated by covariance subtraction. The reference-power regularizer is 1×10^{-4} of its maximum, and RTF magnitude ratios are limited to ± 18 dB. Invalid bins receive unit response: magnitude ratio 1 and phase difference 0.

Magnitude–phase control. J retains the complete complex RTF; A retains magnitude ratios with zero phase differences; P retains phase differences with unit magnitude ratios. Each condition multiplies the corresponding RTF by the ch0 STFT to reconstruct the other channels. The three conditions share an amplitude scaling factor for each recording, limiting waveform peaks to 0.95.

Test/LOSO templates. LOSO denotes Leave-One-Speaker-Out. We exclude all recordings of the target speaker from the genuine Test reference pool before estimating and assigning templates. Reference recordings pass acoustic-validity screening. When same-environment references are unavailable, templates from other environments are used according to predefined rules. The reference pool varies by target speaker, while detectors remain fixed. Reference and target recordings may contain the same utterance.

Train-source control. The control keeps the detector, target recordings, and decision threshold fixed, changing only the template source to genuine Train recordings. Each source determines its own template pool, assignments, and amplitude scaling factors.

Train templates group Env2–Env4 recordings by environment, position, and recording device. In Env1, recordings with unknown position, coded as -1, receive file-level templates. Assignment prioritizes the environment and uses other-environment templates under predefined rules when necessary. Assignment order is determined by `SHA256(seed|bank|file_id)`, whose variables are the fixed seed, template-bank ID, and recording ID.

Table B3: J/A/P conversion rates (%) of Train-source empirical templates. Conditions within each model–array combination share the originally correctly detected replay set.

Model	D1			D2			D3			D4		
	J	A	P	J	A	P	J	A	P	J	A	P
AM	9.87	0.11	10.73	89.39	81.39	80.09	92.33	36.00	93.08	90.67	56.77	91.03
M-ALRAD	3.54	1.85	7.07	77.45	55.54	92.64	46.86	97.94	88.78	29.63	10.96	86.04
MC-RAD	3.93	3.00	6.84	43.11	23.41	46.82	49.68	55.19	42.82	27.00	29.88	30.60
MIAD	0.09	0.00	0.09	58.88	0.08	80.20	33.67	7.58	27.67	48.45	41.03	37.75
VoShield	58.59	71.09	98.73	88.54	88.99	45.73	97.68	97.60	44.76	34.02	19.85	65.32
RawBM.	4.35	4.44	4.08	5.41	5.82	4.18	4.24	4.41	3.63	7.00	6.82	4.88
mch-SSL	1.29	1.03	1.97	3.81	4.97	2.41	2.91	3.23	3.15	8.46	7.63	4.98

Table B4: Overall conversion rates of Train and Test/LOSO empirical templates, computed from pooled original counts. Differences are Test/LOSO minus Train in percentage points (pp), calculated before rounding.

Condition	Train	Emp.	Difference (pp)
J (Joint)	35.96	31.71	-4.25
A (Magnitude)	28.48	27.89	-0.59
P (Phase)	39.45	33.88	-5.57

In Table B4, Train templates have higher overall conversion than Test/LOSO under all three conditions. The difference is largest under phase-only control and smallest under magnitude-only control. Empirical templates thus still induce replay misclassification across multiple models and arrays after changing the reference-data source.

The relative effects of J, A, and P vary across models and arrays. Some configurations attain high conversion by changing only magnitude ratios or phase differences, reflecting different detector dependencies on these channel cues. RawBMamba and mch-SSL retain low conversion under both template sources.

B.3 Channel Copying and Low Detector Responses

To examine whether the two models' low conversion relates to within-channel cues, we enumerate every copying source for RawBMamba and mch-SSL and compare decisions after removing inter-channel differences.

Copying settings. The selected source channel's original time-domain waveform is copied to every input channel, making the channels identical. Copying from the first channel is denoted `copy_ch0`. Evaluation covers full Test for D1–D4, with source indices 0–1, 0–3, 0–5, and 0–6, respectively, retaining native preprocessing and fixed thresholds. Each model–array combination shares an evaluation set. Bidirectional flip rates are computed separately on originally correctly classified genuine and replay recordings.

Figure B1 shows low bidirectional flip rates for both models across copying sources. A single channel waveform retains most original decisions after inter-channel differences are removed. Together with low empirical-template conversion, these results support the interpretation that the models rely more on within-channel acoustic cues.

Empirical templates and channel copying reveal different model responses: several detectors change decisions with inter-channel magnitude–phase relationships, whereas RawBMamba and mch-SSL respond weakly to both operations.

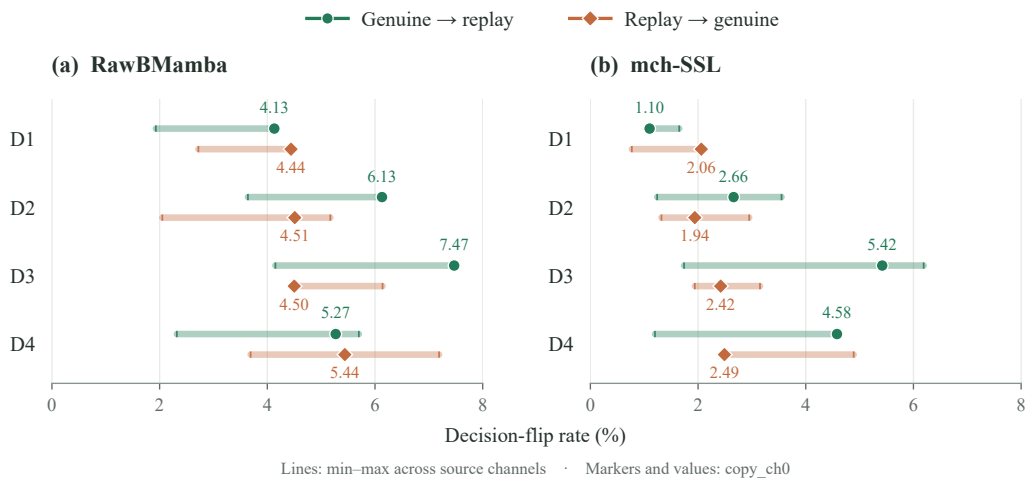


Figure B1: Bidirectional decision-flip rates across copying sources. Segments show minimum–maximum rates; markers and numbers indicate `copy_ch0`. Green circles and orange diamonds denote genuine-to-replay and replay-to-genuine flips. D1–D4 enumerate 2, 4, 6, and 7 sources, respectively. Each direction uses its originally correctly classified genuine or replay count as denominator.

C Black-Box Search and Template Reuse

C.1 Search Settings

Without genuine reference recordings from the evaluation scene, we generate parameterized RTF candidates and select templates using detection scores from a small set of Dev replays. For each of D1–D4, 16 replays are preselected from each of four environments, giving 64 per array. Models within an array share these recordings, but each model selects templates using only its originally correctly classified replays. Search is independent for each model–array–parameterization combination, giving 56 combinations, with fixed checkpoints, Dev thresholds, and recording sets.

After Dev selection, templates remain fixed for final full-Test evaluation. Table C1 lists array construction settings and Test counts.

Table C1: Array settings and Test recording counts.

Array	Channels	Sample rate (Hz)	Assumed geometry	Highest nominal band (Hz)	Full-Test replays
D1	2	44100	Two microphones, 60 mm spacing	4000–8000	1237
D2	4	44100	Linear array, 50 mm adjacent spacing	4000–8000	1367
D3	6	44100	Six-microphone ring, 47 mm radius	4000–8000	1337
D4	7	16000	Six-microphone ring plus center, 75 mm diameter	4000–7600	1365

Candidate parameterization. All arrays use a Hann window, a 1024-point FFT, and a 256-sample hop. Nominal band boundaries are 300, 1500, and 4000 Hz and the upper limit in Table C1, with a 200 Hz transition width. Including outer tapers, window support spans 100–8200 Hz for D1–D3 and 100–7800 Hz for D4. D4 has a nominal upper boundary of 7600 Hz; its taper endpoint also lies below its 8000 Hz Nyquist frequency. Magnitude and phase scales are bounded by $[-2.0000, 2.0000]$, with an 8 dB relative-gain limit. The equivalent delay defined by the phase expression, $\max_{m,f} |(\delta_m - \delta_r)\alpha_\phi(f)/c_0|$, is limited to $24/44100$ s, corresponding to $24 \times 16000/44100$ samples for D4. Group delay depends on the frequency derivative of phase, as defined in Appendix E.3. Identical FFT sizes and hops represent different durations at different sampling rates.

shared2 shares one magnitude scale and one phase scale across all three bands; band6 assigns them separately to each band, giving six parameters. Parameters come from an unscrambled Halton low-discrepancy sequence, with points linearly mapped to $[-2, 2]$. Starting at index 1, we generate 128 consecutive candidates. The six band6 dimensions correspond to $(\alpha_{A,1}, \alpha_{A,2}, \alpha_{A,3}, \alpha_{\phi,1}, \alpha_{\phi,2}, \alpha_{\phi,3})$, with bases 2, 3, 5, 7, 11, and 13. The two shared2 dimensions represent the shared magnitude and phase scales, with bases 2 and 3. Two baseline candidates set all parameters to 0 or 1, giving 130 joint candidates per parameterization. The fixed order is all-zero baseline, all-one baseline, then Halton candidates.

Microphone coordinates for D1–D4 follow the geometries in Table C1 to generate digital candidate responses, preserving channel order. Each array is centered at the coordinate origin with all microphones in the horizontal plane. D1 and D2 run along the x-axis from positive to negative. Ring microphones in D3 and D4 are numbered counterclockwise from the positive x-axis; D4's center microphone is ch6. All arrays use ch0 as reference. The virtual source is 1.5 m from the array center, at azimuth 30° and elevation 45°, with sound speed 343 m/s.

For each recording, all candidate waveforms and the original share an amplitude scaling factor. It is determined by the largest peak across the original and all waveforms constructed from the full response bank, limiting scaled peaks to 0.7080. Scaling is computed separately by model, array, parameterization, and recording.

Template selection. The candidate set is fixed in advance, and detection scores are the only model feedback. We select the J template with the lowest mean Dev replay score and derive A and P directly to compare its magnitude and phase components. Selected parameters remain fixed across recordings.

C.2 Cross-Recording Evaluation and Parameterization Comparison

We compare both parameterizations and all J/A/P conditions on full Test across four arrays to assess fixed-template performance on new recordings. shared2 and band6 have equal candidate counts but different freedom to adjust frequency bands. Each model–array combination shares its originally correctly detected replay set; construction or scoring failures remain in the denominator.

Cross-array comparisons jointly vary channel count, geometry, sampling rate, data, and checkpoints; results reflect their combined effects. Cross-recording rates are computed from original counts, pooling conversion counts and denominators separately for summaries. A recording or speaker may participate in several configurations.

Table C2: Full-Test conversion rates (%) of fixed searched templates. Each parameterization selects its joint template on Dev and derives A/P from it. Conditions within each model–array combination share the evaluation set.

(a) shared2

Model	D1			D2			D3			D4		
	J	A	P	J	A	P	J	A	P	J	A	P
AM	100.00	0.21	100.00	98.81	74.03	98.81	100.00	48.00	100.00	100.00	84.04	100.00
M-ALRAD	91.84	10.52	91.25	96.85	91.75	96.12	100.00	100.00	100.00	100.00	99.35	100.00
MC-RAD	32.68	6.74	32.21	54.06	13.34	51.50	43.72	42.28	43.27	40.68	38.52	39.33
MIAD	0.00	0.00	0.00	14.64	0.00	14.81	77.92	4.00	76.83	100.00	36.52	100.00
VoShield	100.00	100.00	100.00	100.00	100.00	100.00	0.00	0.00	0.00	0.08	0.00	0.00
RawBM.	4.26	3.72	3.90	7.71	8.45	7.47	5.10	5.28	5.02	6.64	6.64	6.64
mch-SSL	2.91	2.83	2.83	2.25	2.49	2.25	2.67	2.50	2.83	4.48	3.15	4.56

(b) band6

Model	D1			D2			D3			D4		
	J	A	P	J	A	P	J	A	P	J	A	P
AM	99.68	0.21	99.68	99.68	73.92	99.68	100.00	48.00	100.00	100.00	84.22	100.00
M-ALRAD	95.88	10.10	96.38	97.90	92.24	97.41	100.00	100.00	100.00	100.00	99.68	100.00
MC-RAD	32.30	5.06	32.30	67.84	10.95	66.70	60.34	42.19	59.35	44.91	37.71	43.92
MIAD	0.00	0.00	0.00	16.07	0.00	15.74	94.00	4.00	93.58	100.00	37.67	100.00
VoShield	100.00	100.00	100.00	100.00	100.00	100.00	98.72	0.00	1.12	0.08	0.00	0.00
RawBM.	5.99	5.08	5.99	9.02	8.61	7.79	4.76	5.19	5.10	7.19	7.19	7.10
mch-SSL	3.17	3.08	3.08	4.43	2.33	4.04	3.07	2.26	3.47	6.22	3.57	5.98

Under both parameterizations, AM and M-ALRAD have high joint conversion across four arrays, whereas RawBMamba and mch-SSL have low conversion. Band-wise control mainly affects MC-RAD, MIAD, and VoShield. MC-RAD has higher band6 than shared2 joint conversion on D2–D4, with similar results on D1. VoShield changes from low shared2 to high band6 conversion on D3. Template effects thus depend on both magnitude–phase components and frequency-band parameterization.

Sensitivity to magnitude–phase changes. AM has identical J and P conversion under both parameterizations, exceeding A; phase differences alone retain these templates' effects. VoShield's band6 templates yield high conversion in all three conditions on D1 and D2 but low conversion on D4. On D3, J/A/P yield 98.72%, 0.00%, and 1.12%, respectively, indicating dependence on joint magnitude-ratio and phase-difference changes.

Dev versus full Test. We use the same joint band6 template on both sets and compare conversion on selection recordings and new recordings.

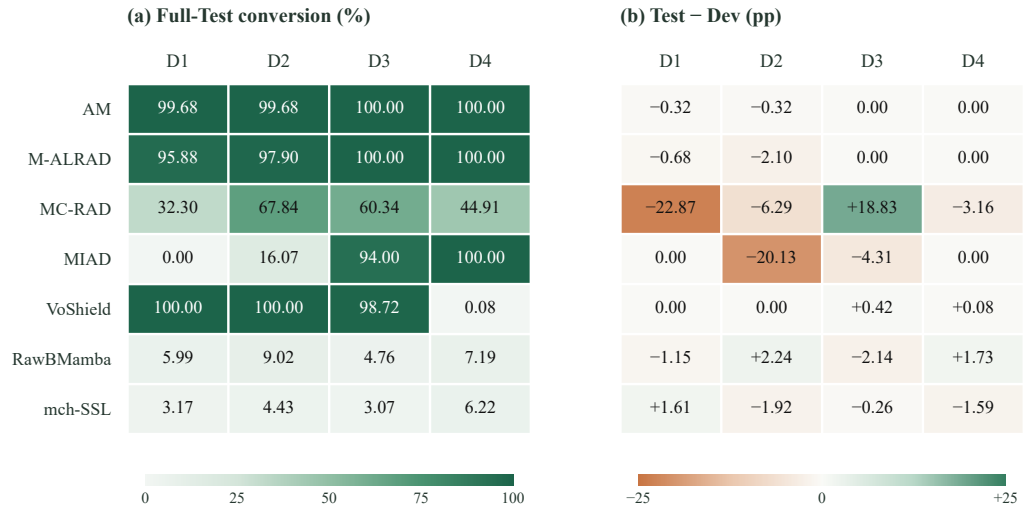


Figure C1: Fixed-template full-Test conversion and its change from Dev. Left: full-Test conversion (%); right: Test minus Dev (percentage points). Rows and columns identify detectors and arrays. Each model–array combination uses the same joint band6 template and threshold on both sets. Green and orange in the right panel indicate increased and decreased conversion.

AM and M-ALRAD reach full-Test conversion of 99.68%–100.00% and 95.88%–100.00%, respectively, close to Dev results, retaining high conversion on new recordings. MC-RAD and MIAD show larger changes on some arrays, with cross-recording performance varying by array configuration. RawBMamba and mch-SSL have low conversion on both sets and respond weakly to these templates.

C.3 Cross-Model Transfer of Fixed Templates

Following cross-recording evaluation, we apply each source model's selected joint band6 template to other detectors on the same array, testing transfer without using target-model feedback for selection. Full-Test evaluation across four arrays yields 196 source–target model pairs, including 168 cross-model pairs.

For each target model, all seven source templates share its originally correctly detected Test replays; construction or scoring failures remain in the denominator. RTF parameters and complex responses remain fixed, and target-model scores are used only for evaluation. Model weights, Dev thresholds, native preprocessing, and per-recording inference settings are unchanged. Amplitude scaling follows Appendix C.1 using the input segment actually processed by the target model.

Transfer results. Mean cross-model conversion for D1–D4 is 33.16%, 31.84%, 40.06%, and 36.91%, respectively. The four-array mean is 35.49%, below the same-model mean of 55.40%. Cross-model and same-model means equally weight unrounded off-diagonal and diagonal rates, respectively. Fixed

Table C3: Full-Test cross-model conversion rates (%) of fixed band6 templates. Rows and columns identify source and target models. Bold entries are same-model results; other entries are cross-model results. All seven templates in a column share the evaluation set. Templates are selected separately for each array, and transfer is evaluated within each array. M-AL: M-ALRAD; MC-R: MC-RAD; VoS: VoShield; RawB: RawBMamba; SSL: mch-SSL.

(a) D1

Source / target	AM	M-AL	MC-R	MIAD	VoS	RawB	SSL
AM	**99.68**	86.28	1.50	0.00	100.00	5.44	2.83
M-AL	92.16	**95.88**	0.37	0.00	100.00	5.99	2.74
MC-R	16.33	0.17	**32.30**	0.00	100.00	5.54	2.83
MIAD	99.68	91.33	0.37	**0.00**	100.00	5.35	3.08
VoS	34.37	37.71	0.47	0.00	**100.00**	6.17	3.68
RawB	72.82	86.70	0.37	0.00	100.00	**5.99**	3.60
SSL	79.16	36.20	4.21	0.00	100.00	5.08	**3.17**

(b) D2

Source / target	AM	M-AL	MC-R	MIAD	VoS	RawB	SSL
AM	**99.68**	76.72	51.41	3.13	0.22	6.73	2.02
M-AL	72.73	**97.90**	3.36	7.28	96.52	8.78	4.20
MC-R	65.69	53.11	**67.84**	6.85	96.75	6.97	2.10
MIAD	37.66	80.03	30.92	**16.07**	21.08	7.88	3.11
VoS	12.77	92.97	4.59	10.58	**100.00**	8.04	3.81
RawB	92.10	94.83	9.54	6.77	72.20	**9.02**	2.56
SSL	28.68	87.63	7.16	4.57	44.96	8.37	**4.43**

(c) D3

Source / target	AM	M-AL	MC-R	MIAD	VoS	RawB	SSL
AM	**100.00**	99.92	44.44	79.33	0.00	4.41	2.67
M-AL	98.83	**100.00**	37.76	94.00	0.00	4.93	3.31
MC-R	89.92	99.92	**60.34**	90.00	0.00	4.41	3.23
MIAD	98.83	100.00	37.76	**94.00**	0.00	4.93	3.31
VoS	68.58	100.00	40.65	41.50	**98.72**	4.50	3.47
RawB	88.17	99.17	25.47	28.08	0.00	**4.76**	4.28
SSL	46.83	100.00	24.75	0.42	0.00	4.67	**3.07**

(d) D4

Source / target	AM	M-AL	MC-R	MIAD	VoS	RawB	SSL
AM	**100.00**	100.00	29.97	98.41	0.08	7.19	4.65
M-AL	93.36	**100.00**	37.89	93.46	0.08	6.73	4.15
MC-R	65.74	11.61	**44.91**	93.81	0.08	7.19	4.48
MIAD	4.75	0.08	34.83	**100.00**	0.08	6.91	4.65
VoS	93.36	100.00	37.89	93.46	**0.08**	6.73	4.15
RawB	59.55	9.50	39.87	99.20	0.08	**7.19**	4.15
SSL	97.67	50.49	42.75	93.81	0.08	7.10	**6.22**

templates also change decisions of detectors not involved in selection, indicating shared sensitivity to selected inter-channel responses among some detectors.

AM and M-ALRAD are generally sensitive to templates from different sources, with cross-model means across arrays ranging from 51.61% to 81.86% and from 45.28% to 99.83%, respectively. MC-RAD, MIAD, and VoShield vary by template source and array; RawBMamba and mch-SSL generally remain low, continuing the model differences observed across recordings.

Directional transfer. Exchanging source and target models can markedly change conversion. On D1, AM templates achieve 100.00% on VoShield, versus 34.37% in reverse. On D4, AM templates achieve 98.41% on MIAD, versus 4.75% in reverse. Transfer therefore depends on the specific source–target combination.

D Speech-Content and Speaker-Representation Preservation

D.1 Evaluation Samples and Template Pairing

Preservation evaluation constructs complete utterances digitally and selects samples by their original speech quality. This set is determined independently of detection experiments, and the two types of evaluation are reported separately.

Original-audio quality screening. Screening starts from 7424 Test recording files across four arrays, comprising 36000 file–channel combinations. Eligible channels must load successfully, contain finite and nonzero waveforms, and pass voice activity detection (VAD) and the quality thresholds in Table D1. Files with at least one eligible channel enter the candidate set.

Speech quality is automatically estimated by DNSMOS P.835 (Reddy et al., 2022), using Microsoft's official implementation and weights. SIG, BAK, and OVRL denote speech, background, and overall quality. Scoring uses 16 kHz audio with a 9.01 s window and a 1 s stride. Short recordings are repeated to fill a window.

Initial quality screening retains 1087 candidate files. Applying the same criteria to replay reference ch0 and non-reference target channels retains 2185 eligible channels in 466 files. Of these, 440 files have eligible non-reference channels, giving 1719 targets. The remaining 26 files have only eligible ch0 and do not contribute evaluation pairs to the main results.

Template coverage and pairing. Test/LOSO uses banks excluding the target speaker, with fixed assignment rules. Files participating in result pairs cover 8 target speakers and 23 array–target-speaker

Table D1: Original-audio quality thresholds. RMS denotes root-mean-square amplitude; dBFS denotes decibels relative to digital full scale. All thresholds apply before waveform construction.

Item	Threshold
Active RMS	≥ -40 dBFS
Clipping definition	Absolute original amplitude ≥ 0.999
Clipped fraction	$< 0.10\%$
Longest clipped run	Fewer than 3 samples at the original sample rate
VAD support	Nonzero speech support
DNSMOS scores	Finite and within [1,5]
SIG / BAK / OVRL	At least 3 / 2.5 / 2.5, respectively

banks. Search includes shared2 and band6; deduplicating 56 configurations by complex response and complete construction settings yields 53 templates. Table D2 lists sample and pairing counts by array.

Table D2: Evaluation samples and template-pair counts. Both template families share evaluation files and target channels. The pairing unit is target channel $\times$ template. Planned pairs are counted per evaluation model and J/A/P condition; searched-template counts denote deduplicated acoustic sources.

Array	Evaluation files	Target channels	Empirical pairs	Searched templates	Searched pairs
D1	71	71	71	13	923
D2	80	214	214	14	2996
D3	187	843	843	13	10959
D4	102	591	591	13	7683
Total	440	1719	1719	53	22561

Each aggregate metric uses its corresponding valid pairs.

RTF construction uses complete recordings at their original sampling rates, followed by resampling to 16 kHz for downstream evaluation. All samples selected before construction participate. Amplitude scaling is recomputed for complete utterances, with searched templates retaining the full-response-bank peak rule. Detection uses cropped inputs, so the two stages may use different scaling factors.

D.2 Transcript Preservation

ASR models and settings. Transcript preservation is evaluated separately with Whisper small.en, Whisper large-v3-turbo (Radford et al., 2023), and Parakeet TDT 0.6B v2 (NVIDIA, 2025). Whisper weights come from small.en and large-v3-turbo; Parakeet uses sherpa-onnx INT8 ONNX weights. The Whisper systems are abbreviated Small.en and Turbo below. Each recognizer is summarized separately.

Both Whisper systems use English greedy decoding. Small.en sets `max_new_tokens=128`; Turbo uses a maximum length of 128 tokens with timestamps disabled. Original and constructed audio share weights and decoding settings within each pair. Parakeet uses the sherpa-onnx 1.13.8 INT8 ONNX CPU backend, extracts 80-dimensional features from 16 kHz audio, and applies greedy decoding without an external language model.

Metrics. Transcripts undergo Unicode NFKC normalization, case folding, punctuation-to-space conversion, and whitespace collapsing. Results are paired and deduplicated by file, channel, template source, model, and condition. Pairs are valid when both transcripts are successfully produced and nonempty. In Test/LOSO evaluation against original target channels, Parakeet has 5 incomparable pairs per J/A/P condition owing to empty transcripts, totaling 15. All metrics use actual valid pairs.

Sentence agreement is the fraction of valid pairs with identical before/after transcripts. Word-edit rate sums Levenshtein word distances across valid pairs and divides by the original-transcript word count. Character-edit rate sums Levenshtein character distances and divides by the original-transcript character count excluding spaces. These measures use original transcripts as references and quantify changes in recognition output. Summaries pool edit counts, agreement counts, and corresponding denominators before computing ratios.

Same-source reconstruction reference. Non-reference channels are reconstructed from ch0, so differences from original targets contain both reference-waveform replacement and RTF control. To examine the RTF contribution, empirical templates use unit-RTF reconstruction as reference, while

searched templates use time-domain ch0 copying. Both references share the reference speech and corresponding amplitude scaling; we collectively call them same-source reconstruction.

Table D3: ASR transcript preservation after RTF control (%). Empirical RTF denotes Test/LOSO templates. Higher sentence agreement and lower edit rates indicate smaller transcript changes. Panels (a) and (b) reference the original target channel and same-source reconstruction, respectively; both pool D1–D4.

(a) Original-target reference: four-array summary

Source	ASR	J			A			P		
		Agree.	Word	Char.	Agree.	Word	Char.	Agree.	Word	Char.
Emp.	Small.en	93.14	1.62	0.87	93.60	1.49	0.80	93.48	1.53	0.78
Emp.	Turbo	95.11	1.09	0.47	95.11	1.17	0.46	95.87	0.99	0.40
Emp.	Parakeet	87.81	3.00	1.74	87.86	2.99	1.73	87.63	3.13	1.89
Search	Small.en	94.49	1.23	0.66	94.53	1.22	0.66	94.49	1.23	0.66
Search	Turbo	96.05	1.10	0.44	96.18	1.11	0.51	96.01	1.12	0.45
Search	Parakeet	88.45	2.80	1.65	88.48	2.78	1.63	88.22	2.86	1.70

(b) Sentence agreement against same-source reconstruction: four-array summary

Source	Arrays	ASR	J	A	P
Emp.	D1–D4	Small.en	96.04	96.57	97.50
Emp.	D1–D4	Turbo	96.10	96.28	97.79
Emp.	D1–D4	Parakeet	91.48	93.17	92.36
Search	D1–D4	Small.en	99.10	99.70	99.11
Search	D1–D4	Turbo	98.80	99.55	98.71
Search	D1–D4	Parakeet	93.77	94.81	94.04

In Table D3(a), each recognizer has similar sentence agreement under J, A, and P, with low word- and character-edit rates.

Sentence agreement increases further against same-source reconstruction (Table D3(b)). Both sides share the reference speech, so differences primarily arise from the selected RTF; high agreement therefore indicates limited RTF-induced transcript changes. Both template families cover D1–D4, using valid pairs under their respective references.

Table D4: Joint-condition ASR sentence agreement (%) by array. Each recognizer references the original target channel and uses its own valid pairs.

Source	Array	Small.en	Turbo	Parakeet
Emp.	D1	91.55	88.73	81.69
Emp.	D2	91.59	96.73	85.51
Emp.	D3	93.48	95.02	88.08
Emp.	D4	93.40	95.43	88.98
Search	D1	94.47	94.37	83.42
Search	D2	92.06	98.40	85.05
Search	D3	94.75	95.93	88.10
Search	D4	95.07	95.50	90.89

In Table D4, all three recognizers retain identical sentences for a high fraction of pairs across four arrays. Parakeet has lower agreement than both Whisper systems, indicating recognizer-dependent transcript changes.

D.3 Speaker-Representation Preservation

Transcript agreement mainly reflects content preservation. To examine identity-related representations, we further evaluate RTF control using three speaker encoders.

ECAPA-TDNN (Desplanques et al., 2020), WeSpeaker ResNet34 (Wang et al., 2023), and X-vector TDNN (Snyder et al., 2018) output 192-, 256-, and 512-dimensional embeddings, respectively. Each model takes a complete single-channel 16 kHz waveform; embeddings are L2-normalized before cosine comparison. Pretrained weights are from SpeechBrain ECAPA, WeSpeaker ResNet34, and SpeechBrain X-vector, respectively. Each encoder uses the same weights, frontend, and data type across all four arrays.

Metrics and references. Evaluation separately references original target channels and same-source reconstruction, reporting mean cosine similarity and the fifth percentile P05. Aggregates use all valid

pairs, with linear interpolation for quantiles. Valid pair counts for each encoder and condition equal the planned counts in Table D2.

Table D5: Speaker-embedding cosine similarity after RTF control. Both panels pool D1–D4 and report J/A/P means and P05 to four decimals. Each evaluation model and condition has 1719 empirical and 22561 searched pairs; sample counts appear in Table D2.

(a) Original-target reference

Source	Encoder	J		A		P	
		Mean	P05	Mean	P05	Mean	P05
Emp.	ECAPA-TDNN	0.9500	0.8889	0.9534	0.8935	0.9573	0.9023
Emp.	WeSpeaker	0.9513	0.8944	0.9551	0.8964	0.9570	0.8979
Emp.	X-vector	0.9938	0.9877	0.9942	0.9885	0.9955	0.9898
Search	ECAPA-TDNN	0.9608	0.9042	0.9611	0.9045	0.9608	0.9043
Search	WeSpeaker	0.9612	0.9030	0.9616	0.9033	0.9613	0.9030
Search	X-vector	0.9958	0.9903	0.9959	0.9904	0.9958	0.9903

(b) Same-source reconstruction reference

Source	Encoder	J		A		P	
		Mean	P05	Mean	P05	Mean	P05
Emp.	ECAPA-TDNN	0.9843	0.9591	0.9894	0.9726	0.9941	0.9792
Emp.	WeSpeaker	0.9838	0.9579	0.9898	0.9747	0.9919	0.9746
Emp.	X-vector	0.9973	0.9936	0.9979	0.9950	0.9994	0.9976
Search	ECAPA-TDNN	0.9988	0.9962	1.0000	1.0000	0.9988	0.9962
Search	WeSpeaker	0.9984	0.9945	1.0000	1.0000	0.9984	0.9945
Search	X-vector	0.9999	0.9998	1.0000	1.0000	0.9999	0.9998

Against original target channels, both template families have high mean embedding similarity and P05 across encoders and conditions. Empirical templates have the highest means and P05 under P, indicating smaller representation changes with phase differences alone. The three searched-template conditions are similar, retaining high similarity under joint control.

Means and P05 increase further against same-source reconstruction, consistent with the transcript comparison: once reference speech is fixed, additional RTF control has a limited effect on speaker representations. Values of 1.0000 are rounded to four decimals.

Similarity distributions. In Figure D1, searched J/A/P distributions are close, while empirical-template distributions shift with condition, consistent with the means and P05 in Table D5. Both template families concentrate at high similarities with tails toward lower values.

Both empirical and searched templates can change some detectors' decisions, while independent complete-utterance evaluation shows generally stable transcripts and speaker representations.

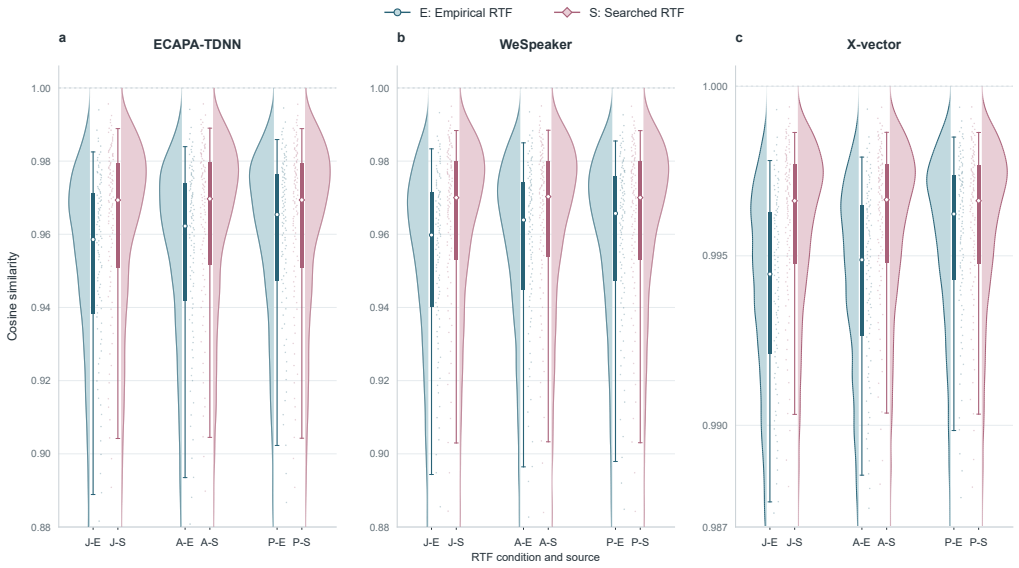


Figure D1: Speaker-embedding similarity distributions after RTF control, relative to original target channels and pooled over D1–D4. J/A/P denote joint, magnitude-ratio-only, and phase-difference-only control; E/S denote empirical and searched RTFs. Half-violins show density; boxes and whiskers show P25–P75 and P05–P95; white markers indicate medians. Points are 100 actual pairs per group, selected at equal intervals after sorting by similarity. Densities and quantiles use all valid pairs. The plots enlarge the high-similarity region, containing at least 96.27% of pairs in each group; encoders use different vertical ranges.

E Method Derivations and Theoretical Analysis

E.1 RTF Estimation and Reference Channels

Covariance subtraction (CS) estimates RTFs from second-order statistical differences between speech-present and noise-dominated frames. Following the main text, m and r denote target and reference channels, f frequency, and k frame index. S is the common source; V_m and V_r are channel noise and modeling residuals; H_m and H_r are effective transfer functions over the estimation interval. At a fixed frequency, assume S and each $V \in \{V_m, V_r\}$ are zero-mean and second-order uncorrelated. Then

$$\begin{aligned} \mathrm{Cov}(S, V) &:= \mathbb{E}[(S - \mathbb{E}[S])(V - \mathbb{E}[V])^*] \\ &= \mathbb{E}[SV^*] - \mathbb{E}[S]\mathbb{E}[V]^* = \mathbb{E}[SV^*] = 0. \end{aligned} \tag{E.1}$$

Here $\mathbb{E}$ denotes expectation, Cov complex covariance, and the asterisk complex conjugation. Conjugate symmetry gives $\mathbb{E}[VS^*] = 0$. Let $\mathcal{T}^{\mathrm{sp}}$ and $\mathcal{T}^{\mathrm{ns}}$ denote speech-present and noise-dominated frames. For a nonempty set $\mathcal{T}^{\ell}$ and any frame statistic $h(f, k)$, define

$$\langle h \rangle_{\ell}(f) := \frac{1}{|\mathcal{T}^{\ell}|} \sum_{k \in \mathcal{T}^{\ell}} h(f, k). \tag{E.2}$$

Here $\ell \in \{\mathrm{sp}, \mathrm{ns}\}$ identifies the frame class, and $|\mathcal{T}^{\ell}|$ is its frame count. Under the observation model in the main text, X_m and X_r are the target and reference STFTs. Cross-power and reference auto-power expand as

$$\begin{aligned}\langle X_m X_r^* \rangle_\ell(f) &= \langle (H_m S + V_m)(H_r S + V_r)^* \rangle_\ell(f) \\ &= H_m(f) H_r^*(f) \langle |S|^2 \rangle_\ell(f) + H_m(f) \langle S V_r^* \rangle_\ell(f) \\ &\quad + H_r^*(f) \langle V_m S^* \rangle_\ell(f) + \langle V_m V_r^* \rangle_\ell(f) \\ &\approx H_m(f) H_r^*(f) \langle |S|^2 \rangle_\ell(f) + \langle V_m V_r^* \rangle_\ell(f).\end{aligned} \tag{E.3}$$

$$\begin{aligned}\langle |X_r|^2 \rangle_\ell(f) &= \langle |H_r S + V_r|^2 \rangle_\ell(f) \\ &= |H_r(f)|^2 \langle |S|^2 \rangle_\ell(f) + H_r(f) \langle S V_r^* \rangle_\ell(f) \\ &\quad + H_r^*(f) \langle V_r S^* \rangle_\ell(f) + \langle |V_r|^2 \rangle_\ell(f) \\ &\approx |H_r(f)|^2 \langle |S|^2 \rangle_\ell(f) + \langle |V_r|^2 \rangle_\ell(f).\end{aligned} \tag{E.4}$$

Equations (E.3) and (E.4) neglect finite-frame source–noise cross-terms. Even under the population condition $\mathbb{E}[SV_m^*] = 0$, finite-frame means can retain residual cross-terms. When these are sufficiently small and noise second-order statistics are approximately stable between frame classes, their statistical differences satisfy

$$\Delta C_{mr} \approx H_m H_r^* \Delta P_s, \qquad \Delta P_r \approx |H_r|^2 \Delta P_s, \tag{E.5}$$

Here ΔC_{mr} and ΔP_r are the cross-power and reference auto-power differences between classes, and $\Delta P_s = \langle |S|^2 \rangle_{\mathrm{sp}} - \langle |S|^2 \rangle_{\mathrm{ns}}$ is the source-power difference; frequency arguments are omitted. Introducing $\varepsilon > 0$ gives the regularized ratio

$$\frac{\Delta C_{mr}}{\Delta P_r + \varepsilon} \approx \frac{H_m H_r^* \Delta P_s}{|H_r|^2 \Delta P_s + \varepsilon}. \tag{E.6}$$

When $\Delta P_s > 0$ and $|H_r|^2 \Delta P_s \gg \varepsilon$, this ratio approximates H_m / H_r, the target-to-reference RTF. Finite-frame error, residual cross-terms, propagation-path stability, and differences in noise statistics between frame classes jointly affect estimation accuracy.

Taking channel-response ratios cancels any shared nonzero complex factor $C(f)$: $(CH_m)/(CH_r) = H_m/H_r$. If the reference changes to r' with nonzero response, then $R_{m,r'} = R_{m,r}/R_{r',r}$, where the second subscript denotes the reference channel.

Multiple recordings of the same acoustic scene can jointly estimate a template to reduce per-recording statistical variability. Let c denote the scene, $\mathcal{U}_c$ its recordings, $h_u(f,k)$ a frame statistic for recording u, and $\mathcal{T}_u^\ell$ its valid frames of class ℓ. Pooling by valid-frame count gives

$$\langle h \rangle_{\ell,c}(f) := \frac{\displaystyle\sum_{u \in \mathcal{U}_c} \sum_{k \in \mathcal{T}_u^\ell} h_u(f,k)}{\displaystyle\sum_{u \in \mathcal{U}_c} |\mathcal{T}_u^\ell|}, \qquad \ell \in \{\mathrm{sp}, \mathrm{ns}\}. \tag{E.7}$$

Here $\ell = \mathrm{sp}$ and $\ell = \mathrm{ns}$ denote speech-present and noise-dominated frames, pooled separately. Choosing $h_u = X_{m,u} X_{r,u}^*$ or $h_u = |X_{r,u}|^2$ gives cross-power or reference auto-power. Pooled recordings must share environment, source position, acquisition devices, array geometry, channel order, and reference channel, with approximately stable propagation relationships and noise statistics. Power statistics pooled by Equation (E.7) are then subtracted and divided to obtain a scene RTF template.

This estimate weights each recording by its valid-frame count. Directly averaging complex RTFs generally changes the weights and yields a different result. Experiments use recording-level or scene-level empirical templates according to predefined rules.

E.2 Template Construction and Magnitude–Phase Control

To compare the roles of inter-channel magnitude ratios and phase differences, we decompose the same template $\widehat{R}_m(f)$ into the following conditions:

$$R_m^{\mathrm{A}} = |\widehat{R}_m|, \qquad R_m^{\mathrm{P}} = e^{\mathrm{j}\arg \widehat{R}_m}, \qquad R_m^{\mathrm{J}} = \widehat{R}_m. \tag{E.8}$$

Let $q \in \{\mathrm{A}, \mathrm{P}, \mathrm{J}\}$ identify the condition. Target spectra are constructed as $\widetilde{Z}_m = R_m^q X_r$, followed by conjugate completion for real signals and inverse STFT with fixed settings; the reference retains its original waveform. Preservation evaluation additionally uses same-source reconstruction: empirical templates reconstruct the reference with unit response $R_m = 1$, while searched templates directly copy the reference waveform. Searched A and P are derived directly from the selected J template to compare the components of that template.

Parameterized templates use relative source–microphone positions as a geometric baseline. Let $\mathbf{p}_m$ denote microphone m, $\mathbf{p}_{\mathrm{src}}$ the virtual source, and c_0 sound speed. Their distance is $\delta_m = \|\mathbf{p}_m - \mathbf{p}_{\mathrm{src}}\|_2 > 0$, where $\|\cdot\|_2$ is the Euclidean norm. Under a free-field point-source model, the nominal propagation response is

$$G_m(f; \mathbf{p}_{\mathrm{src}}) = \frac{\exp[-\mathrm{j}2\pi f \delta_m / c_0]}{4\pi\delta_m}, \tag{E.9}$$

The target-to-reference ratio G_m/G_r cancels the common factor 4π, yielding Equation (10) in the main text. Here δ_r is the source-to-reference distance, with magnitude ratio δ_r/δ_m and relative delay $(\delta_m - \delta_r)/c_0$. These form the geometric baseline for parameterized control; scaled candidates may depart from responses associated with an actual source position. A phase scale of 1 retains geometric phase, while condition A sets the selected template's phase differences to zero.

E.3 Analysis of Speech Semantic Preservation

Transcript and speaker-representation preservation depend on template-induced waveform changes and downstream-model sensitivity to them. A fixed RTF does not vary across frames within a recording, allowing its effects to be analyzed first on constructed spectra before inverse STFT. Let $X_r(f,k)$ be the reference STFT and $R_m(f) = A_m(f)e^{\mathrm{j}\phi_m(f)}$ the template, where A_m and ϕ_m are inter-channel magnitude ratios and phase differences. Then

$$\begin{aligned} \widetilde{Z}_m(f,k) &= R_m(f)X_r(f,k), \\ |\widetilde{Z}_m(f,k)| &= A_m(f)|X_r(f,k)|. \end{aligned} \tag{E.10}$$

Per-frequency temporal changes in log magnitude. When $A_m(f) > 0$ and $|X_r(f,k)|$ is nonzero for the compared frames, taking natural logarithms of Equation (E.10) and differencing across frames gives

$$\begin{aligned} \log|\widetilde{Z}_m(f,k)| &= \log A_m(f) + \log|X_r(f,k)|, \\ \Delta_k \log|\widetilde{Z}_m(f,k)| &= \Delta_k \log|X_r(f,k)|. \end{aligned} \tag{E.11}$$

Here $\Delta_k h(f,k) = h(f,k+1) - h(f,k)$ is the temporal difference. The fixed term $\log A_m(f)$ cancels between adjacent frames, preserving per-frequency log-magnitude temporal differences. A fixed positive gain over the recording introduces another log constant that also cancels.

Magnitude relationships between neighboring frequencies. Suppose $\log A_m(f)$ is Lipschitz within a band with constant L_A. At valid frequencies f_1, f_2,

$$\begin{aligned} &\left| \log \frac{|\widetilde{Z}_m(f_2,k)|}{|\widetilde{Z}_m(f_1,k)|} - \log \frac{|X_r(f_2,k)|}{|X_r(f_1,k)|} \right| \\ &= |\log A_m(f_2) - \log A_m(f_1)| \le L_A|f_2 - f_1|. \end{aligned} \tag{E.12}$$

Equation (E.12) bounds changes in log-magnitude ratios between neighboring frequencies by template smoothness. A smaller $L_A|f_2 - f_1|$ gives a smaller bound. L_A constrains the frequency variation of the log-magnitude response; the magnitude ratio itself may differ from unity. Deep notches, near-zero reference responses, and large estimation errors can violate this smoothness condition.

Phase control and temporal structure. P retains constructed-spectrum magnitudes, $|\widetilde{Z}_m^{\mathrm{P}}| = |X_r|$, but phase control can still change waveform timing. In an ideal continuous-frequency filtering model, differentiable unwrapped phase $\phi_m(f)$ gives group delay $\tau_{g,m}(f)$ as

$$\tau_{g,m}(f) = -\frac{1}{2\pi}\frac{\mathrm{d}\phi_m(f)}{\mathrm{d}f}. \tag{E.13}$$

Linear phase $\phi_m(f) = -2\pi f\tau$ represents a fixed delay τ, whereas nonlinear phase may delay different frequency components differently. Large inter-frequency delay differences can disperse transients, motivating analysis of both phase derivatives and delay ranges.

Multiplying a whole-signal Fourier spectrum by a fixed response can, under the corresponding transform and boundary conditions, be written as convolution $y = \kappa_R * x$. Here x and y are input and output waveforms, $\kappa_R = \mathcal{F}^{-1}\{R\}$ is the inverse Fourier transform of R, and the asterisk denotes convolution. Our framed STFT and inverse STFT also depend on the window, hop, overlap-add, and boundary settings.

Let $\mathcal{S}$ denote STFT analysis and $\mathcal{B}$ a fixed synthesis operator. Assume $\mathcal{B}\mathcal{S} = I$ on the signal space and boundary settings considered, where I is the identity. Let $\mathcal{M}_R$ multiply coefficients frequency-wise by R. The waveform-construction operator T_R before common-gain scaling satisfies

$$\begin{aligned} T_R &= \mathcal{B}\mathcal{M}_R\mathcal{S}, \\ \mathcal{S}(T_R x) &= \Pi\mathcal{M}_R\mathcal{S}x, \\ \Pi &:= \mathcal{S}\mathcal{B}, \quad \Pi^2 = \Pi. \end{aligned} \tag{E.14}$$

With fixed templates, windows, and boundaries, T_R is linear; input-dependent peak scaling can make the full procedure nonlinear. Π maps coefficients to the set of realizable waveform STFTs. Since $\Pi^2 = \Pi$, it is a projection, whose orthogonality depends on the synthesis operator. Modified coefficients may lie outside this set, producing differences between specified spectra and reconstructed-waveform STFTs (Griffin & Lim, 1984).

For specified spectrum $\widetilde{Z} = \mathcal{M}_R\mathcal{S}x$, the difference between the reconstructed-waveform STFT and the specified spectrum is $(\Pi - I)\widetilde{Z}$. Zero residual preserves the constructed magnitude–phase properties; nonzero residual adds changes from consistency projection. Exact magnitude preservation under P thus concerns constructed coefficients, while reconstructed spectra also depend on analysis and synthesis.

Response differences and waveform differences. Suppose R and baseline R_0 share the reference input, analysis and synthesis operators, and fixed positive gain g, omitting its recording index. Write $y = gT_R x$ and $y_0 = gT_{R_0}x$. If $\sup_f |R(f) - R_0(f)| \le \eta$, where η bounds the response difference, finite-dimensional Euclidean and corresponding coefficient norms satisfy

$$\begin{aligned} \|y - y_0\|_2 &= g\|\mathcal{B}(\mathcal{M}_R - \mathcal{M}_{R_0})\mathcal{S}x\|_2 \\ &\le g\|\mathcal{B}\|_{\mathrm{op}}\,\eta\,\|\mathcal{S}\|_{\mathrm{op}}\|x\|_2. \end{aligned} \tag{E.15}$$

Here $\|\cdot\|_{\mathrm{op}}$ is the induced operator norm. The diagonal multiplication operator has norm at most η; submultiplicativity gives Equation (E.15). This directly bounds synthesized-waveform differences and remains applicable when modified spectra have nonzero STFT-consistency residuals.

Baseline y_0 is constructed from the same reference input through R_0. Preservation experiments also reference the original target channel x_m. The triangle inequality gives

$$\|y - x_m\|_2 \le \|y - y_0\|_2 + \|y_0 - x_m\|_2. \tag{E.16}$$

Equation (E.16) bounds the total waveform difference by the sum of two terms: the constructed waveform versus baseline reconstruction, and baseline reconstruction versus the original target. Accordingly, Appendix D uses same-source reconstruction to evaluate RTF control and original target channels to evaluate the complete construction procedure.

Combining these waveform results with downstream-model local stability gives sufficient conditions for transcript and speaker-representation preservation.

Sufficient condition for transcript preservation. Consider deterministic maximum-score decoding. Let $h_w(x)$ score candidate transcript w. Input x_0 has a unique best transcript w_0, exceeding all alternatives by margin $\mu > 0$. Assume all candidate scores share a local Lipschitz constant L, so score changes are bounded by L times input distance. Then

$$\begin{aligned} &\mu = h_{w_0}(x_0) - \sup_{w \neq w_0} h_w(x_0) > 0, \\ &|h_w(y) - h_w(x_0)| \leq L\|y - x_0\|_2. \end{aligned} \tag{E.17}$$

Let $\epsilon_x = \|y - x_0\|_2$ be the distance between constructed waveform y and the original input. For every competing transcript $w \neq w_0$,

$$h_{w_0}(y) - h_w(y) \geq \mu - 2L\epsilon_x. \tag{E.18}$$

The best and competing transcript scores each change by at most $L\epsilon_x$; subtracting these bounds gives Equation (E.18). If $2L\epsilon_x < \mu$, w_0 remains uniquely optimal. This sufficient condition assumes the stated decoder model; actual ASR outputs also depend on steps such as search pruning.

Sufficient condition for speaker-representation preservation. Let $e(x)$ be a nonzero speaker embedding after L2 normalization to unit length, with local Lipschitz constant L_e. If $\|e(y) - e(x_0)\|_2 \leq L_e\epsilon_x$ in the neighborhood considered, then

$$\langle e(y), e(x_0)\rangle = 1 - \frac{1}{2}\|e(y) - e(x_0)\|_2^2 \geq 1 - \frac{1}{2}L_e^2\epsilon_x^2. \tag{E.19}$$

Here $\langle \cdot, \cdot \rangle$ is the vector inner product. Expanding the squared distance between two unit vectors gives the equality; substituting local stability gives the lower bound. Under these conditions, small input changes retain high embedding cosine similarity. The two sufficient conditions connect waveform changes to downstream-model local sensitivity, corresponding to transcript and representation evaluation in Appendix D.

E.4 Template Reuse and Cross-Model Responses

Different detectors' responses to the same RTF template can be analyzed through local changes in mean replay scores with template parameters.

Let d index detectors and s_d denote replay score, with higher values favoring replay. $T_{\boldsymbol{\alpha}}$ is the complete input construction for parameters $\boldsymbol{\alpha}$, including common amplitude scaling. For a fixed, nonempty recording set $\mathcal{D}$, define mean replay score $\mathcal{L}_d$ as

$$\mathcal{L}_d(\boldsymbol{\alpha}; \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} s_d(T_{\boldsymbol{\alpha}}(x)) \,. \tag{E.20}$$

Assume $\mathcal{L}_d$ is differentiable near baseline $\boldsymbol{\alpha}_0$ with an L_d-Lipschitz gradient. For a sufficiently small parameter change $\boldsymbol{\delta}$,

$$\begin{aligned} &\mathcal{L}_d(\boldsymbol{\alpha}_0 + \boldsymbol{\delta}; \mathcal{D}) - \mathcal{L}_d(\boldsymbol{\alpha}_0; \mathcal{D}) \\ &= \nabla\mathcal{L}_d(\boldsymbol{\alpha}_0; \mathcal{D})^\mathsf{T}\boldsymbol{\delta} + \rho_d, \\ &|\rho_d| \leq \frac{L_d}{2}\|\boldsymbol{\delta}\|_2^2. \end{aligned} \tag{E.21}$$

Here $\nabla \mathcal{L}_d$ is the parameter gradient, T denotes transpose, and ρ_d is the first-order remainder. Mean replay score decreases when the first-order term is negative and its magnitude exceeds the remainder bound.

Detectors differ in their first-order terms and original decision margins, so the same parameter change can produce different score and decision responses. Exchanging source and target models can also change transfer, consistent with the directional differences in Table C3. Conversion depends on whether each recording's score crosses the fixed threshold. Gradients and remainders are used only for this theoretical analysis; actual templates are selected using candidate detection scores.

F Algorithms and Implementation

F.1 Empirical-Template Estimation and Reconstruction

Algorithm 1 applies covariance subtraction to power statistics of genuine reference recordings to estimate RTFs and reconstruct target recordings. A nonempty reference group $\mathcal{U}_c$ contains recordings passing validity screening: one recording for a recording-level template, or multiple recordings from the same scene for a scene-level template. Grouping and assignment rule π follows Appendix B.2; Test/LOSO groups exclude the target speaker.

CS assumes a single dominant source and approximately stable propagation over the estimation interval. Overlapping speakers, changing paths, and speech remaining in noise frames can affect estimates; finite-frame cross-terms also produce residuals. When noise-subtracted reference power is low, estimates depend more on regularization. Valid bands, magnitude-ratio limits, and fallback rules are fixed before evaluation.

$\mathrm{FrameSelect}$ denotes frame selection in Appendix B.2. $\mathrm{Stabilize}$ enforces Hermitian symmetry and an eigenvalue floor on the covariance-difference matrix. $\mathcal{P}$ constrains magnitude ratios and sets out-of-band bins or bins with invalid reference power to unit response.

Synth_r denotes multichannel reconstruction in Equation (3), with fixed STFT, real-signal conjugate completion, and inverse STFT settings, retaining the original reference waveform. $\mathrm{CommonGain}$ determines a common positive gain from the peaks of the original, J/A/P, and corresponding controls. It applies to all output channels of a recording to preserve inter-channel magnitude ratios and phase differences.

Algorithm 1. Empirical RTF Estimation and Multichannel Reconstruction

Input: Genuine reference groups $\{\mathcal{U}_c\}_{c\in\mathcal{C}}$; target recordings $\{\mathbf{x}_i\}_{i\in\mathcal{D}}$; reference channel r; assignment rule π; estimation and reconstruction settings (Appendix B.2).

Output: Template bank $\mathcal{K}$ and reconstructed recordings $\mathcal{Y} = \{\mathbf{y}_i^q : i \in \mathcal{D},\ q \in \mathcal{Q}\}$, where $\mathcal{Q} = \{\mathrm{J}, \mathrm{A}, \mathrm{P}\}$.

$\mathcal{K} \leftarrow \varnothing, \quad \mathcal{Y} \leftarrow \varnothing$
for $c \in \mathcal{C}$ **do**
  for $u \in \mathcal{U}_c$ **do**
    $\mathbf{X}_u \leftarrow \mathrm{STFT}(u), \quad (\mathcal{T}_u^{\mathrm{sp}}, \mathcal{T}_u^{\mathrm{ns}}) \leftarrow \mathrm{FrameSelect}(X_{r,u})$
  end for
  $\mathbf{C}_c^{\ell} \leftarrow \dfrac{\sum_{u\in\mathcal{U}_c}\sum_{k\in\mathcal{T}_u^{\ell}} \mathbf{X}_u(f,k)\mathbf{X}_u(f,k)^{\mathrm{H}}}{\sum_{u\in\mathcal{U}_c}|\mathcal{T}_u^{\ell}|}, \quad \ell \in \{\mathrm{sp}, \mathrm{ns}\}$
  $\mathbf{D}_c \leftarrow \mathrm{Stabilize}(\mathbf{C}_c^{\mathrm{sp}} - \mathbf{C}_c^{\mathrm{ns}})$
  $\widehat{R}_{m,c} \leftarrow \mathcal{P}\left(\dfrac{D_{mr,c}}{D_{rr,c} + \varepsilon_c}\right)\ (m \neq r), \quad \widehat{R}_{r,c} \leftarrow 1$
  $\mathcal{K}[c] \leftarrow \widehat{R}_c$
end for
for $i \in \mathcal{D}$ **do**
  $c_i \leftarrow \pi(i, \mathcal{K}), \quad R_i^{\mathrm{J}} \leftarrow \mathcal{K}[c_i]$
  $R_i^{\mathrm{A}} \leftarrow |R_i^{\mathrm{J}}|, \quad R_i^{\mathrm{P}} \leftarrow \exp(\mathrm{j}\arg R_i^{\mathrm{J}})$
  for $q \in \mathcal{Q}$ **do**
    $\widetilde{\mathbf{z}}_i^q \leftarrow \mathrm{Synth}_r(\mathbf{x}_i, R_i^q)$
  end for
  $g_i \leftarrow \mathrm{CommonGain}(\mathbf{x}_i, R_i^{\mathrm{J}})$
  $\mathcal{Y}[i,q] \leftarrow g_i\widetilde{\mathbf{z}}_i^q, \quad q \in \mathcal{Q}$
end for
return $\mathcal{K}, \mathcal{Y}$

Here $\mathbf{X}_u(f,k)$ is a column vector of channel STFTs, and superscript H denotes conjugate transpose. $\mathbf{C}_c^{\ell}$ is the power matrix pooled by valid-frame counts, $D_{mr,c}$ is entry (m,r) of $\mathbf{D}_c$, and ε_c is the reference-power-based regularizer. Subsequent steps omit frequency indices. Template estimation and assignment do not query detectors; conversion after construction is computed as in Appendix B.1.

F.2 Parameterized-Template Selection

Algorithm 2 selects the J template with the lowest mean Dev replay score from a fixed candidate set. Θ is generated as in Appendix C.1 and passes numerical-validity, relative-gain, and equivalent-delay checks. $\mathcal{D}_{\mathrm{sel}}$ is a fixed, nonempty Dev selection set containing only replays originally correctly classified by the source detector. $R(\boldsymbol{\alpha})$ follows Equation (12), with band weights from Appendix F.3.

The algorithm uses Synth_r and $\mathrm{CommonGain}$ from Algorithm 1. Define $\mathrm{Eval}_d(\mathbf{x}, R, g) = s_d(g\,\mathrm{Synth}_r(\mathbf{x}, R))$, where s_d includes native detector preprocessing. Construction or scoring failure returns $+\infty$. The common gain is determined from the complete candidate bank, corresponding A/P responses, and controls, and fixed before candidate scoring.

Algorithm 2. Black-Box Selection of Parameterized RTFs

Input: Ordered candidate set Θ; Dev selection set $\mathcal{D}_{\mathrm{sel}}$; source detector d; microphone and virtual source coordinates; reference channel r; construction settings (Appendix C.1).

Output: Selected parameters $\boldsymbol{\alpha}_d^\star$ and responses $\{R_d^{\star,q} : q \in \mathcal{Q}\}$.

$\mathcal{R}_\Theta \leftarrow \{R(\boldsymbol{\alpha}) : \boldsymbol{\alpha} \in \Theta\}$
for $i \in \mathcal{D}_{\mathrm{sel}}$ **do**
  $g_i \leftarrow \mathrm{CommonGain}(\mathbf{x}_i, \mathcal{R}_\Theta)$
end for
$\mathcal{L}^\star \leftarrow +\infty, \quad \boldsymbol{\alpha}_d^\star \leftarrow \perp$
for $\boldsymbol{\alpha} \in \Theta$ (in fixed order) **do**
  $\mathcal{L} \leftarrow 0$
  for $i \in \mathcal{D}_{\mathrm{sel}}$ **do**
    $\mathcal{L} \leftarrow \mathcal{L} + \mathrm{Eval}_d(\mathbf{x}_i, R(\boldsymbol{\alpha}), g_i)$
  end for
  $\mathcal{L} \leftarrow \mathcal{L}/|\mathcal{D}_{\mathrm{sel}}|$
  if $\mathcal{L} < \mathcal{L}^\star$ **then**
    $\mathcal{L}^\star \leftarrow \mathcal{L}, \quad \boldsymbol{\alpha}_d^\star \leftarrow \boldsymbol{\alpha}$
  end if
end for
if $\boldsymbol{\alpha}_d^\star = \perp$ **then return** FAILURE
$R_d^{\star,\mathrm{J}} \leftarrow R(\boldsymbol{\alpha}_d^\star)$
$R_d^{\star,\mathrm{A}} \leftarrow |R_d^{\star,\mathrm{J}}|, \quad R_d^{\star,\mathrm{P}} \leftarrow \exp(\mathrm{j}\arg R_d^{\star,\mathrm{J}})$
return $\boldsymbol{\alpha}_d^\star, \{R_d^{\star,q} : q \in \mathcal{Q}\}$

The optimum is updated only when the score strictly decreases, retaining the first candidate in a tie. After selection, template parameters and detection thresholds remain fixed on independent Test, and A/P are directly derived from the selected J template. Cross-recording and cross-model evaluation reuse this template, with per-recording common gains determined from the complete response bank.

F.3 Tukey Windows and Band Interpolation

Tukey windows smoothly connect control parameters across adjacent bands. For band v, let f_v^{L} and f_v^{H} be the nominal lower and upper boundaries and $\delta_f > 0$ the transition width. The raw window is

$$w_v(f) = \begin{cases} \dfrac{1}{2}\left[1 - \cos\left(\pi \dfrac{f - f_v^{\mathrm{L}} + \delta_f}{\delta_f}\right)\right], & f_v^{\mathrm{L}} - \delta_f \le f < f_v^{\mathrm{L}}, \\ 1, & f_v^{\mathrm{L}} \le f \le f_v^{\mathrm{H}}, \\ \dfrac{1}{2}\left[1 + \cos\left(\pi \dfrac{f - f_v^{\mathrm{H}}}{\delta_f}\right)\right], & f_v^{\mathrm{H}} < f \le f_v^{\mathrm{H}} + \delta_f, \\ 0, & \text{otherwise.} \end{cases} \tag{F.1}$$

Adjacent bands satisfy $f_v^{\mathrm{H}} = f_{v+1}^{\mathrm{L}}$, with band widths $f_v^{\mathrm{H}} - f_v^{\mathrm{L}} \ge 2\delta_f$, so at most two adjacent windows are nonzero at any frequency.

Let $\alpha_{A,v}$ and $\alpha_{\phi,v}$ be magnitude and phase parameters of band v, with control functions $\alpha_A(f)$ and $\alpha_\phi(f)$. When only window v is nonzero, its raw weight is retained: $\alpha_A(f) = w_v(f)\alpha_{A,v}$ and $\alpha_\phi(f) = w_v(f)\alpha_{\phi,v}$. When two adjacent windows are nonzero, weights are normalized as

$$\widetilde{w}_v(f) = \frac{w_v(f)}{w_v(f) + w_{v+1}(f)}, \qquad \widetilde{w}_{v+1}(f) = \frac{w_{v+1}(f)}{w_v(f) + w_{v+1}(f)} \tag{F.2}$$

Magnitude and phase scales are weighted sums of the corresponding parameters. Both scales are zero when all windows are zero. The outermost single-window regions retain raw weights so that scales taper smoothly to zero. The main-text $\omega_v(f)$ denotes the resulting final weight. Numerical settings and model interfaces appear in Appendix B; candidate parameters and selection appear in Appendix C.1.