EVIDENCE OF A RECURRENT SCAM-CALL SPEAKER ACROSS THE KITBOGA VIDEO ARCHIVE

A three-model acoustic retrieval study using diarization, same-video controls, archive-wide search, and a known-voice benchmark

RESEARCH THESIS

Independent archival audio study · July 28, 2026

Research proposition

The same scam-call speaker recurs across a substantial share of the processed archive.

Abstract

This thesis tests the hypothesis that the same scam-call speaker recurs across a substantial share of Kitboga's public video archive. The study begins with 244 successfully processed videos. Speaker diarization divided each recording into video-local voices, and multiple speech turns from each local voice were combined into speaker profiles. A staged archive search, independent local verification, stronger speaker reconstruction, and listening review produced a frozen 91-video recurrence cohort containing one nominated voice from each source video. SpeechBrain, WeSpeaker, and Resemblyzer then evaluated the same cohort and the same archive universe.

The design separates four questions that resemblance alone tends to blur. First, do the 91 nominated voices resemble one another across different videos? Second, do their pairwise scores exceed the wider archive background? Third, within each source video, does the nominated voice outrank the other diarized speakers recorded under the same episode conditions? Fourth, can the same machinery recover a recognizable recurring voice that is known from archive context—the Kitboga Grandma character?

The frozen cohort produced 4,095 cross-video pairs per model. With multi-segment speaker profiles, the probability that a cohort pair outranked an external archive pair was 0.942 for SpeechBrain, 0.905 for WeSpeaker, and 0.927 for Resemblyzer. The three models therefore placed the cohort far above the broader cross-video background on their own score scales. In the condition-matched analysis, the nominated voice was compared with 720 other voices from the same 91 videos. At least two models ranked the nominated voice first in 66 of 91 videos and within the top two in 85 of 91. Model-specific target-versus-alternative win rates ranged from 92.5% to 97.1%. A standardized three-model test yielded a target-control separation of 1.896 standard deviations, with a video-blocked 95% confidence interval of 1.779–2.012 and an empirical sign-flip p-value below 0.0001.

Archive-wide retrieval added 37 videos under the strict three-model rule. The strict evidence set therefore contains 128 of 244 processed videos, or 52.5% of the archive. Broader ranking rules produced listening pools of 152 and 173 videos. Four profile-aggregation rules yielded almost identical cohort-versus-background separation, showing that the principal finding is stable across plausible ways of combining speech turns. The known-voice benchmark recovered the title-supported Grandma label in the top two for all 15 held-out videos under every model. In 12 episodes containing both the nominated scam-call voice and a separate Grandma voice, all three models selected the matching voice in both directions, producing 12 correct bidirectional switches out of 12. A separate 46-video long-profile analysis compared the nominated speakers with 20,000 random same-video cohorts; the observed mean three-model percentile was 0.9197 versus 0.5456 under the null ($p < 0.0001$).

Taken together, the results favor a recurrent-speaker explanation. The same acoustic pattern persists across years and recording conditions, separates from speakers in the same episodes, replicates across three model families, extends beyond the discovery cohort, and remains distinct from a second recurring voice profile. Across all 4,095 cohort pairs, similarity declined moderately with publication-year separation while remaining highly elevated even at the longest observed gaps. The strict archive estimate concerns processed video uploads. Authenticated identity, production intent, and independence among underlying calls remain separate questions for future forensic and provenance work.

Keywords: speaker retrieval; recurrent speaker; speaker embeddings; diarization; same-video controls; archive search; SpeechBrain; WeSpeaker; Resemblyzer

Contents

C.6 Global three-model coherence diagnostics

Table C5. Global coherence diagnostics across the three model systems.

Evidence universe	Reliability / concordance	First component
Complete 2,051-speaker profile universe	Standardized reliability 0.805; Kendall W 0.720	72.2% variance
46-video long-profile subset	Kendall W 0.754	69.3% variance

The diagnostics summarize shared ordering after standardization. They do not convert the three systems into independent replications or make their raw cosine values directly comparable.

C.7 Long-profile same-video null-cohort robustness

Table C6. Rebuilt long-profile cohort against video-matched null cohorts.

Statistic	Observed cohort	Random same-video cohort	Inference
Mean three-model pair percentile	0.9197	0.5456	14.79 null SD; $p < 0.0001$
All-pair p95 edge rate	36.0%	6.0%	$p < 0.0001$
All three models rank target first	24/46	Expected 3.98	Exact $p < 0.0001$
At least two models rank target first	36/46	Expected 5.74	Exact $p < 0.0001$

Statistic	Observed cohort	Random same-video cohort	Inference
At least two models place target top two	44/46	Video-label exchangeability	Near-complete split-aware recovery

C.8 Illustrative outside-cohort candidates

The rows below illustrate the auditable speaker-level content of the wider q05 and q01 listening bands. They remain separate from the stricter q10 archive count.

Table C7. Selected speaker-level candidates from the wider ranked listening bands.

Band	Video / local speaker	Mean percentile	Weakest model	Evidence note
q05-q10	20230708 c9YpR2EHsGs / SPEAKER_05	0.915	0.088	24 clips; 229.94 s
q05-q10	20250831 21Cu-ulwTqY / SPEAKER_02	0.914	0.099	24 clips; 172.78 s
q05-q10	20220524 nn6eq_k2_4g / SPEAKER_08	0.912	0.099	17 clips; 135.12 s
q05-q10	20220213 R3P8BM2C2qI / SPEAKER_06	0.912	0.088	Listening matched labels 05 and 06
q05-q10	20211124 _rN7QDWAYl4 / SPEAKER_08	0.883	0.088	Listening matched labels 07 and 08
q01-q05	20211113 Qd21_Xy3kDc / SPEAKER_05	0.898	0.011	17 clips; 128.72 s

Across outside-video rankings, the model correlations were 0.835 for SpeechBrain-WeSpeaker, 0.636 for SpeechBrain-Resemblyzer, and 0.596 for WeSpeaker-Resemblyzer; all p-values were below 5 x 10^-16.

C.9 Temporal robustness statistics

Table C8. Similarity across publication-year gaps in the 91-video cohort.

Measure	Observed result	Network-preserving permutation
Three-model consensus rho	-0.318	p = 0.0002
SpeechBrain rho	-0.310	p = 0.0002
WeSpeaker rho	-0.302	p = 0.0002

Measure	Observed result	Network-preserving permutation
Resemblyzer rho	-0.255	p = 0.0002
Same-year mean consensus percentile	0.933	670 pairs
Nine-year-gap mean consensus percentile	0.868	16 pairs

C.10 Known-voice benchmark by model

Table C9. Held-out Grandma retrieval across 235 non-seed videos.

Model	Video AUC	Exact rank first	Top two
SpeechBrain	0.860	14/15	15/15
WeSpeaker	0.879	14/15	15/15
Resemblyzer	0.817	15/15	15/15
Three-model consensus	0.850	-	15/15

C.11 Hard matched-alternative review queue

A separate 40-row review queue contains the highest-scoring, duration-screened same-video alternatives. The queue records the source video, nominated and alternative local labels, total speech, retained clip counts, model scores, and bundle locations. These rows are preserved as ambiguous hard comparisons because some may represent the recurrent speaker split across more than one local label.

- Appendix D. Evidence-package guide
- Appendix E. Prospective blinded listening protocol

1. Introduction

Kitboga videos present long, edited conversations with scam callers, hosts, character voices, music, and telephone-channel artifacts. Across this archive, one male scam-call voice appears repeatedly enough to motivate a formal question: does the archive contain the same recurring speaker across many different uploads? Human listening can generate that hypothesis, yet a scientific answer requires an explicit evidence unit, a reproducible comparison rule, matched alternatives, and a clear distinction between acoustic recurrence and authenticated identity.

This thesis treats the problem as archive-scale speaker retrieval. The target is a recurring acoustic signature, the search space is every video-local speaker profile, and the decisive comparisons occur both across videos and inside the same videos. The study therefore moves beyond a montage of similar-sounding excerpts. It asks whether the nominated voice behaves as a coherent retrieval target across three models and whether it systematically outranks other speakers exposed to the same episode-level recording conditions.

The archive is especially demanding because the audio was created for entertainment rather than research. Calls span multiple years, phone connections, compression levels, emotional states, speaking rates, and background environments. Diarization can split one human into several local labels or merge overlapping voices. Video editing can place multiple calls in one upload. These conditions make a simple global score threshold fragile. They also make the archive a useful natural test bed: a genuinely recurring speaker should survive substantial acoustic variation while still separating from unrelated voices.

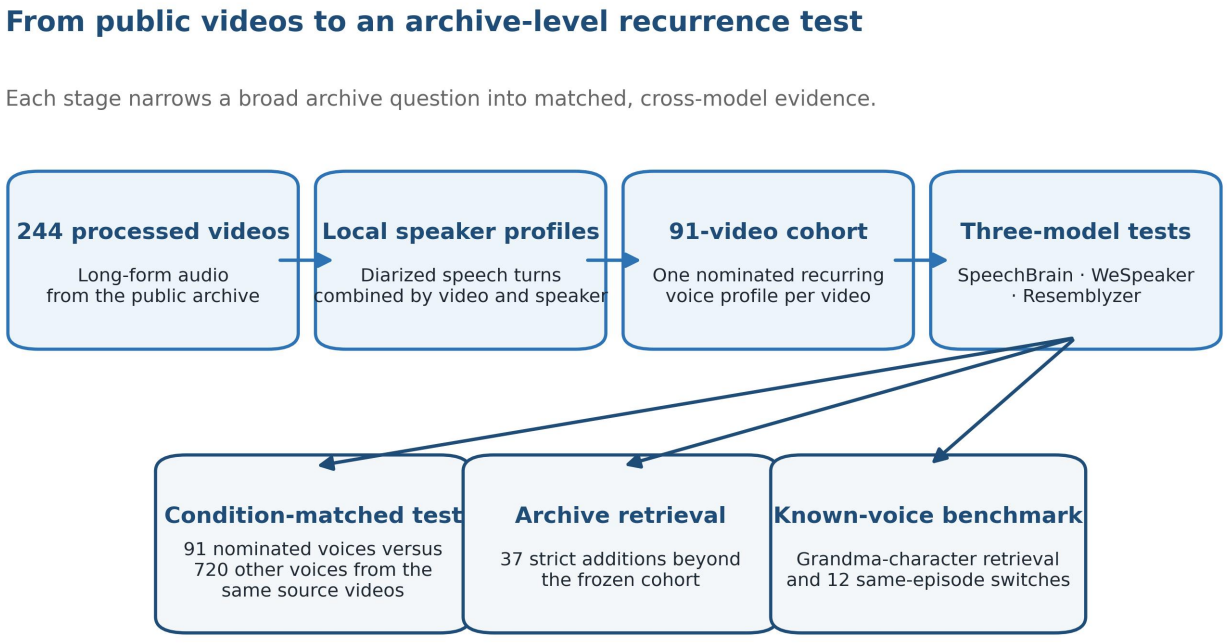

Figure 1. Reader-facing study design. The principal evidence combines cross-video recurrence, same-video alternatives, archive extension, and a known-voice benchmark.

1.1 Research questions

Question	Operational test	Evidence expected under recurrence
RQ1. Cross-video cohesion	Compare all 91 nominated voices with one another.	Scores form a consistently elevated distribution across videos.
RQ2. Archive separation	Compare cohort pairs with external cross-video pairs.	Cohort pairs outrank archive-background pairs within each model.
RQ3. Same-video identification	Rank each nominated voice among all local speakers in its source video.	The nominated voice ranks first or near first under the common cohort profile.
RQ4. Archive extension	Search outside the frozen cohort with model-specific percentile rules.	Additional videos satisfy the same three-model recurrence rule.
RQ5. Method validity	Retrieve the recurring Grandma character and switch profiles in shared episodes.	Each profile selects its corresponding voice under matched recording conditions.

1.2 Hypotheses

H1 predicts that a randomly selected pair from the recurrence cohort will outrank a randomly selected external cross-video pair more often than chance. H2 predicts that the nominated voice will occupy the leading positions inside its own source video when scored against a profile made from the other cohort videos. H3 predicts positive rank agreement among SpeechBrain, WeSpeaker, and Resemblyzer despite their different architectures and raw score scales. H4 predicts that strict archive retrieval will add videos beyond the frozen cohort. H5 predicts that the same retrieval design will recover the separately defined Grandma voice and distinguish it from the nominated scam-call voice in shared episodes.

These hypotheses form a ladder of evidence. Internal cohesion establishes measurable resemblance. Archive separation establishes relative unusualness. Same-video ranking controls episode conditions. Cross-model agreement establishes reproducibility across implementations. The known-voice benchmark shows that the pipeline can separate two recurring voices under the same source conditions. The strict archive-extension rule translates the acoustic pattern into a video-level coverage result.

1.3 Thesis claim

The thesis advances a specific archive-level claim: the processed Kitboga archive contains a recurrent scam-call acoustic signature that is best explained by repeated presence of the same speaker across many uploads. The strict evidence set covers 128 of 244 processed videos. The inference concerns acoustic recurrence across uploads. A small group of highly similar speakers, recurring source material, and editorial

reuse remain competing production histories, and Chapter 7 defines the evidence needed to distinguish among them.

1.4 Contributions

- A transparent reconstruction of a 91-video recurrence cohort from long-form public audio.
- Three-model replication using SpeechBrain, WeSpeaker, and Resemblyzer on identical evidence units.
- A condition-matched comparison against 720 other voices from the same 91 source videos.
- An archive-wide strict retrieval result expressed at the video level.
- A duration-aggregation sensitivity analysis using every retained speaker segment.
- A Kitboga-native known-voice benchmark centered on the recurring Grandma character.
- A clear separation between acoustic recurrence, upload coverage, call independence, and authenticated identity.

2. Literature review and conceptual framework

Speaker-recognition research contains several related tasks whose distinction determines the correct interpretation of this archive. Diarization asks who spoke when inside one recording. Verification asks whether two samples are compatible with the same claimed speaker. Closed-set identification assigns a sample to one enrolled identity. Open-set identification adds the possibility that the speaker is absent from the enrollment set. Speaker retrieval searches a large archive for all files or local speakers that resemble a query and ranks the results [1–3].

The present study is closest to speaker retrieval in the wild. The speaker's civil identity is outside the model, archive metadata provide weak labels, and acoustic conditions vary naturally. The scientific target is therefore a recurring voice pattern and its ranking behavior across a large collection. That framing favors profile aggregation, within-model percentiles, same-video controls, and information-retrieval style outputs over a single universal similarity cutoff.

2.1 Speaker embeddings and modern comparison systems

Modern systems convert variable-length speech into fixed-dimensional embeddings designed to preserve speaker-discriminative information while reducing linguistic and channel variation. Cosine similarity between normalized embeddings then provides a convenient measure of proximity. ECAPA-TDNN extends the x-vector family with channel attention, multi-layer feature aggregation, and attentive statistics pooling [4]. SpeechBrain packages an ECAPA model trained on VoxCeleb data [5]. WeSpeaker supplies a separate production-oriented embedding stack, here represented by a ResNet34-LM model [6]. Resemblyzer provides a d-vector style encoder whose architecture and training lineage differ from the two newer systems [7].

The three systems share the broad embedding paradigm and exposure to unconstrained speech corpora, so their outputs represent convergent measurements rather than statistically independent votes. Their differences still matter. Agreement across models reduces dependence on one implementation, one score scale, or one neighborhood geometry. Disagreement identifies cases where channel, duration, phonetic content, or model-specific inductive bias alters the ranking.

VoxCeleb and VoxCeleb2 established large-scale audiovisual corpora for speaker recognition under naturally variable conditions [8,9]. Their real-world interview audio offers a useful training domain for telephone and edited-video material, although the Kitboga archive adds stronger telephone compression, voice acting, adversarial emotion, and long-form editing. This domain gap motivates archive-native controls.

2.2 Retrieval in unconstrained media archives

Campbell and Singer [1] formulated query-by-example speaker search as a content graph: speech samples become nodes, speaker similarity becomes edges, and multiple examples can guide retrieval through the graph. That formulation supplies an important conceptual point for the present study. A recurring speaker

need not form a perfect all-to-all clique. Some samples may connect strongly through central examples while distant-year or distorted samples produce weaker direct edges.

Loweimi et al. [2] provide the closest recent analogue. Their BBC Rewind study combines preprocessing, diarization, segment embeddings, speaker-level aggregation, and archive ranking under uncontrolled acoustic conditions. They distinguish segment-level and speaker-level retrieval, describe how diarization errors propagate into the embedding stage, and test duration-based aggregation. Their framework motivates the present use of video-local speaker profiles and the sensitivity analysis across equal, linear, square-root, and capped duration rules.

The two archives differ in scale, metadata, and target labels. BBC Rewind contains known public figures and supports conventional retrieval metrics such as precision at K, mean average precision, mean reciprocal rank, and normalized discounted cumulative gain. The Kitboga archive begins with an unlabeled recurring-voice hypothesis. Its strongest evidence therefore comes from matched alternatives, frozen-cohort replication, and a known-character benchmark rather than from identity-labeled ground truth.

2.3 Open-set search and the false-alarm problem

VoxWatch formalizes a central open-set difficulty: as the number of enrolled or searched speakers grows, the largest score among out-of-set speakers also tends to grow [3]. This watchlist-size effect makes raw maxima increasingly vulnerable to false alarms. Score calibration and fusion can improve performance, while adaptive normalization offers no universal guarantee.

The present study addresses the same structural risk in three ways. First, scores are ranked within each model rather than averaged across incompatible raw scales. Second, strict archive additions require concordant placement across all three models. Third, the principal local test compares the nominated voice with every alternative speaker in the same video. These choices make the conclusion depend on repeated, structured ranking behavior instead of one extreme score found in a large search.

2.4 Diarization and speaker-profile aggregation

Pyannote.audio provides neural building blocks and end-to-end pipelines for speaker diarization [10,11]. Diarization assigns local labels within a recording; the label SPEAKER_03 in one video carries no identity relationship to SPEAKER_03 in another. Cross-video recurrence therefore requires a second stage that compares the audio assigned to those local labels.

Local labels are imperfect evidence units. Rapid turns, overlap, voice changers, music, and editing can split one person across labels or combine material from multiple sources. Speaker-profile aggregation reduces dependence on any single short turn by averaging embeddings from multiple retained segments. Equal weighting treats each segment as one observation. Duration weighting emphasizes longer speech, while a cap prevents a single long segment from dominating. Stability across these rules indicates that the conclusion is supported by the speaker's broader segment collection.

2.5 Channel, duration, and domain mismatch

Forensic speaker-comparison research shows that device, transmission, net speech duration, and relevant-population choice influence discrimination and calibration [12–15]. The NFI-FRIDA work demonstrates a practical methodology for studying device variability with simultaneous recordings from forensically relevant channels [12]. Other studies document degradation under mismatched recording conditions and language or channel variation [13,14].

The archive offers partial natural control over these factors. Other speakers in the same upload share much of the editing chain, loudness treatment, and episode-level environment. They may also share telephone channels and broad accent categories. Same-video ranking therefore provides a stronger comparison than a random voice from another year. At the same time, exact handset, call route, accent, age, and speech duration remain unannotated. The interpretation accordingly emphasizes relative retrieval and matched-video separation.

2.6 Evidential interpretation and forensic validation

A similarity score measures proximity in a model's representation space. Its evidential meaning depends on the distribution of scores from both same-speaker and different-speaker comparisons drawn from a relevant population. Forensic guidance therefore emphasizes domain-matched validation, explicit propositions, known same-speaker and different-speaker data, and calibrated likelihood ratios [16–18].

This study remains an archival retrieval analysis. Descriptive AUC answers a ranking question: how often does a cohort pair outrank an external pair? Same-video tests answer a conditional identification question: how often does the nominated voice beat local alternatives? These quantities provide strong recurrence evidence while remaining distinct from a calibrated forensic likelihood ratio. Chapter 7 concentrates the resulting inferential boundary in one place.

Human-assisted recognition also requires controlled design. The NIST 2010 pilot selected difficult trials and found limited evidence that human assistance improved the systems in that small evaluation [19]. That result supports blinding, multiple raters, balanced trial types, and predeclared decision rules. Appendix E specifies a prospective listening protocol aligned with those principles.

2.7 Ethics, privacy, and proportionality

Voice embeddings can reveal identity-related and demographic information, creating privacy and governance concerns even when the source audio is public [20–22]. A proportionate archival study minimizes redistribution of long recordings, reports aggregate findings, separates acoustic recurrence from civil identity, and preserves an auditable chain from results to source videos. The evidence package follows that structure: derived clips, manifests, score tables, and checksums support replication, while the main thesis focuses on the scientific argument.

2.8 Position of the present study

Research line	Typical question	Connection to this thesis
Query-by-example graphs [1]	Which samples connect to a query through speaker similarity?	Motivates multiple exemplars and distributional cohesion rather than a perfect clique.
Speaker retrieval in the wild [2]	How can diarization and embeddings search uncontrolled archives?	Provides the closest end-to-end methodological analogue.
Open-set recognition [3]	How does search size affect false alarms?	Motivates strict multi-model ranking and same-video alternatives.
Forensic validation [12,16]	How should similarity be calibrated for evidential claims?	Defines the boundary between retrieval evidence and identity evidence.
Present study	Does one scam-call voice recur across a public video archive?	Combines a frozen cohort, three models, same-video controls, archive extension, and a known-voice benchmark.

3. Development of the 91-video recurrence cohort

The cohort originated from an acoustic observation rather than a preexisting identity label. A male scam-call voice appeared similar across several videos. The research process converted that observation into a reproducible archive target through successive stages: long-recording search, video-local speaker extraction, independent local comparison, stronger speaker reconstruction, and video-level freezing.

This chapter presents that chain in reader-facing terms. Detailed counts, physical file origins, and artifact locations appear in Appendices A and D. The central principle is simple: every cohort entry represents one nominated local speaker from one distinct source video.

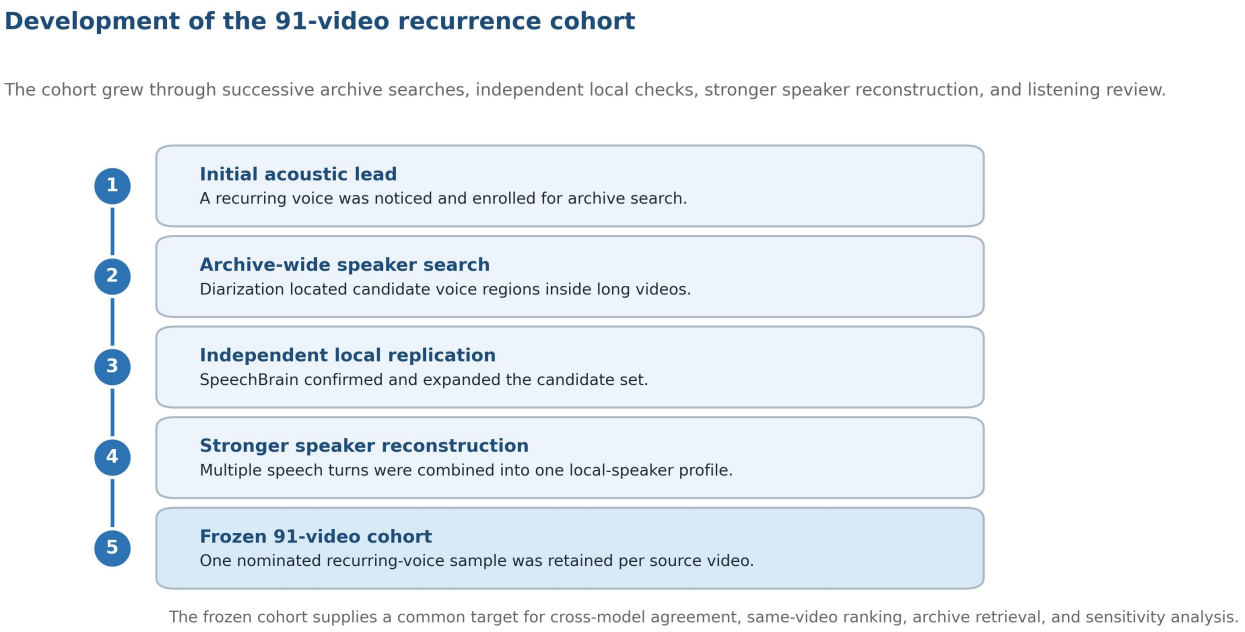

Figure 2. Development of the frozen 91-video recurrence cohort.

3.1 The source archive

The retained collection contains 296 long-form FLAC recordings. Historical cloud processing succeeded for 244 source videos, and those 244 define the denominator for the principal archive-coverage results. Video identifiers embedded in filenames provide stable source keys across manifests, derived clips, and analysis outputs.

Each source recording can contain several calls, multiple scammer voices, Kitboga character voices, musical interludes, sound effects, and overlapping speech. The study therefore treats the source video as a container and the diarized local speaker as the primary acoustic unit. A local speaker label indicates a grouping inside one video; cross-video identity emerges only through embedding comparison.

3.2 Initial archive search

The initial recurring-voice sample was enrolled in a pyannote identification workflow. Long recordings were diarized, and each local speaker was scored against the enrolled voice. This stage served two functions: it located candidate speakers inside hours of material, and it preserved timestamps for the speech turns assigned to each local label. The historical search nominated 41 rows at the working confidence threshold and resolved 40 physical stitched clips.

The historical confidence score was a discovery signal. It identified where to look and which local speaker label to export. Independent local systems later supplied the principal comparative evidence. This division between discovery and evaluation prevents one cloud score from carrying the final conclusion.

3.3 Building the all-speaker universe

The timestamp-bearing diarization results were reused to export one representative profile for every local speaker label in the 244 processed videos. That pass produced 2,051 video-local speaker representatives. A later completeness pass indexed 117,663 diarization turns and retained 9,474 additional turns of at least five seconds. Combined with the representative set, the canonical local corpus contains 11,525 FLAC clips organized by video and local speaker.

These retained segments preserve the evidence beneath each speaker profile. A profile is an average of clip embeddings; every contributing clip remains listed in the metadata. This design permits analysis at three levels: clip against clip, local speaker against local speaker, and video-level retrieval based on the best local speaker.

3.4 Independent local expansion

SpeechBrain provided the first local verification and expansion stage. Candidate speaker clips were compared with a reference voice, then with a growing group of strong matches. Pairwise analysis revealed a coherent center alongside weaker peripheral samples. The center supplied a working speaker profile, and that profile was compared with every exported local speaker.

This expansion changed the scientific question from “does one clip resemble another?” to “which local speakers consistently resemble a multi-video speaker profile?” Candidates were tracked at the video level so that multiple local labels from one source could be inspected without inflating the count of covered videos.

3.5 Stronger speaker reconstruction

Short discovery snippets can undersample a speaker and amplify noise, overlap, or atypical phonetic content. For promising videos, the nominated local label was therefore rebuilt from additional diarized turns. Multiple clean speech regions were combined into a longer local-speaker sample and re-evaluated against the defining voice, the central profile, and the strongest cohort examples.

The reconstructed files represent the same nominated video-local speakers with more speech evidence. Forty-six such entries entered the final cohort alongside 45 earlier representatives from distinct videos. The resulting 91 entries therefore span 91 unique source videos.

3.6 Freezing the cohort

Cohort membership was fixed before the final three-model analyses. Each row records the source video, local speaker label, selected audio path, origin class, speech duration, and supporting scores available at the time of selection. A listening pass judged the retained samples compatible with the recurring voice hypothesis. The listening judgments informed cohort construction and remain distinct from the later model-based tests.

Freezing creates a common denominator. SpeechBrain, WeSpeaker, and Resemblyzer evaluate the same 91 videos, the same nominated local speakers, and the same external archive. Pair counts and rank outcomes therefore align exactly across models.

3.7 Evidence units

Unit	Definition	Role in inference
Speech clip	One retained diarized speech segment or stitched sample.	Feeds model embeddings and supports listening audit.
Local speaker profile	Average of all retained clip embeddings for one label inside one video.	Primary acoustic comparison unit.
Source video	One processed public upload identified by video ID.	Primary counting and resampling unit.
Cohort	Ninety-one nominated local speakers from 91 distinct videos.	Frozen recurrence target.
Strict evidence set	The 91 cohort videos plus 37 strict archive extensions.	Principal archive-coverage result.

4. Analytical methods

4.1 Overall design

The final analysis uses a repeated-measures retrieval design. Each model embeds the same clip inventory. Clip embeddings are aggregated into video-local speaker profiles. The 91 nominated profiles define the recurrence cohort. Pairwise similarity describes internal cohesion, external archive pairs provide a background distribution, and leave-one-video-out profiles rank the nominated speaker against every alternative label in the same source video.

Raw score scales remain model-specific. SpeechBrain and WeSpeaker produce cosine ranges centered differently from Resemblyzer. The analysis therefore compares ranks, percentiles, AUCs, and standardized effects within each model. Cross-model consensus is computed from those normalized positions.

4.2 Speaker-comparison models

Model	Representation	Role
SpeechBrain	ECAPA-TDNN speaker embedding trained on VoxCeleb data.	Primary local verification and final comparison model.
WeSpeaker	ResNet34-LM ONNX speaker embedding.	Modern independent implementation and replication model.
Resemblyzer	Generalized end-to-end d-vector style voice encoder.	Older architectural cross-check with a different score geometry.

For each model m, normalized clip embeddings e were compared with cosine similarity:

$s_m(i,j) = e_{m,i} \cdot e_{m,j}$.

A higher score places two clips closer in that model's speaker space. The analysis interprets this value relative to model-specific background scores and video-local ranks.

4.3 Speaker-profile aggregation

For local speaker k with retained clip embeddings $e_1 \ldots e_n$, the reference profile is the normalized arithmetic mean:

$p_k = \text{normalize}[(1/n) \Sigma e_i]$.

Equal clip weighting is the frozen primary rule. Three sensitivity rules modify the contribution of clip i through its duration d_i: linear duration, square-root duration, and duration capped at 15 seconds. The same retained clip inventory and model embeddings feed every rule. Similar outcomes across rules

indicate that the result is distributed across the speaker's available speech rather than carried by a particular weighting choice.

4.4 Exact 91-by-91 comparison

The exact selected sample for each cohort video was embedded by every model. The 91 entries yield 91 × 90 / 2 = 4,095 unique cross-video pairs per model. Pair means, medians, quantiles, leave-one-out cohesion, nearest-neighbor overlap, and model-to-model rank correlations summarize this frozen set.

Pair rows share clips and therefore form a dependent network. Their descriptive distributions show cohesion; inferential claims rely on video-level resampling and matched alternatives.

4.5 Cohort versus the wider archive

Multi-segment speaker profiles were compared across the 2,051 video-local speaker universe. Cohort–cohort pairs were contrasted with external cross-video pairs. Descriptive AUC estimates the probability that a randomly selected cohort pair receives a higher model score than a randomly selected external pair. The fraction of cohort pairs above the external 95th percentile measures upper-tail enrichment.

This comparison asks whether the cohort occupies an unusually coherent region of each model's speaker space. It complements the same-video test by providing a broad archive background.

4.6 Same-video ranking

Each cohort video was evaluated in leave-one-video-out form. A recurrence profile was made from the other 90 nominated speakers. Every local speaker in the held-out video was scored against that profile. The nominated label's rank, score margin, global percentile, and pairwise wins against local alternatives were recorded. Across the 91 videos, this produced 720 nominated-versus-alternative comparisons per model.

A second standardized analysis transformed target and control scores within each model, formed a three-model mean, and resampled by video. It reports the standardized target-control difference, a video-blocked bootstrap confidence interval, a sign-flip test, and the number of nominated speakers above the local control 95th percentile.

The highest-scoring same-video alternatives were retained in a dedicated hard-alternative review queue rather than automatically treated as known-different speakers. This preserves cases in which one human may have been divided between multiple diarization labels.

4.7 Archive-wide extension

Every local speaker outside the cohort was scored against the recurrence profile. Scores were converted to within-model percentile ranks. The principal strict rule retained a video when its best local speaker reached the top decile in all three models. This rule added 37 videos. Two broader rules retained all-three-model top-5% and top-1% style rank bands defined in the historical analysis; in the reader-facing results they serve as progressively broader listening pools rather than calibrated prevalence estimates.

Video-level deduplication ensures that several high-scoring labels inside one upload count once. The selected local speaker and runner-up remain available for split-aware listening review.

4.8 Cross-model agreement

Spearman correlation compares the ordering of the 4,095 exact cohort pairs between models. Nearest-neighbor overlap compares whether each target retrieves the same top neighbors. Agreement is interpreted as shared measurement structure. Model-specific evidence remains visible in every table so that a strong result from one system cannot conceal disagreement from another.

4.9 Known-voice benchmark

A separate profile was developed for Kitboga's recurring Grandma/Granny/Edna character using nine fixed seed videos. Fifteen held-out title-supported videos supplied known archive labels. The benchmark measures whether each model ranks the expected Grandma label first or within the top two.

Twelve cohort videos contain both a nominated scam-call label and a separate Grandma-cluster label. These episodes create a bidirectional switch test. The recurrence profile should select the nominated scam-call label, while the Grandma profile should select the Grandma label. The same audio container, editing chain, and episode context therefore remain fixed as the intended recurring voice changes.

4.10 Temporal robustness

A recurrent speaker observed across a multi-year archive should remain acoustically elevated despite aging, changes in performance, telephone transmission, codecs, and production practices. Temporal robustness was therefore evaluated across all 4,095 unique cross-video cohort pairs by relating absolute publication-year separation to each model's similarity percentile and to the three-model consensus percentile. Because the pair observations share videos and are not statistically independent, significance was estimated through 5,000 node-label permutations that preserved the pair network while randomly reassigning publication years.

4.11 Long-profile null-cohort robustness

A secondary robustness analysis used the 46 videos for which longer multi-turn speaker profiles had been reconstructed. The observed cohort was compared with 20,000 null cohorts formed by randomly selecting one available local speaker from each of the same 46 videos. This preserved the exact source-video set and its recording conditions while changing only which local speaker represented each video.

4.12 Statistical interpretation

The primary reported quantities are descriptive AUC, model-specific rank, target-versus-alternative win rate, standardized target-control separation, video-blocked bootstrap intervals, sign-flip p-values, and cross-model Spearman correlation. The source video is the resampling and coverage unit. Pairwise

matrices remain available for network and clustering analyses, while video-level summaries carry the main inferential argument.

Statistical significance evaluates the declared comparison design. It measures how unusual the observed target-control ordering is under exchangeability or sign reversal. The same-speaker interpretation then draws on the convergence of pairwise cohesion, archive separation, same-video ranking, model replication, archive extension, and the known-voice benchmark.

5. Results

5.1 Exact cohesion within the 91-video cohort

The exact selected clips produced 4,095 cross-video comparisons per model. Mean similarities were 0.428 for SpeechBrain, 0.354 for WeSpeaker, and 0.814 for Resemblyzer. Their different raw ranges reflect model geometry. Within every model, the cohort forms a broad elevated distribution rather than a set of uniformly extreme pairings.

Model	Pairs	Mean	Median	5th percentile	95th percentile
SpeechBrain	4,095	0.428	0.425	0.264	0.596
WeSpeaker	4,095	0.354	0.351	0.153	0.565
Resemblyzer	4,095	0.814	0.818	0.719	0.893

The lower tail matters conceptually. Telephone degradation, year gap, emotion, pitch manipulation, and diarization leakage can weaken individual links. A recurring speaker across an uncontrolled archive is therefore expected to resemble a connected distribution more than a strict all-pairs clique. The subsequent controls determine whether that distribution is unusual.

5.2 Separation from the wider archive

Reconstructing each local speaker from all retained turns increased the amount of speech supporting each comparison. Under the equal-clip profile rule, cohort pair means were 0.441 for SpeechBrain, 0.369 for WeSpeaker, and 0.826 for Resemblyzer. External cross-video means were 0.138, 0.114, and 0.678.

Descriptive AUCs were 0.942, 0.905, and 0.927. In plain language, a randomly selected cohort pair outranked a randomly selected external cross-video pair about 91%–94% of the time, depending on the model. Cohort upper-tail enrichment was also substantial: 57.9% of SpeechBrain pairs, 31.6% of WeSpeaker pairs, and 56.3% of Resemblyzer pairs exceeded the respective external 95th-percentile threshold.

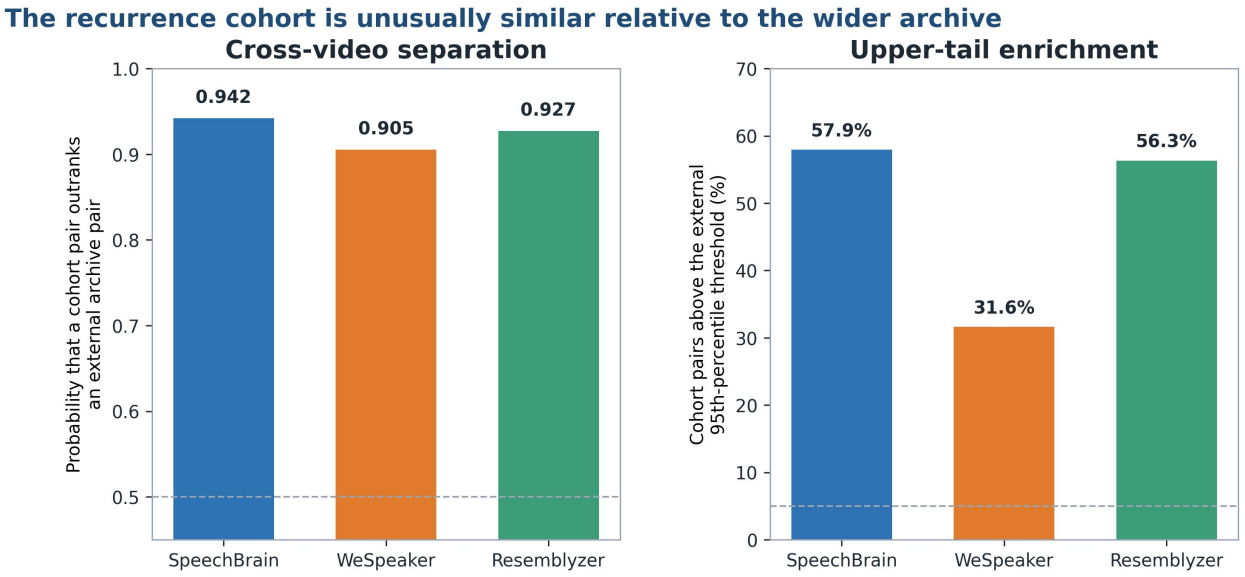

Figure 3. Cohort–archive separation under equal-clip speaker profiles. AUC and upper-tail enrichment are computed separately on each model's score scale.

5.3 Cross-model agreement

The models agreed positively on the ordering of all 4,095 exact cohort pairs. SpeechBrain and WeSpeaker showed the strongest rank correlation, ρ = 0.766. SpeechBrain and Resemblyzer reached ρ = 0.521, while WeSpeaker and Resemblyzer reached ρ = 0.465. The pattern is consistent with two modern systems sharing more geometry and Resemblyzer supplying a looser architectural cross-check.

Top-neighbor overlap followed the same ordering. The models therefore identify related structure without producing identical rankings. This is the desired form of redundancy: a common recurrence signal remains visible while model-specific uncertainty is retained.

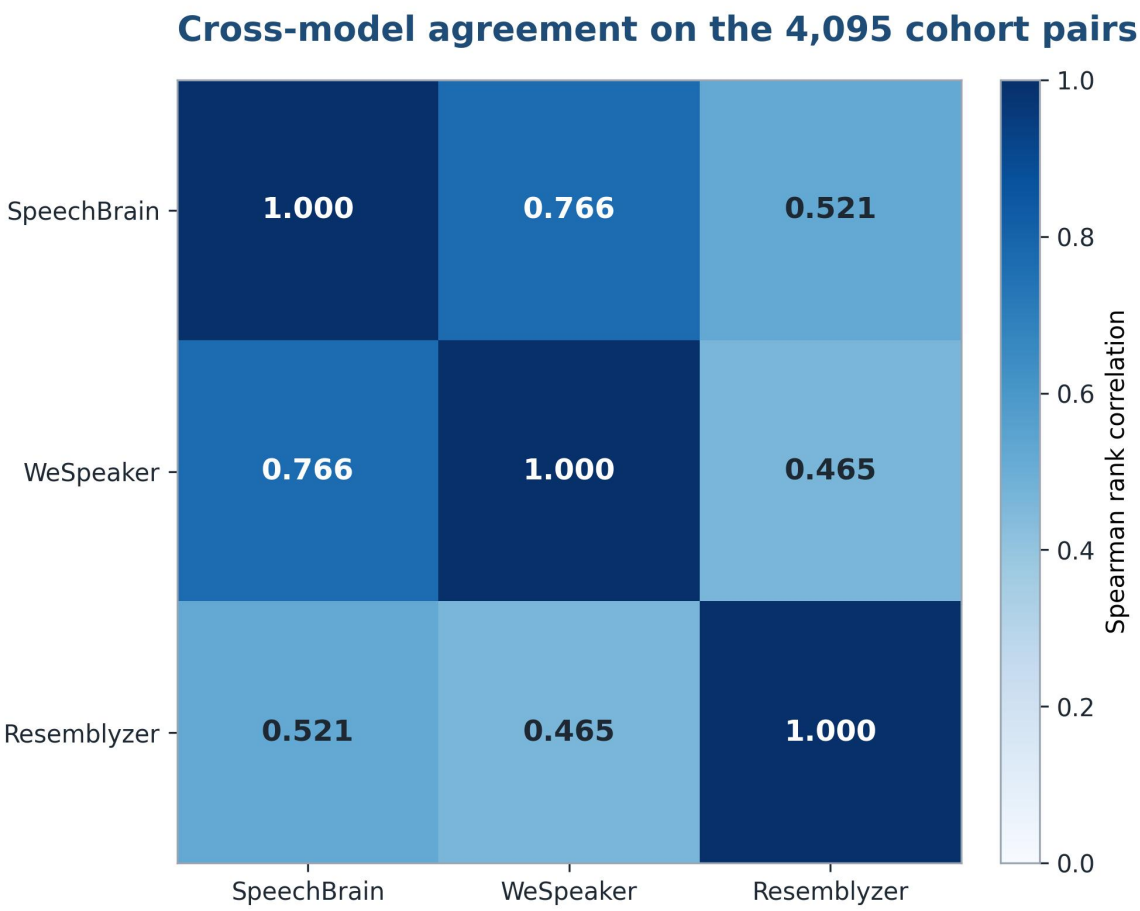

Figure 4. Spearman rank agreement across the 4,095 exact cohort pairs.

The three systems also shared a substantial archive-wide similarity structure. Across the complete 2,051-speaker profile universe, standardized three-model reliability was 0.805, Kendall's coefficient of concordance was 0.720, and the first principal component explained 72.2% of the variance in standardized model scores. Within the 46-video long-profile robustness subset, Kendall's concordance was 0.754 and the first component explained 69.3%. These statistics show that the models recover a strong common ordering of acoustic relationships without treating their raw score scales as interchangeable.

5.4 Same-video discrimination

Same-video ranking provides the clearest response to the idea that the models simply prefer telephone voices or scammer speech. Each nominated speaker competed against the other local speakers in the identical source upload. SpeechBrain ranked the nominated voice first in 73 of 91 videos and within the top two in 89. WeSpeaker produced 67 first-place and 82 top-two results. Resemblyzer produced 60 first-place and 78 top-two results. At least two models ranked the nominated voice first in 66 videos and within the top two in 85.

Across all 720 local alternatives, the nominated voice won 97.1% of SpeechBrain comparisons, 95.1% of WeSpeaker comparisons, and 92.5% of Resemblyzer comparisons. The standardized three-model target-

control difference was 1.896 standard deviations. The video-blocked 95% confidence interval was 1.779–2.012, and the empirical sign-flip p-value was below 0.0001. Seventy-five nominated voices exceeded their local three-model control 95th percentile.

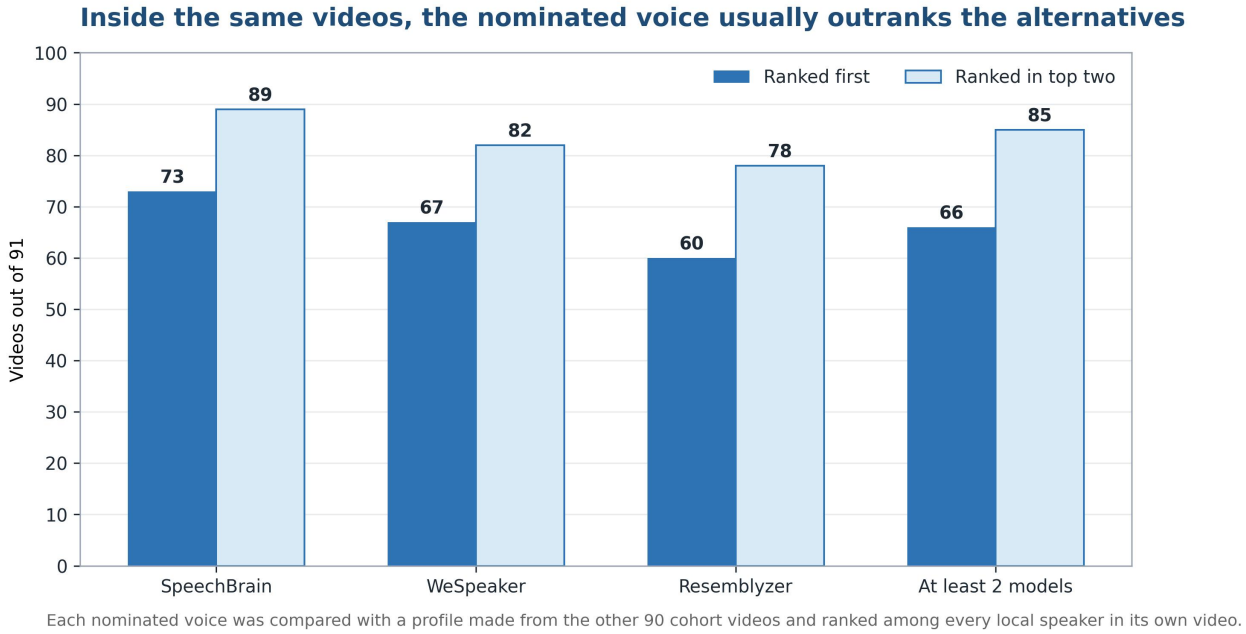

Figure 5. Rank of the nominated recurring voice among all local speakers in each source video.

Interpretation

The recurring-voice profile usually selects the nominated local speaker even when every competing voice comes from the same video. This result ties the cross-video pattern to a specific local voice rather than to the upload as a whole.

The longer-profile robustness analysis reached the same conclusion under a different statistical design. The nominated 46-video cohort achieved a mean three-model pair percentile of 0.9197, compared with 0.5456 among random same-video cohorts. The observed result was 14.79 null-distribution standard deviations above the random mean ($p < 0.0001$). Its high-similarity edge rate was 36.0%, compared with 6.0% under the null. At least two models ranked the nominated speaker first in 36 of 46 videos and within the top two in 44 of 46. The finding therefore survives the use of longer speaker profiles and a video-matched random-selection test.

5.5 Archive extension

The 91 cohort videos account for 37.3% of the processed archive. Applying the strict three-model rule to the remaining videos added 37 outside videos, producing a strict evidence set of 128 of 244 videos, or 52.5%. The broader ranking bands contained 152 videos (62.3%) and 173 videos (70.9%).

The 52.5% figure is the principal archive-coverage result because every strict addition achieved the declared high rank in all three models. The broader sets are valuable for listening review, diarization-split

inspection, and prospective validation. Their nested structure shows how coverage changes as the retrieval rule expands.

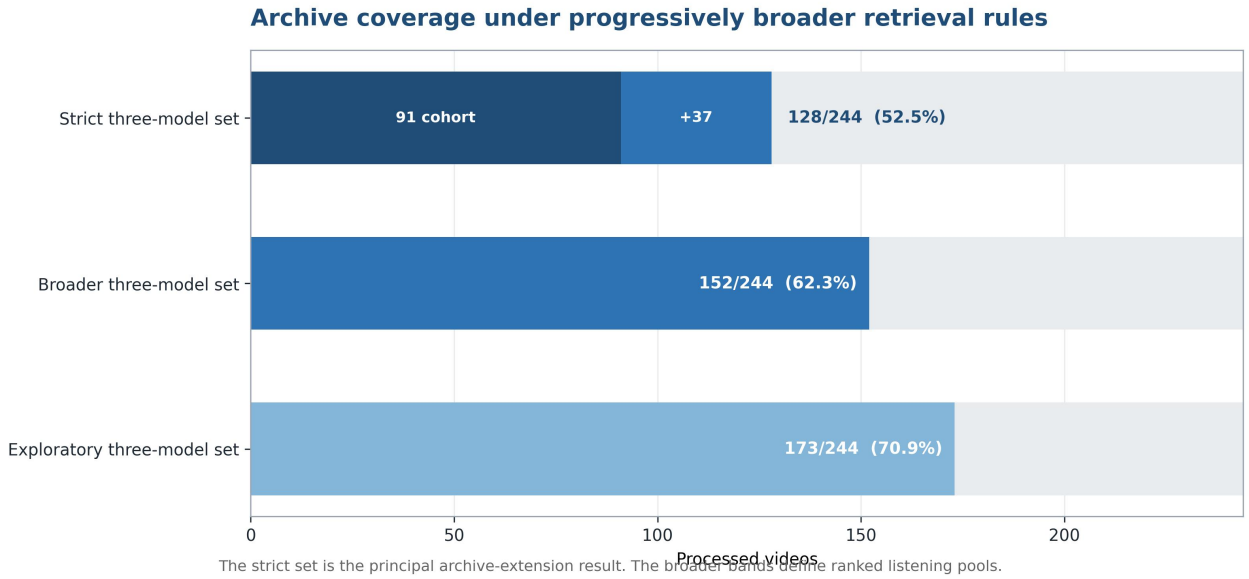

Figure 6. Video-level archive coverage. The strict set combines 91 frozen cohort videos and 37 strict archive extensions.

Agreement remained substantial during archive-wide retrieval. Among the best local speakers selected from outside-cohort videos, the Spearman correlation was 0.835 between SpeechBrain and WeSpeaker, 0.636 between SpeechBrain and Resemblyzer, and 0.596 between WeSpeaker and Resemblyzer; all three associations had p < 5 x 10^-16. The outside-video ranking was therefore shared across the model families rather than being carried by one unusually permissive system.

5.6 Sensitivity to profile aggregation

Equal clip weighting, linear duration weighting, square-root duration weighting, and duration capped at 15 seconds produced mean three-model cohort-versus-background AUCs from 0.925 to 0.928. At least two models ranked the nominated voice first in 66–67 of 91 videos and within the top two in 85 of 91 under every rule.

Pair order also remained stable. Mean model-specific Spearman correlation with the equal-weight result ranged from 0.983 to 0.995 for the alternative rules. The recurrence finding therefore persists when longer clips receive more weight and when their influence is capped.

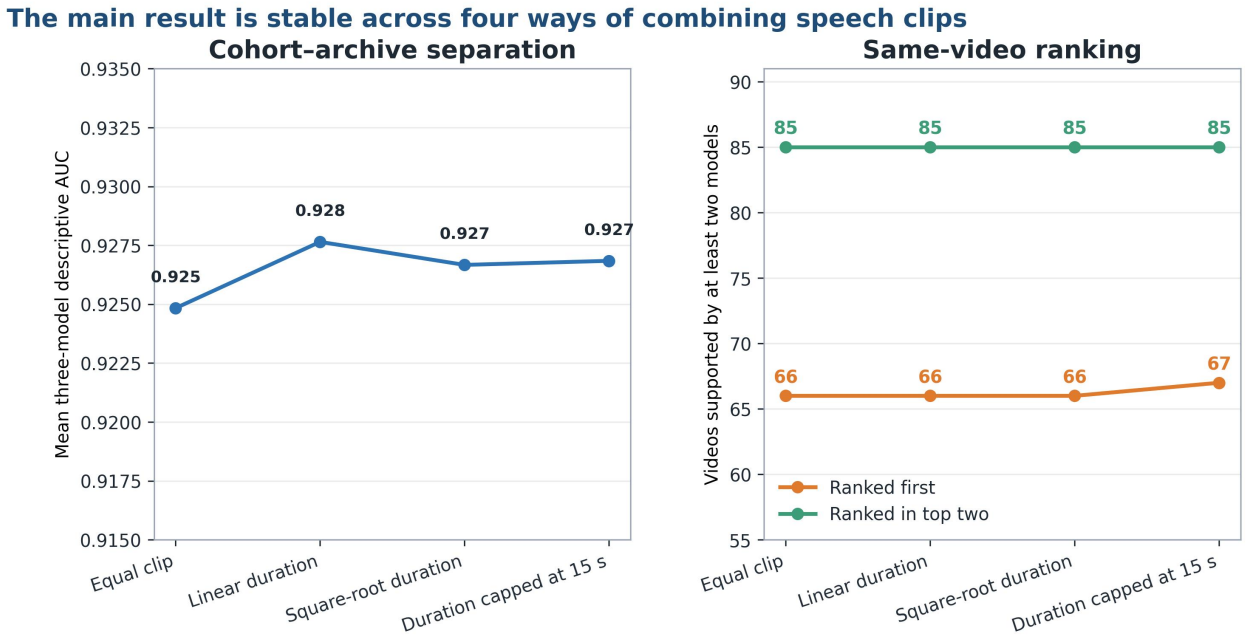

Figure 7. Sensitivity of cohort separation and same-video ranking to four profile-aggregation rules.

5.7 Known-voice Grandma benchmark

The Grandma benchmark supplies an archive-native positive control. In 15 held-out title-supported videos, SpeechBrain ranked the expected Grandma label first in 14 and within the top two in all 15. WeSpeaker produced the same 14/15 first-place and 15/15 top-two result. Resemblyzer ranked the expected label first in all 15.

The bidirectional same-episode switch was perfect across the 12 shared episodes. With the recurrence profile, every model preferred the nominated scam-call voice. With the Grandma profile, every model preferred the separate Grandma label. The result held under all four aggregation rules. For each model and rule, 12 correct choices out of 12 correspond to a one-sided sign-test probability of 0.000244.

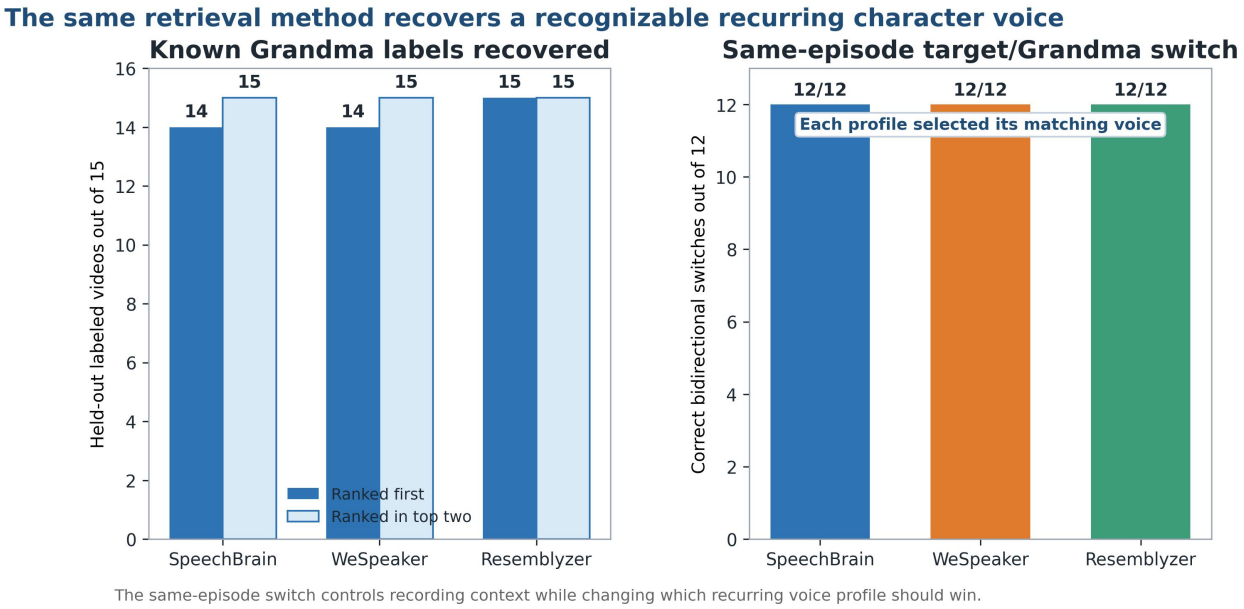

Figure 8. Known-voice retrieval and the bidirectional target/Grandma same-episode switch.

The fixed nine-video Grandma profile was evaluated across 235 non-seed videos. Video-level retrieval AUC was 0.860 for SpeechBrain, 0.879 for WeSpeaker, and 0.817 for Resemblyzer, with a three-model consensus AUC of 0.850. SpeechBrain and WeSpeaker ranked the exact Grandma label first in 14 of 15 held-out title-supported episodes, while Resemblyzer ranked it first in all 15; every model placed all 15 labels within its top two. The benchmark therefore supplies both positive retrieval evidence and same-episode discriminability.

5.8 Temporal persistence

Similarity declined moderately as publication years became more widely separated, but remained strongly elevated throughout the archive. The three-model consensus correlation between publication-year gap and similarity was -0.318 under the network-preserving permutation test ($p = 0.0002$). The model-specific correlations were -0.310 for SpeechBrain, -0.302 for WeSpeaker, and -0.255 for Resemblyzer, each with $p = 0.0002$.

Same-year cohort pairs had a mean consensus percentile of 0.933 across 670 comparisons. Even the 16 pairs separated by nine years retained a mean percentile of 0.868. The result is consistent with gradual acoustic drift rather than the disappearance of the recurrent pattern: larger temporal gaps make the comparison harder, yet the cross-video relationship remains unusually strong.

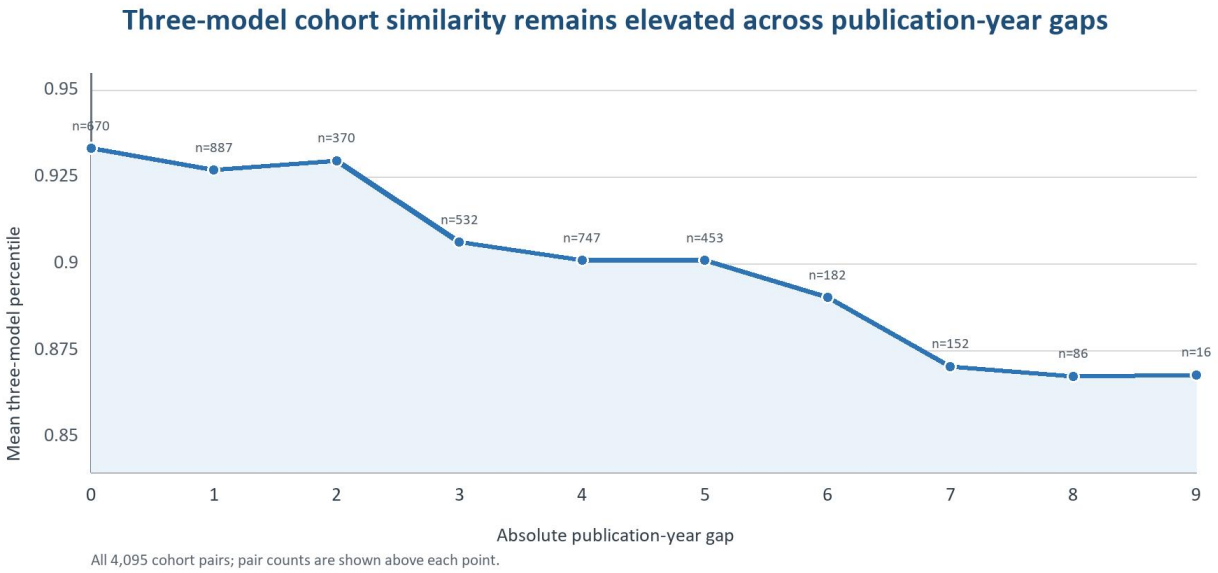

Figure 9. Mean three-model similarity percentile by publication-year gap.

5.9 Result synthesis

Evidence layer	Observed result	Scientific contribution
Exact cohort pairs	4,095 pairs per model with positive cross-model rank agreement.	Shows measurable cross-video cohesion.
Cohort versus archive	AUC 0.942 / 0.905 / 0.927.	Shows that the cohort occupies an unusually coherent region.
Same-video alternatives	66/91 first and 85/91 top two by at least two models; $p < 0.0001$ standardized test.	Identifies the nominated voice under matched episode conditions.
Strict archive extension	37 additions; 128/244 videos total.	Shows recurrence beyond the discovery cohort.
Duration sensitivity	Mean three-model AUC 0.925–0.928; 85/91 top two throughout.	Shows stability across profile construction rules.
Grandma benchmark	All 15 known labels in top two; 12/12 bidirectional switches.	Shows that the method retrieves and separates a second recurring voice.

6. Discussion

6.1 Answer to the primary question

The evidence supports the hypothesis that the same scam-call speaker recurs across a substantial share of the processed Kitboga archive. The conclusion rests on a pattern of convergence. The frozen 91-video cohort is internally coherent, its pair scores outrank the wider archive, the nominated voice usually wins inside its own source video, three models reproduce the pattern, strict retrieval adds 37 outside videos, aggregation choices leave the result stable, and a separate Grandma profile behaves correctly under the same machinery.

The same-video result is especially important. A high target–target score can reflect shared channel, accent, emotion, or selection. A nominated voice that repeatedly outranks other speakers in the same upload has survived a more relevant comparison. The 720 alternatives share episode context yet rarely displace the nominated voice.

6.2 From acoustic recurrence to a same-speaker explanation

Several production histories can generate recurrence. One person may appear in many independent calls. A small group may contain acoustically similar voices. A compilation may reuse source audio. Editing may repeat segments. The present pattern favors the first explanation because it extends across years, many titles, many local speaker labels, and multiple recording conditions. The result also survives replacement of short discovery snippets with longer multi-turn profiles.

The strongest defensible wording is therefore comparative: the balance of archive evidence favors a recurring-speaker explanation over generic telephone or scammer similarity. Civil identity and production intent require evidence outside speaker embeddings, such as authenticated call records, original media provenance, or known-speaker enrollment.

6.3 Competing explanations

The same-speaker interpretation provides the simplest unified account of the cross-video cohesion, same-video discrimination, temporal persistence, model convergence, and archive-wide retrieval results. Several alternative explanations nevertheless make distinct predictions that can be compared with the observed evidence.

Table 8. Competing explanations and their predictions.

Explanation	Prediction	Relationship to the evidence
One recurring human speaker	Similarity persists across uploads, years, and models.	Consistent with all principal results.

Explanation	Prediction	Relationship to the evidence
Small homogeneous speaker group	Several stable acoustic subgroups coexist.	Not excluded; the broad score distribution permits this possibility.
Generic accent or speaking style	Many unrelated archive speakers score similarly.	Weakened by same-video alternatives and the archive background.
Shared telephone or production channel	Multiple speakers in one episode rise together.	Weakened because the nominated voice usually outranks co-occurring voices.
Diarization fragmentation	Target speech is divided across neighboring labels.	Supported in a substantial minority and preserved by split-aware outputs.
Repeated excerpts or compilations	Near-identical source material recurs across uploads.	Plausible for individual uploads and testable through fingerprinting.
Host or character leakage	A known recurring channel voice contaminates the target group.	Weakened by the separate Grandma profile and 12/12 bidirectional switch.

A small acoustically homogeneous group remains the strongest alternative to a single recurring human. Generic accent, shared channel, and host-character leakage explain the combined same-video, temporal, and profile-switch results less economically. The evidence therefore favors recurrence of one speaker while preserving a narrower alternative involving more than one unusually similar speaker.

6.4 Why the cohort behaves like a distribution rather than a clique

Strict clustering requires every member to exceed an edge threshold with every other member. That rule is poorly matched to multi-year telephone audio. A weak early recording and a heavily compressed later recording may both connect strongly to central samples while sharing only moderate direct similarity with one another. Complete-link clustering therefore fragments the cohort even when retrieval and same-video ranking remain strong.

The distributional approach is more informative. It measures how the entire cohort shifts relative to external voices and whether each held-out video retrieves its nominated speaker. The cohort can contain heterogeneous acoustic conditions while still expressing a stable central identity signal. This is precisely the query-by-example graph intuition described by Campbell and Singer [1].

6.5 Cross-model convergence

SpeechBrain and WeSpeaker show the strongest pair-order agreement, as expected from two contemporary speaker-embedding systems. Resemblyzer produces a compressed high-score range and lower neighbor agreement, yet it independently reproduces archive separation and same-video dominance.

Agreement across these systems is strongest when expressed through ranks and AUC rather than raw cosine values.

The three-model design also exposes uncertain cases. A candidate supported by all three systems is less sensitive to one model's geometry. A candidate supported by two systems and ranked lower by the third becomes a natural listening-review item. The final reports retain each score separately so that consensus remains auditable.

6.6 Interpreting the 52.5% archive estimate

The strict evidence set contains 128 processed uploads. Ninety-one entered through the frozen cohort, and 37 entered through the all-three-model strict archive rule. This establishes that the recurrent acoustic signature is widespread within the processed collection.

The source video is the counting unit. A video can contain one call, several calls, or edited excerpts. The 52.5% result therefore describes upload coverage. The number of independent calls and the number of unique recording sessions may be lower or higher depending on the underlying production history. A future provenance audit can group shared audio, publication chronology, and call metadata.

6.7 What the Grandma benchmark contributes

The Grandma result demonstrates more than general pipeline activity. It shows that the same embeddings and aggregation rules can retrieve a recognizable recurring character voice and distinguish it from the nominated scam-call voice within the same episodes. The perfect 12/12 switch means that the system responds to which recurring profile is supplied, rather than merely preferring one telephone channel, one upload, or one diarization label.

The benchmark also clarifies the role of archive-native validation. A real identity-labeled control corpus would offer stronger calibration, yet the Kitboga archive itself contains a known recurring voice with relevant editing and channel characteristics. That makes Grandma a useful positive control and a strong design check.

6.8 Scientific and practical implications

Scientifically, the study demonstrates how an unlabeled recurring-voice hypothesis can be converted into an auditable archive-retrieval design. Diarization locates local evidence; speaker profiles stabilize short segments; model-specific percentiles control incompatible scales; same-video alternatives supply relevant comparison voices; and a second recurring character profile tests discriminability.

Practically, the resulting evidence package supports several forms of review. Researchers can reproduce the pairwise matrices, inspect the local speaker selected in each video, listen to strict archive extensions, compare runner-up labels where diarization may have split one voice, and run a blinded rating study from the frozen manifests. The method is reusable for other public media archives containing recurrent unlabeled speakers.

7. Limitations and responsible interpretation

7.1 Cohort selection

The 91-video cohort developed through exploratory retrieval and listening review. Its internal cohesion therefore reflects both the underlying archive and the selection process. The strongest confirmatory leverage comes from analyses performed after the freeze: three-model replication, same-video alternatives, strict outside-video extensions, aggregation sensitivity, and the Grandma benchmark. A prospectively frozen second archive would provide the next level of confirmation.

7.2 Diarization uncertainty

Local speaker labels are machine-generated clusters. Fragmentation can place one person in two labels, particularly during overlap, emotion, pitch shifting, or abrupt channel changes. Merging can place speech from more than one person in a single profile. The split-aware outputs preserve the top local alternatives and their margins so that ambiguous videos remain reviewable.

Forty particularly strong same-video alternatives were preserved for targeted review. A high-scoring alternative can represent either a genuinely similar non-target voice or the recurrent speaker divided between multiple diarization labels. Treating every alternative label as a confirmed different person would overstate control performance. The retained queue preserves both labels, their speech totals, clip counts, model scores, and source audio so that future listening or human-assisted label merging can resolve the ambiguity.

7.3 Channel and relevant-population coverage

Same-video alternatives provide strong episode-level matching, while detailed handset, codec, call route, language variety, age, and accent annotations remain unavailable. The archive background is broad rather than a curated relevant population of acoustically similar Indian-accented English male callers. A future benchmark should sample hard non-target voices with matched accent, sex, approximate age, telephone channel, speech duration, and emotional style.

7.4 Model dependence and calibration

All three systems use neural speaker embeddings and public speech corpora, producing partial methodological dependence. Their agreement is treated as convergent measurement. Raw cosine values remain descriptive. A forensic identity conclusion would require domain-matched known same-speaker and known different-speaker trials, calibrated likelihood ratios, and an explicit relevant population.

7.5 Human listening

Existing listening review assisted cohort construction and candidate interpretation. Prospective human validation should use randomized identifiers, balanced same- and different-speaker trials, multiple raters, and predeclared exclusion and decision rules. Appendix E provides a concrete protocol.

7.6 Unit of inference

Archive coverage is measured in processed uploads. A source upload may contain several calls or repeated material, and several uploads may draw from one call. The present thesis therefore establishes recurrence across videos. Call-level prevalence and production chronology require audio-duplicate analysis and source provenance.

7.7 Identity and intent

The acoustic evidence favors the same-speaker explanation while remaining compatible with a small acoustically homogeneous group. Authenticated identity requires a trusted enrollment source or external records. Production intent, contractual relationships, and editorial decisions require documentary evidence. These questions sit beyond acoustic similarity and deserve separate investigation.

7.8 Ethics and publication language

Responsible publication centers aggregate evidence, minimizes personal identification, and provides direct access to auditable clips and score tables. The clearest public statement is: “Three independent local speaker-comparison implementations found a recurrent acoustic signature across more than half of the processed videos, with strong separation from other voices in the same episodes. The pattern is consistent with repeated use of the same speaker.”

This wording communicates the strength of the archive result and preserves the distinction between an acoustic inference and a civil identity claim.

8. Conclusion

This thesis began with a simple listening observation and converted it into an archive-scale retrieval study. A 91-video recurrence cohort was frozen from a broad set of diarized local speakers. SpeechBrain, WeSpeaker, and Resemblyzer evaluated the same cohort, the same 720 same-video alternatives, and the same wider archive. A 46-video long-profile null-cohort test and a network-preserving temporal analysis supplied independent robustness checks using different units of analysis.

Every major evidence layer points in the same direction. The cohort forms an elevated cross-video similarity distribution. A cohort pair outranks an external archive pair with descriptive probabilities of 0.942, 0.905, and 0.927 across the three models. The nominated voice usually ranks first or second inside its own source video. The standardized same-video target-control separation reaches 1.896 standard deviations with $p < 0.0001$. Thirty-seven strict outside-video additions expand the evidence set to 128 of 244 processed uploads, or 52.5%. Four profile-aggregation rules preserve the result. The known Grandma voice is recovered cleanly, and the target-versus-Grandma profile switch succeeds in all 12 shared episodes.

The balance of evidence favors recurrence of the same scam-call speaker across a substantial portion of the Kitboga archive. The conclusion is stronger than clip resemblance because it combines video-level controls, model replication, archive extension, long-profile null cohorts, temporal persistence, and a second recurring-voice benchmark. The next scientific step is prospective validation with blinded listeners and a relevant population of closely matched non-target speakers, followed by provenance work that distinguishes independent calls from repeated source material.

Thesis conclusion

A recurrent scam-call acoustic signature appears in 128 of 244 processed videos under the strict three-model rule. The full evidence pattern is most consistent with repeated presence of the same speaker across the archive.

References

[1] Campbell, W. M., & Singer, E. (2012). Query-by-example using speaker content graphs. Interspeech 2012.

[2] Loweimi, E., Qian, M., Knill, K., & Gales, M. (2025). Speaker Retrieval in the Wild: Challenges, Effectiveness and Robustness. arXiv:2504.18950.

[3] Peri, R., Sadjadi, S. O., & Garcia-Romero, D. (2023). VoxWatch: An open-set speaker recognition benchmark on VoxCeleb. arXiv:2307.00169.

[4] Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. Interspeech 2020, 3830–3834. doi:10.21437/Interspeech.2020-2650.

[5] Ravanelli, M., Parcollet, T., Plantinga, P., et al. (2021). SpeechBrain: A general-purpose speech toolkit. arXiv:2106.04624.

[6] Wang, H., Liang, C., Wang, S., et al. (2022). WeSpeaker: A research and production oriented speaker embedding learning toolkit. arXiv:2210.17016.

[7] Resemble AI. Resemblyzer: A Python package to analyze and compare voices with deep learning. Software repository.

[8] Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: A large-scale speaker identification dataset. Interspeech 2017.

[9] Chung, J. S., Nagrani, A., & Zisserman, A. (2018). VoxCeleb2: Deep speaker recognition. Interspeech 2018.

[10] Bredin, H., Yin, R., Coria, J. M., et al. (2020). pyannote.audio: Neural building blocks for speaker diarization. ICASSP 2020, 7124–7128.

[11] Bredin, H. (2023). pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. Interspeech 2023.

[12] van der Vloed, D., Kelly, F., & Alexander, A. (2020). Exploring the Effects of Device Variability on Forensic Speaker Comparison Using VOCALISE and NFI-FRIDA. Odyssey 2020, 402–407.

[13] Alexander, A., Botti, F., Dessimoz, D., & Drygajlo, A. (2004). The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications. Forensic Science International, 146(S1), S95–S99.

[14] Silnova, A., Stafylakis, T., Mošner, L., et al. (2022). Analyzing speaker verification embedding extractors and back-ends under language and channel mismatch. Odyssey 2022, 9–16.

[15] Gerlach, L., Kelly, F., McDougall, K., & Alexander, A. (2024). Exploring speaker similarity based selection of relevant populations for forensic automatic speaker recognition. Odyssey 2024.

[16] Morrison, G. S., Enzinger, E., Hughes, V., et al. (2021). Consensus on validation of forensic voice comparison. Science & Justice, 61, 299–309. doi:10.1016/j.scijus.2021.02.002.

[17] Campbell, J. P., Shen, W., Campbell, W. M., Schwartz, R., Bonastre, J.-F., & Matrouf, D. (2009). Forensic speaker recognition. IEEE Signal Processing Magazine, 26(2), 95–103.

[18] Morrison, G. S., Ochoa, F., & Thiruvaran, T. (2012). Database selection for forensic voice comparison. Odyssey 2012.

[19] Martin, A. F., & Greenberg, C. S. (2010). Human Assisted Speaker Recognition in NIST 2010 Speaker Recognition Evaluation. Speech and Language Processing Technical Committee Newsletter.

[20] Rusti, C., Leschanowsky, A., Quinlan, C., Pnacekova, M., Gorce, L., & Hutiri, W. (2023). Benchmark dataset dynamics, bias and privacy challenges in voice biometrics research. IEEE IJCB.

[21] Leschanowsky, A., Rusti, C., Quinlan, C., Pnacek, M., Gorce, L., & Hutiri, W. (2025). A data perspective on ethical challenges in voice biometrics research. IEEE Transactions on Biometrics, Behavior, and Identity Science, 7(1), 118–131.

[22] Kröger, J. L., Lutz, O. H.-M., & Raschke, P. (2020). Privacy implications of voice and speech analysis—information disclosure by inference. Privacy and Identity Management, 242–258.

[23] Efron, B., & Tibshirani, R. J. (1993). An Introduction to the Bootstrap. Chapman & Hall.

[24] Zhang, C., Morrison, G. S., Ochoa, F., & Enzinger, E. (2013). Effects of telephone transmission on formant-trajectory-based forensic voice comparison. Speech Communication, 55, 796–813.

[25] Hughes, V., Llamas, C., & Kettig, T. (2022). Eliciting and evaluating likelihood ratios for speaker recognition by human listeners under forensically realistic channel-mismatched conditions. Interspeech 2022.

[26] Liu, T., & Yu, K. (2022). BER: Balanced error rate for speaker diarization. arXiv:2211.04304.

[27] Paturi, R., Srinivasan, S., & Li, X. (2023). Lexical speaker error correction: Leveraging language models for speaker diarization error correction. arXiv:2306.09313.

[28] Wang, B. X., & Hughes, V. (2021). System performance as a function of calibration methods, sample size and sampling variability in likelihood ratio-based forensic voice comparison. Interspeech 2021.

Appendix A. Cohort-construction audit

This appendix records the evidence genealogy behind the reader-facing description in Chapter 3. Counts refer to surviving artifacts in the organized analysis workspace.

Stage	Completed evidence	Analytical meaning
Long-form source collection	296 retained FLAC recordings	Available archive audio.
Successful historical processing	244 source videos	Denominator for archive coverage.
Initial acoustic search	41 result rows; 40 resolved stitched clips	Discovery seed and timestamps.
All-speaker representative export	2,051 video-local speakers	One representative per local diarization label.
Turn census	117,663 indexed turns	Complete timestamp inventory.
Five-second retained turns	9,474 additional FLACs	Longer clip-level evidence.
Canonical clip corpus	11,525 FLACs	Input inventory for three-model profile construction.
Frozen recurrence cohort	91 speakers from 91 videos	Common target evaluated by all three models.
Same-video alternatives	720 local voices	Condition-matched comparison set.
Strict archive extension	37 additional videos	Outside-cohort recurrence evidence.

A.1 Physical origin of the 91 cohort entries

Forty-five cohort entries use the earlier all-speaker representative files. Forty-six entries use rebuilt multi-turn files created from the nominated local speaker's diarization segments in 46 additional videos. The two origin classes together cover 91 distinct source videos. The rebuilt files provide longer evidence for the same local-speaker unit used in archive comparisons.

Every cohort row retains a source-video filename, video identifier, local speaker label, physical clip path, origin class, duration, and links to the model outputs. These fields make it possible to trace each score back to the contributing audio.

A.2 Historical discovery labels

Historical pyannote speaker labels reset inside every source video. A label such as SPEAKER_03 therefore identifies one local cluster in one recording. The archive-level speaker identity arises from the later embedding comparisons. The historical 41-row screen supplied candidate labels and timestamps; the 91-video cohort emerged through local expansion, stronger reconstruction, and video-level review.

A.3 Early cross-system verification and negative reference

The first local verification compared the 40 surviving historical candidates with the original reference using SpeechBrain. All 40 received the model's same-speaker decision, with scores ranging from 0.3774 to 0.8748 and a mean of 0.5169. A deliberately different 30-second speaker reference was then compared with the same 40 candidates. All 40 received different-speaker decisions; scores ranged from -0.0978 to 0.1445, with a mean of 0.0115. This early check established that the local verifier distinguished the candidate set from at least one clearly unrelated voice before the larger all-speaker analysis was built.

A.4 Cohort freeze

The final manifest fixes one nominated local speaker per source video. Subsequent analyses preserve those 91 rows as a common evidence set. Outside-video retrieval adds new videos to separate candidate tables while retaining the frozen cohort unchanged.

Appendix B. Statistical definitions and decision rules

Term	Definition
Cosine similarity	Dot product of two unit-normalized speaker embeddings.
Speaker profile	Normalized mean of retained clip embeddings for one video-local speaker.
Exact cohort pair	One of the 4,095 unique cross-video pairs among the 91 selected samples.
External pair	A cross-video speaker-profile pair outside the recurrence cohort.
Descriptive AUC	Probability that a cohort pair outranks an external pair within one model.
Global percentile	Position of a score inside the corresponding model's archive distribution.
Same-video rank	Position of the nominated local speaker among all labels in its own source video.
Strict archive rule	Best local speaker reaches the declared top-decile position in all three models.
Video-blocked bootstrap	Resampling by source video so that dependent local comparisons remain grouped.
Sign-flip test	Random reversal of target-control directions under a symmetric null.

B.1 Decision hierarchy

Tier	Rule	Use
Frozen cohort	One nominated voice from each of 91 preselected videos.	Primary recurrence target.
Strict extension	All three models satisfy the strict archive rule.	Principal outside-cohort evidence.

Tier	Rule	Use
Broader extension	All three models satisfy wider percentile bands.	Ranked listening and validation pool.
Split-aware review	Top two or three local labels remain close.	Inspection of diarization fragmentation.
Known-voice benchmark	Expected Grandma label ranks first or top two.	Archive-native validation.

B.2 Model consensus

Consensus uses ranks, percentiles, standardized effects, or agreement counts. Raw cosine scores stay on their original model scales. A strict all-three-model result requires concordant high placement. “At least two models” summarizes robust majority agreement while preserving the third model’s rank in the detailed tables.

Appendix C. Supplementary results

C.1 Same-video ranking under equal clip weighting

System	Ranked first	Top two	Target wins vs 720 alternatives
SpeechBrain	73/91	89/91	97.1%
WeSpeaker	67/91	82/91	95.1%
Resemblyzer	60/91	78/91	92.5%
All three models	55/91	74/91	—
At least two models	66/91	85/91	—

C.2 Duration-weighting sensitivity

Aggregation rule	Mean 3-model AUC	≥2 models rank first	≥2 models top two	Grandma switches
Equal clip	0.9248	66/91	85/91	12/12
Linear duration	0.9276	66/91	85/91	12/12
Square-root duration	0.9267	66/91	85/91	12/12
Duration capped at 15 s	0.9268	67/91	85/91	12/12

C.3 Exact cohort model agreement

Model pair	Spearman ρ	Mean top-10 overlap	Top-neighbor agreement
SpeechBrain–WeSpeaker	0.766	0.564	0.418
SpeechBrain–Resemblyzer	0.521	0.359	0.187
WeSpeaker–Resemblyzer	0.465	0.329	0.143

C.4 Standardized same-video control test

Model	Standardized target-control difference	Video-bootstrap 95% CI	Above local p95
SpeechBrain	2.067 SD	1.940–2.189	82/91
WeSpeaker	1.983 SD	1.831–2.137	72/91
Resemblyzer	1.553 SD	1.432–1.668	69/91
Three-model consensus	1.896 SD	1.779–2.012	75/91

Every model-specific and consensus sign-flip test produced $p < 0.0001$. The source video is the resampling unit.

C.5 Contextual transcript analysis

Faster-Whisper transcripts were generated for the suspected recurring-speaker samples, and redeem-family terms were indexed for review. The phrase evidence supplies a small contextual case study associated with the well-known “do not redeem” motif. Acoustic comparisons remain the principal evidence because lexical overlap can reflect scripts, call type, or editing.

Appendix D. Evidence-package guide

The evidence archive is organized so that extraction beneath C:\KitbogaAudio preserves the paths used by the reports. A portable package can contain derived clips, manifests, matrices, workbooks, figures, checksums, and this thesis while the large long-form source recordings remain available through their public video identifiers.

The following English labels identify the principal evidence families. Detailed intermediate runs remain indexed in the organized evidence workbook.

Evidence family	Path	Contents
Frozen 91-video cohort	C:\KitbogaAudio\experiment_all_speakers_20260707\NO_BILLING_CURRENT_CLUSTER_REPORT.xlsx	Sheet: final_91_strong_likely
Exact three-model cohort pairs	C:\KitbogaAudio\experiment_all_speakers_20260707\LOCAL_VOICE_COMPARERS_20260722\3model_91_analysis_20260722\ALL_3_MODEL_91_PAIRWISE_REPORT.xlsx	4,095 unique pairs per model
Same-video alternative controls	C:\KitbogaAudio\experiment_all_speakers_20260707\LOCAL_VOICE_COMPARERS_20260722\91_same_video_controls_20260722\ALL_3_MODEL_91_SAME_VIDEO_CONTROL_REPORT.xlsx	91 targets and 720 alternatives
Cohort versus all representatives	C:\KitbogaAudio\experiment_all_speakers_20260707\91_vs_all_2051_3model_20260723\91_VS_ALL_2051_3MODEL_REPORT.xlsx	Three-model archive retrieval
Complete statistical suite	C:\KitbogaAudio\experiment_all_speakers_20260707\THESIS_SPEAKER_REUSE_ANALYSIS_20260724\COMPREHENSIVE_PYTHON_STATS_20260724_V3	Matched cohorts, model coherence, candidates, clusters
Grandma and temporal analysis	C:\KitbogaAudio\experiment_all_speakers_20260707\GRANDMA_CONTROL_AND_YEAR_ANALYSIS_20260727	Known-voice benchmark and year diagnostics
Aggregation sensitivity and matched switches	C:\KitbogaAudio\Web Work\CODEX_ADDITIVE_REVISION_20260728\analysis\OTHER_METHODS_20260728_V2\OTHER_METHODS_AND_MATCHED_CONTROLS_20260728.xlsx	Four profile rules and Grandma switches
Organized evidence index	C:\KitbogaAudio\experiment_all_speakers_20260707\ORGANIZED_EVIDENCE_20260721\ORGANIZED_EVIDENCE_INDEX.xlsx	Artifact and analysis-run index

D.1 Historical multi-reference cloud supplement

Eight independently saved cohort voiceprints were jointly evaluated against the full long-form recordings of the remaining 83 cohort videos. The completed run produced 5,960 local-speaker/reference scores, 45,168 timestamped turns, and 745 complete eight-score local-speaker vectors. The preserved report records the winning and runner-up labels, individual reference confidences, aggregate scores, score gaps, and matched timestamps. This material supplies supplemental multi-reference and diarization evidence as a separate auditable evidence family.

D.2 Reproduction sequence

- Open the frozen cohort manifest and confirm 91 rows and 91 unique video identifiers.
- Open the exact pairwise report and confirm 4,095 unique non-self pairs for each model.
- Open the same-video report and confirm 720 local alternatives plus video-level statistics.
- Open the archive-retrieval report and trace strict outside candidates to their local speaker labels.
- Open the aggregation workbook and compare equal, linear, square-root, and capped-duration results.
- Use the evidence index to locate clips, manifests, score matrices, and QA artifacts.

Appendix E. Prospective blinded listening protocol

E.1 Objective

The listening study estimates whether human raters distinguish cohort same-speaker pairs from matched different-speaker pairs and whether their ratings agree with the three-model consensus.

E.2 Raters and blinding

- Recruit at least five raters with normal self-reported hearing.
- Use randomized clip identifiers that conceal video title, date, model score, and candidate tier.
- Present trial order independently for every rater.
- Collect confidence and a brief reason code after every judgment.

E.3 Trial construction

Trial class	Example construction	Purpose
Cohort same-speaker hypothesis	Two cohort videos separated by year and channel.	Tests perceived recurrence under hard variation.
Same-video different voice	Nominated voice paired with a strong local alternative.	Controls episode and channel context.
Hard archive non-target	Cohort sample paired with a high-scoring outside voice below the strict rule.	Tests model-near negatives.
Grandma positive control	Two held-out Grandma samples.	Checks rater engagement and known recurring-voice recognition.
Target/Grandma contrast	Nominated scam-call voice paired with Grandma from a shared episode.	Tests separation of two recurring profiles.

E.4 Rating instrument

- Same speaker
- Probably same speaker
- Unsure
- Probably different speakers
- Different speakers

E.5 Analysis

The primary endpoint is the difference in “same” or “probably same” response rate between cohort trials and matched control trials. Secondary endpoints are ordinal mixed-effects estimates, inter-rater reliability, calibration against confidence, and association with the three-model consensus percentile. Video and rater enter as crossed random effects. Exclusion rules, sample size, trial counts, and the primary contrast should be preregistered before ratings begin.