INTERBIAS-SV: COMPOUND CONDITIONS IN SPEAKER VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Speaker verification systems encounter combinations of noise, channel distortion, and changes in speech. Evaluating each condition separately does not establish whether their effects add. InterBias-SV organises this question around a four-term comparison: joint error, two marginal errors, and a common reference. Its results artefact contains 4,068 scored records across 17 experiments, 12 encoder labels, and six speech corpora, totalling 1.20×10^7 trial evaluations. Three experiment families contain the same-corpus termsneeded to compute additive contrasts. For labels assigned to speaker-trained encoders, their mean contrasts are $+0.0026$, $+0.0088$, and $+0.0024$ in equal error rate (EER), with larger variation across settings. These descriptive averages do not establish equivalence to additivity: trial matching, checkpoint identity, and parts of the condition metadata remain unverified. We also examine two interpretation problems. Near-chance EER can make additive predictions difficult to interpret, but chance performance is not a hard EER ceiling, and correlation with the prediction does not identify a saturation mechanism. Ratios of demographic gaps are unstable when their clean reference is near zero; absolute gaps provide a more direct summary. The benchmark provides condition definitions, analysis scripts, and explicit requirements for interpretable compound-condition comparisons, while separating recomputable summaries from claims that require further experimental validation.

1 INTRODUCTION

A speaker verification (SV) system may receive noisy, compressed speech from a caller whose speaking style differs from enrolment. A benchmark can measure the error associated with each condition, but those measurements alone do not show how the conditions combine. The practical question is whether a joint condition produces the error predicted by adding the two single-condition changes to a common reference. Answering it requires measuring the joint condition and its reference terms under comparable sampling and scoring procedures.

Existing resources establish useful but different comparisons. VoxCeleb and the VoxCeleb Speaker Recognition Challenges evaluate recognition in unconstrained recordings (Nagrani et al., 2017; Chung et al., 2018; Huh et al., 2024). SVeritas evaluates a broad set of stressors, including compounded noise, reverberation, and codec conditions (Baali et al., 2025). Such joint-condition results describe difficulty; attributing a departure from additivity additionally requires the corresponding marginals and an explicit contrast. Our focus is this comparison, rather than a claim that prior benchmarks evaluate only isolated stressors.

InterBias-SV combines condition definitions, an evaluation harness, and an analysis of an existing results artefact. Its proposed evaluation unit is a four-term comparison for one encoder and a specified factor setting. We distinguish acoustic manipulations from population strata: an age-group contrast measures how codec effects differ between populations, not the effect of changing a speaker's age. We also distinguish the presence of four numerical terms from evidence that they form a controlled experiment. That distinction matters here: the artefact permits three families of same-corpus contrasts, but the supplied code and metadata do not establish all the matching and provenance requirements.

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.

The contributions are:

1. A benchmark organisation for compound-condition SV evaluation, with 17 recorded experiments, 4,068 scored records, and scripts that reconstruct its tabulated summaries (Section 4).
2. Explicit requirements for interpreting additive contrasts: comparable populations and trial sampling, measured reference terms, and a clearly defined metric scale (Section 3). The accompanying audit identifies which requirements the current artefact does and does not establish.
3. Descriptive evidence on compound-condition errors and demographic gaps (Section 5). Small positive mean contrasts coexist with large setting-level deviations. Near-chance performance and small reference gaps limit interpretations of additive and ratio summaries.

The paper contributes an evaluation framework and empirical analysis, rather than a new encoder or a newly collected speech dataset. Its usefulness depends on making both the measurements and their limits reproducible.

2 Related Work and Positioning

Robustness evaluation. VoxCeleb provides in-the-wild speech and verification protocols (Nagrani et al., 2017; Chung et al., 2018); VoxSRC extends evaluation across training and adaptation settings (Huh et al., 2024). Noise and reverberation augmentation are also established components of speaker-embedding pipelines (Ko et al., 2017; Snyder et al., 2018). SVeritas is the closest benchmark: it evaluates duration, speaking conditions, noise, reverberation, channels, codecs, age, spoofing, and adversarial perturbations, with demographic breakdowns (Baali et al., 2025). It includes compound conditions. InterBias-SV's specific emphasis is the contrast between a joint error and the additive prediction from its own reference terms. This is an application of a factorial contrast, not a new statistical definition of interaction.

Representations and demographic evaluation. SUPERB evaluates speech representations across downstream tasks (Yang et al., 2021). WavLM, HuBERT, and UniSpeech-SAT support speaker-related applications (Chen et al., 2022a; Hsu et al., 2021; Chen et al., 2022b), but a frozen mean-pooled representation is a different system from one with a trained speaker head. Our comparison retains that distinction. Work on demographic bias in speech recognition (Koenecke et al., 2020) and speaker recognition (Hutiri & Ding, 2022; Fenu et al., 2021; Peri et al., 2023) motivates subgroup reporting. Here we examine how reported gaps vary across conditions, without treating subgroup EERs as a fairness assessment at a shared deployment threshold.

Table 1: Evaluation emphasis of related resources. The comparison describes reported analyses, not an exhaustive claim about what their data could support.

Resource	Evaluation emphasis
VoxCeleb / VoxSRC	In-the-wild verification and challenge protocols
SUPERB	Downstream evaluation of speech representations
Fair Voice Biometrics	Demographic imbalance and group fairness
SVeritas	Broad stressor coverage, including compound conditions
InterBias-SV	Additive contrasts and their sampling and metric requirements

3 Task, Estimand and Validity Conditions

Task and scoring. The intended task is text-independent SV. A frozen encoder maps an utterance to an L_2-normalised embedding, and cosine similarity scores each genuine or impostor trial. The artefact reports equal error rate (EER), normalised minimum detection cost (minDCF; target prior 0.01, miss and false-accept costs both 1), and area under the receiver operating characteristic curve (AUC). The supplied EER routine selects the ROC point with the smallest absolute difference between false-positive and false-negative rates and averages those rates; it does not interpolate the

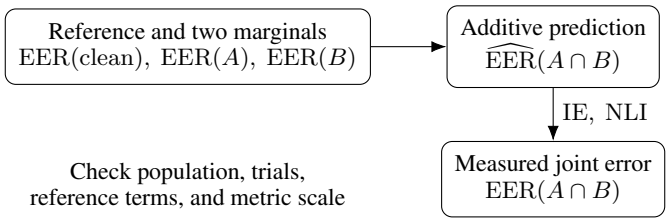

Figure 1: Four terms define an additive comparison. Their numerical availability permits recomputation; comparable sampling and verified condition provenance are needed to interpret the contrast.

crossing. All EERs here are fractions, so 0.01 corresponds to one percentage point. No score calibration, score normalisation, fine-tuning, or domain adaptation is specified. A common scoring rule removes one source of variation, but does not by itself control trial sampling or establish checkpoint identity.

The estimand. For conditions A and B and reference clean, define

$$\widehat{\mathrm{EER}}(A{\cap}B) = \mathrm{EER}(A){+}\mathrm{EER}(B){-}\mathrm{EER}(\mathrm{clean}), \qquad \mathrm{IE}(A,B) = \mathrm{EER}(A{\cap}B){-}\widehat{\mathrm{EER}}(A{\cap}B). \tag{1}$$

The non-linearity index is $\mathrm{NLI}(A,B) = \mathrm{EER}(A{\cap}B)/\widehat{\mathrm{EER}}(A{\cap}B)$ when the denominator exceeds 10^{-9}; zero or negative predictions retain their IE values but have undefined NLI. Small positive denominators can still make the ratio unstable. Positive IE indicates error above the additive prediction on the EER scale; negative IE indicates error below it. Neither sign alone identifies a causal mechanism. Additivity is scale-dependent and does not imply statistical independence of the factors. We report means and standard deviations (SDs) over encoder–setting combinations; these SDs are not confidence intervals.

Requirements for interpretation. **C1 (comparable source population):** all terms should refer to the same corpus and a documented target population. A shared corpus alone does not control speaker, session, or content composition. **C2 (comparable trials):** retain trial identities across acoustic transformations, and match speaker and content composition when a factor requires different recordings. Report sampling, class counts, exclusions, and uncertainty. Independent trial draws can estimate population contrasts under an appropriate design, but their variation must be accounted for. **C3 (measured terms):** both marginals and the reference must be measured; substituting the reference for a missing marginal changes the estimand. **C4 (metric domain):** check the attainable metric range and the stability of ratio denominators. These are requirements for the intended analysis, not conditions certified by the presence of result rows.

Three experiment families have the same-corpus terms needed for numerical recomputation: E13 (age group $\times$ codec), E14 (emotion $\times$ noise), and E17 (recorded drift label $\times$ codec). The updated E13 and E14 exports identify a common protocol cohort within each experiment and include checkpoint hashes. They are labelled paired-age and matched-RAVDESS runs, respectively; the aggregate CSV itself contains no trial identities, so C2 cannot be independently checked from this release. E13 compares an age stratum with the overlapping all-age population, rather than measuring within-speaker ageing. E14 uses all seven non-neutral RAVDESS emotions crossed with three noise labels and three SNRs. E17 still requires recording-level dates that the bundled legacy loader does not provide (Section 7).

E15 and E18 are reported as within-corpus clean/codec summaries, not isolated pathology or accent interactions: their reference populations belong to different corpora. E16 (gender $\times$ noise) is analysed through subgroup gaps. E20's recorded condition values are retained descriptively. The run audit found measured reference terms, but temporal and age proxies limit their interpretation. The analysis therefore distinguishes recomputable four-term contrasts from verified controlled experiments.

Chance performance and mathematical coupling. With fixed score orientation, EER lies in $[0, 1]$; 0.5 is the chance reference, not a hard upper bound. The supplied routine neither reverses

nor clips scores, and the artefact contains EERs above 0.5. An additive prediction above chance therefore does not force a negative interaction. More generally, if an error metric is bounded above by U, then $\mathrm{IE} \leq U - \widehat{\mathrm{EER}}$; a prediction above U necessarily gives a negative contrast. For this EER implementation, $U = 1$.

Correlation between IE and $\widehat{\mathrm{EER}}$ is also not an independent test of saturation: the latter is subtracted in the former. Writing $Y = \mathrm{EER}(A \cap B)$ and $X = \widehat{\mathrm{EER}}(A \cap B)$ gives $\mathrm{Cov}(Y - X, X) = \mathrm{Cov}(Y, X) - \mathrm{Var}(X)$. A negative correlation can consequently arise without a ceiling mechanism. We report the additive range and this correlation as descriptive checks, without using them to certify or exclude an encoder family.

4 Benchmark Construction

Corpora and encoder configurations. The records name RAVDESS (Livingstone & Russo, 2018), TORGO (Rudzicz et al., 2012), EARS (Richter et al., 2024), Common Voice (Ardila et al., 2020), L2-ARCTIC (Zhao et al., 2018), and VoxCeleb1 (Nagrani et al., 2017). RIRS_NOISES is the configured external-noise source (Ko et al., 2017). The audio front end resamples to 16 kHz, rejects clips shorter than 1 s, and truncates those longer than 10 s. The artefact includes speaker counts for some records, but does not provide complete corpus-release, audio-duration, or sampling manifests. Its trial total sums scoring operations across conditions and models, not distinct utterances or independent speaker pairs.

Twelve model labels appear. Seven are assigned to speaker encoders: X-Vector (Snyder et al., 2018), ECAPA-TDNN and ECAPA-TDNN-Large (Desplanques et al., 2020), RawNet3 (Jung et al., 2022), CAM++ (Wang et al., 2023b), TitaNet (Koluguri et al., 2022), and ReDimNet (Yakovlev et al., 2024). The wrappers use SpeechBrain (Ravanelli et al., 2021), WeSpeaker (Wang et al., 2023a), or model-specific loaders, with unrecorded fallback paths. Thus these labels do not verify seven distinct resolved speaker checkpoints.

Five self-supervised learning (SSL) configurations use WavLM-Base, WavLM-Base+, WavLM-Large (Chen et al., 2022a), UniSpeech-SAT (Chen et al., 2022b), and HuBERT-Large (Hsu et al., 2021). Their intended utterance representation is the temporal mean of the final hidden layer, without a trained speaker head. We retain the recorded family split for summarisation; it is not a controlled comparison of training paradigms. Original model-label spellings are retained in tables to permit exact CSV lookup.

Operators and provenance. The harness specifies telephony encode–decode operations with `ffmpeg`, noise at 5, 15, and 25 dB signal-to-noise ratio (SNR), 20 ms packet-loss masks, and time-scale perturbations. The run audit identifies E04 as recorded EARS styles and E05 as SBC, AAC, aptX and aptX-HD encode–decode conditions. The DSP/AAC proxy paths in the bundled legacy harness do not describe those retained runs. Codec round trips do not model a wireless radio channel. Other transformations also need verification. The environmental and crosstalk operators draw from the same noise directory, and revert to Gaussian noise when files are unavailable. Codec failures can return unchanged audio. Synthesis and conversion wrappers name VoiceCraft (Peng et al., 2024), E2 TTS (Eskimez et al., 2024), kNN-VC (Baas et al., 2023), and FreeVC (Li et al., 2023), but also return original audio on failures. The updated export includes augmentation and decode-failure counts for some records, but these fields do not cover every experiment. Appendix A.1 separates configured operations from verified execution.

Trials, splits, and leakage. The general sampler requests balanced genuine and impostor trials with a cap of 10^4, using a module-level random generator initialised with seed 42. Repeated calls advance that generator; a fixed initial seed does not give the same trial list to every model or condition. Scoring failures can change class balance. The general genuine-pair sampler uses different recordings, but this constraint is not enforced by every specialised sampler. Audio augmentation randomness and cache state are also not fully captured in the result rows.

No training or validation split is defined for benchmark-specific learning; the intended protocol freezes pre-trained encoders. This does not establish that evaluation speakers were unseen in pre-training. Several checkpoints use VoxCeleb, and E07, E17, and E20 reference VoxCeleb1. Speaker-

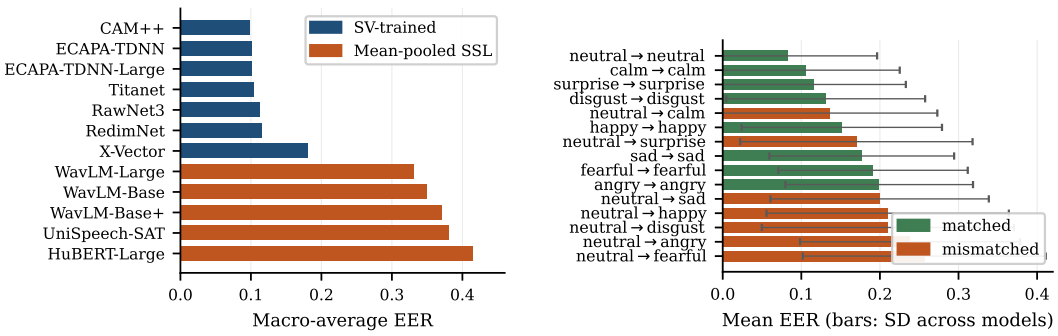

Figure 2: **Left:** macro EER by recorded model label; all 12 encoders cover the same 17 experiments. **Right:** E01 on RAVDESS, averaging available encoders for each condition. Bars show between-encoder SD, not uncertainty in a mean. Both panels use EER fractions. Family assignments and condition labels follow the artefact.

level overlap has not been audited. Future model selection on these public conditions would also require a separate held-out protocol to assess generalisation.

Coverage and availability. The submission CSV contains 4,068 scored records from 17 experiments and 12 model labels, totalling 11,975,026 trial evaluations. It includes ten single-factor, six two-factor, and one nominal three-factor experiment. E13 and E14 use only the replacement cohorts, each covering 12 models; the single-model legacy E13 cohort is excluded. E21 is removed because the updated cohort covers only one of 13 model labels, leaving the rest as legacy results. E09, E12, and E19 remain absent. E15 and E18 are restored from the historical export as descriptive results, with 240 and 120 scored rows respectively. Their corpus-level counts and aggregate metrics agree with the available run audit; their cross-corpus contrasts are excluded from the interaction analysis.

The retained roster contains the same 12 models in every experiment, with complete coverage of each experiment's observed condition grid. ERes2Net is excluded because it lacks four of the retained experiments. No missing measurements are imputed. Appendix A.2 reports coverage. E05, E10, and E11 contain operator labels in the dataset field rather than identifiable corpus names; these are shown as unrecorded.

The anonymised supplement contains the selected results, analysis scripts, and available legacy harness. It permits recomputation of aggregate summaries, but the harness is not established as the implementation of the updated runs. No corpus audio is included. Corpus acquisition and reuse remain subject to original distributor terms; an artefact licence and permanent repository have not yet been specified.

Intended use. The intended uses are condition-level comparisons, additive contrasts with verified reference terms, and stress-conditioned subgroup analysis. A low interaction effect is not itself a robustness score: a system can have low accuracy under every condition and still have a small contrast. Repeated tuning on public cells can also overfit the benchmark. Absolute error, balanced coverage, uncertainty, and held-out conditions are needed alongside interaction summaries.

5 Results

Encoder summaries. Macro-averaging first averages condition values within experiments and then weights experiments equally. CAM++ has macro EER 0.098, ECAPA-TDNN 0.101, and ECAPA-TDNN-Large 0.102; nominal speaker-trained labels span 0.098–0.181. Mean-pooled SSL labels span 0.331–0.415 (Figure 2; Appendix B). All 12 labels cover the same 17 experiments and each observed condition grid is complete. The macro ordering remains descriptive: common coverage alone does not establish matched trials or comparable sampling across encoders. Checkpoint hashes are present for a subset of the results; sampling and provenance still limit comparisons.

Table 2: Descriptive additive contrasts from available same-corpus terms. Means and SDs are across model–setting combinations; n counts those combinations, not independent replicates. EER uses fractional units; "IE > 0" is a percentage. SV denotes nominal speaker-trained labels and SSL denotes frozen mean-pooled configurations. E13 compares an age stratum with an overlapping all-age reference; E17's temporal provenance is unresolved. For E14 SV, NLI averages 198 positive-denominator cells; $n = 441$ includes all contrasts, including 243 with nonpositive predictions.

	SV labels					**SSL labels**				
Experiment	IE	SD	NLI	IE > 0	n	IE	SD	NLI	IE > 0	n
E13 age × codec	+0.0026	0.0204	1.089	52	84	-0.0063	0.0209	0.973	35	60
E14 emotion × noise	+0.0088	0.0301	1.549	34	441	-0.0465	0.0540	0.925	22	315
E17 drift × codec	+0.0024	0.0094	1.137	62	84	+0.0049	0.0198	1.014	57	60

Table 3: Additive-prediction summaries. $\widehat{\overline{\mathrm{EER}}}$ and $\max \widehat{\mathrm{EER}}$ are the mean and maximum prediction; r is Pearson correlation between the prediction and IE. These are descriptive diagnostics, not a test that distinguishes saturation from mathematical coupling.

	SV labels			**SSL labels**		
Exp.	$\widehat{\overline{\mathrm{EER}}}$	$\max \widehat{\mathrm{EER}}$	r	$\widehat{\overline{\mathrm{EER}}}$	$\max \widehat{\mathrm{EER}}$	r
E13	0.053	0.125	+0.070	0.254	0.399	+0.012
E14	0.039	0.351	+0.504	0.438	0.670	-0.706
E17	0.048	0.208	-0.220	0.392	0.467	-0.293

Emotion and enrolment–test mismatch. On RAVDESS, the neutral/neutral mean is 0.083 EER and the neutral→fearful mean is 0.256, a ratio of $3.11\times$. Matched fearful speech has mean EER 0.191 ($2.32\times$ the reference). Each neutral-to-emotion mismatch has a higher mean than the corresponding same-emotion condition (Appendix C). This pattern is consistent with a contribution from emotional mismatch, but does not isolate a causal mechanism or rank emotion against stressors on other corpora. All 15 conditions cover the same 12 model labels; matched and mismatched trials have different budgets. Between-encoder SDs are 0.11–0.16, not confidence intervals for the condition differences. The updated E01 export contains RAVDESS only.

Same-corpus additive contrasts. For E13, E14, and E17 we reconstruct Equation (1) from the CSV, retaining settings with all four finite terms for the same model label. This verifies the arithmetic, not the sampling or condition provenance (Section 3). Table 2 retains all such settings without a post-hoc family-level validity filter.

The nominal speaker-trained group has mean contrasts $+0.0026$, $+0.0088$, and $+0.0024$ for E13, E14, and E17; 52%, 34%, and 62% of combinations have positive signs. Mean NLIs are 1.089, 1.549, and 1.137. These average per-cell ratios over positive predictions only; in E14, 243 of the 441 SV predictions are nonpositive and have no NLI, while their contrasts remain in the IE summary. Small denominators make NLI particularly unstable in this experiment. Signed mean departures are small, but SDs of 0.009–0.030 exceed them. Individual E14 SV contrasts range from -0.059 to $+0.139$. Cancellation and zero contrasts can therefore obscure substantial setting-level deviations. The table does not establish equivalence to additivity or justify replacing joint-condition tests with marginal measurements. Appendices E.1 and E.2 report this variation.

Near-chance predictions. For mean-pooled SSL labels in E14, the mean contrast is -0.0465, and 22% of combinations have positive contrasts. The additive prediction averages 0.438 and reaches 0.670; its correlation with IE is $r = -0.706$ (Table 3, Figure 3). For nominal speaker-trained labels, the mean prediction is 0.039 and $r = +0.504$. These differences describe the supplied configurations, without identifying a saturation mechanism: EER can exceed chance and the correlation shares a term with the contrast. E13's SSL predictions remain below 0.5, so a chance-ceiling criterion would not exclude that family either.

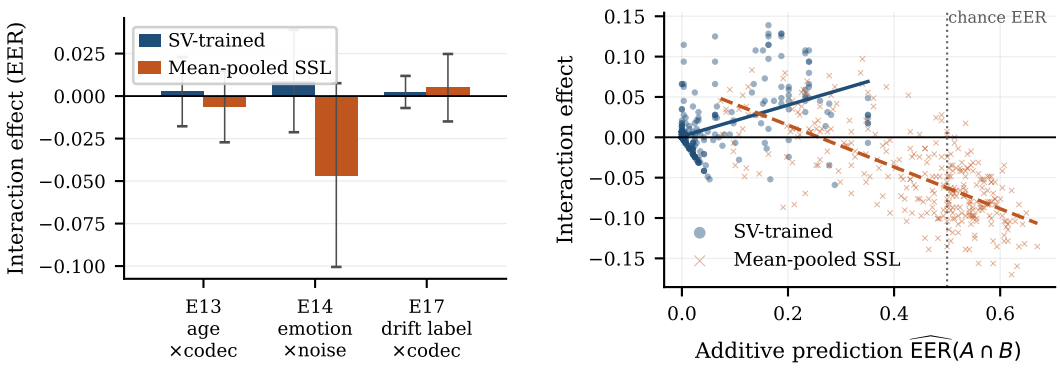

Figure 3: **Left:** mean additive contrast by experiment and recorded family; bars show SD across model–setting combinations. **Right:** E14 contrasts versus additive predictions, with descriptive least-squares fits. The dotted line marks chance EER, not an upper bound. The shared prediction term contributes to the negative correlation.

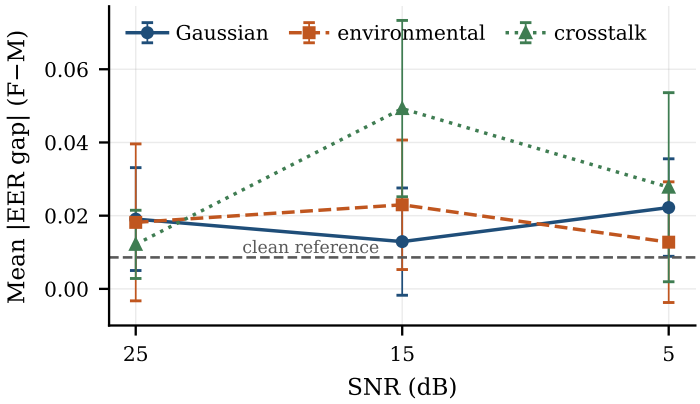

Figure 4: E16 mean absolute gender EER gap by labelled SNR. Bars show SD across 12 model labels. Noise names follow the CSV; source-class provenance is unverified. The dashed line is the clean mean. Independently sampled trial lists and unequal clean/noisy budgets limit causal interpretation.

Across the five SSL labels in E14, mean predictions range from 0.397 to 0.457, observed means from 0.350 to 0.414, and mean contrasts from -0.073 to -0.028. This within-family variation is descriptive, rather than a control that separates competing explanations for negative contrasts.

Demographic gaps under stress. E16 reports gender-stratified Common Voice conditions. We average the absolute per-model gap $|\mathrm{EER}_F - \mathrm{EER}_M|$ over 12 model labels. The mean gap is 0.0086 in the clean condition. At 25, 15, and 5 dB, Gaussian-labelled noise gives 0.0191, 0.0129, and 0.0222; environmental-labelled noise gives 0.0182, 0.0230, and 0.0128; and crosstalk gives 0.0122, 0.0493, and 0.0278. All nine noisy means exceed the clean mean, but none of the three sequences increases monotonically as SNR falls (Figure 4). This does not imply an effect for every encoder or identify one population as consistently worse served. Appendix D.1 lists the subgroup summaries.

The signed clean gap ranges in magnitude from 0.0004 to 0.0352 across models. Dividing noisy signed gaps by these small values produces ratios from -157.6 to $+193.7$ (Appendix D.2). Absolute gaps avoid that denominator, but still need sampling uncertainty: E16 uses 5,000 clean versus 555 noisy trials per subgroup, and absolute differences can increase with estimator variability. Three clean signed gaps are positive and nine negative, so there is no uniform direction across encoders.

6 DISCUSSION

An additive baseline makes the comparison of compound and single conditions explicit. The small signed means in the artefact motivate testing its predictive accuracy, but do not establish it. A useful validation would predefine an acceptable error, measure prediction errors on held-out factor settings, and retain their per-model distribution. A benchmark should report where the prediction fails, not only whether positive and negative deviations cancel.

Metric interpretation is equally consequential. EER evaluates a condition's own equal-error operating point; it does not directly measure risk at a fixed production threshold. Likewise, demographic EER gaps compare separately optimised operating points. Deployment-oriented analysis needs subgroup false-accept and false-reject rates at a common threshold selected on separate development data. These additions would test the practical meaning of the observed contrasts without treating benchmark averages as deployment guarantees.

7 LIMITATIONS AND THREATS TO VALIDITY

Result provenance. The selected CSV includes run IDs for 2,508 records, revision identifiers for 2,130, checkpoint hashes for 1,878, and genuine-speaker counts for 924. Coverage is incomplete; trial lists, per-trial scores, hardware, and full software environments are not supplied. The legacy harness can substitute checkpoints under other labels, whereas the updated E13/E14 cohorts carry explicit hashes. Aggregate agreement alone cannot authenticate an encoder (Appendix F).

Code–artefact discrepancies. The bundled legacy harness differs from producing revisions. The run audit found positive-gap execution, two-class E11 scoring, recorded E04 styles, real E05 codecs, and measured E20 reference terms. These findings supersede failures inferred solely from old code paths. Exact scored-pair manifests and full checkpoint binding are still unavailable; publication-year and age proxies, and the unrestricted nominal zero-gap reference, limit temporal interpretation (Appendix G).

Uncertainty and comparability. Each retained record has one aggregate measurement and no linked score list. Repeated speakers and shared reference terms make the contrasts dependent. Between-model SDs do not quantify sampling uncertainty. Trial budgets vary, and corpus membership alone does not establish matched population composition.

Scope. Per-trial noise, codec, synthesis, and conversion execution is not completely documented. E15/E18 mix source populations across corpora; E20 uses temporal and age proxies. Pre-training overlap, corpus versions, full subgroup sizes, and metadata provenance remain unaudited. These limits prevent broad claims about real-world compound robustness or demographic fairness.

8 CONCLUSION

InterBias-SV organises compound-condition SV evaluation around measured reference terms and an explicit additive contrast. The supplied artefact permits recomputable summaries for three same-corpus experiment families, with small positive signed means for nominal speaker-trained labels and substantial setting-level variation. Those averages remain descriptive until trial matching and experimental provenance are established. Chance EER is not a hard ceiling, and a near-zero demographic reference makes gap ratios unstable. The framework and audit identify what must be measured and recorded before compound-condition results can support stronger conclusions.

Ethics Statement

The benchmark uses existing speech corpora and introduces no new participant recruitment or annotation. The supplementary package contains aggregate results and code, without corpus audio. Public availability does not itself establish consent for every downstream use: corpus versions, licence terms, and permitted uses must be checked with the original distributors. No claim of institutional ethics exemption follows from the absence of new data collection.

Demographic categories are coarse, subgroup metadata are incomplete, and the retained gender and age comparisons do not establish fairness for specific populations. Speaker counts in some updated rows do not resolve label provenance or sampling uncertainty. The results should not be used to certify fairness or to justify deployment of biometric surveillance or access-control systems. Speech and linked demographic metadata can reveal identity even when speaker IDs are pseudonymous; any later release of manifests or audio-derived information needs a separate privacy and licence review. Spoofing and conversion experiments concern impersonation risks, but their execution must be verified before drawing vulnerability conclusions. No generated attack audio is included.

Reproducibility Statement

Section 3 defines scoring and contrasts; Section 4 and Appendix A.1 describe the available implementation. The supplementary CSV and two analysis scripts reconstruct aggregate tables and plots, with detailed listings in Appendices E and H. The protocol diagram is defined in the LaTeX source. This is summary-level reproducibility: Section 7 and Appendix G identify missing run provenance and code–artefact discrepancies that prevent full experimental reproduction.

AI Use Statement

Generative AI assisted implementation and analysis scripting, results-file formatting, interpretation of numerical summaries, literature retrieval, reference formatting, and manuscript drafting and editing. It also provided methodological criticism, suggested follow-up experiments, and helped revise mathematical statements about EER and interaction contrasts during the submission-readiness review. The supplied retained observations were not altered; this editorial review did not generate experimental observations or run new benchmark experiments. Checks in this revision covered aggregate recomputation, source-code consistency, selected primary references, and document compilation. These checks do not establish the provenance of the underlying experiments. The authors remain responsible for verifying the scientific content and for ensuring that the submission-form disclosure accurately reflects the full research workflow.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proc. Language Resources and Evaluation Conf. (LREC)*, pp. 4218–4222, 2020.

Massa Baali, Sarthak Bisht, Francisco Teixeira, Kateryna Shapovalenko, Rita Singh, and Bhiksha Raj. SVeritas: Benchmark for robust speaker verification under diverse conditions. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 9714–9731, 2025.

Matthew Baas, Benjamin van Niekerk, and Herman Kamper. Voice conversion with just nearest neighbors. In *Proc. Interspeech*, pp. 2053–2057, 2023.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Topics Signal Process.*, 16(6):1505–1518, 2022a.

Sanyuan Chen, Yu Wu, Chengyi Wang, Zhengyang Chen, Zhuo Chen, Shujie Liu, Jian Wu, Yao Qian, Furu Wei, Jinyu Li, and Xiangzhan Yu. UniSpeech-SAT: Universal speech representation learning with speaker aware pre-training. In *Proc. IEEE ICASSP*, pp. 6152–6156, 2022b.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. VoxCeleb2: Deep speaker recognition. In *Proc. Interspeech*, pp. 1086–1090, 2018.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Proc. Interspeech*, pp. 3830–3834, 2020.

Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, Yanqing Liu, Sheng Zhao, and Naoyuki Kanda. E2 TTS: Embarrassingly easy fully non-autoregressive zero-shot TTS. In *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2024.

Gianni Fenu, Mirko Marras, Giacomo Medda, and Giacomo Meloni. Fair voice biometrics: Impact of demographic imbalance on group fairness in speaker recognition. In *Proc. Interspeech*, pp. 1892–1896, 2021.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech, Language Process.*, 29: 3451–3460, 2021.

Jaesung Huh, Joon Son Chung, Arsha Nagrani, Andrew Brown, Jee-weon Jung, Daniel Garcia-Romero, and Andrew Zisserman. The VoxCeleb speaker recognition challenge: A retrospective. *IEEE/ACM Trans. Audio, Speech, Language Process.*, 32:3850–3866, 2024.

Wiebke Toussaint Hutiri and Aaron Yi Ding. Bias in automated speaker recognition. In *Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT)*, pp. 230–247, 2022.

Jee-weon Jung, Youjin Kim, Hee-Soo Heo, Bong-Jin Lee, Youngki Kwon, and Joon Son Chung. Pushing the limits of raw waveform speaker recognition. In *Proc. Interspeech*, pp. 2228–2232, 2022.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev Khudanpur. A study on data augmentation of reverberant speech for robust speech recognition. In *Proc. IEEE ICASSP*, pp. 5220–5224, 2017.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proc. Natl. Acad. Sci. U.S.A.*, 117(14):7684–7689, 2020.

Nithin Rao Koluguri, Taejin Park, and Boris Ginsburg. TitaNet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context. In *Proc. IEEE ICASSP*, pp. 8102–8106, 2022.

Jingyi Li, Weiping Tu, and Li Xiao. FreeVC: Towards high-quality text-free one-shot voice conversion. In *Proc. IEEE ICASSP*, 2023.

Steven R. Livingstone and Frank A. Russo. The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13(5):e0196391, 2018.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. VoxCeleb: A large-scale speaker identification dataset. In *Proc. Interspeech*, pp. 2616–2620, 2017.

Puyuan Peng, Po-Yao Huang, Abdelrahman Mohamed, and David Harwath. VoiceCraft: Zero-shot speech editing and text-to-speech in the wild. In *Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL)*, pp. 12442–12462, 2024.

Raghuveer Peri, Krishna Somandepalli, and Shrikanth Narayanan. A study of bias mitigation strategies for speaker recognition. *Computer Speech & Language*, 79:101481, 2023.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. SpeechBrain: A general-purpose speech toolkit. *arXiv:2106.04624*, 2021.

Julius Richter, Yi-Chiao Wu, Steven Krenn, Simon Welker, Bunlong Lay, Shinji Watanabe, Alexander Richard, and Timo Gerkmann. EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation. In *Proc. Interspeech*, pp. 4873–4877, 2024.

Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, 46 (4):523–541, 2012.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust DNN embeddings for speaker recognition. In *Proc. IEEE ICASSP*, pp. 5329–5333, 2018.

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 *Proc. IEEE ICASSP*, 2023a.

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. CAM++: A fast and efficient network for speaker verification using context-aware masking. In *Proc. Interspeech*, pp. 5301–5305, 2023b.

Ivan Yakovlev, Rostislav Makarov, Andrei Balykin, Pavel Malov, Anton Okhotnikov, and Nikita Torgashov. Reshape dimensions network for speaker recognition. In *Proc. Interspeech*, pp. 3235–3239, 2024.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. SUPERB: Speech processing universal performance benchmark. In *Proc. Interspeech*, pp. 1194–1198, 2021.

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. L2-ARCTIC: A non-native english speech corpus. In *Proc. Interspeech*, pp. 2783–2787, 2018.

A BENCHMARK DOCUMENTATION

A.1 MOTIVATION, COMPOSITION AND INTENDED USE

Purpose. InterBias-SV specifies additive comparisons between joint conditions and their measured reference terms. The results below reproduce the supplied aggregate artefact. Numerical availability is distinct from validated condition provenance and matched sampling (Section 3).

Instances. A condition combines a source population, transformation, and trial-construction rule. The selected file contains 4,068 records with EER, minDCF, AUC, and trial counts across 17 experiments, totalling 11,975,026 trial evaluations. This sum includes reuse across conditions and models. All included records have finite scores; absent model–condition cells are not imputed.

Degradation operators. Table 4 documents the available implementation, not proof that each operation succeeded in the runs that produced the CSV. Failure counts are supplied for some updated records, but execution provenance is incomplete.

Table 4: Legacy harness operators (not the producing revisions for all retained rows). Proxy operators are marked p. Execution and checkpoint provenance are not verified by the result rows.

Operator	Settings	Implementation
Telephony codec	G.711, GSM 06.10, AMR-NB, Opus	`ffmpeg` encode–decode round trip; G.711 μ-law, GSM and AMR-NB at 8 kHz, Opus at 16 kHz / 12 kbit/s; failures return original audio
Additive noise	Gaussian, environmental, crosstalk; 5/15/25 dB SNR	environmental and crosstalk labels share a RIRS_NOISES directory (Ko et al., 2017); missing files trigger Gaussian fallback
Packet loss	0.5%, 2%, 5%, 10%	20 ms frames zeroed independently; no packet-loss concealment or burst-loss model
Speaking rate	$0.75\times$, $1.0\times$, $1.25\times$	time-scale perturbation
Wireless codec^p	32–990 kbit/s	every condition encoded with AAC at a varying bitrate under labels nominally denoting SBC, aptX and LDAC; measures perceptual-codec bitrate sensitivity, not codec identity
Voice mode^p	whisper, shout	signal-level DSP proxy, not recorded phonation-mode speech; EARS contains genuine whispered and shouted recordings, which this operator does not use
Synthesis	VoiceCraft (Peng et al., 2024), E2-TTS (Eskimez et al., 2024)	named wrappers; failures can return original audio
Voice conversion	kNN-VC (Baas et al., 2023), FreeVC (Li et al., 2023)	named wrappers; failures can return original audio
Audio front end	16 kHz; 1–10 s	clips < 1 s discarded, clips > 10 s truncated
Trials	$\leq 10^4$ per condition	balanced sampling requested; module-level seed 42; condition-specific resampling and exclusions; specialised samplers need separate checks
Scoring	cosine on L_2-normalised embeddings	minDCF at $P_{\text{target}} = 0.01$, $C_{\text{miss}} = C_{\text{fa}} = 1$; no score normalisation, calibration or adaptation

Splits. There is no benchmark-specific training split or fine-tuning. Pre-trained encoders may nevertheless have seen evaluation speakers. Corpus versions, sampling manifests, and pre-training overlap need to be established before any condition can be described as held out.

Intended use. The artefact supports descriptive condition summaries and recomputation of same-corpus contrasts. Controlled interaction, demographic, spoofing, and temporal-drift claims require the provenance and experimental checks described in Appendix G.

A.2 FULL EXPERIMENT INVENTORY WITH AGGREGATE SCORES

Table 5 is the per-experiment inventory: corpus, encoder and condition counts, scored records, trial counts, and the descriptive use of each experiment under Section 3. Table 6 adds the aggregate EER, minDCF and AUC per experiment. Those aggregates pool heterogeneous conditions and encoders: they describe the difficulty of a condition set and are not a controlled ranking of stressors.

Table 5: Experiment inventory. "Scored" counts model–condition records with a computed EER. "Use" is what the experiment is admitted for: † cross-corpus populations: descriptive only; ‡ temporal/age proxy interpretation; § crosses a population attribute with a stressor, analysed as a subgroup gap; p legacy proxy implementation only. E05, E10, and E11 use operator labels rather than corpus identifiers in their dataset fields. All 12 retained encoders have complete coverage of each observed condition grid.

T	ID	Factor(s)	Corpus	Enc.	Cond.	Scored	Trials	Use
1	E01	Emotional speech	RAVDESS	12	15	180	330,000	marginal
1	E02	Dysarthric speech	TORGO	12	4	48	153,706	marginal
1	E03	Speaking rate	CommonVoice, EARS	12	18	216	2,159,910	marginal
1	E04	Recorded voice style	EARS	12	9	108	117,036	marginal
1	E05	Codec round trips	*not recorded*	12	15	180	1,799,550	marginal
1	E06	Packet loss	CommonVoice, EARS	12	18	216	2,159,901	marginal
1	E07	Temporal drift	VoxCeleb1	12	4	48	480,000	marginal
1	E08	L2 accent	CommonVoice, L2-ARCTIC	12	18	216	378,840	marginal
1	E10	TTS spoofing	*not recorded*	12	7	84	815,448	marginal
1	E11	Voice conversion	*not recorded*	12	3	36	180,000	marginal
2	E13	Age × codec	CommonVoice	12	20	240	1,199,940	contrast (audit)
2	E14	Emotion × noise	RAVDESS	12	80	960	276,480	contrast (audit)
2	E15	Pathology × codec†	EARS, TORGO	12	20	240	468,067	descriptive†
2	E16	Gender × noise	CommonVoice	12	30	360	419,868	subgroup gap§
2	E17	Drift × codec	VoxCeleb1	12	20	240	479,760	contrast (audit)
2	E18	Accent × codec†	CommonVoice, L2-ARCTIC	12	10	120	239,648	descriptive†
3	E20	Age × codec × drift‡	VoxCeleb1	12	48	576	316,872	descriptive‡

Historical descriptive experiments (E15 and E18). E15 contributes 240 scored rows (12 encoders, 20 conditions): TORGO dysarthric clean/codec results and EARS healthy-reference results. E18 contributes 120 rows (12 encoders, 10 conditions): L2-ARCTIC and Common Voice English clean/codec results. All condition grids are complete. Compare clean and codec conditions within each corpus; differences between corpora cannot isolate pathology or accent. E18's literal `native` condition labels are retained for lookup, but English locale does not establish native-speaker status. These historical rows enter descriptive summaries only, not the E13/E14/E17 interaction table.

Table 6: Experiment inventory with aggregate scores. EER is the unweighted mean over scored records with its standard deviation across those records. † cross-corpus populations, reported descriptively; ‡ temporal/age proxy interpretation.

Tier	ID	Factor(s)	Enc.	Cond.	Scored	Trials	EER	minDCF	AUC
1	E01	Emotional speech	12	15	180	330,000	0.172 ± 0.137	0.617	0.880
1	E02	Dysarthric speech	12	4	48	153,706	0.318 ± 0.109	0.964	0.737
1	E03	Speaking rate	12	18	216	2,159,910	0.351 ± 0.132	0.887	0.688
1	E04	Recorded voice style	12	9	108	117,036	0.182 ± 0.183	0.531	0.850
1	E05	Codec round trips	12	15	180	1,799,550	0.252 ± 0.161	0.686	0.791
1	E06	Packet loss	12	18	216	2,159,901	0.188 ± 0.162	0.582	0.853
1	E07	Temporal drift	12	4	48	480,000	0.175 ± 0.183	0.524	0.854
1	E08	L2 accent	12	18	216	378,840	0.136 ± 0.169	0.413	0.891
1	E10	TTS spoofing	12	7	84	815,448	0.407 ± 0.147	0.876	0.620
1	E11	Voice conversion	12	3	36	180,000	0.274 ± 0.128	0.737	0.774
2	E13	Age × codec	12	20	240	1,199,940	0.130 ± 0.110	0.505	0.913
2	E14	Emotion × noise	12	80	960	276,480	0.178 ± 0.195	0.439	0.846
2	E15	Pathology × codec†	12	20	240	468,067	0.311 ± 0.119	0.883	0.739
2	E16	Gender × noise	12	30	360	419,868	0.182 ± 0.156	0.540	0.863
2	E17	Drift × codec	12	20	240	479,760	0.188 ± 0.177	0.560	0.847
2	E18	Accent × codec†	12	10	120	239,648	0.145 ± 0.144	0.529	0.893
3	E20	Age × codec × drift‡	12	48	576	316,872	0.186 ± 0.176	0.524	0.850

B Encoder Results

Table 7 gives macro-average results: conditions are averaged within each experiment, then experiments are weighted equally. The last column is the number of experiments contributing: seventeen for every retained encoder. Condition coverage is complete, but trial matching and provenance limitations keep the ordering descriptive.

Table 7: Macro-average encoder results. "Exp." is the number of experiments contributing to that encoder's macro average.

Encoder	Family	EER	minDCF	AUC	Exp.
CAM++	SV	0.098	0.347	0.933	17
ECAPA-TDNN	SV	0.101	0.360	0.932	17
ECAPA-TDNN-Large	SV	0.102	0.357	0.931	17
Titanet	SV	0.104	0.341	0.928	17
RawNet3	SV	0.113	0.409	0.917	17
RedimNet	SV	0.115	0.381	0.919	17
X-Vector	SV	0.181	0.637	0.877	17
WavLM-Large	SSL	0.331	0.949	0.720	17
WavLM-Base	SSL	0.350	0.927	0.695	17
WavLM-Base+	SSL	0.371	0.961	0.673	17
UniSpeech-SAT	SSL	0.380	0.967	0.661	17
HuBERT-Large	SSL	0.415	0.986	0.617	17

C Single-Factor Results: E01 Emotional Speech

Table 8: E01, RAVDESS. Mean EER over available encoders, with the standard deviation across encoders; the last column is the ratio to the neutral/neutral reference. "Matched" conditions use the same emotion for enrolment and test. The SD is between-encoder variability, not a confidence interval. All conditions cover the same 12 model labels. Ratios divide the condition mean by the neutral-reference mean; they do not quantify sampling uncertainty.

Condition	Type	EER	SD	minDCF	AUC	Ratio
neutral/neutral	matched	0.083	0.114	0.424	0.951	1.00×
calm/calm	matched	0.106	0.119	0.378	0.936	1.28×
surprise/surprise	matched	0.116	0.117	0.501	0.927	1.40×
disgust/disgust	matched	0.131	0.127	0.513	0.918	1.59×
neutral→calm	mismatched	0.136	0.137	0.527	0.905	1.65×
happy/happy	matched	0.151	0.128	0.580	0.899	1.84×
neutral→surprise	mismatched	0.170	0.148	0.709	0.879	2.06×
sad/sad	matched	0.177	0.118	0.603	0.881	2.14×
fearful/fearful	matched	0.191	0.120	0.589	0.869	2.32×
angry/angry	matched	0.199	0.119	0.671	0.870	2.41×
neutral→sad	mismatched	0.200	0.139	0.704	0.856	2.42×
neutral→happy	mismatched	0.210	0.154	0.768	0.842	2.54×
neutral→disgust	mismatched	0.210	0.160	0.711	0.843	2.55×
neutral→angry	mismatched	0.238	0.140	0.777	0.819	2.89×
neutral→fearful	mismatched	0.256	0.155	0.798	0.799	3.11×

D Demographic Results

D.1 Absolute gender gap by condition

Table 9: E16 gender gap on Common Voice. F and M are the recorded subgroup labels; columns give mean EERs over 12 model labels; |gap| is the mean absolute per-encoder difference, and max|gap| its largest value over encoders.

Condition	EER (F)	EER (M)	|gap|	max|gap|	Enc.
clean (reference)	0.110	0.117	0.0086	0.0352	12
Gaussian, 25 dB	0.139	0.141	0.0191	0.0595	12
Gaussian, 15 dB	0.139	0.146	0.0129	0.0432	12
Gaussian, 5 dB	0.159	0.180	0.0222	0.0450	12
environmental, 25 dB	0.131	0.147	0.0182	0.0811	12
environmental, 15 dB	0.171	0.187	0.0230	0.0523	12
environmental, 5 dB	0.201	0.200	0.0128	0.0631	12
crosstalk, 25 dB	0.167	0.162	0.0122	0.0324	12
crosstalk, 15 dB	0.270	0.236	0.0493	0.0829	12
crosstalk, 5 dB	0.347	0.344	0.0278	0.0883	12

D.2 Instability of ratio-form bias amplification

Table 10: Instability of ratio-form bias amplification on E16. The clean gender gap is the denominator of the ratio; "mean |gap|" is the mean absolute gap over the nine noise conditions. The ratio range is reported to show its sensitivity to a near-zero denominator, not as a reliable amplification estimate.

Encoder	Clean gap	Mean |gap|	Ratio range
CAM++	-0.0022	0.0208	-37.7 to +8.2
ECAPA-TDNN	-0.0076	0.0158	-5.2 to +3.8
ECAPA-TDNN-Large	-0.0106	0.0194	-6.3 to +1.9
HuBERT-Large	+0.0036	0.0198	-17.5 to +3.0
RawNet3	-0.0048	0.0164	-6.4 to +6.8
RedimNet	+0.0012	0.0188	-25.5 to +43.5
Titanet	-0.0020	0.0146	-13.5 to +15.3
UniSpeech-SAT	-0.0216	0.0178	-1.5 to +1.4
WavLM-Base	-0.0126	0.0499	-7.0 to +6.4
WavLM-Base+	-0.0352	0.0252	-1.7 to +1.5
WavLM-Large	-0.0004	0.0218	-157.6 to +193.7
X-Vector	+0.0014	0.0226	-28.3 to +19.3

E Interaction Analysis in Detail

E.1 Per-encoder interaction effects

Table 11 decomposes the contrasts by model label. In E14, all five SSL means are negative, ranging from -0.073 to -0.028 EER. Six of seven speaker-trained means are positive; RawNet3 is negative (-0.003). The table includes all finite four-term contrasts, while NLI means use only positive additive predictions. No family-level chance screen is applied; sampling and provenance limitations remain.

Table 11: Interaction effect by encoder and experiment, over the settings for which all four factorial terms are available. n is the number of settings.

Exp.	Encoder	Fam.	$\overline{\text{IE}}$	SD	min	max	$\overline{\text{NLI}}$	n
E13	CAM++	SV	+0.0028	0.0103	-0.0127	+0.0248	1.226	12
E13	ECAPA-TDNN	SV	+0.0019	0.0141	-0.0154	+0.0218	1.020	12
E13	ECAPA-TDNN-Large	SV	-0.0050	0.0158	-0.0467	+0.0112	0.940	12
E13	HuBERT-Large	SSL	+0.0003	0.0334	-0.0668	+0.0520	1.007	12

Table 11 – continued

Exp.	Encoder	Fam.	$\overline{\text{IE}}$	SD	min	max	$\overline{\text{NLI}}$	n
E13	RawNet3	SV	-0.0046	0.0178	-0.0491	+0.0240	0.949	12
E13	RedimNet	SV	+0.0087	0.0116	-0.0119	+0.0297	1.356	12
E13	Titanet	SV	+0.0005	0.0129	-0.0151	+0.0263	1.016	12
E13	UniSpeech-SAT	SSL	-0.0003	0.0154	-0.0430	+0.0126	1.000	12
E13	WavLM-Base	SSL	-0.0136	0.0210	-0.0730	+0.0020	0.931	12
E13	WavLM-Base+	SSL	-0.0100	0.0131	-0.0480	+0.0033	0.962	12
E13	WavLM-Large	SSL	-0.0077	0.0149	-0.0453	+0.0060	0.965	12
E13	X-Vector	SV	+0.0140	0.0407	-0.0231	+0.1011	1.118	12
E14	CAM++	SV	+0.0084	0.0225	+0.0000	+0.1076	1.275	63
E14	ECAPA-TDNN	SV	+0.0134	0.0252	-0.0243	+0.1250	1.355	63
E14	ECAPA-TDNN-Large	SV	+0.0119	0.0358	-0.0069	+0.1389	1.284	63
E14	HuBERT-Large	SSL	-0.0364	0.0445	-0.1250	+0.0799	0.927	63
E14	RawNet3	SV	-0.0031	0.0331	-0.0521	+0.1076	0.564	63
E14	RedimNet	SV	+0.0161	0.0320	-0.0069	+0.1146	6.753	63
E14	Titanet	SV	+0.0138	0.0325	+0.0000	+0.1285	1.484	63
E14	UniSpeech-SAT	SSL	-0.0728	0.0531	-0.1701	+0.0347	0.860	63
E14	WavLM-Base	SSL	-0.0469	0.0551	-0.1562	+0.0833	0.954	63
E14	WavLM-Base+	SSL	-0.0279	0.0502	-0.1215	+0.0972	0.962	63
E14	WavLM-Large	SSL	-0.0484	0.0572	-0.1285	+0.0799	0.920	63
E14	X-Vector	SV	+0.0014	0.0230	-0.0590	+0.0451	0.911	63
E17	CAM++	SV	+0.0007	0.0047	-0.0071	+0.0096	1.034	12
E17	ECAPA-TDNN	SV	+0.0044	0.0059	-0.0034	+0.0133	1.204	12
E17	ECAPA-TDNN-Large	SV	+0.0071	0.0093	-0.0103	+0.0194	1.183	12
E17	HuBERT-Large	SSL	-0.0030	0.0224	-0.0484	+0.0284	0.995	12
E17	RawNet3	SV	+0.0063	0.0093	-0.0064	+0.0309	1.296	12
E17	RedimNet	SV	+0.0020	0.0106	-0.0129	+0.0244	1.022	12
E17	Titanet	SV	+0.0019	0.0048	-0.0055	+0.0087	1.265	12
E17	UniSpeech-SAT	SSL	+0.0186	0.0156	-0.0174	+0.0362	1.049	12
E17	WavLM-Base	SSL	+0.0074	0.0200	-0.0217	+0.0299	1.019	12
E17	WavLM-Base+	SSL	-0.0012	0.0201	-0.0252	+0.0365	0.997	12
E17	WavLM-Large	SSL	+0.0028	0.0155	-0.0276	+0.0286	1.009	12
E17	X-Vector	SV	-0.0055	0.0138	-0.0255	+0.0175	0.957	12

E.2 SETTING-LEVEL INTERACTION LISTING

Table 12 gives the pooled observed joint EER and additive prediction, followed by separate family-mean contrasts for every setting with four available terms. The dispersion visible here is what the pooled means of Section 5 average over.

Table 12: Per-setting interaction summary. "Obs." and "Add." pool all available encoder labels; the final two columns give family-specific mean contrasts.

Exp.	Setting	Obs. (all)	Add. (all)	IE (SV)	IE (SSL)
E13	`middle amr`	0.153	0.149	0.0025	0.0053
E13	`middle g711`	0.148	0.143	0.0016	0.0089
E13	`middle gsm`	0.158	0.153	0.0047	0.0054
E13	`middle opus`	0.134	0.137	-0.0019	-0.0055
E13	`older amr`	0.141	0.146	0.0117	-0.0265
E13	`older g711`	0.135	0.140	0.0016	-0.0132
E13	`older gsm`	0.143	0.149	0.0220	-0.0444
E13	`older opus`	0.132	0.134	-0.0048	0.0009
E13	`young amr`	0.123	0.126	-0.0063	-0.0004
E13	`young g711`	0.121	0.120	0.0003	0.0016
E13	`young gsm`	0.127	0.130	0.0006	-0.0065
E13	`young opus`	0.114	0.115	-0.0005	-0.0006
E14	`ravdess angry crosstalk 15dB`	0.267	0.269	0.0253	-0.0396

Table 12 – continued

Exp.	Setting	Obs. (all)	Add. (all)	IE (SV)	IE (SSL)
E14	ravdess angry crosstalk 25dB	0.214	0.229	0.0045	-0.0410
E14	ravdess angry crosstalk 5dB	0.380	0.367	0.0615	-0.0563
E14	ravdess angry environmental 15dB	0.182	0.202	0.0005	-0.0500
E14	ravdess angry environmental 25dB	0.160	0.194	-0.0015	-0.0792
E14	ravdess angry environmental 5dB	0.196	0.219	0.0015	-0.0590
E14	ravdess angry gaussian 15dB	0.112	0.102	-0.0050	0.0312
E14	ravdess angry gaussian 25dB	0.094	0.096	-0.0055	0.0028
E14	ravdess angry gaussian 5dB	0.136	0.114	0.0005	0.0528
E14	ravdess calm crosstalk 15dB	0.239	0.262	-0.0050	-0.0486
E14	ravdess calm crosstalk 25dB	0.194	0.222	-0.0010	-0.0674
E14	ravdess calm crosstalk 5dB	0.371	0.361	0.0437	-0.0375
E14	ravdess calm environmental 15dB	0.175	0.196	-0.0010	-0.0479
E14	ravdess calm environmental 25dB	0.156	0.187	-0.0045	-0.0687
E14	ravdess calm environmental 5dB	0.196	0.213	0.0020	-0.0424
E14	ravdess calm gaussian 15dB	0.097	0.096	-0.0040	0.0076
E14	ravdess calm gaussian 25dB	0.067	0.090	-0.0040	-0.0479
E14	ravdess calm gaussian 5dB	0.098	0.108	-0.0025	-0.0187
E14	ravdess disgust crosstalk 15dB	0.255	0.285	0.0263	-0.1090
E14	ravdess disgust crosstalk 25dB	0.209	0.245	0.0040	-0.0917
E14	ravdess disgust crosstalk 5dB	0.387	0.383	0.0685	-0.0868
E14	ravdess disgust environmental 15dB	0.178	0.218	0.0005	-0.0979
E14	ravdess disgust environmental 25dB	0.165	0.209	-0.0035	-0.1014
E14	ravdess disgust environmental 5dB	0.204	0.235	0.0040	-0.0812
E14	ravdess disgust gaussian 15dB	0.122	0.118	-0.0040	0.0153
E14	ravdess disgust gaussian 25dB	0.124	0.112	-0.0060	0.0361
E14	ravdess disgust gaussian 5dB	0.120	0.130	-0.0025	-0.0215
E14	ravdess fearful crosstalk 15dB	0.256	0.263	0.0322	-0.0632
E14	ravdess fearful crosstalk 25dB	0.213	0.223	0.0129	-0.0431
E14	ravdess fearful crosstalk 5dB	0.401	0.362	0.0908	-0.0340
E14	ravdess fearful environmental 15dB	0.172	0.197	0.0000	-0.0590
E14	ravdess fearful environmental 25dB	0.149	0.188	-0.0010	-0.0917

Table 12 – continued

Exp.	Setting	Obs. (all)	Add. (all)	IE (SV)	IE (SSL)
E14	ravdess fearful environmental 5dB	0.194	0.214	-0.0000	-0.0479
E14	ravdess fearful gaussian 15dB	0.108	0.097	-0.0015	0.0299
E14	ravdess fearful gaussian 25dB	0.095	0.091	-0.0015	0.0132
E14	ravdess fearful gaussian 5dB	0.119	0.109	0.0030	0.0215
E14	ravdess happy crosstalk 15dB	0.247	0.280	0.0079	-0.0889
E14	ravdess happy crosstalk 25dB	0.216	0.240	0.0055	-0.0653
E14	ravdess happy crosstalk 5dB	0.369	0.378	0.0382	-0.0757
E14	ravdess happy environmental 15dB	0.177	0.213	-0.0010	-0.0868
E14	ravdess happy environmental 25dB	0.159	0.205	-0.0020	-0.1069
E14	ravdess happy environmental 5dB	0.201	0.230	-0.0015	-0.0674
E14	ravdess happy gaussian 15dB	0.105	0.113	-0.0045	-0.0132
E14	ravdess happy gaussian 25dB	0.096	0.107	-0.0045	-0.0201
E14	ravdess happy gaussian 5dB	0.122	0.125	-0.0020	-0.0042
E14	ravdess sad crosstalk 15dB	0.254	0.281	0.0099	-0.0785
E14	ravdess sad crosstalk 25dB	0.203	0.241	0.0010	-0.0931
E14	ravdess sad crosstalk 5dB	0.381	0.380	0.0506	-0.0674
E14	ravdess sad environmental 15dB	0.175	0.214	-0.0055	-0.0861
E14	ravdess sad environmental 25dB	0.161	0.206	-0.0050	-0.1007
E14	ravdess sad environmental 5dB	0.194	0.231	-0.0045	-0.0847
E14	ravdess sad gaussian 15dB	0.115	0.114	-0.0079	0.0139
E14	ravdess sad gaussian 25dB	0.107	0.108	-0.0079	0.0083
E14	ravdess sad gaussian 5dB	0.114	0.126	-0.0050	-0.0215
E14	ravdess surprise crosstalk 15dB	0.276	0.284	0.0536	-0.0951
E14	ravdess surprise crosstalk 25dB	0.229	0.244	0.0228	-0.0674
E14	ravdess surprise crosstalk 5dB	0.397	0.383	0.0873	-0.0889
E14	ravdess surprise environmental 15dB	0.188	0.218	0.0005	-0.0708
E14	ravdess surprise environmental 25dB	0.172	0.209	-0.0020	-0.0868
E14	ravdess surprise environmental 5dB	0.205	0.235	0.0055	-0.0785
E14	ravdess surprise gaussian 15dB	0.120	0.117	-0.0020	0.0090
E14	ravdess surprise gaussian 25dB	0.115	0.111	-0.0050	0.0146
E14	ravdess surprise gaussian 5dB	0.132	0.129	0.0064	-0.0035

Table 12 – continued

Exp.	Setting	Obs. (all)	Add. (all)	IE (SV)	IE (SSL)
E17	1yr amr	0.202	0.193	-0.0007	0.0235
E17	1yr g711	0.196	0.187	-0.0030	0.0254
E17	1yr gsm	0.206	0.203	-0.0031	0.0120
E17	1yr opus	0.183	0.177	-0.0053	0.0220
E17	3yr amr	0.198	0.197	0.0098	-0.0105
E17	3yr g711	0.192	0.191	0.0052	-0.0069
E17	3yr gsm	0.209	0.207	0.0155	-0.0160
E17	3yr opus	0.178	0.181	0.0006	-0.0089
E17	5yr amr	0.199	0.193	0.0048	0.0075
E17	5yr g711	0.190	0.187	-0.0007	0.0075
E17	5yr gsm	0.209	0.203	0.0089	0.0021
E17	5yr opus	0.176	0.177	-0.0029	0.0013

F Encoder-Agreement Audit

Several model wrappers can replace a requested encoder with a SpeechBrain ECAPA-TDNN checkpoint, without recording the resolved identity in the CSV. Table 13 compares aggregate rows. No pair agrees exactly on more than 20.9% of the compared conditions, but this cannot exclude fallback: models can receive different sampled trials and augmentation realisations. Agreement rates include exactly zero errors in the updated matched conditions and cannot authenticate a checkpoint. The provided hashes offer additional provenance for the subset of records that contains them.

Table 13: The twelve most similar encoder pairs of 66, by exact agreement over co-observed conditions. "Cond." is the number of co-observed conditions, "Agree" the percentage on which the two EERs are identical, ρ the correlation of their EERs, and MAD the mean absolute EER difference.

Encoder A	**Encoder B**	**Cond.**	**Agree (%)**	ρ	**MAD**
CAM++	ECAPA-TDNN-Large	339	20.9	0.993	0.0075
CAM++	Titanet	339	20.6	0.984	0.0126
ECAPA-TDNN-Large	Titanet	339	19.5	0.982	0.0139
ECAPA-TDNN-Large	RedimNet	339	18.0	0.919	0.0184
ECAPA-TDNN	Titanet	339	16.5	0.846	0.0270
CAM++	RedimNet	339	16.2	0.921	0.0187
RedimNet	Titanet	339	16.2	0.926	0.0192
CAM++	RawNet3	339	15.0	0.969	0.0191
CAM++	ECAPA-TDNN	339	14.7	0.829	0.0296
RawNet3	Titanet	339	14.7	0.973	0.0171
ECAPA-TDNN	RedimNet	339	14.2	0.812	0.0326
ECAPA-TDNN	ECAPA-TDNN-Large	339	13.9	0.816	0.0309

G Extended Limitations

Uncertainty. A single run does not in principle prevent uncertainty estimation from retained scores. Here, only aggregate measurements are supplied. Recover per-trial scores and identities, then use a sampling scheme that respects repeated speakers and shared references across the four terms. Independent speaker/trial and augmentation seeds would separately test construction sensitivity. Treating all model–setting combinations as independent replicates would understate dependence. An equivalence claim additionally needs a predefined practical margin; a small signed mean is insufficient.

Checkpoint and run provenance. Preserve the resolved model identifier, checkpoint hash, software environment, and successful operator configuration for each run. Disable silent substitution in future measurements. The updated summary includes run identifiers, revision fields, and checkpoint

hashes for subsets of records, but no score manifests linking all observations to the bundled legacy implementation. The agreement table cannot fill this gap.

Temporal labels and producing revisions. The run audit found per-video years, positive-gap execution and measured E20 reference terms. The old bundled loader is not the producing revision. Publication dates and age estimates remain proxies; the nominal zero-gap reference is unrestricted, not a verified same-session or same-year baseline. Complete pair manifests are needed to verify sampling and overlap.

Noise and attack operators. The audit found two-class E11 scoring and stored synthesis/conversion outputs. This resolves contradictions inferred from the legacy one-class branch, but does not supply exact scored-pair lists or establish real-time latency. Noise-source corrections and failure counts cover some runs; complete operator and checkpoint binding remains unavailable.

Demographic and emotion labels. E13 compares an age stratum with an overlapping all-age reference, not two disjoint age populations. Common Voice age bins are mapped to representative integers rather than measured ages. Report speaker counts, missing labels, label sources, and mapping rules for all strata. The retained E01 and replacement E14 rows use RAVDESS. The paired and matched cohort labels do not replace trial manifests needed to verify population and content matching.

Contamination and coverage. Check pre-training/evaluation speaker overlap for each resolved checkpoint, especially on VoxCeleb1. All twelve retained labels cover all seventeen experiments and every observed condition. Comparable speaker composition and successful trials still need verification; complete coverage does not establish matched sampling.

Excluded experiments and cohorts. E21 is omitted: only one model has an updated cohort, with the other twelve represented by legacy rows. The single-model legacy E13 cohort is omitted in favour of the 12-model paired replacement. E09, E12, and E19 have no valid replacement results. ERes2Net is excluded because it lacks four retained experiments. E15 and E18 are restored only as historical descriptive results. E20 remains descriptive and is not used as a three-way interaction estimate.

Scale and release metadata. The trial total counts scoring operations, not independent data. The release lacks complete corpus versions, speaker-count coverage, audio hours, run hardware, exact library versions, and execution costs. Some records include checkpoint hashes and speaker counts. E05, E10, and E11 lack usable corpus identifiers. An artefact licence, third-party attribution, acquisition instructions, and stable protocol manifests are needed for reuse.

H Condition-Level Results for Every Experiment

The tables below list every scored condition in the selected 17 experiments, sorted by mean EER. "SD" is the standard deviation across contributing model rows, not a confidence interval. There are no unscored rows in the selected file. Every observed condition has all twelve retained encoders; the encoder count in each row makes this coverage explicit. CV denotes Common Voice, and L2-ARCTIC/CV denotes a condition involving both corpora.

The E13 and E14 identifiers correspond to the paired-age and matched-RAVDESS replacement exports. Their double-underscore condition strings are preserved for exact lookup. The run audit identifies recorded EARS styles for E04 and actual codec round trips for E05; the proxy paths in the bundled legacy harness do not describe those retained observations.

H.1 E01 – EMOTIONAL SPEECH

Table 14: E01 – Emotional speech: condition-level means over the contributing encoder rows. SD is over contributing encoder rows, not a confidence interval.

Condition	Corpus	Enc.	EER	SD	minDCF	AUC	Trials
ravdess/neutral_neutral	ravdess	12	0.083	0.114	0.424	0.951	120,000
ravdess/calm_calm	ravdess	12	0.106	0.119	0.378	0.936	15,000
ravdess/surprise_surprise	ravdess	12	0.116	0.117	0.501	0.927	15,000
ravdess/disgust_disgust	ravdess	12	0.131	0.127	0.513	0.918	15,000
ravdess/neutral_calm	ravdess	12	0.136	0.137	0.527	0.905	15,000
ravdess/happy_happy	ravdess	12	0.151	0.128	0.580	0.899	15,000
ravdess/neutral_surprise	ravdess	12	0.170	0.148	0.709	0.879	15,000
ravdess/sad_sad	ravdess	12	0.177	0.118	0.603	0.881	15,000
ravdess/fearful_fearful	ravdess	12	0.191	0.120	0.589	0.869	15,000
ravdess/angry_angry	ravdess	12	0.199	0.119	0.671	0.870	15,000
ravdess/neutral_sad	ravdess	12	0.200	0.139	0.704	0.856	15,000
ravdess/neutral_happy	ravdess	12	0.210	0.154	0.768	0.842	15,000
ravdess/neutral_disgust	ravdess	12	0.210	0.160	0.711	0.843	15,000
ravdess/neutral_angry	ravdess	12	0.238	0.140	0.777	0.819	15,000
ravdess/neutral_fearful	ravdess	12	0.256	0.155	0.798	0.799	15,000

H.2 E02 – DYSARTHRIC SPEECH

Table 15: E02 – Dysarthric speech: condition-level means over the contributing encoder rows. SD is over contributing encoder rows, not a confidence interval.

Condition	Corpus	Enc.	EER	SD	minDCF	AUC	Trials
TORGO/dysarthria_moderate_matched	torgo	12	0.302	0.117	0.941	0.755	37,404
TORGO/control_matched	torgo	12	0.318	0.120	0.962	0.735	39,689
TORGO/dysarthria_mild_matched	torgo	12	0.326	0.105	0.978	0.730	38,154
TORGO/dysarthria_severe_matched	torgo	12	0.327	0.106	0.977	0.726	38,459

H.3 E03 – SPEAKING RATE

Table 16: E03 – Speaking rate: condition-level means over the contributing encoder rows. SD is over contributing encoder rows, not a confidence interval.

Condition	Corpus	Enc.	EER	SD	minDCF	AUC	Trials
common_voice/clean	CV	12	0.108	0.118	0.454	0.933	120,000
common_voice/slow_slow	CV	12	0.167	0.077	0.767	0.899	120,000

Table 16 – continued

Condition	Corpus	Enc.	EER	SD	minDCF	AUC	Trials
common_voice/fast_fast	CV	12	0.192	0.070	0.798	0.881	120,000
ears/clean	ears	12	0.244	0.174	0.658	0.796	119,990
common_voice/fast_normal	CV	12	0.336	0.091	0.927	0.718	120,000
common_voice/normal_fast	CV	12	0.337	0.090	0.928	0.718	120,000
common_voice/normal_slow	CV	12	0.350	0.093	0.931	0.699	120,000
common_voice/slow_normal	CV	12	0.351	0.093	0.930	0.698	120,000
ears/slow_slow	ears	12	0.360	0.044	0.938	0.691	119,990
common_voice/fast_slow	CV	12	0.379	0.088	0.936	0.661	120,000
common_voice/slow_fast	CV	12	0.381	0.086	0.936	0.661	120,000
ears/fast_fast	ears	12	0.381	0.032	0.966	0.664	119,990
ears/fast_normal	ears	12	0.450	0.067	0.966	0.566	119,990
ears/normal_fast	ears	12	0.451	0.067	0.962	0.566	119,990
ears/slow_normal	ears	12	0.452	0.059	0.957	0.565	119,990
ears/normal_slow	ears	12	0.452	0.060	0.956	0.565	119,990
ears/slow_fast	ears	12	0.461	0.038	0.974	0.555	119,990
ears/fast_slow	ears	12	0.463	0.041	0.977	0.552	119,990

H.4 E04 – Recorded voice style

Table 17: E04 – Recorded voice style: condition-level means over the contributing encoder rows. SD is over contributing encoder rows, not a confidence interval.

Condition	Corpus	Enc.	EER	SD	minDCF	AUC	Trials
shout_shout	ears	12	0.148	0.183	0.428	0.870	13,332
modal_modal	ears	12	0.148	0.182	0.424	0.875	13,332
whisper_whisper	ears	12	0.157	0.165	0.477	0.880	13,332
modal_shout	ears	12	0.162	0.189	0.441	0.864	12,840
shout_modal	ears	12	0.171	0.201	0.449	0.853	12,840
whisper_modal	ears	12	0.209	0.194	0.626	0.827	12,840
modal_whisper	ears	12	0.209	0.193	0.657	0.831	12,840
whisper_shout	ears	12	0.212	0.191	0.632	0.826	12,840
shout_whisper	ears	12	0.217	0.194	0.645	0.824	12,840

H.5 E05 – Codec round trips

Table 18: E05 – Codec round trips: condition-level means over the contributing encoder rows. SD is over contributing encoder rows, not a confidence interval.

Condition	Corpus	Enc.	EER	SD	minDCF	AUC	Trials
aac_high_matched	–	12	0.246	0.172	0.669	0.797	119,970
clean	–	12	0.246	0.172	0.660	0.796	119,970
aac_medium_matched	–	12	0.246	0.171	0.671	0.797	119,970
aptx_high_matched	–	12	0.246	0.172	0.662	0.796	119,970
sbc_high_matched	–	12	0.246	0.172	0.669	0.796	119,970
aptx_vs_aac_high	–	12	0.246	0.171	0.671	0.796	119,970
aac_vs_aptx_medium	–	12	0.246	0.173	0.666	0.796	119,970
aptx_low_matched	–	12	0.248	0.170	0.661	0.795	119,970

Table 18 – continued

Condition	Corpus	Enc.	EER	SD	minDCF	AUC	Trials
sbc_vs_aac_medium	–	12	0.250	0.168	0.686	0.793	119,970
sbc_vs_aptx_medium	–	12	0.250	0.168	0.681	0.793	119,970
aptx_vs_sbc_high	–	12	0.251	0.168	0.684	0.793	119,970
aac_vs_sbc_medium	–	12	0.251	0.168	0.682	0.793	119,970
sbc_medium_matched	–	12	0.252	0.166	0.687	0.791	119,970
aac_low_matched	–	12	0.254	0.165	0.725	0.792	119,970
sbc_low_matched	–	12	0.308	0.121	0.819	0.743	119,970

H.6 E06 – Packet loss

Table 19: E06 – Packet loss: condition-level means over the contributing encoder rows. SD is over contributing encoder rows, not a confidence interval.

Condition	Corpus	Enc.	EER	SD	minDCF	AUC	Trials
common_voice/clean	CV	12	0.108	0.117	0.463	0.933	120,000
common_voice/clean_x_0.5pct	CV	12	0.109	0.118	0.462	0.931	120,000
common_voice/0.5pct_0.5pct	CV	12	0.110	0.118	0.463	0.930	120,000
common_voice/clean_x_2.0pct	CV	12	0.112	0.121	0.473	0.929	120,000
common_voice/2.0pct_2.0pct	CV	12	0.112	0.121	0.474	0.928	120,000
common_voice/5.0pct_5.0pct	CV	12	0.116	0.123	0.482	0.925	120,000
common_voice/clean_x_5.0pct	CV	12	0.116	0.124	0.478	0.924	120,000
common_voice/10.0pct_10.0pct	CV	12	0.120	0.125	0.495	0.922	120,000
common_voice/clean_x_10.0pct	CV	12	0.124	0.131	0.500	0.917	120,000
ears/0.5pct_0.5pct	ears	12	0.245	0.172	0.662	0.798	119,989
ears/2.0pct_2.0pct	ears	12	0.245	0.172	0.666	0.798	119,989
ears/5.0pct_5.0pct	ears	12	0.246	0.172	0.667	0.798	119,989
ears/10.0pct_10.0pct	ears	12	0.248	0.170	0.677	0.796	119,989
ears/clean	ears	12	0.257	0.165	0.667	0.787	119,989
ears/clean_x_0.5pct	ears	12	0.278	0.178	0.704	0.762	119,989
ears/clean_x_2.0pct	ears	12	0.278	0.177	0.711	0.762	119,989
ears/clean_x_5.0pct	ears	12	0.278	0.178	0.709	0.762	119,989
ears/clean_x_10.0pct	ears	12	0.280	0.177	0.719	0.760	119,989

H.7 E07 – Temporal drift

Table 20: E07 – Temporal drift: condition-level means over the contributing encoder rows. SD is over contributing encoder rows, not a confidence interval.

Condition	Corpus	Enc.	EER	SD	minDCF	AUC	Trials
voxceleb1/gap_0yr	voxceleb1	12	0.168	0.184	0.501	0.862	120,000
voxceleb1/gap_5yr	voxceleb1	12	0.176	0.188	0.542	0.855	120,000
voxceleb1/gap_1yr	voxceleb1	12	0.177	0.190	0.522	0.852	120,000
voxceleb1/gap_3yr	voxceleb1	12	0.180	0.193	0.533	0.848	120,000

H.8 E08 – L2 ACCENT

Table 21: E08 – L2 accent: condition-level means over the contributing encoder rows. SD is over contributing encoder rows, not a confidence interval.

Condition	Corpus	Enc.	EER	SD	minDCF	AUC	Trials
l2_arctic/level_3_vs_native	L2-ARCTIC/CV	12	0.077	0.103	0.372	0.956	19,944
l2_arctic/level_1_vs_native	L2-ARCTIC/CV	12	0.079	0.105	0.386	0.954	19,908
l2_arctic/level_4_vs_native	L2-ARCTIC/CV	12	0.082	0.117	0.350	0.948	19,968
l2_arctic/l2_all_vs_native	L2-ARCTIC/CV	12	0.084	0.116	0.371	0.951	19,932
l2_arctic/level_2_vs_native	L2-ARCTIC/CV	12	0.088	0.116	0.376	0.944	19,932
l2_arctic/level_5_vs_native	L2-ARCTIC/CV	12	0.091	0.120	0.388	0.944	19,932
cv_en/native_matched	CV	12	0.110	0.118	0.411	0.935	19,992
l2_arctic/level_3_matched	L2-ARCTIC	12	0.142	0.171	0.430	0.888	23,892
l2_arctic/level_4_matched	L2-ARCTIC	12	0.150	0.184	0.416	0.872	23,964
l2_arctic/arabic_matched	L2-ARCTIC	12	0.159	0.191	0.418	0.863	19,968
l2_arctic/level_2_matched	L2-ARCTIC	12	0.165	0.196	0.440	0.861	23,916
l2_arctic/level_1_matched	L2-ARCTIC	12	0.168	0.205	0.433	0.851	23,904
l2_arctic/spanish_matched	L2-ARCTIC	12	0.168	0.201	0.430	0.855	19,932
l2_arctic/korean_matched	L2-ARCTIC	12	0.169	0.209	0.421	0.845	19,980
l2_arctic/level_5_matched	L2-ARCTIC	12	0.170	0.198	0.462	0.852	23,904
l2_arctic/vietnamese_matched	L2-ARCTIC	12	0.176	0.203	0.463	0.848	19,920
l2_arctic/mandarin_matched	L2-ARCTIC	12	0.181	0.216	0.428	0.840	19,944
l2_arctic/hindi_matched	L2-ARCTIC	12	0.182	0.209	0.445	0.834	19,908

H.9 E10 – TTS SPOOFING

Table 22: E10 – TTS spoofing: condition-level means over the contributing encoder rows. SD is over contributing encoder rows, not a confidence interval.

Condition	Corpus	Enc.	EER	SD	minDCF	AUC	Trials
real_real_baseline	–	12	0.239	0.170	0.649	0.804	120,000
voicecraft/cross_utt_spoof	–	12	0.311	0.094	0.781	0.750	113,664
e2_tts/cross_utt_spoof	–	12	0.368	0.050	0.913	0.691	118,104
voicecraft/synth_vs_real	–	12	0.433	0.127	0.904	0.584	113,664
voicecraft/real_vs_synth	–	12	0.436	0.130	0.901	0.581	113,568
e2_tts/synth_vs_real	–	12	0.530	0.080	0.996	0.465	118,104

Table 22 – continued

Condition	Corpus	Enc.	EER	SD	minDCF	AUC	Trials
e2_tts/real_vs_synth	–	12	0.530	0.079	0.992	0.466	118,344

H.10 E11 – VOICE CONVERSION

Table 23: E11 – Voice conversion: condition-level means over the contributing encoder rows. SD is over contributing encoder rows, not a confidence interval.

Condition	Corpus	Enc.	EER	SD	minDCF	AUC	Trials
freevc/spoof_vs_genuine	–	12	0.225	0.096	0.620	0.832	60,000
real_real_baseline	–	12	0.241	0.166	0.700	0.801	60,000
knn_vc/spoof_vs_genuine	–	12	0.357	0.060	0.890	0.690	60,000

H.11 E13 – AGE × CODEC

Table 24: E13 – Age × codec: condition-level means over the contributing encoder rows. SD is over contributing encoder rows, not a confidence interval.

Condition	Corpus	Enc.	EER	SD	minDCF	AUC	Trials
all__clean	CV	12	0.108	0.119	0.462	0.933	120,000
young__clean	CV	12	0.111	0.117	0.449	0.929	39,996
all__opus	CV	12	0.112	0.118	0.471	0.931	120,000
young__opus	CV	12	0.114	0.116	0.462	0.928	39,996
all__g711	CV	12	0.117	0.111	0.528	0.929	120,000
young__g711	CV	12	0.121	0.111	0.498	0.924	39,996
young__amr	CV	12	0.123	0.109	0.537	0.924	39,996
all__amr	CV	12	0.124	0.108	0.555	0.927	120,000
all__gsm	CV	12	0.127	0.108	0.581	0.925	120,000
young__gsm	CV	12	0.127	0.104	0.542	0.922	39,996
older__clean	CV	12	0.130	0.129	0.420	0.907	39,996
older__opus	CV	12	0.132	0.130	0.438	0.902	39,996
middle__clean	CV	12	0.134	0.118	0.449	0.907	39,996
middle__opus	CV	12	0.134	0.117	0.472	0.905	39,996
older__g711	CV	12	0.135	0.116	0.509	0.895	39,996
older__amr	CV	12	0.141	0.110	0.542	0.893	39,996
older__gsm	CV	12	0.143	0.092	0.545	0.899	39,996
middle__g711	CV	12	0.148	0.115	0.508	0.900	39,996
middle__amr	CV	12	0.153	0.109	0.548	0.895	39,996
middle__gsm	CV	12	0.158	0.110	0.582	0.890	39,996

H.12 E14 – EMOTION × NOISE

Table 25: E14 – Emotion × noise: condition-level means over the contributing encoder rows. SD is over contributing encoder rows, not a confidence interval.

Condition	Corpus	Enc.	EER	SD	minDCF	AUC	Trials
neutral__clean__None	ravdess	12	0.054	0.079	0.168	0.972	3,456
calm__gaussian__25	ravdess	12	0.067	0.099	0.236	0.955	3,456
calm__clean__None	ravdess	12	0.072	0.088	0.226	0.964	3,456

Table 25 – continued

Condition	Corpus	Enc.	EER	SD	minDCF	AUC	Trials
neutral__gaussian_ _25	ravdess	12	0.072	0.107	0.257	0.953	3,456
fearful__clean__None	ravdess	12	0.073	0.092	0.228	0.962	3,456
angry__clean__None	ravdess	12	0.078	0.095	0.260	0.948	3,456
neutral__gaussian_ _15	ravdess	12	0.078	0.120	0.247	0.946	3,456
happy__clean__None	ravdess	12	0.089	0.108	0.263	0.942	3,456
neutral__gaussian__5	ravdess	12	0.090	0.133	0.282	0.938	3,456
sad__clean__None	ravdess	12	0.090	0.109	0.260	0.949	3,456
surprise__clean_ _None	ravdess	12	0.093	0.117	0.292	0.945	3,456
disgust__clean__None	ravdess	12	0.094	0.120	0.308	0.946	3,456
angry__gaussian__25	ravdess	12	0.094	0.120	0.308	0.926	3,456
fearful__gaussian_ _25	ravdess	12	0.095	0.124	0.267	0.937	3,456
happy__gaussian__25	ravdess	12	0.096	0.123	0.298	0.929	3,456
calm__gaussian__15	ravdess	12	0.097	0.136	0.300	0.933	3,456
calm__gaussian__5	ravdess	12	0.098	0.144	0.286	0.930	3,456
happy__gaussian__15	ravdess	12	0.105	0.146	0.324	0.917	3,456
sad__gaussian__25	ravdess	12	0.107	0.135	0.315	0.917	3,456
fearful__gaussian_ _15	ravdess	12	0.108	0.146	0.288	0.920	3,456
angry__gaussian__15	ravdess	12	0.112	0.153	0.339	0.912	3,456
sad__gaussian__5	ravdess	12	0.114	0.148	0.295	0.918	3,456
surprise__gaussian_ _25	ravdess	12	0.115	0.146	0.306	0.909	3,456
sad__gaussian__15	ravdess	12	0.115	0.156	0.316	0.911	3,456
fearful__gaussian__5	ravdess	12	0.119	0.160	0.313	0.911	3,456
disgust__gaussian__5	ravdess	12	0.120	0.148	0.336	0.916	3,456
surprise__gaussian_ _15	ravdess	12	0.120	0.160	0.341	0.908	3,456
disgust__gaussian_ _15	ravdess	12	0.122	0.158	0.337	0.904	3,456
happy__gaussian__5	ravdess	12	0.122	0.167	0.339	0.908	3,456
disgust__gaussian_ _25	ravdess	12	0.124	0.157	0.332	0.908	3,456
surprise__gaussian_ _5	ravdess	12	0.132	0.159	0.396	0.902	3,456
angry__gaussian__5	ravdess	12	0.136	0.174	0.359	0.897	3,456
fearful_ _environmental__25	ravdess	12	0.149	0.181	0.406	0.874	3,456
calm__environmental_ _25	ravdess	12	0.156	0.192	0.388	0.865	3,456
happy_ _environmental__25	ravdess	12	0.159	0.190	0.400	0.861	3,456
angry_ _environmental__25	ravdess	12	0.160	0.189	0.403	0.859	3,456
sad__environmental_ _25	ravdess	12	0.161	0.193	0.404	0.862	3,456
disgust_ _environmental__25	ravdess	12	0.165	0.199	0.410	0.859	3,456
neutral_ _environmental__25	ravdess	12	0.170	0.207	0.401	0.853	3,456
surprise_ _environmental__25	ravdess	12	0.172	0.205	0.411	0.850	3,456
fearful_ _environmental__15	ravdess	12	0.172	0.209	0.411	0.845	3,456
sad__environmental_ _15	ravdess	12	0.175	0.212	0.411	0.841	3,456
calm__environmental_ _15	ravdess	12	0.175	0.211	0.398	0.840	3,456

Table 25 – continued

Condition	Corpus	Enc.	EER	SD	minDCF	AUC	Trials
happy_environmental__15	ravdess	12	0.177	0.212	0.407	0.839	3,456
disgust_environmental__15	ravdess	12	0.178	0.209	0.407	0.836	3,456
neutral_environmental__15	ravdess	12	0.178	0.217	0.403	0.835	3,456
angry_environmental__15	ravdess	12	0.182	0.215	0.415	0.838	3,456
surprise_environmental__15	ravdess	12	0.188	0.222	0.457	0.827	3,456
calm__crosstalk__25	ravdess	12	0.194	0.215	0.462	0.825	3,456
sad__environmental_5	ravdess	12	0.194	0.227	0.439	0.812	3,456
fearful_environmental__5	ravdess	12	0.194	0.230	0.444	0.811	3,456
neutral_environmental__5	ravdess	12	0.195	0.233	0.411	0.812	3,456
angry_environmental__5	ravdess	12	0.196	0.226	0.443	0.812	3,456
calm__environmental_5	ravdess	12	0.196	0.228	0.435	0.810	3,456
happy_environmental__5	ravdess	12	0.201	0.238	0.428	0.808	3,456
sad__crosstalk__25	ravdess	12	0.203	0.218	0.447	0.815	3,456
disgust_environmental__5	ravdess	12	0.204	0.233	0.448	0.805	3,456
neutral__crosstalk_25	ravdess	12	0.205	0.230	0.451	0.812	3,456
surprise_environmental__5	ravdess	12	0.205	0.232	0.501	0.805	3,456
disgust__crosstalk_25	ravdess	12	0.209	0.225	0.483	0.811	3,456
fearful__crosstalk_25	ravdess	12	0.213	0.226	0.466	0.816	3,456
angry__crosstalk__25	ravdess	12	0.214	0.231	0.448	0.807	3,456
happy__crosstalk__25	ravdess	12	0.216	0.234	0.472	0.805	3,456
surprise__crosstalk_25	ravdess	12	0.229	0.231	0.506	0.794	3,456
calm__crosstalk__15	ravdess	12	0.239	0.243	0.520	0.776	3,456
neutral__crosstalk_15	ravdess	12	0.245	0.249	0.516	0.769	3,456
happy__crosstalk__15	ravdess	12	0.247	0.238	0.552	0.770	3,456
sad__crosstalk__15	ravdess	12	0.254	0.242	0.545	0.766	3,456
disgust__crosstalk_15	ravdess	12	0.255	0.228	0.561	0.770	3,456
fearful__crosstalk_15	ravdess	12	0.256	0.225	0.555	0.772	3,456
angry__crosstalk__15	ravdess	12	0.267	0.236	0.612	0.760	3,456
surprise__crosstalk_15	ravdess	12	0.276	0.215	0.661	0.754	3,456
neutral__crosstalk_5	ravdess	12	0.343	0.158	0.910	0.695	3,456
happy__crosstalk__5	ravdess	12	0.369	0.135	0.933	0.673	3,456
calm__crosstalk__5	ravdess	12	0.371	0.138	0.835	0.678	3,456
angry__crosstalk__5	ravdess	12	0.380	0.113	0.936	0.649	3,456
sad__crosstalk__5	ravdess	12	0.381	0.129	0.921	0.653	3,456
disgust__crosstalk_5	ravdess	12	0.387	0.119	0.969	0.644	3,456
surprise__crosstalk_5	ravdess	12	0.397	0.107	0.951	0.637	3,456

fearful__crosstalk_5	ravdess	12	0.401	0.110	0.933	0.642	3,456

H.13 E15 – Pathology and codec (descriptive)

Table 26: E15 – Pathology and codec (descriptive): condition-level means over the contributing encoder rows. SD is over contributing encoder rows, not a confidence interval.

Condition	Corpus	Enc.	EER	SD	minDCF	AUC	Trials
EARS/healthy_clean	ears	12	0.246	0.175	0.663	0.797	119,988
EARS/healthy_opus	ears	12	0.259	0.160	0.679	0.787	29,997
EARS/healthy_amr	ears	12	0.260	0.165	0.702	0.786	29,996
EARS/healthy_gsm	ears	12	0.260	0.163	0.718	0.788	29,996
EARS/healthy_g711	ears	12	0.279	0.155	0.755	0.772	29,998
TORGO/dysarthria_moderate_clean	torgo	12	0.303	0.106	0.932	0.752	37,411
TORGO/dysarthria_moderate_g711	torgo	12	0.304	0.100	0.880	0.758	9,376
TORGO/dysarthria_moderate_gsm	torgo	12	0.305	0.113	0.921	0.744	9,367
TORGO/dysarthria_moderate_amr	torgo	12	0.308	0.112	0.916	0.744	9,359
TORGO/dysarthria_severe_g711	torgo	12	0.309	0.121	0.946	0.744	9,645
TORGO/dysarthria_moderate_opus	torgo	12	0.316	0.089	0.930	0.731	9,356
TORGO/dysarthria_severe_clean	torgo	12	0.327	0.099	0.975	0.725	38,450
TORGO/dysarthria_severe_opus	torgo	12	0.333	0.098	0.967	0.717	9,607
TORGO/dysarthria_severe_gsm	torgo	12	0.334	0.099	0.964	0.717	9,611
TORGO/dysarthria_severe_amr	torgo	12	0.337	0.097	0.951	0.717	9,622
TORGO/dysarthria_mild_clean	torgo	12	0.339	0.096	0.972	0.708	38,187
TORGO/dysarthria_mild_opus	torgo	12	0.342	0.088	0.954	0.706	9,521
TORGO/dysarthria_mild_gsm	torgo	12	0.350	0.092	0.949	0.696	9,548
TORGO/dysarthria_mild_amr	torgo	12	0.350	0.096	0.938	0.697	9,482
TORGO/dysarthria_mild_g711	torgo	12	0.351	0.099	0.951	0.697	9,550

H.14 E16 – Gender × noise

Table 27: E16 – Gender × noise: condition-level means over the contributing encoder rows. SD is over contributing encoder rows, not a confidence interval.

Condition	Corpus	Enc.	EER	SD	minDCF	AUC	Trials
CV/all_clean	CV	12	0.109	0.118	0.446	0.932	60,000
CV/F_clean	CV	12	0.110	0.113	0.432	0.932	60,000
CV/M_clean	CV	12	0.117	0.116	0.466	0.927	60,000
CV/all_gaussian_25dB	CV	12	0.131	0.135	0.453	0.912	13,332

Table 27 – continued

Condition	Corpus	Enc.	EER	SD	minDCF	AUC	Trials
CV/F_environmental_25dB	CV	12	0.131	0.138	0.448	0.910	6,660
CV/all_gaussian_15dB	CV	12	0.137	0.143	0.496	0.900	13,332
CV/F_gaussian_25dB	CV	12	0.139	0.134	0.450	0.910	6,660
CV/F_gaussian_15dB	CV	12	0.139	0.137	0.503	0.906	6,660
CV/all_environmental_25dB	CV	12	0.140	0.144	0.437	0.901	13,332
CV/M_gaussian_25dB	CV	12	0.141	0.149	0.410	0.899	6,660
CV/M_gaussian_15dB	CV	12	0.146	0.132	0.509	0.897	6,660
CV/M_environmental_25dB	CV	12	0.147	0.142	0.457	0.896	6,660
CV/all_environmental_15dB	CV	12	0.154	0.154	0.522	0.889	13,332
CV/F_gaussian_5dB	CV	12	0.159	0.148	0.579	0.884	6,660
CV/M_crosstalk_25dB	CV	12	0.162	0.146	0.430	0.885	6,660
CV/all_crosstalk_25dB	CV	12	0.163	0.154	0.495	0.880	13,332
CV/F_crosstalk_25dB	CV	12	0.167	0.145	0.444	0.882	6,660
CV/all_gaussian_5dB	CV	12	0.167	0.160	0.574	0.875	13,332
CV/F_environmental_15dB	CV	12	0.171	0.162	0.527	0.876	6,660
CV/M_gaussian_5dB	CV	12	0.180	0.153	0.565	0.866	6,660
CV/M_environmental_15dB	CV	12	0.187	0.169	0.523	0.856	6,660
CV/all_environmental_5dB	CV	12	0.193	0.182	0.554	0.836	13,332
CV/M_environmental_5dB	CV	12	0.200	0.171	0.535	0.847	6,660
CV/F_environmental_5dB	CV	12	0.201	0.181	0.525	0.834	6,660
CV/all_crosstalk_15dB	CV	12	0.227	0.165	0.604	0.821	13,332
CV/M_crosstalk_15dB	CV	12	0.236	0.188	0.551	0.803	6,660
CV/F_crosstalk_15dB	CV	12	0.270	0.172	0.616	0.784	6,660
CV/all_crosstalk_5dB	CV	12	0.333	0.121	0.928	0.720	13,332
CV/M_crosstalk_5dB	CV	12	0.344	0.104	0.855	0.712	6,660
CV/F_crosstalk_5dB	CV	12	0.347	0.108	0.854	0.714	6,660

H.15 E17 – Temporal drift × codec

Table 28: E17 – Temporal drift × codec: condition-level means over the contributing encoder rows. SD is over contributing encoder rows, not a confidence interval.

Condition	Corpus	Enc.	EER	SD	minDCF	AUC	Trials
VoxCeleb1/gap_0_clean	voxceleb1	12	0.168	0.186	0.502	0.862	120,000
VoxCeleb1/gap_0_opus	voxceleb1	12	0.169	0.180	0.496	0.862	30,000
VoxCeleb1/gap_5yr_opus	voxceleb1	12	0.176	0.183	0.523	0.860	9,984
VoxCeleb1/gap_1yr_clean	voxceleb1	12	0.176	0.189	0.516	0.853	39,984
VoxCeleb1/gap_5yr_clean	voxceleb1	12	0.176	0.187	0.516	0.856	39,984

Table 28 – continued

Condition	Corpus	Enc.	EER	SD	minDCF	AUC	Trials
VoxCeleb1/gap_3yr_opus	voxceleb1	12	0.178	0.181	0.496	0.854	9,984
VoxCeleb1/gap_0_g711	voxceleb1	12	0.179	0.177	0.567	0.856	30,000
VoxCeleb1/gap_3yr_clean	voxceleb1	12	0.180	0.192	0.527	0.849	39,984
VoxCeleb1/gap_1yr_opus	voxceleb1	12	0.183	0.197	0.499	0.847	9,984
VoxCeleb1/gap_0_amr	voxceleb1	12	0.185	0.177	0.608	0.853	30,000
VoxCeleb1/gap_5yr_g711	voxceleb1	12	0.190	0.182	0.602	0.850	9,984
VoxCeleb1/gap_3yr_g711	voxceleb1	12	0.192	0.177	0.572	0.846	9,984
VoxCeleb1/gap_0_gsm	voxceleb1	12	0.195	0.182	0.645	0.842	30,000
VoxCeleb1/gap_1yr_g711	voxceleb1	12	0.196	0.195	0.542	0.839	9,984
VoxCeleb1/gap_3yr_amr	voxceleb1	12	0.198	0.174	0.588	0.839	9,984
VoxCeleb1/gap_5yr_amr	voxceleb1	12	0.199	0.180	0.590	0.842	9,984
VoxCeleb1/gap_1yr_amr	voxceleb1	12	0.202	0.193	0.576	0.834	9,984
VoxCeleb1/gap_1yr_gsm	voxceleb1	12	0.206	0.193	0.613	0.828	9,984
VoxCeleb1/gap_5yr_gsm	voxceleb1	12	0.209	0.180	0.609	0.837	9,984
VoxCeleb1/gap_3yr_gsm	voxceleb1	12	0.209	0.174	0.620	0.835	9,984

H.16 E18 – Accent and codec (descriptive)

Table 29: E18 – Accent and codec (descriptive): condition-level means over the contributing encoder rows. SD is over contributing encoder rows, not a confidence interval.

Condition	Corpus	Enc.	EER	SD	minDCF	AUC	Trials
CV_EN/native_gsm	CV	12	0.105	0.117	0.388	0.936	15,000
CV_EN/native_amr	CV	12	0.106	0.119	0.381	0.934	15,000
CV_EN/native_clean	CV	12	0.109	0.119	0.438	0.933	60,000
CV_EN/native_opus	CV	12	0.113	0.120	0.424	0.930	15,000
CV_EN/native_g711	CV	12	0.115	0.111	0.456	0.929	15,000
L2-ARCTIC/accent_gsm	L2-ARCTIC	12	0.175	0.171	0.649	0.859	14,951
L2-ARCTIC/accent_amr	L2-ARCTIC	12	0.177	0.174	0.551	0.854	14,951
L2-ARCTIC/accent_clean	L2-ARCTIC	12	0.179	0.169	0.650	0.856	59,828
L2-ARCTIC/accent_opus	L2-ARCTIC	12	0.180	0.166	0.673	0.853	14,957
L2-ARCTIC/accent_g711	L2-ARCTIC	12	0.188	0.168	0.682	0.850	14,961

H.17 E20 – AGE × CODEC × DRIFT (AGE PROXIED)

Table 30: E20 – Age × codec × drift (age proxied): condition-level means over the contributing encoder rows. SD is over contributing encoder rows, not a confidence interval.

Condition	Corpus	Enc.	EER	SD	minDCF	AUC	Trials
VoxCeleb1/age_middle_clean	voxceleb1	12	0.163	0.180	0.463	0.868	15,000
VoxCeleb1/AC_older_opus	voxceleb1	12	0.163	0.178	0.488	0.869	7,500
VoxCeleb1_3way/older_gap5yr_opus	voxceleb1	12	0.166	0.178	0.491	0.862	3,312
VoxCeleb1/gap0_opus	voxceleb1	12	0.167	0.180	0.471	0.867	9,996
VoxCeleb1/AC_middle_opus	voxceleb1	12	0.167	0.177	0.469	0.866	7,500
VoxCeleb1/age_older_clean	voxceleb1	12	0.167	0.183	0.473	0.865	15,000
VoxCeleb1/clean	voxceleb1	12	0.168	0.184	0.488	0.863	30,000
VoxCeleb1/BC_older_gap5yr	voxceleb1	12	0.168	0.181	0.476	0.864	7,488
VoxCeleb1/age_young_clean	voxceleb1	12	0.168	0.183	0.467	0.865	15,000
VoxCeleb1/BC_middle_gap5yr	voxceleb1	12	0.169	0.186	0.459	0.860	7,488
VoxCeleb1/AC_young_opus	voxceleb1	12	0.170	0.183	0.482	0.864	7,500
VoxCeleb1/gap_5yr_clean	voxceleb1	12	0.171	0.180	0.529	0.863	9,984
VoxCeleb1_3way/middle_gap5yr_opus	voxceleb1	12	0.173	0.192	0.437	0.863	3,312
VoxCeleb1/AC_older_g711	voxceleb1	12	0.174	0.170	0.512	0.867	7,500
VoxCeleb1/gap_3yr_clean	voxceleb1	12	0.175	0.181	0.485	0.857	9,984
VoxCeleb1_3way/middle_gap5yr_g711	voxceleb1	12	0.177	0.183	0.502	0.858	3,312
VoxCeleb1/AC_young_g711	voxceleb1	12	0.177	0.170	0.539	0.861	7,500
VoxCeleb1/gap_5yr_opus	voxceleb1	12	0.178	0.182	0.457	0.860	3,312
VoxCeleb1/BC_older_gap3yr	voxceleb1	12	0.179	0.191	0.485	0.853	7,488
VoxCeleb1/BC_young_gap3yr	voxceleb1	12	0.180	0.189	0.483	0.851	7,488
VoxCeleb1/gap0_g711	voxceleb1	12	0.183	0.183	0.541	0.854	9,996
VoxCeleb1_3way/older_gap5yr_g711	voxceleb1	12	0.183	0.169	0.599	0.859	3,312
VoxCeleb1/AC_middle_g711	voxceleb1	12	0.184	0.181	0.534	0.852	7,500
VoxCeleb1_3way/young_gap3yr_opus	voxceleb1	12	0.184	0.193	0.515	0.845	3,312
VoxCeleb1_3way/middle_gap5yr_gsm	voxceleb1	12	0.184	0.175	0.525	0.853	3,312
VoxCeleb1/gap_5yr_g711	voxceleb1	12	0.185	0.183	0.521	0.856	3,312
VoxCeleb1/BC_middle_gap3yr	voxceleb1	12	0.187	0.195	0.505	0.844	7,488
VoxCeleb1/gap_3yr_opus	voxceleb1	12	0.187	0.200	0.478	0.845	3,312
VoxCeleb1_3way/young_gap5yr_opus	voxceleb1	12	0.188	0.186	0.493	0.845	3,312
VoxCeleb1/gap_3yr_g711	voxceleb1	12	0.188	0.185	0.597	0.841	3,312
VoxCeleb1/BC_young_gap5yr	voxceleb1	12	0.188	0.199	0.486	0.842	7,488

Table 30 – continued

Condition	Corpus	Enc.	EER	SD	minDCF	AUC	Trials
VoxCeleb1/AC_older_gsm	voxceleb1	12	0.189	0.180	0.574	0.852	7,500
VoxCeleb1/gap0_gsm	voxceleb1	12	0.190	0.174	0.597	0.845	9,996
VoxCeleb1_3way/older_gap5yr_gsm	voxceleb1	12	0.193	0.173	0.632	0.850	3,312
VoxCeleb1_3way/young_gap3yr_g711	voxceleb1	12	0.194	0.183	0.502	0.841	3,312
VoxCeleb1_3way/middle_gap3yr_opus	voxceleb1	12	0.194	0.200	0.479	0.835	3,312
VoxCeleb1/AC_middle_gsm	voxceleb1	12	0.197	0.176	0.585	0.843	7,500
VoxCeleb1/AC_young_gsm	voxceleb1	12	0.197	0.177	0.598	0.843	7,500
VoxCeleb1/gap_5yr_gsm	voxceleb1	12	0.199	0.173	0.557	0.847	3,312
VoxCeleb1_3way/older_gap3yr_opus	voxceleb1	12	0.199	0.192	0.542	0.836	3,312
VoxCeleb1_3way/young_gap5yr_g711	voxceleb1	12	0.204	0.185	0.556	0.833	3,312
VoxCeleb1/gap_3yr_gsm	voxceleb1	12	0.205	0.173	0.572	0.831	3,312
VoxCeleb1_3way/middle_gap3yr_g711	voxceleb1	12	0.208	0.194	0.535	0.829	3,312
VoxCeleb1_3way/young_gap3yr_gsm	voxceleb1	12	0.208	0.183	0.560	0.834	3,312
VoxCeleb1_3way/older_gap3yr_g711	voxceleb1	12	0.211	0.187	0.592	0.829	3,312
VoxCeleb1_3way/young_gap5yr_gsm	voxceleb1	12	0.214	0.171	0.601	0.828	3,312
VoxCeleb1_3way/older_gap3yr_gsm	voxceleb1	12	0.215	0.183	0.626	0.820	3,312
VoxCeleb1_3way/middle_gap3yr_gsm	voxceleb1	12	0.230	0.194	0.601	0.814	3,312