Recapp

Technology Landscape

What it currently takes to turn an in-person meeting, heard by a device lying on the table, into a transcript, a summary and a list of action items, ...

2026-09-01

Prepared by Copundit

AI for Advanced Ideation™

What it currently takes to turn an in-person meeting, heard by a device lying on the table, into a transcript, a summary and a list of action items, and how well each step actually measures. Two constraints run underneath the whole chain, and neither is a question about the microphones a phone contains: whether the device keeps hold of the microphone for the length of a meeting, and how many audio channels its operating system releases to an application at all.

Contents

Contents

Abbreviations

Abbr	Stands for	What it actually is (plain English)
AEC	Acoustic Echo Cancellation	Removes the device's own loudspeaker output from what its microphone hears
AGC	Automatic Gain Control	Automatically raises quiet audio and squashes loud audio to a steady level
AI	Artificial Intelligence	Used here only as the category name products in this field give themselves
AMI	Augmented Multi-party Interaction	A 2005 corpus of recorded meetings, captured on headsets and on a table array at once
ANE	Apple Neural Engine	The dedicated machine-learning chip inside an iPhone, separate from its main processor
API	Application Programming Interface	The published way one piece of software calls another
ASR	Automatic Speech Recognition	The software that turns recorded speech into written words
ATF	Acoustic Transfer Function	How a room changes a sound between the mouth that made it and the microphone that hears it
BLEU	Bilingual Evaluation Understudy	An old word-overlap score borrowed from machine translation and still used on summaries
CDD	Compatibility Definition Document	Google's rulebook for what a device must do to be allowed to call itself Android
CER	Character Error Rate	The word error rate's equivalent for languages written without spaces, such as Mandarin
CPU	Central Processing Unit	The phone's general-purpose processor, as opposed to its neural or graphics chips
CSS	Continuous Speech Separation	A model that splits a continuous mixed recording into separate non-overlapping streams
DASR	Distant Automatic Speech Recognition	Speech recognition when the microphone is across the room instead of at the mouth
DER	Diarization Error Rate	Percentage of speaking time attributed to the wrong person, or missed, or invented
DSP	Digital Signal Processing	The fixed audio clean-up a device applies before any application sees the sound
EEND	End-to-End Neural Diarization	A single network that decides who spoke when, instead of a chain of separate steps
FGS	Foreground Service	The Android mechanism that lets an app keep working while it is not on screen
FOA	First Order Ambisonics	A four-channel format that records the direction a sound came from, not just the sound
GB	Gigabyte	A unit of storage; a compact on-device speech or language model occupies one to two
GDPR	General Data Protection Regulation	The European Union's data protection law; a recording of a person is personal data

Abbr	Stands for	What it actually is (plain English)
GPU	Graphics Processing Unit	The server chip that machine-learning models run on; rented by the hour
GSS	Guided Source Separation	Given who spoke when, it uses several microphones to isolate one talker from the mixture
HAL	Hardware Abstraction Layer	The manufacturer's own code between Android and the actual chips in a specific handset
ICSI	International Computer Science Institute	The Berkeley institute whose 2003 corpus of research-group meetings is still a benchmark
JER	Jaccard Error Rate	A diarization score that weights every speaker equally, so missing a quiet one is costly
LALM	Large Audio-Language Model	A text model fed audio directly, so it hears the recording instead of reading a transcript
LLM	Large Language Model	A text-generating model; here, the part that writes the summary and pulls out action items
MB	Megabyte	A unit of storage; one recorded hour of compressed speech is tens of these
MC	Multi-Channel	Audio recorded by several microphones at once, which preserves where each sound came from
MEMS	Micro-Electro-Mechanical System	The chip-scale microphone used in every phone and every pocket recorder
ML Kit	Machine Learning Kit	Google's set of ready-made on-device model interfaces for Android applications
MVDR	Minimum Variance Distortionless Response	A classic beamformer: combines microphones to favour one direction, suppress the rest
NAIST	Nara Institute of Science and Technology	A Japanese research institute; a team name in the meeting-transcription challenges
NIST	National Institute of Standards and Technology	The United States agency that ran the early speech-transcription evaluations
NPU	Neural Processing Unit	The Android equivalent of Apple's neural chip; runs models without the main processor
NTT	Nippon Telegraph and Telephone	The Japanese telecommunications company whose research labs enter these challenges
OS	Operating System	The phone's own software layer that decides which app may hold the microphone
PCC	Pearson Correlation Coefficient	How closely two measurements move together on a straight line, from -1 to +1
PCM	Pulse-Code Modulation	Raw uncompressed digital audio, the format a microphone's samples arrive in
RIR	Room Impulse Response	A recording of how one room smears a single click; used to fake far-field audio
ROUGE	Recall-Oriented Understudy for Gisting Evaluation	Counts word overlap between a generated summary and a human one
RTF	Real-Time Factor	Seconds of computing per second of audio; below 1 means faster than the recording itself
SC	Single-Channel	Audio recorded by one microphone, which throws away all directional information

Abbr	Stands for	What it actually is (plain English)
SNR	Signal-to-Noise Ratio	How loud the speech is against the room noise, in decibels
SOT	Serialized Output Training	Teaching one recognizer to write out several overlapping talkers one after another
STCON	Team name, expansion not stated in the sources read	The group with the best macro-averaged score in the CHiME-8 distant-speech challenge
TRL	Technology Readiness Level	A 1-to-9 maturity scale: 1 is an idea, 4 is a lab result, 9 is shipping to real users
TSE	Target Speaker Extraction	Pulling one named voice out of a mixture, given a sample of that voice
VAD	Voice Activity Detection	The step that decides which parts of a recording contain speech at all
WER	Word Error Rate	Percentage of words a transcript gets wrong; the standard accuracy number for speech
WPE	Weighted Prediction Error	A filter that removes the echo a room adds, before anything else reads the audio
WWDC	Worldwide Developers Conference	Apple's annual June event where it announces what applications will be able to do

Executive Summary

The task under survey is **distant automatic speech recognition (DASR)** of conversational speech: several people talking, interrupting each other, into a microphone that is on the table rather than at anybody's mouth, followed by **speaker attribution** (who said which words) and then by a language model that writes a summary and extracts action items. Six facts define the frontier. First, **the binding constraint on a general-purpose phone is not acoustic but architectural**: neither mobile operating system offers a third-party application a documented route to the raw per-microphone signals its hardware collects, and no handset has been shown to expose one, so such an application competes in the single-channel condition whatever microphones the device contains. Second, **the size of what that forfeits is measured**: on the 170-meeting NOTSOFAR-1 blind evaluation set of real office meetings, the winning system scored 22.2 percent time-constrained minimum-permutation word error rate reading one distant microphone and 10.8 percent reading a seven-microphone tabletop array, an 11.4-point gap over the same meetings, which every device recorded simultaneously (arXiv 2501.17304, arXiv 2409.02041); that gap is single-device-versus-conference-array, and a pocket recorder pays it too. Third, **the spatial half of the gap is not recoverable by the method that produces it**, because that front end fits a spatial covariance matrix built from phase differences between microphones and one channel has none; the field's named alternative, a single-channel separator trained on real rather than simulated meeting audio, is untried. Fourth, **real recorded audio, not simulation, is the scarce input**: multi-talker models trained on synthetic mixtures alone reached 16.0 percent on the AMI single-distant-microphone condition and 20.1 percent on NOTSOFAR-1 single-channel, and adding a small fraction of real in-domain conversational data moved them to 15.2 and 16.3 (arXiv 2605.15442). Fifth, **the summarisation layer is remarkably tolerant of transcription error and that tolerance is itself the hazard**: transcripts above 50 percent word error produced summaries scoring roughly on par with transcripts near 11 percent, and in the only human-annotated study of meeting-summary metrics about a third of the metric-and-error combinations either ignore the error or reward it, with hallucination detected by none of the nine metrics tested (arXiv 2507.18161, arXiv 2404.11124). Sixth, **the platform layers are moving in opposite directions**: every documented background-audio change on either platform restricted capture further, through Android 12 to 15 and iOS 15 to 17, while the same vendors have since newly exposed on-device recognition and summarisation to third-party applications in iOS 26 and in Android's on-device gener-

ative stack, so any argument that something recently changed has to be made at the recognition layer and not at the capture layer. Three dead zones stand out. **No public benchmark measures any of this on a smartphone**, so the phone-versus-dedicated-recorder question has no published answer in either direction; **no benchmark supplies labelled action items with owners**, so the third output of this category's standard product has no accuracy number anyone can quote; and **nobody has measured a named-entity word error rate on far-field meeting audio**, which is the number on which the whole summarisation-tolerance argument actually rests.

The Problem

What it is. A meeting held in a room is a physical event: several people produce speech, the sound bounces off walls, tables and whiteboards, and some of it arrives at a microphone. Turning that into a written record requires solving several things at once, and each one degrades the next. The recording device must actually capture the whole meeting without stopping. The signal arriving at it is a mixture, because people talk over each other and the room adds echo and noise. How much of that mixture reaches the software depends on how many microphone channels the device's own software layer is willing to release. The words then have to be recognised despite the mixture, and attributed to the right person, because a decision or a commitment is meaningless without knowing who made it. Speech technology treats this as a distinct discipline, **distant automatic speech recognition (DASR)**, separate from the near-field dictation problem that a phone held to the mouth solves well; the field has run dedicated challenge series on it since the early 2000s, from the National Institute of Standards and Technology (NIST) Rich Transcription evaluations through the CHiME series (arXiv 2507.18161).

Whom it affects and at what scale. Everyone who sits in meetings, which is why the commercial pull is large: over 1.5 million dedicated recording devices sold by one vendor by January 2026 (https://techcrunch.com/2026/01/04/plaud-launches-a-new-ai-pin-and-a-desktop-meeting-notetaker), one software incumbent reported at roughly 100 million US dollars of annual recurring revenue in March 2025 (https://sacra.com/c/otter/), and another raised at a 1.5 billion dollar valuation in March 2026 (https://techcrunch.com/2026/03/25/granola-raises-125m-hits-1-5b-valuation-as-it-expands-from-meeting-notetaker-to-enterprise-ai-app/). The research effort is proportionate but narrower: the two most recent challenge rounds drew 9 teams and 32 systems on the geometry-agnostic task (arXiv 2507.18161) and 5 further teams on the fixed-geometry meeting task (arXiv 2501.17304). The scientific record is thin in one specific place. The reference corpora are recorded on headsets, on ceiling and table arrays, on smart glasses and on purpose-built conference devices; a recent survey's 36-row table of notable meeting corpora contains no smartphone-captured entry at all.

How it is created. The chain from speech to a wrong action item is ordered, and every stage is measured by a different number.

- **Capture.** The device must be recording, and keep recording. On a general-purpose phone this is the earliest failure and the only one that is unrecoverable: an interrupted or suspended recording cannot be repaired downstream, however good the models are. It is measured, when anyone measures it, as the fraction of sessions captured end to end, and no such measurement has been published on either platform.
- **Acoustic propagation.** Sound leaves a mouth, reflects, and arrives at the microphone attenuated, reverberated and mixed with noise and with other talkers. Distance from talker to microphone is the dominant variable, and it is fixed by where the device sits. Measured as signal-to-noise ratio and reverberation time. What distance costs has been measured on one corpus built for the purpose: LOTUSDIS recorded the same Thai conversations simultaneously on nine independent single-channel devices of six microphone types spanning 0.12 to 10 metres, and an off-the-shelf Whisper model scored 64.3 percent word error rate pooled across all nine positions against 81.6 percent on the far-field ones alone, a 17.3-point penalty for distance within one recording session (https://arxiv.org/abs/2509.18722). Fine-tuning on distance-diverse audio moved the same pair to 38.3 and 49.5 percent, so the penalty narrows to 11.2 points and does not close. For a product whose only placement control is asking the user where to put the phone, this is the most actionable single-channel number in the field, and it exists in Thai only.

- **Transduction and channel release.** How many microphones hear the event, and, separately, how many of their signals survive the device's own audio stack to reach the application. Several spatially separated microphones let a system infer direction, which is what makes separating overlapping talkers tractable. One channel throws that away, and on both mobile platforms the number of channels released to a third-party application is one, or two that are phase-corrupted by the platform's own voice processing.
- **Segmentation and speaker attribution.** Deciding when speech is present, how many people are in the room, and which of them is talking. Errors here compound catastrophically, because everything downstream is conditioned on them. Measured as diarization error rate and Jaccard error rate.
- **Recognition.** Turning the separated, attributed audio into words. Measured as word error rate, and in the meeting setting as a speaker-attributed variant of it. Within that total, the errors that fall on names, numbers and decisions matter disproportionately, and are measured separately as a named-entity word error rate.
- **Abstraction.** A language model compresses the transcript into a summary. Measured, badly, by overlap with a human summary or by a second language model's judgment.
- **Extraction.** Commitments are pulled out as action items with owners. This is the stage where a fluent model most easily invents a commitment nobody made, and it is the stage with the least measurement of all.

The earliest step in that chain, the acoustic path from a mouth to a microphone across a room, is the one no later stage can undo, and it is the step a device's physical design controls; the second step, how much of what was captured reaches the software, is the one a platform's software design controls, and it is where this field's decisive asymmetry now sits. The parameters used to measure each stage are defined next.

Landscape at a Glance

Every technique profiled below, grouped by mechanism family, and inside each family the one that carries most weight for producing the three outputs first, which is not the same as the most mature. Technology Readiness Level (TRL) runs 1 (basic principle) to 9 (shipping with real-world use); 4 is lab-validated, 6 is a first full-scale demonstration, and it is a separate axis from position in this table.

#	Technique	Modality	TRL
1	iOS background audio session and the background-restart prohibition	Handset capture and channel release	9
2	Android microphone foreground service under manufacturer process killing	Handset capture and channel release	9
3	Chunked write-through to persistent storage	Handset capture and channel release	9
4	iOS input routing and virtual polar patterns	Handset capture and channel release	9
5	Android unprocessed audio source and manufacturer hardware layer	Handset capture and channel release	9
6	Hosted cloud speech recognition on a general model	Recognition back end	9
7	Streaming recognition during the meeting	Recognition back end	9
8	Off-the-shelf Whisper on far-field meeting audio	Recognition back end	9
9	Contextual biasing on personal entities	Recognition back end	9
10	Whisper fine-tuned with self-supervised speech features	Recognition back end	6

#	Technique	Modality	TRL
11	Multi-talker training on synthetic mixtures plus real in-domain audio	Recognition back end	5
12	Language-model correction of a completed transcript	Recognition back end	4
13	Large audio-language models for meeting transcription	Recognition back end	4
14	Prompted language-model meeting summarisation with speaker tags	Summarisation and extraction	9
15	Action-item and commitment extraction	Summarisation and extraction	4
16	Language-model-as-judge summary evaluation	Summarisation and extraction	5
17	Apple SpeechAnalyzer and SpeechTranscriber	On-device inference on phone hardware	9
18	Core ML Whisper inference on the Apple Neural Engine	On-device inference on phone hardware	9
19	Apple Foundation Models on-device language model	On-device inference on phone hardware	9
20	Gemini Nano and ML Kit on-device generative inference	On-device inference on phone hardware	9
21	Map-reduce chunked on-device summarisation	On-device inference on phone hardware	6
22	Speaker-embedding extraction with spectral clustering	Speaker attribution	8
23	Quantised on-device diarization	Speaker attribution	6
24	Target-speaker voice activity detection refinement	Speaker attribution	6
25	Speaker enrolment and target-speaker conditioning	Speaker attribution	6
26	Continuous speech separation and single-channel spectral masking	Far-field front end	6
27	Guided source separation	Far-field front end	7
28	Joint diarization and separation with neural target-speaker extraction	Far-field front end	5
29	Multichannel neural beamforming for speaker-attributed recognition	Far-field front end	4
30	Parallel-capture device comparison with a headset reference	Device-comparison measurement	9
31	Capture-integrity telemetry instrumentation	Device-comparison measurement	6

Parameter Primer

Modalities are measurement methods; what they produce is parameters. This section defines every quantity used in the profiles below, in the order of the chain in The Problem, so a reader can map any number in this document onto the stage of the pipeline it judges.

Stage	What can go wrong	Parameter	How it is measured
Capture	The recording stops	Capture survival	Fraction of sessions captured end to end; no published figure exists
Propagation	Distance, echo, noise	Signal-to-noise ratio	Decibels, per recording condition
Transduction	Directional information is lost	Channels released to software	Number of usable phase-coherent channels an application receives
Attribution	Wrong or miscounted speakers	DER, JER	Percentage of speaking time misassigned
Recognition	Wrong words	WER, tcpWER, tcorcWER	Percentage of words wrong, with or without speaker credit
Recognition	The wrong words are the names	Named-entity WER	Word error rate counted only over names, numbers and acronyms
Abstraction	A fluent but wrong summary	G-Eval, ROUGE, UniEval	Model or overlap judgment against the transcript
Extraction	A commitment nobody made	No accepted metric	Not established
Delivery	The user is still waiting	Wall-clock latency, speed factor	Seconds from tap-stop to finished output
Delivery	The phone is hot or flat	Battery draw per recorded hour, thermal throttle factor	Percentage of charge per hour of audio; percentage of throughput lost under sustained load

Word error rate (WER). The percentage of words a transcript gets wrong, counting substitutions, deletions and insertions against a human reference. It is the field's universal accuracy number and it is only comparable between two systems measured on the same test set under the same audio condition. Near-field read speech scores in the low single digits; far-field conversational meeting speech scores between roughly 10 and 80 percent. It is produced by every technique in the Recognition back end family.

Concatenated minimum-permutation word error rate (cpWER) and its time-constrained form (tcpWER). A word error rate for meetings, where the system must also decide who said what. It concatenates each speaker's words, tries every matching of system speakers to real speakers, and reports the best one, so a system that transcribes perfectly but attributes badly is still penalised. The time-constrained version, introduced in the MeetEval toolkit, additionally requires the words to land in roughly the right place in time, which makes it much more sensitive to segmentation errors (arXiv 2507.18161). It is the ranking metric of the CHiME-7, CHiME-8 and NOTSOFAR-1 challenges and is the single most useful number in this document. Because cpWER and tcpWER are different numbers on the same audio, and both differ from plain word error rate, the three are never compared across a single column in this document.

Time-constrained optimal reference combination word error rate (tcorcWER). The same measurement with speaker identity removed: it scores only whether the words were recognised, not who was credited with them. Reading tcpWER and tcorcWER together separates a recognition failure from an attribution failure. On the NOTSOFAR-1 evaluation set the winning single-channel system scored 22.2 percent tcpWER against 17.7 percent tcorcWER, so roughly a fifth of its errors were attribution rather than recognition (arXiv 2501.17304).

Speaker-dependent character error rate (SD-CER). The same idea as cpWER for Mandarin, counted in characters because Chinese is written without word spaces. It is the metric of the AliMeeting corpus results quoted below and is not comparable with any English word error rate.

Named-entity word error rate. Word error rate restricted to the tokens a summary actually needs: personal names, company names, numbers, dates and acronyms. It matters because plain word error rate weights a filler word and a customer's name equally, while a summary built on a transcript that lost the name is wrong in a way a fluent summary hides. No published measurement of it on far-field meeting audio was found in this survey.

Diarization error rate (DER) and Jaccard error rate (JER). DER is the percentage of speaking time given to the wrong speaker, plus missed speech, plus speech invented where there was none. JER computes the same kinds of error but weights every speaker equally rather than by talking time, so failing to notice a quiet participant costs as much as failing on a talkative one. JER tracks final transcript quality better than DER because it exposes speaker-counting errors, which are the catastrophic ones: across 22 challenge systems, JER against tcpWER gave a Pearson correlation coefficient (PCC) of 0.92 and DER against tcpWER 0.88 (arXiv 2507.18161).

Signal-to-noise ratio and scale-invariant signal-to-noise ratio. How loud the wanted speech is relative to everything else, in decibels; the scale-invariant form is the standard training objective for separation models. A warning attaches to it: in one meeting system, joint training improved recognition by 34 percent relative while making the scale-invariant ratio worse, so a cleaner-sounding separated signal does not necessarily produce a better transcript (arXiv 2211.00511).

Real-time factor and speed factor. Real-time factor is seconds of computing per second of audio; a value below 1 means the system runs faster than the recording itself. Speed factor is its inverse, the seconds of audio processed per second of wall clock, so a speed factor of 60 means one minute of audio per second of processing. Vendors and benchmark blogs routinely print speed factors while calling them real-time factors; every such figure in this document has been converted to a speed factor and labelled as one. Neither number is the wait a user experiences, which also includes queueing, model loading, uploading and the summarisation pass.

Battery draw per recorded hour and thermal throttle factor. How much of the phone's charge one hour of audio costs, and how much of the processor's throughput is lost once the device has been running warm for a while. It is the parameter that decides whether an on-device architecture is shippable at all, and it is the worst-measured quantity in this document. Recording alone is close to free, at about 0.66 percent of charge an hour in one self-report with no device model or method behind it. A fully offline continuous pipeline is put at 25 to 40 percent an hour and recognition alone at 10 to 15 percent, both from single blog posts; the sustained throttling penalty is reported anecdotally at 30 to 50 percent of throughput and has been measured by nobody. A widely circulated figure of 3 percent of battery per minute of audio is arithmetically impossible, since it consumes 180 percent of a battery over the one-hour meeting its own source describes, and must never be quoted.

Context window. The number of tokens (roughly, word pieces) a language model can hold at once, counting instructions, transcript and output together. One hour of speech at conversational pace is about 9,000 words, which two independent sources put at roughly 11,000 to 13,000 tokens, so the context window is the parameter that decides whether a whole-meeting summary can be produced on the phone in one pass at all.

G-Eval, UniEval and ROUGE. The three families of automatic summary score. ROUGE counts overlapping word sequences against a human-written summary. UniEval is a fine-tuned model that rates coherence, consistency, relevance and fluency. G-Eval prompts a strong language model to rate the same four dimensions given only the transcript and the summary, needing no human reference. Their reliability is itself contested and is covered in Dead Ends.

Techniques by Modality

Seven mechanism families, ordered by how much weight each carries for turning a room full of talking into the three outputs: first whether the phone holds the microphone for the whole meeting and what it releases to an application, then what was said, then what is made of it, then where the computation runs, then who said it, then what happens to the sound before a recognizer sees it, and last how any of

it can be measured against a competing device. The order is not the order of the pipeline and not the order of maturity. The front end sits low because none of it runs on one channel; summarisation sits high because it produces two of the three outputs.

Handset capture and channel release

What this modality is. Before any signal processing matters, a general-purpose phone has to hold the microphone, keep holding it for the length of a meeting, and hand the resulting audio to the application in a usable form. Both mobile operating systems treat microphone access as a privileged, revocable, interruptible resource, and both interpose their own audio clean-up between the physical microphones and the application: an application declares a background capability, holds a session, receives a processed virtual channel rather than the raw sensor signals, and may have the session taken away for a phone call, another application, or power management. The family is a platform rulebook rather than a signal-processing method. Its parameters are the number of phase-coherent channels released to software and capture survival, the fraction of sessions recorded end to end.

State of the modality. The mechanisms are documented on both platforms and neither vendor publishes a reliability figure; an exhaustive search of developer post-mortems, engineering blogs from companies shipping recorder applications, academic mobile-systems literature and public bug trackers found no published, statistically meaningful capture-survival measurement on either platform. What is not contested is the direction of travel: every background-audio change documented on either platform, across Android 12 to 15 and iOS 15 to 17, tightened capture, and neither platform has opened a new channel since. What is contested, and unmeasured, is how often a correctly written application actually loses a meeting.

iOS background audio session and the background-restart prohibition | TRL 9

- **What it is:** The iOS audio-session model for long recordings: an application declares the `audio` background mode in its property list, activates an `AVAudioSession` (in practice the play-and-record category), and registers for interruption notifications so it can stop and restart around events that seize the microphone.
- **Key labs / groups:** Apple (platform audio).
- **Audio condition and test set:** none; this is platform documentation plus developer reports.
- **Best published performance:** not published, and the substantive finding is a deterministic failure rather than a rate. When a cellular or voice-over-internet call interrupts a recording and then ends, the system sends an interruption-ended notification carrying `AVAudioSessionInterruptionOptionShouldResume`, but the privacy subsystem refuses to let a backgrounded application reactivate the microphone; the attempt returns OSStatus 561145187, `AVAudioSessionErrorCodeCannotStartRecording`, logged as `AUIOClient_StartIO failed`. Unless the user unlocks the phone and foregrounds the application by hand, the rest of the meeting is silently lost. Everything that is not a call, a competing microphone request or a memory kill continues: screen lock, switching applications, low power mode and ordinary thermal throttling all leave the recording running, and a system memory reclaim kills the process outright with no notification. Developer forums additionally report an input tap that keeps firing after a background anomaly while delivering zero-filled buffers, so the application writes a valid-looking file containing digital silence; that report is a forum tag index rather than an identified thread and is the weakest-sourced of the load-bearing failure modes here.
- **Translation status:** shipping; every recording application on iOS uses it.
- **Blockers to use:** interruption handling is a recovery mechanism and not a prevention one; the recovery path is blocked in precisely the state a meeting recorder is in; the failure is silent until the recording is inspected; App Store Review Guideline 2.5.4 is separately reported to drive rejections of applications that declare the background audio mode without a visible user-facing feature that needs it, which pushes developers toward a persistent on-screen recording surface.

Android microphone foreground service under manufacturer process killing | TRL 9

- **What it is:** The Android mechanism that lets an application keep recording while it is not

on screen. From Android 14 (API level 34) the foreground service (FGS) must declare `android:foregroundServiceType="microphone"`, pass `FOREGROUND_SERVICE_TYPE_MICROPHONE` at `startForeground()`, hold the `FOREGROUND_SERVICE_MICROPHONE` manifest permission and the `RECORD_AUDIO` runtime permission, and show a persistent notification.

- **Key labs / groups:** Google (Android platform); Samsung, Xiaomi, Huawei and Oppo system software teams.
- **Audio condition and test set:** none; platform documentation plus manufacturer behaviour reports.
- **Best published performance:** not published. Google documents the type as "Continue microphone capture from the background, such as voice recorders or communication apps" and states the binding restriction, that a microphone foreground service cannot be created while the application is in the background (https://developer.android.com/develop/background-work/services/fgs/service-types). Omitting the declaration is a hard `SecurityException` rather than a graceful failure, and Android 12 and above throw `ForegroundServiceStartNotAllowedException` on a background microphone-service start. The one restriction that does not apply is the length of the meeting: Android 15 introduced a six-hour cumulative timeout for foreground services and exempted the microphone type from it, while tightening background starts further, so a correctly declared service is not cut off by a clock.
- **Translation status:** shipping on every Android 14 or later device.
- **Blockers to use:** the operating system is not the main adversary; four named manufacturers are. Samsung's sleeping and deep-sleeping application tiers and its RAM Plus swap eviction, Xiaomi's and Huawei's termination shortly after screen lock even with the persistent notification visible, and Oppo's instant freezing of background services on screen lock each kill recordings regardless of correct interface use, and each requires the user to walk a different multi-step settings path, with major updates observed silently reverting those settings. No measured failure rate exists for any of this; see Contested Evidence for the two illustrative percentages that circulate and must not be quoted.

Chunked write-through to persistent storage | TRL 9

- **What it is:** Encoding and flushing captured audio to disk in short increments, typically every 5 to 60 seconds, rather than holding one continuous buffer in volatile memory and writing a single container when the user taps stop. On iOS the documented pattern taps the `AVAudioEngine` input node, accumulates raw samples and encodes them in roughly 60-second blocks; on Android the parallel mitigation is returning `START_STICKY` from the service so the system restarts it after a memory kill, plus requesting an explicit exemption from battery optimisation.
- **Key labs / groups:** the shipping recorder-application developer community; documented in open code such as the BabyApp ambient recorder and in the help pages of RecForge II and similar tools.
- **Audio condition and test set:** none; this is an engineering mitigation with no benchmark.
- **Best published performance:** not published as a rate. What it demonstrably bounds is loss on a hard kill: the user loses the final unwritten seconds rather than the whole session.
- **Translation status:** shipping, standard practice across the category.
- **Blockers to use:** it covers process termination and covers nothing else. It does not detect or repair silent microphone revocation, where the audio callback keeps firing with empty buffers, and it cannot restart a session that the platform refuses to let a backgrounded application restart. A related compatibility mitigation, forcing a universally supported 44.1 kHz capture rate and resampling in software, exists because requesting unusual sample rates causes silent failures on fragmented Android hardware.

iOS input routing and virtual polar patterns | TRL 9

- **What it is:** The interface through which an application on an iPhone obtains audio. `AVAudioSession` owns input routing; the application selects a data source (a physical microphone group) and a polar pattern, and receives a virtualized microphone signal: an omnidirectional channel, or a cardioid beam the operating system forms from two microphones in tandem. Since iOS 14 an application may request two channels by setting `preferredInputOrientation` and `preferredInputNumberOfChannels` and checking `supportedPolarPatterns`. iOS 26 adds First Order Ambisonics (FOA) capture through `AVAssetWriter`, a four-component representation encoding the direction sound

arrived from.
- **Key labs / groups:** Apple platform audio.
- **Audio condition and test set:** none. This is platform documentation and conference material, not a measurement; no test set exists and Apple publishes no accuracy figure for any capture path (https://developer.apple.com/videos/play/wwdc2020/10226/, https://developer.apple.com/library/archive/qa/qa1799/_index.html, https://developer.apple.com/videos/play/wwdc2025/251/).
- **Best published performance:** not published. The load-bearing fact is a capability boundary rather than a score: a third-party application cannot request a raw, synchronised four-channel stream from the four built-in microphones of a recent iPhone, and the stereo path that does exist is reported by developers to be forced to 16 kHz, with a requested 48 kHz silently overridden unless external Made-for-iPhone hardware is attached. A 16 kHz sample rate caps the represented band at 8 kHz. That downsampling report is a developer forum post rather than Apple documentation and is carried as such.
- **Translation status:** shipping on every current iPhone.
- **Blockers to use:** declaring a voice-communication mode causes CoreAudio to insert Voice Processing I/O units applying automatic gain control (AGC), noise suppression and acoustic echo cancellation (AEC), which destroys the inter-channel phase differences that spatial separation depends on; the application is otherwise handed the platform's own beamformer output, tuned for a talker at roughly 30 centimetres for handset use or 1 metre for speakerphone, with the rest of the room treated as noise to remove. Whether the new ambisonic capture path carries enough spatial information to drive separation is unmeasured by anyone.

Android unprocessed audio source and manufacturer hardware layer | TRL 9

- **What it is:** The Android constant `MediaRecorder.AudioSource.UNPROCESSED` exists to deliver audio that bypasses the platform's front-end processing, that is, automatic gain control, noise reduction and high-pass filtering, and is the only documented route to something resembling raw sensor audio on the platform.
- **Key labs / groups:** Google (Android platform); every handset manufacturer's own hardware abstraction layer (HAL) team.
- **Audio condition and test set:** none; this is platform specification, not a measurement.
- **Best published performance:** not published. The Compatibility Definition Document (CDD), read at its Android 10 and Android 12 revisions, guarantees support for the unprocessed source only for the built-in system camera application; third-party access is left to each manufacturer's hardware abstraction layer. Where a third-party application does request multi-channel unprocessed audio, device-specific signal processing is reported to deliver 1 mono channel mirrored across 2 tracks, or to apply undocumented limiters, so the application never receives phase-accurate multi-channel data.
- **Translation status:** shipping, with per-manufacturer behaviour that no vendor documents.
- **Blockers to use:** the guarantee that matters is scoped to an application the developer does not own; there is no published enumeration of which handsets or manufacturers expose true unprocessed multichannel audio to a third party, so the fleet-wide answer is unknown rather than negative, and establishing it would require device-by-device testing.

Recognition back end

What this modality is. The recognizer converts an audio stream into words. Modern systems are large neural sequence models trained on tens of thousands of hours of speech, either weakly supervised on transcribed audio (the Whisper family) or self-supervised on untranscribed audio and then fine-tuned (the WavLM and wav2vec 2.0 families). They read the audio the front end produces, one separated stream at a time, and output text with optional word-level timestamps. The parameter they produce is word error rate, and with attribution, tcpWER.

State of the modality. The field has converged: end-to-end models built on large pre-trained backbones have displaced the hybrid systems that dominated earlier CHiME rounds, mainly because the pre-trained backbone lowers the amount of matched training data a team needs (arXiv 2507.18161). Whisper is the

de facto starting point, present in most challenge submissions and in the baselines. Two things are contested. How much ensembling and language-model rescoring can still add is one: in CHiME-8 the single team that tried a large-language-model rescoring track gained 0.5 percentage points absolute macro tcpWER over its own main submission, and a fine-tuned Llama-2-7B rescorer bought 1.42 to 2.94 percent relative across the CHiME-8 datasets. How fast the field is improving is the other, and the trend tables in circulation mix three different metrics down one column, so the direction is credible and the slope is not quotable.

Hosted cloud speech recognition on a general model | TRL 9

- **What it is:** A metered network service that accepts an audio file or stream and returns a transcript, with optional speaker labels, from a general-purpose model the vendor trains and operates.
- **Key labs / groups:** Deepgram, AssemblyAI, OpenAI, ElevenLabs, Groq, among others.
- **Audio condition and test set:** none published for meeting conditions. Vendor accuracy claims in this category name neither a test set nor an audio condition, which under this document's evidence rule makes them unusable as evidence; the streaming figures the vendors do publish, in the range of 8 to 9 percent word error rate, are quoted for "real-world scenarios featuring cross-talk and moderate accents" with no named corpus.
- **Best published performance:** the reliably published figures are prices, not accuracies. Deepgram Nova-3 at 0.0043 US dollars a minute is 0.258 dollars per audio hour; AssemblyAI lists 0.37; self-hosted Whisper large-v3 on provisioned GPUs falls to about 0.12; a summarisation call on a 12,000-token transcript is about 0.002, two orders of magnitude below the transcription half of the bill. Two independent assemblies put the all-in marginal cost of a recorded hour at 0.26 to 0.38 and at 0.21 to 0.62 dollars including storage and egress.
- **Translation status:** shipping, and the default build for products in this category.
- **Blockers to use:** a single uploaded channel means the single-channel condition and no spatial separation; the per-hour meter runs against a flat subscription, which inverts the usual software margin as usage grows; audio leaves the device, which is the point at which recording-consent and data-protection duties attach; and what a user waits for is queueing and transport rather than inference, so a batch cloud architecture is measured at 10 to 20 minutes to a finished summary while the same models run hundreds of times faster than real time on the server.

Streaming recognition during the meeting | TRL 9

- **What it is:** Sending audio to the recognizer while the meeting is still running, over a connection held open for its duration, and receiving partial hypotheses that are revised as later context arrives, instead of uploading a finished file once the user taps stop. The transcript is therefore complete, to within an endpointing delay of a second or two, at the moment the recording ends, and the only work left for the walk back is the summarisation and extraction pass. It is an engineering architecture rather than a model: the same recognizers appear in the streaming and batch paths.
- **Key labs / groups:** the hosted vendors that also serve the batch path (Deepgram, AssemblyAI, OpenAI, Google, Microsoft, among others); no research community owns it, and no challenge in this field scores it.
- **Audio condition and test set:** none published for meeting conditions. The streaming accuracy figures the vendors publish, in the 8 to 9 percent word error rate range, are quoted for "real-world scenarios featuring cross-talk and moderate accents" and name no corpus, so under this document's evidence rule they are not evidence in either direction. No streaming system has been scored on NOTSOFAR-1, AMI, CHiME or any other named meeting corpus.
- **Best published performance:** the load-bearing figure is a delivery time rather than an accuracy, and it is derived rather than measured. A live-streaming cloud pipeline has the transcript finished when the meeting finishes and adds roughly ten seconds of summarisation on top, against the 10 to 20 minutes a batch architecture is measured at for the same recording. No accuracy penalty for streaming against batch decoding has been published on any meeting corpus.
- **Translation status:** shipping as a documented mode on every hosted recognition interface in this category. Which products actually use it, and what they deliver, is the competitive survey's question rather than this one's.
- **Blockers to use:** it trades the capture problem for a connectivity problem. The connection has

to survive the whole meeting, in the conference rooms and basements where it is least likely to, and a drop mid-meeting is a new failure mode sitting beside the platform's own ones. It forecloses the on-device privacy position, because the audio leaves the phone continuously rather than once and deliberately. The meter runs for the full duration whether or not anyone reads the result. A recognizer denied future context is worse than the same model run offline, and how much worse on far-field meeting audio is measured nowhere. And nothing in the field's separation or diarization stack runs online at challenge quality: the most efficiency-oriented system in either challenge still needs more than two seconds of computation per second of audio, so a streaming product is choosing the general single-stream recognizer and giving up the front end entirely.

Off-the-shelf Whisper on far-field meeting audio | TRL 9

- **What it is:** Using a general-purpose weakly supervised recognizer, unmodified, on distant conversational audio. It is the default engineering choice and therefore the honest baseline for anything built quickly.
- **Key labs / groups:** OpenAI (the model); every product team that ships a wrapper over it.
- **Audio condition and test set:** far-field distant microphones on LOTUSDIS, a Thai far-field conversational meeting corpus with distance-diverse overlapping speech (arXiv 2509.18722); separately, silence and near-silence conditions in throughput benchmarking.
- **Best published performance:** the finding is a collapse rather than a score. An off-the-shelf Whisper model scored 81.6 percent word error rate on the LOTUSDIS distant microphones, falling to 49.5 percent only after fine-tuning on distance-diverse overlapping conversational data (arXiv 2509.18722). The failure modes are structural: Whisper variants have no intrinsic overlap awareness and either delete the quieter speaker or enter hallucination loops. A reported 44 percent hallucination rate for Whisper Large-v3-Turbo when exposed to silence comes from a single benchmarking blog rather than a peer-reviewed source and is carried with that caveat; it matters here because any single microphone in a meeting hears mostly silence.
- **Translation status:** shipping everywhere, including inside on-device wrappers.
- **Blockers to use:** the model was never trained for overlapped far-field speech and no amount of prompting changes that; conditioning each 30-second window on the previous window's text improves coherence in batch and propagates hallucination when a window goes bad; and the assertion in circulation that commercial interfaces average 30 to 40 percent tcpWER in true single-channel meeting conditions carries no citation anywhere and is recorded as unverified.

Contextual biasing on personal entities | TRL 9

- **What it is:** Injecting a user-specific vocabulary into the recognizer's attention or shallow-fusion path, so that names from the address book, the calendar invite and recent project titles are more likely to be recognised than acoustically similar common words.
- **Key labs / groups:** the PROCTER line of work on personalized recognition of contextual text by entity rescoring.
- **Audio condition and test set:** the reported evaluations are entity-recognition tasks with personalized and rare entities; the audio condition is not stated in the material available for this survey, and the result is recorded as condition-unstated.
- **Best published performance:** a reported 44 percent improvement in named-entity word error rate overall and 57 percent on rare personalized entities, with only a marginal parameter increase. Because the condition is unstated, the figure is carried as a direction rather than as a transferable number.
- **Translation status:** shipping. The published entity-adapter form is research, but the same mechanism is exposed by every commercial recognition interface in this category as custom vocabulary, keyterm boosting or speech adaptation, and switching it on requires no research and no model of one's own.
- **Blockers to use:** it needs access to personal data (contacts, calendar) that is exactly the data a privacy-positioned product would rather not read; it helps names it was told about and not names it was not; and the number it is supposed to move, named-entity word error rate on far-field meeting audio, has never been measured, so there is no baseline to improve against.

Whisper fine-tuned with self-supervised speech features | TRL 6

- **What it is:** The Whisper encoder-decoder recognizer, modified and re-trained for meeting audio: WavLM features added to the front, architectural changes such as rotary position encodings and mixture-of-experts layers, multi-task training and noise augmentation.
- **Key labs / groups:** USTC-NERCSLIP with iFlytek Research; NPU (Whisper large-v2 fine-tuned with AdaLoRA); NAIST (a Zipformer recognizer with WavLM features instead).
- **Audio condition and test set:** far-field, NOTSOFAR-1 and the four CHiME-8 scenarios, decoding separated streams produced by the front end.
- **Best published performance:** the recognizer inside both NOTSOFAR-1 winners. The single most repeated win in the challenge was adapting the recognition model to real far-field separated audio: the winning team's own ablation cut development-set tcpWER from 16.57 to 9.87 percent, a 40 percent relative improvement, purely by adding real multi-channel training audio and simulated LibriSpeech audio processed with oracle guided source separation to the fine-tuning set (arXiv 2501.17304). A separate team cut tcorcWER from 37.6 to 31.0 percent on the development set by filtering empty segments, merging short segments and prompting the recognizer for verbatim rather than tidied output.
- **Translation status:** research; the base models are open and the modifications are published but not packaged.
- **Blockers to use:** the architectural improvements are small relative to the front-end and diarization gaps; the best results come from ensembles of several recognizers, which multiplies compute; and the largest single gain comes from training data nobody outside the challenge holds.

Multi-talker training on synthetic mixtures plus real in-domain audio | TRL 5

- **What it is:** Training a multi-talker recognizer mostly on synthetic overlapping mixtures built by convolving clean speech with measured room responses, then adding a small fraction of real in-domain conversational recordings to close the residual gap.
- **Key labs / groups:** the DiCoW line of work (arXiv 2605.15442); the NOTSOFAR-1 organisers, who shipped a 1,000-hour simulated training set built from 15,000 physically measured acoustic transfer functions.
- **Audio condition and test set:** AMI single distant microphone and NOTSOFAR-1 single-channel.
- **Best published performance:** synthetic mixtures alone gave 16.0 percent tcpWER on AMI single distant microphone and 20.1 percent on NOTSOFAR-1 single-channel; adding a small fraction of real in-domain conversational data gave 15.2 and 16.3 percent, and the macro average across benchmarks moved from 9.8 to 8.8 percent (arXiv 2605.15442). The stated reason simulation alone is not enough is that randomly mixing two tracks does not reproduce turn-taking, backchannelling or the Lombard reflex, the involuntary raising of the voice against competing noise. The 16.3 percent figure would beat the challenge winner on the same condition, which is most plausibly a development-set-versus-evaluation-set mismatch; it is recorded here as contested and is not quoted as a state of the art.
- **Translation status:** research, published with reproducible mixtures.
- **Blockers to use:** the real fraction is the expensive part and no public corpus supplies it for a phone; the simulation itself needed acoustic transfer functions physically measured on the target hardware, so simulating a new device is not a shortcut around recording on it.

Language-model correction of a completed transcript | TRL 4

- **What it is:** Passing a noisy multi-talker transcript, with its speaker labels, through a strong language model that repairs phonetic misspellings and speaker leakage using long-range linguistic context.
- **Key labs / groups:** the DiarizationLM and MT-LLM lines.
- **Audio condition and test set:** benchmark subsets characterised by high speaker leakage; the underlying audio condition is far-field meeting material, and the specific test sets are not identified in the source available for this survey.
- **Best published performance:** relative concatenated word error rate reductions of up to 29 percent on high-leakage subsets, from a record that could not be resolved to a published paper and is therefore carried as a weakly sourced ceiling from one favourable subset. For contrast, the peer-reviewed record on rescoring, a different operation, is far smaller: a fine-tuned Llama-2-7B rescorer bought 1.42 to 2.94 percent relative tcpWER across CHiME-8 datasets and 1.87 percent on

the NOTSOFAR-1 development set, and the challenge winner used no language-model rescoring at all (arXiv 2501.17304, arXiv 2507.18161).
- **Translation status:** research, and in practice inside commercial post-processing nobody documents.
- **Blockers to use:** the failure mode has a name, the word error rate trap: once the acoustic evidence is gone the model stops correcting and starts generating fluent sentences nobody said, and the rate at which it crosses that line is measured nowhere.

Large audio-language models for meeting transcription | TRL 4

- **What it is:** Feeding audio embeddings directly into a large language model backbone, so that one model hears the recording and writes attributed text, using semantic reasoning to disambiguate acoustically overlapping words instead of separating them signal-first.
- **Key labs / groups:** the TagSpeech, GLSC-SDR and Qwen-Omni lines.
- **Audio condition and test set:** AliMeeting far-field single-channel (Mandarin, character-based) and AMI single distant microphone (English).
- **Best published performance:** 13.10 percent cpWER reported on AliMeeting far-field and 14.90 percent cpWER on AMI single distant microphone, both 2026 figures. These come from a benchmark table whose CHiME-8 row is wrong by 13.7 percentage points against the peer-reviewed record (see Contested Evidence), so every row of it including these two carries a reduced trust tier and none should be quoted as a settled state of the art.
- **Translation status:** research; the models are 7 to 30 billion parameters.
- **Blockers to use:** size. Continuous audio through a model of that scale exceeds mobile thermal and memory budgets, so this path is cloud-only for the foreseeable term; and a model that reasons its way to a plausible word is the same model that writes a fluent sentence nobody said.

Summarisation and extraction

What this modality is. Given a transcript, produce a short prose account of what happened and a list of commitments with owners. Two approaches exist: fine-tune a sequence-to-sequence model on meeting transcripts paired with human summaries, or prompt a general large language model (LLM) and let it read the transcript directly. The input is text, so every acoustic error upstream arrives here as a wrong or misattributed word, and the output is judged either by overlap with a human summary or by a second model's opinion.

State of the modality. Prompted general models have displaced the fine-tuned meeting summarisers in practice, and they are strikingly robust to transcription error, which is both the field's most encouraging and its most dangerous finding. The measurement layer has not kept up: overlap metrics correlate weakly with human judgment of meeting summaries, no vendor in the category publishes any quantitative factuality figure, and the one thing this category's products promise beyond a summary, a list of action items with owners, has no accepted benchmark at all.

Prompted language-model meeting summarisation with speaker tags | TRL 9

- **What it is:** The transcript, with each utterance tagged by speaker, is placed in a prompt asking a general language model for a summary of a given length that preserves who said what, who proposed what and who took on which action item.
- **Key labs / groups:** the CHiME-7 and CHiME-8 review team ran the largest published controlled study of it, using Gemini 2.0 Flash over the transcripts of every challenge system (arXiv 2507.18161).
- **Audio condition and test set:** far-field, NOTSOFAR-1 evaluation sessions of about six minutes, roughly 1,600 reference words each, summarised to about 200 words; 1,760 system-session samples.
- **Best published performance:** the striking result is a non-result. Summary quality barely tracks transcript quality: systems above 50 percent tcpWER produced summaries scoring roughly on par with systems around 11 percent, and the correlation between tcpWER and the best summary metric was moderate at best (G-Eval overall PCC -0.51, consistency -0.54, ROUGE-1 -0.34, UniEval overall -0.15). Only severe artificial corruption moved the scores. The authors also report that

removing explicit speaker-attribution instructions from the prompt made summaries generic and destroyed any correlation at all (arXiv 2507.18161). A separate line of evidence puts a condition on the tolerance: a transcript can carry 25 to 30 percent word error and still yield an accurate summary only if the errors fall on fillers, false starts and conjunctions rather than on named entities and decisions, and far-field acoustics smear consonants so proper nouns fail first.

- **Translation status:** shipping in every product in this category, and free on both phone platforms.
- **Blockers to use:** the robustness cuts both ways. A summary that reads well over a broken transcript is a summary whose errors the user cannot detect, and consistency, the dimension that penalises invented content, is the one that degrades most with transcription error. The summary side of the -0.51 correlation is itself a model judge with documented self-preference and position bias, and no human read those summaries. Long transcripts also exceed small context windows, which forces chunking.

Action-item and commitment extraction | TRL 4

- **What it is:** Identifying, in a multi-party transcript, the utterances that commit someone to do something, and rendering them as a task with an owner and where possible a deadline. It has been studied as a dialogue-act classification problem since Purver and colleagues' work on detecting action items in multi-party dialogue in 2006, and is currently done as one more instruction inside a summarisation prompt.
- **Key labs / groups:** the early dialogue-act line of work; no current benchmark community.
- **Audio condition and test set:** clean human transcripts only. The supervision that exists is the AMI corpus's Actions annotation: one analysis puts 101 of the 137 AMI scenario meetings as carrying a non-empty action-item annotation, for 381 items in total, an average of 3.7 per meeting, with ICSI released without publicly available action-item annotation. That count is sourced to a record with no identifier, a second source declines to give a count for either corpus, and the project's own reading of a third contradicts both, so the size of the world's supervision for this task is itself unsettled somewhere below 400 labelled items.
- **Best published performance:** not established for meeting conditions. No accuracy, precision, recall or hallucination rate for action-item extraction from far-field meeting audio was found anywhere in this survey, and the one precision-and-recall pair reported for the AMI test sets was published as an unreadable image rather than as numbers. No vendor in the category publishes a quantitative factuality figure for this output.
- **Translation status:** shipping in commercial products, with no published evaluation behind it.
- **Blockers to use:** no labelled test set at scale, therefore no number; the failure mode is a fluent, plausible task attributed to a named person who never agreed to it, driven by the model binding a task to whoever produced most of the tokens describing it; the summary metrics that would catch it either ignore hallucination or reward it (arXiv 2404.11124); and the failure modes are asymmetric in a way a balanced score hides, since a missed commitment leaves the user where they started while an invented one actively misleads them, so the evaluation has to weight precision far above recall.

Language-model-as-judge summary evaluation | TRL 5

- **What it is:** A strong language model is prompted to score a summary for coherence, consistency, relevance and fluency, given the source transcript and no human reference summary. G-Eval is the standard instance; MESA, CREAM and FineSurE are variants.
- **Key labs / groups:** the G-Eval authors; applied to meeting transcription in arXiv 2507.18161.
- **Audio condition and test set:** applied to NOTSOFAR-1 far-field transcripts, 1,760 samples, with eight independently seeded summaries per session to obtain error bounds.
- **Best published performance:** it is the only metric family that responds to transcription error at all (PCC -0.51 overall, -0.54 for consistency), against -0.15 for UniEval and -0.34 for ROUGE-1 (arXiv 2507.18161). Against human error annotation it improves on legacy metrics by roughly 0.25 in point-biserial correlation on specific error-detection tasks.
- **Translation status:** research practice, increasingly used as a default in papers and, per the same sources, inside commercial evaluation nobody publishes.
- **Blockers to use:** it introduces a second opaque model into the measurement; fluency scores sat-

urate because the judge rates model-written text highly, including summaries of randomly corrupted transcripts; it carries self-preference bias toward models with similar backbones and a documented “lost in the middle” position bias that misses omissions buried in the body of a summary.

On-device inference on phone hardware

What this modality is. Where the computation physically runs. A recognizer or a language model can execute on the phone’s own neural accelerator, or on rented server hardware reached over the network. The choice sets three things at once: the wall-clock wait after the recording stops, the marginal cost per recorded hour, and whether the audio ever leaves the device. On-device inference is constrained by memory, sustained thermal budget and, for language models, by context window; cloud inference is constrained by upload time, by a per-hour meter, and by multi-tenant queueing, which is what a user actually waits for.

State of the modality. Both platform vendors now expose on-device speech and language models to third-party applications, and both did so recently: Apple’s frameworks in iOS 26, Google’s alpha speech interface with its generative mode gated to two Pixel generations. The same shipment put the same capability into the vendors’ own free applications, so the why-now and the absorption risk are one fact with two signs. The contested part is capability rather than availability. Server-class meeting pipelines are ensembles running many times slower than real time on data-centre GPUs, so what fits on a phone is a single general model with no separation front end, and the on-device language model’s context window is roughly a third of an hour-long transcript.

Apple SpeechAnalyzer and SpeechTranscriber | TRL 9

- **What it is:** An on-device speech framework introduced at Apple’s Worldwide Developers Conference (WWDC) in 2025 and available to third-party applications from iOS 26, superseding `SFSpeechRecognizer`. It is modular: `SpeechTranscriber` for long-form audio, `DictationTranscriber` for short utterances, `SpeechDetector` for voice-activity gating, all driven through Swift structured concurrency and running entirely on the device.
- **Key labs / groups:** Apple.
- **Audio condition and test set:** two conditions are published and they are far apart. On clear read-aloud speech, a 2,620-sample clean subset of a 5,559-utterance independent benchmark, it scores 2.12 percent word error rate, against Whisper Small at 3.74 and Whisper Large V3 Turbo at 3.01 on the same data. On earnings22, roughly twelve hours of real corporate earnings-call conversation, it scores 14.0 percent. No evaluation on in-person meeting audio captured by a phone exists at any condition.
- **Best published performance:** 2.12 percent word error rate on clear read-aloud speech and 14.0 percent on earnings-call conversation, at a speed factor of about 70 on Apple silicon. The legacy interface it replaces scored 9.02 percent on the same clean read-aloud set. The cloud comparator quoted alongside the earnings22 figure, Whisper Large v3 at about 6.0 percent, is explicitly labelled an estimated delta by its source and carries no citation, so the on-device number is sourced and the number it is compared against is not.
- **Translation status:** shipping in iOS 26, iPadOS 26 and macOS 26 to any third-party application, at no per-hour cost.
- **Blockers to use:** it is a general recognizer with no separation, no multichannel front end and no speaker diarization, so on a table in a four-person meeting it works in the hardest condition with the weakest tooling; the minimum operating system requirement truncates the addressable install base; and earnings-call audio is not in-person meeting audio captured by a phone, so a product built on it still cannot state a number for the condition it will actually run in.

Core ML Whisper inference on the Apple Neural Engine | TRL 9

- **What it is:** The Whisper recognizer compiled to Apple’s Core ML and executed on the Apple Neural Engine (ANE) and GPU, packaged as an open Swift library so any application can transcribe without a network call; the Android parallel is whisper.cpp.

- **Key labs / groups:** Argmax (WhisperKit), on OpenAI's Whisper models.
- **Audio condition and test set:** earnings22 for the meeting-adjacent figure and clean read-aloud speech for the other; a widely circulated 2.2 percent word error rate figure for the same library names no test set, no audio condition and no device, which is itself the finding (https://arxiv.org/abs/2507.10860).
- **Best published performance:** 12.8 percent word error rate on earnings22 for the Small model, marginally better than the platform framework on the same messy audio, and 3.74 percent on clean read-aloud speech, marginally worse. Throughput is a speed factor of about 35 for Small; for the 809-million-parameter Large-v3-Turbo model the reported speed factor on an iPhone 15 Pro is 10 to 16, which is the only phone-side figure available and comes from the library's own vendor. Memory footprints are about 1.6 GB for Large-v3-Turbo, 466 MB for Small, 140 MB for Base and 40 MB for Tiny. For scale, the same Large-v3-Turbo model reaches a peak speed factor of 597 on a data-centre H100 GPU, so cloud silicon is roughly 40 to 60 times faster at raw inference.
- **Translation status:** shipping, open-source.
- **Blockers to use:** model size against phone storage and memory; sustained thermal throttling over a long batch pass, reported anecdotally at a 30 to 50 percent throughput penalty and measured by nobody; and the recognizer has no separation or diarization front end, so the meeting-condition accuracy of this whole path remains unmeasured.

Apple Foundation Models on-device language model | TRL 9

- **What it is:** Direct third-party access, through `SystemLanguageModel`, to the on-device Apple Intelligence language model for summarisation, extraction and structured output, running on the phone's own silicon.
- **Key labs / groups:** Apple.
- **Audio condition and test set:** not applicable; this reads a transcript, not audio. No meeting-summary benchmark result for it exists.
- **Best published performance:** the binding number is not an accuracy but a capacity. The model is about 3 billion parameters, quantised to under 4 bits per parameter on average, and enforces a hard limit of 4,096 tokens per session covering system prompt, transcript and generated output together, throwing a context-size-exceeded error past it; the server model it was distilled and pruned from was itself trained at sequence length 4,096 (Apple's own documentation and technote TN3193; arXiv 2407.21075). Apple states the inference is free of cost and works offline on any Apple-Intelligence-capable device (https://www.apple.com/newsroom/2025/09/apples-foundation-models-framework-unlocks-new-intelligent-app-experiences/).
- **Translation status:** shipping to third-party applications.
- **Blockers to use:** 4,096 tokens is on the order of 3,000 words, while one hour of meeting speech is roughly 11,000 to 13,000 tokens, so a whole-meeting summary cannot be produced in one pass and must be chunked and merged; developer reports on an iOS 27 beta describe a device entering tool-correction loops and failing outright as chunking pushes the accumulated context past the limit, and those reports are single-developer accounts on beta software.

Gemini Nano and ML Kit on-device generative inference | TRL 9

- **What it is:** Google's on-device generative stack. The AICore system service distributes and hardware-accelerates the Gemini Nano small language model so third-party application packages stay small; ML Kit exposes on-device summarisation, and a separate ML Kit generative speech-recognition interface is in alpha.
- **Key labs / groups:** Google (Android platform, Pixel).
- **Audio condition and test set:** none stated for the speech interface. Its capability is described only as "better overall quality" than the legacy interface, with no word error rate under any condition, which is recorded as condition-unstated.
- **Best published performance:** operational figures only. The speech interface has a Basic Mode on API level 31 and above across fifteen locales and an Advanced Mode hardware-gated to Pixel 10 and Pixel 11. The language model ships as 1.8-billion and 3.25-billion-parameter variants at 4-bit quantisation for a footprint near 1 GB, with an initial token latency under 100 milliseconds, generation of 1 to 5 tokens per second and roughly 60 percent neural-processing-unit utilisation on

flagship hardware; those throughput figures come from a developer blog rather than Google documentation. The context window is reported at 4,096 tokens total with a stricter 1,024 tokens per prompt and no memory between sessions, from the same weak source. Google's own recorder application, which uses this model, has been demonstrated summarising a 41-minute transcript, which exceeds the stated window several times over; either the first-party application chunks internally or the published context figure is wrong, and that contradiction is unresolved.
- **Translation status:** shipping on Pixel devices and, in fragmented form, across a dozen other manufacturers.
- **Blockers to use:** fragmentation is the whole story. An application must check at run time which model generation a device carries and fall back to the cloud when it carries none, and the generative speech mode is currently available on two handset models from one manufacturer.

Map-reduce chunked on-device summarisation | TRL 6

- **What it is:** The forced workaround for a 4,096-token window: split the transcript into three or four chronological segments, summarise each with an independent model pass, then pass the segment summaries back through the model for a final synthesis.
- **Key labs / groups:** application developers on both platforms; no research community owns it.
- **Audio condition and test set:** not applicable; it operates on text. No published quality evaluation of chunk-and-merge meeting summaries against single-pass ones was found.
- **Best published performance:** not established. What is established is the arithmetic that forces it: one hour of speech at roughly 150 words a minute is about 9,000 words, put at 11,000 to 13,000 tokens by two independent sources, against a 4,096-token window that must also hold the instructions and the output.
- **Translation status:** shipping in products, undocumented and unevaluated.
- **Blockers to use:** it multiplies the number of generation passes, and each pass is a fresh opportunity to invent a commitment; a decision negotiated across a chunk boundary is exactly the content a chunk-local summary loses; the repeated passes generate sustained heat on a device that has just finished recording; and no metric in the field detects the specific failure it introduces, which is a cross-chunk reference quietly dropped.

Speaker attribution

What this modality is. Diarization answers "who spoke when" by cutting the recording into speech segments, turning each segment into a numerical fingerprint of the voice, and grouping those fingerprints into as many clusters as there were people in the room. It reads the same audio the recognizer reads and outputs time intervals labelled with anonymous speaker identities; naming them is a separate step, and a legally consequential one, because a vector tied to a named identity is treated as a voiceprint in several jurisdictions while an anonymous cluster inside one recording generally is not. The parameters it produces are diarization error rate and Jaccard error rate.

State of the modality. The consensus pipeline is a clustering pass followed by a neural refinement pass, and every best system in the two most recent CHiME rounds refines with target-speaker voice activity detection (arXiv 2507.18161). Two things are settled and unhelpful for a phone. Speaker counting, not acoustics, is the dominant residual error, and nothing in a two-tap interaction tells the application how many people are in the room. And neither mobile platform exposes diarization to third-party developers at all, while Google's own recorder application has labelled speakers since its version 4.2, so an application on the phone must ship its own model to do what the phone's own application already does.

Speaker-embedding extraction with spectral clustering | TRL 8

- **What it is:** A speaker-recognition network (TitaNet, ECAPA-TDNN, ResNet-221 and similar) converts each speech segment into a fixed-length vector that depends on the voice and not the words; the vectors are then grouped by normalized maximum eigengap spectral clustering, which also estimates how many speakers there are.
- **Key labs / groups:** NVIDIA NeMo (TitaNet, the CHiME-8 NeMo baseline); Brno University of Technology (ECAPA-TDNN); the pyannote pipeline used in the ESPnet baseline.

- **Audio condition and test set:** far-field arrays, CHiME-6, DiPCo, Mixer 6, NOTSOFAR-1.
- **Best published performance:** measured as a whole-pipeline component. The baselines built on it score 28.3 percent tcpWER (NOTSOFAR-1 multi-channel) and 41.4 percent (single-channel), against 10.8 and 22.2 for the winners (arXiv 2501.17304). Diarization error rates on the CHiME-8 evaluation splits run from 10.3 percent (Mixer 6) to 60.0 percent (CHiME-6) for one baseline and 13.4 to 56.7 percent for the other (arXiv 2407.16447).
- **Translation status:** shipping in open toolkits and in commercial transcription services.
- **Blockers to use:** the clustering step must guess the number of speakers, and the two published CHiME-8 baselines mis-count in opposite directions depending on which scenario they were tuned for; many segmentation front ends assume at most three speakers in any local audio window, so a crowded room degrades them structurally; and performance falls with short segments, laughter and transient noise, which the NOTSOFAR-1 organisers identify as the conditions that hurt both tracks most.

Quantised on-device diarization | TRL 6

- **What it is:** A third-party diarization model, in practice the open pyannote pipeline, converted to Core ML and run on Apple silicon inside the application, because neither platform framework supplies speaker labels. Shipping wrappers include Argmax SpeakerKit and FluidInference FluidAudio.
- **Key labs / groups:** pyannote (the model); Argmax and FluidInference (the on-device ports).
- **Audio condition and test set:** not stated for the on-device measurements. The published speed figures name no corpus and no audio condition, and no on-device accuracy figure under meeting conditions was found.
- **Best published performance:** the only figure available is throughput, and it is measured on the wrong hardware: a speed factor of roughly 110 to 190, that is, 110 to 190 seconds of audio per second of wall clock, on an Apple M4 Pro, which is a laptop and desktop class chip. No phone-side diarization measurement of speed, memory or accuracy exists anywhere in this survey.
- **Translation status:** shipping as third-party libraries; not a platform capability.
- **Blockers to use:** memory. Parsing an hour of audio needs memory-mapped streaming to avoid materialising the whole tensor array, or the application is killed; the model weights are heavily quantised with an unmeasured accuracy cost; and it duplicates, at the application's expense, a capability the platform already runs in its own recorder.

Target-speaker voice activity detection refinement | TRL 6

- **What it is:** A second diarization pass that takes the first pass's speaker estimates as input and, for each frame of audio and each candidate speaker, predicts whether that specific speaker is talking. Because it is conditioned on speaker identity, it handles overlapping speech that a clustering pass cannot.
- **Key labs / groups:** USTC-NERCSLIP (the NSD-MS2S model); STCON; NTT. All three best CHiME-8 systems use it.
- **Audio condition and test set:** far-field arrays, CHiME-6, DiPCo, Mixer 6, NOTSOFAR-1.
- **Best published performance:** in the winning NOTSOFAR-1 system, refinement plus re-clustering took the reported development-set result from 24.611 to 22.989 percent tcpWER, and staged diarization error rates fell from 32.65 percent at initialization to about 14 percent after decoding (arXiv 2409.02041).
- **Translation status:** research; the challenge review names it as universal among the best systems.
- **Blockers to use:** it inherits the first pass's speaker count, so it corrects boundaries and not miscounting; and it adds a second full pass over the audio, which costs latency a walk-back deadline does not have.

Speaker enrolment and target-speaker conditioning | TRL 6

- **What it is:** Giving the system a stored voice embedding for a known person, so that attribution becomes supervised verification against a known vector rather than unsupervised clustering of anonymous ones. Related techniques condition the separation or recognition network on the same embedding.
- **Key labs / groups:** the target-speaker voice activity detection community; the Neuro-TM diarizer;

voice-conversion and PixIT-based systems in NOTSOFAR-1 that skip clustering entirely.
- **Audio condition and test set:** conversational corpora rather than meeting-room far-field: VoxConverse and VoxCeleb for the enrolment comparison.
- **Best published performance:** a reported diarization error rate improvement of 12.60 percent on VoxConverse and 14.01 percent on VoxCeleb against clustering-based approaches, from a single source; neither corpus is a far-field meeting recording, so the transfer to the condition that matters here is asserted rather than measured.
- **Translation status:** research, and deliberately avoided by parts of the commercial market.
- **Blockers to use:** the technique that fixes attribution is the technique that creates a biometric record. Anonymous ephemeral clustering inside one recording is the described safe harbour; tying a vector to an identity, by voice enrolment, a calendar invite or an email address, is the transition into named biometric identification and the compliance duties that attach to it. At least one incumbent states publicly that it creates no voice biometric profiles, so the market's revealed preference is anonymous labels and no enrolment.

Far-field front end

What this modality is. When someone speaks across a room, the microphone receives the direct sound plus dozens of delayed reflections, plus room noise, plus whoever else is talking. The front end is everything that happens to that signal before a recognizer reads it: choosing which microphones to use, removing the room's echo, and separating the mixture into one stream per talker. The physical basis of most of it is that a sound arriving from a particular direction reaches spatially separated microphones at slightly different times, so combining several microphones with the right delays reinforces one talker and cancels the others. This is beamforming, and it is only available to a system that can see more than one microphone signal. The family yields the separated audio that the recognition and attribution stages consume, and its quality is measured downstream in word error rate rather than directly.

State of the modality. The consensus front end for meeting audio is guided source separation, a statistically driven spatial method from 2018 that every top system in the two most recent CHiME rounds still uses; the reviewers of those challenges state plainly that current neural separation and enhancement "are still unable to reliably deal with complex scenarios and different recording setups" (arXiv 2507.18161). What is contested is whether purely neural separation can replace it. What is effectively closed is the question of whether one channel can match several: the NOTSOFAR-1 challenge measured that gap directly on identical meetings and the multi-channel track roughly halves the error of the single-channel track.

Continuous speech separation and single-channel spectral masking | TRL 6

- **What it is:** A neural model that runs over a continuous recording in a short sliding window (three seconds in the NOTSOFAR-1 baseline) and emits a fixed number of output streams arranged so that no two talkers ever overlap within one stream. Unlike guided source separation it needs no prior diarization, and single-channel variants exist; without inter-channel phase to work from, those variants must separate voices on pitch, timbre and harmonic structure alone.
- **Key labs / groups:** Microsoft (the conformer-based NOTSOFAR-1 baseline); NPU-TEA and NAIST built on it in CHiME-8 Task 2; the PixIT joint separation-and-diarization scheme is used the same way.
- **Audio condition and test set:** far-field, NOTSOFAR-1, both the single-channel condition (one channel from one device on a conference-room table) and the known-geometry multi-channel condition.
- **Best published performance:** the best separation-first system reached 18.7 percent tcpWER on the NOTSOFAR-1 multi-channel evaluation set against 10.8 percent for the guided-separation winner and 28.3 percent for the baseline; adding a weighted prediction error (WPE) dereverberation stage in front of it took one system from 30.3 to 28.9 percent tcpWER on the development set (arXiv 2501.17304). A separation-first single-device system using zero diarization scored 41.2 percent tcpWER on the NOTSOFAR-1 evaluation set.
- **Translation status:** research, with an openly released baseline model.

- **Blockers to use:** the submitted separation models were trained only on simulated audio, and the challenge organisers name the remaining simulation-to-reality gap as the likely reason separation-first systems lost to guided source separation; the three-second window denies the model longer context the organisers believe would help; and purely spectral separation fails hardest exactly where meetings are hardest, when two overlapping speakers sound alike or reverberation smears the spectral envelope. What the best single-channel spectral front end buys over none is quantified nowhere.

Guided source separation | TRL 7

- **What it is:** A statistical method that takes an existing estimate of who spoke when, then fits a complex angular central Gaussian mixture model whose parameter is a spatial covariance matrix built from the phase differences between microphones, runs expectation-maximisation to convergence to obtain per-talker masks, and applies mask-based minimum variance distortionless response (MVDR) beamforming to extract each talker. It is guided in the sense that it needs the diarization first.
- **Key labs / groups:** Paderborn University (Boeddeker and colleagues, the original implementation); used by the CHiME-7 and CHiME-8 baselines and by every top-ranked team in both.
- **Audio condition and test set:** far-field arrays: CHiME-6 (six linear arrays, 24 microphones, dinner-party sessions of 120 to 150 minutes), DiPCo (five circular arrays, 35 microphones), Mixer 6 (10 heterogeneous devices, two-speaker interviews) and NOTSOFAR-1 (one circular seven-microphone tabletop array, office meetings of about six minutes).
- **Best published performance:** pipelines built on it hold the top places in both challenges: 10.8 percent tcpWER on the NOTSOFAR-1 multi-channel evaluation set, and the best macro-averaged results across all four CHiME-8 scenarios at 33.6 percent tcpWER for STCON and 35.3 percent for NTT, against published baselines of 56.5 and 62.6 percent (arXiv 2501.17304, arXiv 2507.18161).
- **Translation status:** research code in production research pipelines; not a product component. It is computationally heavy and is run offline.
- **Blockers to use:** it is mathematically undefined on one channel, because a single stream has no spatial covariance matrix and zero inter-channel phase, so it does not apply at all in the condition a phone application occupies; it requires a diarization pass first, so its output inherits speaker-counting errors; and it is expensive enough that challenge systems using it report multi-hour test runs on server graphics processing units (GPUs).

Joint diarization and separation with neural target-speaker extraction | TRL 5

- **What it is:** A single trained model that decides who is speaking and extracts that speaker's voice at the same time, so that separation is conditioned on attribution rather than following it as a separate stage. In the winning CHiME-8 Task 2 system it is combined with, rather than replacing, guided source separation, the two providing complementary estimates.
- **Key labs / groups:** University of Science and Technology of China with iFlytek Research (USTC-NERCSLIP); related work at IACAS-Thinkit.
- **Audio condition and test set:** far-field, NOTSOFAR-1, seven-microphone circular tabletop array (multi-channel track) and one channel from one device (single-channel track), 170 blind evaluation meetings of about six minutes with four to eight speakers.
- **Best published performance:** 10.8 percent tcpWER multi-channel and 22.2 percent single-channel on the evaluation set, both first place, against a published single-channel baseline of 41.4 percent; the same system reported 14.265 and 22.989 percent respectively on the development set (arXiv 2501.17304, arXiv 2409.02041). The single-channel winner is an ensemble of three modified Whisper models and used no language-model rescoring at all.
- **Translation status:** challenge submission with a published technical report; not a product.
- **Blockers to use:** cost and complexity. The team reports separation-model training of about four days and recognition training of about 20 hours on A100 GPUs, and development-set testing of about one hour for diarization, one for separation and six for recognition on V100 or A40 GPUs (arXiv 2409.02041). Nothing in that pipeline runs on a phone, and its ensemble structure makes it far from real time even on server hardware.

Multichannel neural beamforming for speaker-attributed recognition | TRL 4

- **What it is:** A neural network that computes beamformer filter weights directly from the multichannel input and applies them to isolate a target speaker, trained end to end with the recognizer rather than designed as a separate signal-processing block.
- **Key labs / groups:** the AliMeeting line of work on multichannel speaker-attributed recognition (arXiv 2211.00511).
- **Audio condition and test set:** AliMeeting, Mandarin meetings of 15 to 30 minutes with two to four participants, recorded simultaneously as eight-channel far-field array audio and as single-channel near-field headset audio, with an overlap ratio of 42.27 percent in training and 34.76 percent in evaluation. Results are speaker-dependent character error rates.
- **Best published performance:** on the same recordings, moving from one beamformed far-field channel to the eight-channel array cut average SD-CER from 34.4 to 28.3 percent (17.7 percent relative) for the target-speaker system, from 36.8 to 30.7 for the word-level system, and from 41.2 to 33.5 for the frame-level system (arXiv 2211.00511). This is the field's second independent measurement, in another language on another corpus, that channel count and not model quality is the dominant lever.
- **Translation status:** research prototype.
- **Blockers to use:** it requires the raw multichannel signals and a known array geometry; joint training improved recognition while degrading measured separation quality, so the components cannot be tuned independently; and character error rates are not comparable with English word error rates.

Device-comparison measurement

What this modality is. Everything above produces numbers, and a claim that one capture device beats another is only meaningful if the two were recorded on the same conversation, decoded by the same models and compared with a test that accounts for the fact that speech errors are not normally distributed. This family is the accepted experimental apparatus of the field: how a parallel-capture corpus is recorded, how independent devices with no shared clock are aligned after the fact, what counts as a reference transcript, what simulated audio may and may not be used for, and how software reliability is measured in the field rather than in a lab. It produces no audio and no transcript; it produces the conditions under which a comparison is admissible.

State of the modality. The protocol is settled and unglamorous, and its constraints are what make the central question of this document expensive rather than impossible. The field accepts simulated far-field audio for training acoustic models and rejects it for evaluating a hardware claim. It rejects machine-aided reference transcription. And it rules out every repurposable public archive for a modern device comparison: podcasts are near-field, video-call recordings are irreversibly processed by vendor noise suppression, body-camera audio is the wrong environment, parliamentary recordings use push-to-talk gooseneck microphones, and the AMI and ICSI corpora were recorded on 2000s hardware.

Parallel-capture device comparison with a headset reference | TRL 9

- **What it is:** The accepted way to compare two capture devices. Both devices sit adjacent at the centre of the table; every participant wears a close-talking headset or lapel microphone whose isolated stream provides the ground truth; the two device streams are aligned to the reference after the fact by cross-correlation, in practice with the `sox` toolchain inside the CHiME-6 recipe, because independent devices share no hardware word clock and drift measurably over a 30-minute session; the identical pre-trained recognizer and front end decode both streams, because a different recognizer on either device fatally confounds the hardware test; and scoring uses concatenated minimum-permutation word error rate with significance from the NIST matched-pairs sentence-segment word error test, cross-checked non-parametrically with McNemar or Wilcoxon signed-rank.
- **Key labs / groups:** the AliMeeting topology (an eight-channel array plus per-participant headsets, 118.75 hours, 13 rooms of 8 to 55 square metres, reverberation times of 0.3 to 0.6 seconds); the NOTSOFAR-1 and CHiME-6 protocol lines; NIST, for the scoring toolkit.
- **Audio condition and test set:** far-field tabletop against near-field headset reference, by construction.

- **Best published performance:** the protocol's own published thresholds rather than an accuracy. Stable comparison is reported to need on the order of 10 to 15 hours of test material across 20 to 30 unique speakers to accumulate enough matched segments for a 95 percent confidence interval; a credible pilot is specified as at least 20 sessions of at least 15 minutes, at least 20 speakers in groups of three to five, across at least four acoustic environments. The achievable positive result is bounded: a pass is statistical superiority or a non-significant difference read as functional equivalence, and a failure is the competing device scoring significantly lower error at 95 percent confidence.
- **Translation status:** the standard method behind every published meeting corpus; directly reusable.
- **Blockers to use:** reference transcripts must be produced by humans listening de novo, because the NOTSOFAR-1 creators found that annotators accept plausible but incorrect machine guesses in noisy segments and no authoritative corpus has validated the machine-assisted shortcut; annotation is correspondingly expensive, with a labour ratio above 50 hours per audio hour asserted in one source with no citation behind it; and on a phone the operating system's own gain control and noise suppression cannot be bypassed, so the test unavoidably compares the phone as a shipped hardware-and-software ecosystem against the other device rather than capsule against capsule.

Capture-integrity telemetry instrumentation | TRL 6

- **What it is:** Instrumenting a shipping application to log the state of its own audio pipeline, so that capture reliability can be measured in the field where it actually fails. The logged records are session start and end timestamps, buffer overrun and underrun counts, operating-system interruption events, voice-activity and endpointing timings, and watchdog firings when the audio callback thread goes quiet, all as structured records with the audio payload never leaving the device.
- **Key labs / groups:** no research community; the design is assembled from real-time keyword-detection practice.
- **Audio condition and test set:** not applicable; it measures software behaviour, not acoustics.
- **Best published performance:** none, because the measurement has never been published by anyone. The specified shape for a credible study is at least 500 sessions from at least 50 real users across both platforms, with a pass at above 99.0 percent defect-free completion, a defect being any session with a buffer underrun, an unhandled interruption or a watchdog timeout that loses audio frames. No shipped meeting recorder has published a capture defect rate, so there is no category baseline to judge that bar against.
- **Translation status:** buildable today; nothing about it is research.
- **Blockers to use:** the watchdog thresholds in the one available specification are internally inconsistent, given as 2,000 milliseconds in one place and 1,000 in another, and are drawn from a keyword-detection application rather than a meeting recorder; and the residual failure it is most needed for, the zero-filled buffer that looks like a healthy recording of a silent room, may not be distinguishable from inside the application at all, which is an open question nobody has answered.

Dead Ends, Techniques That Will Not Translate

A platform limit: the channels exist and the application cannot have them

Multi-channel front-end processing inside a third-party phone application | Blocked by the operating system, not by the hardware

- **What it promises:** that because a modern phone contains three or four high-quality micro-electro-mechanical system (MEMS) microphones, an application on it could run the same spatial front end that gives a purpose-built array its advantage.
- **Why it will not translate:** neither platform releases the raw per-microphone signals to a third-party application. On iOS, input routing belongs to `AVAudioSession` and there is no request for a synchronised four-channel stream from the built-in microphones; the two-channel path that exists is reported forced to 16 kHz. On Android, the unprocessed audio source is guaranteed by the Compatibility Definition Document only for the built-in camera application, with third-party access

left to each manufacturer's hardware abstraction layer. Declaring a voice-communication mode inserts voice-processing units whose automatic gain control and noise suppression destroy the inter-channel phase that spatial separation reads. The consequence is structural: the application competes in the single-channel condition permanently, and the microphones being good does not change which condition it is in.
- **What would unstick it:** a platform release exposing raw per-microphone channels to third parties, or a named handset whose hardware abstraction layer already does so, or a measurement showing that the ambisonic capture path added in iOS 26 carries enough spatial information to drive separation. None of the three is established either way, and the second is a device-by-device empirical question nobody has answered.
- **Should we watch?** Yes. The trigger is any of those three, and it is the highest-value trigger in this document.

A physics limit: one channel cannot recover what a room mixed together

Single-channel separation of overlapping meeting speech | Stuck below the multi-channel result by roughly a factor of two

- **What it promises:** that clever software could let a device with one usable audio channel reach the transcript quality of a purpose-built microphone array.
- **Why it will not translate:** guided source separation, the technique every winning system depends on, fits a spatial covariance matrix built from the phase differences between spatially separated microphones, and one channel has no such matrix and zero inter-channel phase. Measured on the same 170 meetings, the winning single-channel system scored 22.2 percent tcpWER against 10.8 percent for the same team's multi-channel system, and the challenge organisers describe this as "a significant gap that single-channel systems are currently unable to bridge" (arXiv 2501.17304). The gap is worst exactly where meetings are hardest: debate-style overlapping speech is "particularly challenging for SC systems, but MC systems can handle it effectively". Note carefully what the gap is and is not. It is a single-device-versus-conference-array gap, measured against a seven-microphone tabletop array; a card-sized pocket recorder with two or four microphones does not have that array either, and pays some version of the same penalty.
- **What would unstick it:** a purely neural single-channel separator trained on real rather than simulated meeting audio. The challenge organisers explicitly name this as untried, because every separation model submitted was trained on simulated data alone.
- **Should we watch?** Yes. The trigger is a single-channel system reaching within a few points of the multi-channel track on NOTSOFAR-1 or its successor.

An economics limit: server-grade pipelines do not fit a phone or a walk back to the desk

Ensemble diarization-separation-recognition pipelines | Too heavy for the deadline this category promises

- **What it promises:** the best transcript quality available, around 10 to 11 percent tcpWER on office meetings.
- **Why it will not translate:** the winning system is a stack of a neural diarizer, a target-speaker refinement pass, a joint diarization-separation model, guided source separation, and an ensemble of three modified Whisper models. Its own report gives development-set testing times of about one hour for diarization, one for separation and six for recognition on V100 or A40 GPUs (arXiv 2409.02041), and the most efficiency-oriented system in either challenge, with no ensembling and no diarization refinement, still ran at a real-time factor above 2, taking more than two seconds of computation per second of audio (arXiv 2507.18161).
- **What would unstick it:** distillation of the ensemble into a single pass, or the accuracy gap narrowing enough that the cheap pipeline is good enough.

- **Should we watch?** Yes. The trigger is a challenge-competitive system published with a real-time factor below 1 on commodity hardware.

A capability limit: the platform keeps a feature for its own application

Third-party speaker diarization from a platform framework | Not exposed on either platform

- **What it promises:** that an application could get speaker labels from the same on-device stack that gives it transcription, at no cost in memory, battery or accuracy.
- **Why it will not translate:** Apple's `SpeechAnalyzer` supplies long-form transcription and voice-activity detection and no diarization; Android's recognition interfaces are tuned for command-and-control and emit no robust speaker labels for long-form overlapping conversation. Meanwhile Google's own Pixel Recorder has labelled speakers since its version 4.2. The capability exists on the phone and is withheld from third parties, so an application must ship and manage a quantised third-party model to match what the phone's own application already does.
- **What would unstick it:** either platform adding a diarization module to its public speech framework.
- **Should we watch?** Yes. The trigger is a diarization module in a public platform speech interface, which would simultaneously remove a cost and remove a differentiator.

A method dead end: simulating the device instead of recording on it

Simulated far-field audio as the evidence for a hardware claim | Accepted for training, rejected for evaluation

- **What it promises:** that a device's array geometry could be simulated, converting cheap clean speech into the far-field audio that device would have captured, and settling a microphone comparison with no new recordings.
- **Why it will not translate:** the field accepts simulated audio for training acoustic models and rejects it universally for evaluating a hardware claim, because linear mixing does not reproduce the Lombard reflex, turn-taking or non-linear hardware effects. The shortcut is closed from both ends: NOTSOFAR-1's own simulator rests on 15,000 acoustic transfer functions physically measured on the target hardware in real rooms, so building a simulator for a new device requires the recording sessions the simulation was meant to avoid.
- **What would unstick it:** a validated simulation-to-reality transfer result for device comparison specifically, which no authoritative corpus has published.
- **Should we watch?** No. Watch the parallel-capture protocol instead.

A measurement dead end: judging a transcript by the summary it produces

Downstream summarisation as a proxy for transcription quality | Measures the language model, not the audio

- **What it promises:** a way to skip word error rate and evaluate the thing users actually read.
- **Why it will not translate:** the tolerance of language models to transcription error breaks the link. Systems above 50 percent tcpWER produced summaries scoring roughly on par with systems around 11 percent, and the authors conclude that "meeting summarization may not be a good proxy evaluation task due to its high robustness to transcription errors" (arXiv 2507.18161). The summary side of that comparison is itself a model judge with self-preference and position bias, and no human read the summaries.
- **What would unstick it:** dialogue-specific summarisation metrics that respond to attribution errors, which the same authors call for, validated against human annotation.
- **Should we watch?** Yes, but as a hazard rather than an opportunity: it means a product can look good to its own users while its transcript is poor.

Overlap metrics as a measure of meeting-summary quality | They mask, and sometimes reward, the worst error

- **What it promises:** cheap automatic scoring of summaries against a human reference.
- **Why it will not translate:** expert annotation of 175 machine summaries over an error taxonomy found correlations between the standard metrics and real errors to be weak to moderate, with about a third of metric-and-error combinations either ignoring or rewarding the error. Perplexity rewards wrong speaker references at +0.44 (p <= 0.01) and LENS rewards structural disorganisation at +0.45 (p <= 0.01); in the severity analysis BLEU and QuestEval correlate positively with hallucination at +0.35 and +0.34, both significant at p <= 0.05. The only metric-error pair that behaves as a product team would want is ROUGE-1 against missing information at -0.40, and it detects only gross omission (arXiv 2404.11124).
- **What would unstick it:** a benchmark with human-annotated hallucination and misattribution labels on meeting summaries, of which the cited study is the closest thing that exists, at 175 annotated samples.
- **Should we watch?** Yes. The trigger is a public leaderboard scoring commitment fidelity rather than overlap.

Contested Evidence

Each entry below is reproduced without paraphrase from the evidence base's registry of debunked claims, minus only its internal filing reference, because the point of such an entry is to stop a later reader from citing a number already judged wrong.

- **"There is no iOS 26."** Circulated in a web-search summary during the kickoff field-map pass, while that summary was in fact describing iOS 18 results. Debunked by Apple's own developer documentation, which states iOS 26, iPadOS 26 and macOS 26 as the minimum operating system for the `SpeechAnalyzer` framework and the Foundation Models framework. A search summary is the weakest evidence this project admits and this is the clean instance of why.
- **"Apple's on-device foundation model has a 65,000-token context window."** The real figure is 4,096 tokens, a hard per-session limit covering system prompt, transcript and generated output together, from Apple's own documentation ("Managing the context window" and technote TN3193). Corroborated independently from Apple's Foundation Language Models paper (arXiv 2407.21075), where the server model is trained at sequence length 4,096 and the on-device model is distilled and pruned to match. No source supports 65,000.
- **"The acoustic penalty of a table position is paid identically by the phone and by the card stuck to the phone."** Wrong in the direction that flatters the pitch. The two devices do occupy the same point in the room, but a third-party application cannot obtain raw per-microphone channels on either platform, while the recorder has full low-level access to its own array. Same position, different number of usable channels, therefore a different achievable error rate.
- **"No vendor in the category states a wall-clock time from the end of a recording to a finished summary."** Fireflies publishes 15 to 20 minutes explicitly, and Otter publishes that for a 30-minute meeting the finalized transcript is typically available within 1.5 to 3 minutes and the summary within 2.5 to 6 minutes. What none of them publishes is a guaranteed maximum, and that narrower claim does hold across all eight vendors examined.
- **"A product built on Apple's `SpeechAnalyzer` cannot state an accuracy number."** An independent benchmark over 5,559 test utterances measured `SpeechAnalyzer` at 2.12 percent word error rate under the stated condition "clear read-aloud speech", on the 2,620-sample clean subset, against Whisper Small at 3.74 percent and Whisper Large V3 Turbo at 3.01 percent on the same data. The correction is narrow: a number can be stated, for read-aloud speech only, and the companion claim that no meeting-condition evaluation exists still stands.
- **"The platform vendors' built-in applications cannot label speakers."** False on Android: Google Pixel Recorder has shipped speaker labels since application version 4.2. The true statement is narrower and worse for us: neither platform exposes speaker diarization to a third-party developer, while Google's own first-party recorder does it.
- **"The best CHiME-8 DASR macro-averaged time-constrained speaker-attributed word error rate is 19.9 percent."** Contradicted by Cornell et al. 2025, "Trends in Distant Conversational Speech Recognition: A Review of CHiME-7 and 8 DASR Challenges", read in full: the best macro-averaged

tcpWER over the four CHiME-8 scenarios is 33.6 percent for STCON and 35.3 percent for NTT, against published baselines of 56.5 and 62.6 percent. The claim is wrong by 13.7 percentage points on a benchmark the project holds the review paper for, and every other row of that table inherits a lower trust tier as a result.

- **"The 22.2 percent single-channel NOTSOFAR-1 figure is a macro baseline produced by a 'Baseline Mamba' system."** 22.2 percent is the winning system's score on the 170-meeting blind evaluation set, the published baseline is 41.4 percent, and the winner is an ensemble of three modified Whisper models. The number survives, the label and the attribution do not.
- **"A phone application's theoretical performance ceiling is mathematically bound to roughly 22.2 percent tcpWER."** The single most dangerous sentence for this project to quote. 22.2 percent is what the world's best challenge system achieved on a purpose-built single distant microphone in NOTSOFAR-1. A consumer phone, with automatic gain control running, near-field front-end tuning, a forced 16 kHz path on the two-channel route, and an unknown placement, has no measured number at all and would plausibly be worse. 22.2 percent is a ceiling on the phone, not a prediction of it.
- **"If the phone's array geometry can be simulated, part of the microphone question needs no new recording."** The field accepts simulated far-field audio for training acoustic models and rejects it universally for evaluating a hardware claim, and NOTSOFAR-1's own simulator was built on 15,000 acoustic transfer functions physically measured on the target hardware in real rooms. The shortcut is closed in both directions.
- **"The AMI and ICSI corpora carry 212 summarised meetings."** The underlying numbers give 137 AMI scenario meetings with an abstractive summary plus 61 of ICSI's 75, which is 198 meetings carrying an abstractive summary of any kind. The count with a non-empty Actions section is necessarily lower and is measured nowhere. Retire 212.
- **"On-device transcription costs roughly 3 percent of battery per minute of audio processed."** Its own worked example does not survive: 3 percent per minute across the 60-minute meeting it names is 180 percent of the battery. The figure is impossible and the conclusion built on it cannot be carried. The real battery cost of on-device transcription is an open measurement, not a settled one; the independent figures of 25 to 40 percent per hour for a full live pipeline and 10 to 15 percent per hour for speech recognition alone are the ones to work from, and they are themselves single-blog sourced.
- **"Capture failure rates run near 0 percent on a stock Pixel and above 40 percent on a heavily optimised Xiaomi or Samsung handset."** Read in context it is an illustration of why an aggregate rate is meaningless, attached to a discussion thread, with no study, no sample size, no device list and no recording length behind it. Treat both numbers as invented for illustration and never quote them.

Four further contested items were identified inside the sources read for this document.

- **Vendor accuracy claims in this category name no test set and no audio condition.** Figures such as "95%+ accuracy" (https://otter.ai/blog/ai-notetaker-for-in-person-meetings), a claimed 98.86 percent across 58 languages, and the "62 percent average accuracy on real business audio" and "up to 30 percent speaker misattribution" figures circulating in 2026 review write-ups all name neither a corpus nor a microphone condition. Under the evidence standard used throughout this document they are not evidence of anything, in either direction.
- **One published table transposes its own labels.** In arXiv 2501.17304, Table 6 lists the single-channel winner at 10.8 percent and the multi-channel winner at 22.2 percent, which contradicts Tables 2 and 3 of the same paper and the accompanying text stating a 51 percent relative improvement for multi-channel. This document uses Tables 2 and 3: single-channel 22.2 percent, multi-channel 10.8 percent.
- **Two different accountings of the same NOTSOFAR-1 results circulate.** The challenge summary averages tcpWER across sessions; the CHiME review accumulates error statistics instead, arguing that averaging biases the figure toward easier sessions. Rankings agree; absolute values differ by 1 to 2 percentage points (arXiv 2507.18161). Numbers from the two sources must not be mixed in one comparison.
- **The circulating improvement curve for far-field recognition mixes three metrics down one col-**

umn. The series in circulation reads 43.9 to 47.7 percent on AMI single distant microphone in 2020, roughly 28 to 32 in 2022, 19.51 in 2024, 15.2 in 2025 and 14.90 in 2026, and is summarised as a 15 to 20 percent relative error reduction per year. The 2020, 2022 and 2024 points are plain or concatenated word error rates, the 2025 point is a time-constrained speaker-attributed rate and the 2026 point is a concatenated one; two of the five endpoints are sourced to records that do not read like benchmark papers. The direction of travel is credible; the slope is not quotable, and this document states no annual improvement rate.

Emerging Patterns

The channel a platform releases, not the microphone a device contains, sets the ceiling. The phone's components are not the constraint; four good micro-electro-mechanical microphones sit behind an interface that hands an application one processed channel, or two forced to 16 kHz whose inter-channel phase the platform's own voice processing destroys on request. Every downstream technique that produces the field's best numbers consumes exactly the information that interface removes. The operational consequence is that any comparison of two capture devices is meaningless unless it states how many phase-coherent channels each one released to the software that decoded it, and that a specification sheet listing microphone count is not that statement.

Spatial information is worth more than any model improvement. The multi-channel-versus-single-channel gap on identical meetings is 11.4 points absolute and 51 percent relative in speaker-attributed error and 46 percent relative in speaker-agnostic error (arXiv 2501.17304); a second measurement in another language on another corpus moves character error rate from 34.4 to 28.3 percent purely by adding channels (arXiv 2211.00511). Against that, the best model-level refinements in the same challenges move results by fractions of a point, and a large-language-model rescoring pass gained 0.5 points absolute (arXiv 2507.18161). The consequence is that where a device sits and how many of its microphones reach the software dominates everything downstream.

Real recorded audio, not more simulation, is the scarce input. Every top system in two consecutive challenge rounds used a classical spatial front end rather than a learned separator, and the organisers attribute this to the gap between simulated training audio and real rooms. The same conclusion arrives from the training side: synthetic mixtures alone gave 16.0 and 20.1 percent on two single-channel conditions, and a small fraction of real in-domain data gave 15.2 and 16.3 (arXiv 2605.15442). The consequence for anyone building here is that the bottleneck is a recording programme, not a modelling idea.

The capture layer is closing while the recognition layer opens. Every documented background-audio change on either platform tightened it, and they are worth naming rather than counting: Android 12 blocked background microphone-service starts, Android 13 forced active foreground services into a unified task manager that prevents stealth capture, Android 14 made the microphone foreground-service type mandatory with a hard exception on omission, Android 15 tightened background starts further while exempting microphone services from its six-hour cumulative timeout, and iOS 15 and 16 cemented the background-restart refusal that iOS 17 left in place. The capability that arrived later arrived somewhere else: both vendors newly exposed on-device recognition and generative summarisation to third parties, and shipped the same capability into their own free applications. The consequence is that any claim that something recently changed has to be made at the recognition layer, and that the same shipment which enables a third-party product also arms its most dangerous competitor.

Latency inverts the intuition: cloud batch is faster at thinking and slower at delivering. Data-centre silicon reaches a speed factor near 600 on the same recognizer that reaches 10 to 16 on a recent phone, roughly 40 to 60 times faster at raw inference; yet measured end-to-end delivery runs the other way, with cloud batch products timed at 10 to 20 minutes to a finished summary while on-device paths are derived at 1.5 to 5.5 minutes, because multi-tenant queueing and transport, not compute, are what the user waits for. The inversion is a property of batching rather than of the cloud: a streaming pipeline recognises while the meeting is still running and leaves only the summarisation pass to run at tap-stop. The consequence is that a latency claim is a claim about architecture and queueing, and any figure

quoted without the recording length, the processing location and whether the recognizer ran during or after the meeting is uninterpretable.

On-device inference moved from throughput-bound to context-bound. The speech half fits comfortably on phone hardware at speed factors well above real time; the language half is capped at 4,096 tokens on both platforms' on-device models, against an hour-long transcript of 11,000 to 13,000 tokens. The consequence is that the interesting engineering question about on-device output is chunking and merging without losing a cross-chunk commitment, not raw speed, and that the workaround this forces has no published quality evaluation of any kind.

The language-model layer decouples the product from the transcript, and the decoupling is measured only by machines. Summaries barely track transcript quality, so a product can deliver a readable summary over a poor transcript and its users will not notice (arXiv 2507.18161). But the tolerance is conditional: it holds when errors fall on fillers and conjunctions, and far-field acoustics smear consonants so proper nouns fail first. The load-bearing number is therefore a named-entity word error rate on far-field meeting audio, and nobody has measured one. The consequence is double: the bar for shipping something that feels good is lower than the speech literature suggests, and the bar for shipping something trustworthy is higher, because the surviving errors are invisible to the reader and concentrate in exactly the attribution the action items depend on.

Nobody publishes a guaranteed maximum, and nobody publishes a quality figure at all. Several vendors do publish typical processing times, from 1.5 to 3 minutes to a transcript up to 15 to 20 minutes to a summary, each hedged with peak-load language, and none publishes a bound. On the quality side the silence is total: no vendor in the category publishes a quantitative factuality metric for its summary or its action items, and the plausible reason is that no flattering and defensible automatic metric exists to publish. The consequence is that purchase decisions in this market are being made on something other than measured output quality, and that a new entrant cannot prove it is better any more than an incumbent can.

Every headline number in this field comes from a short meeting. The NOTSOFAR-1 sessions behind the 22.2 and 10.8 percent figures run about six minutes each, and the winning system's own paper tests no meeting longer than that; AliMeeting runs 15 to 30 minutes and AMI about half an hour. The meeting this category records runs forty minutes to two hours. Nothing in a six-minute benchmark exercises what an hour does: speaker clusters drift as voices warm and people move, a recognizer conditioned on its own previous window can carry one hallucination across a hundred windows, and the memory, storage and thermal budgets that decide whether a phone finishes the job never come under load. The consequence is that the error rates quoted throughout this document are a floor for the product's condition rather than an estimate of it, and that duration, like recording distance and channel count, has to be stated before two numbers may be compared.

The evidence base has a smartphone-shaped hole. Every corpus behind every number in this document was recorded on a headset, a ceiling or table array, smart glasses, a body-worn binaural rig or a purpose-built conference device. None was recorded on a phone, and the field names three structural reasons it avoids them: platform gain control breaks the linear inter-channel amplitude relationship blind source separation assumes, fusing streams from several phones drifts without a hardware word clock, and phone firmware is tuned for a single talker at 30 centimetres to 1 metre and treats the rest of the room as noise. The consequence is that the phone-versus-dedicated-recorder comparison cannot currently be settled from published data by anybody, in either direction, and that the only route to settling it is the parallel-capture protocol profiled above.

Also Found, Not Profiled

Every technique the sweep surfaced that did not earn a profile, alphabetical, carrying what was already established.

Technique	What it measures or does	What we know
AdaLoRA fine-tuning	Low-rank adaptation of a large recognizer on a small parameter budget	Used by NPU to fine-tune Whisper large-v2 for NOTSOFAR-1 (arXiv 2501.17304)
Atomic Content Units annotation	Decomposes a reference summary into binary atomic claims for human scoring	The most robust published human protocol; the RoSE benchmark needed over 150 hours of annotation for 22,000 summary-level judgments
Automatic speech recognition ensembling	Combines several recognizers' hypotheses	Used by most challenge submissions; one small system matched near-top results without it (arXiv 2507.18161)
cACGMM rectification	Statistical spatial mixture model refining separation masks	Used with a 120-second window in the winning NOTSOFAR-1 system (arXiv 2501.17304)
Channel selection by envelope variance	Picks which microphones to feed the front end	Introduced in the CHiME-7 baseline; the channel with lowest correlation is excluded (arXiv 2507.18161)
DOVER-Lap	Fuses several diarization outputs by voting over time	Used by the two best CHiME-8 systems for speaker counting (arXiv 2507.18161)
End-to-end neural diarization with vector clustering	One network labels active speakers per frame, then stitches windows by embedding	The Brno and Johns Hopkins NOTSOFAR-1 system reached 24.9 percent tcpWER multi-channel with no explicit separation, and its speaker-agnostic 15.6 percent beat every separation-based system on that track (arXiv 2501.17304)
Fine-tuned long-context abstractive summarisers	Encoder-decoder models adapted to thousands of input tokens	Best reported AMI ROUGE-1 56.26 and ICSI 60.7 against human summaries on clean transcripts; superseded in practice by prompted general models
Forced alignment for word timestamps	Aligns recognized words to audio to refine speaker boundaries	Improved one system's diarization but hurt its time-constrained word error rate (arXiv 2507.18161)
Global mapping files for scoring	Standardises contractions and disfluencies before a word error rate is computed	Universal expansion of contractions double-counts errors; the NIST convention requires explicit alternation syntax instead
Large-language-model rescoring of recognizer output	Rewrites recognition hypotheses using a text model	The only CHiME-8 attempt gained 0.5 points absolute macro tcpWER; a fine-tuned Llama-2-7B rescorer gave 1.42 to 2.94 percent relative (arXiv 2507.18161, arXiv 2501.17304)

Technique	What it measures or does	What we know
Matched-pairs sentence-segment word error test	Decides whether two systems differ significantly on the same audio	The field standard, implemented in the NIST scoring toolkit; cross-checked with McNemar or Wilcoxon signed-rank
MeetEval time-constrained scoring toolkit	Computes tcpWER and tcorcWER	The scoring basis of both current challenge series (arXiv 2507.18161)
Minimum variance distortionless response beamforming	Classical fixed-target beamformer	Used after continuous speech separation in the NOTSOFAR-1 baseline (arXiv 2501.17304)
Overlapping speech detection	Flags the regions where two people talk at once	Best reported 82.76 percent F1 on the AMI test set with self-supervised models and progressive training; error rate inside overlapped regions is reported nowhere
PixIT	Joint separation and diarization training scheme	Named among single-channel separation options tried in CHiME-8 Task 2; a system using it with zero diarization scored 41.2 percent tcpWER on NOTSOFAR-1 evaluation
Pragmatic action tagging	Adds dialogue-act tags such as propose or ask-clarification alongside speaker labels	Reported to cut model perplexity from 20.37 to 6.64 on civic-meeting transcripts; single source, not a meeting-product evaluation
pyannote diarization pipeline	Open segmentation and clustering toolkit	The basis of the CHiME-8 ESPnet baseline diarization, and the model most on-device ports wrap (arXiv 2507.18161)
Query-focused meeting summarisation	Answers a specific question about a meeting rather than summarising it	QMSum: 1,808 query-summary pairs over 232 meetings (arXiv 2212.08206)
ROUGE, BERTScore, METEOR, BLANC, LENS, perplexity	Automatic summary scores	Correlate weakly with annotated meeting-summary errors; a third mask or reward errors (arXiv 2404.11124)
Serialized output training	One recognizer writes several overlapping talkers in sequence with speaker-change symbols	30.7 percent average SD-CER word-level and 33.5 frame-level with multichannel fusion on AliMeeting, against 36.8 and 41.2 single-channel (arXiv 2211.00511); brittle above two or three simultaneous speakers

Technique	What it measures or does	What we know
Simulated far-field audio from measured transfer functions	Convolves clean speech with room responses measured on the target device to manufacture training audio	The NOTSOFAR-1 organisers measured 15,000 real acoustic transfer functions for a 1,000-hour simulated set; synthetic-only training gave 16.0 and 20.1 percent tcpWER against 15.2 and 16.3 with real data added (arXiv 2605.15442)
Silero voice activity detection	A 2 MB model that gates the recognizer during silence	Reported to cut total on-device compute by 40 to 60 percent in a continuous pipeline; single-source figure
Sortformer diarization	A sorted-output neural diarizer with noise and reverberation augmentation	Reported diarization error rates of 25.9 percent on AMI at 180 seconds, 22.9 on AliMeeting, 16.3 on DIHARD-III and 24.3 on MSDWild; from a table whose other rows are contested, so the tier is reduced
TF-GridNet	Neural time-frequency separation network	Named among single-channel separation options in CHiME-8 Task 2 (arXiv 2501.17304)
UniEval	Fine-tuned multi-dimensional summary evaluator	Correlation with tcpWER near zero (overall PCC -0.15) (arXiv 2507.18161)
Weighted prediction error dereverberation	Removes room echo before separation	Took one NOTSOFAR-1 system from 30.3 to 28.9 percent tcpWER on the development set (arXiv 2501.17304)
wav2vec 2.0 frame-level speaker embeddings	Speaker fingerprints from a self-supervised speech model	Part of the STCON CHiME-8 diarization improvement (arXiv 2507.18161)
Whisper word-level timestamps for segmentation	Uses the recognizer itself to segment before clustering	The NOTSOFAR-1 baseline diarization approach (arXiv 2501.17304)
Zipformer recognizer with WavLM features	Alternative recognizer architecture	NAIST reported a large inference speed-up over the baseline with better accuracy (arXiv 2507.18161)

Watchlist, Early Techniques to Monitor

Dated 2026-09-01. A fired trigger means this document is owed a refresh.

- **Third-party access to per-microphone channels on a phone.** Status: not available on either platform, with the Android unprocessed source guaranteed only for the built-in camera application and iOS releasing a virtualized microphone with a chosen polar pattern. Trigger: a platform release exposing raw per-microphone channels, a named handset whose hardware abstraction layer already does, or a measurement showing that iOS 26's ambisonic capture carries enough spatial information to drive separation. Why it matters: it decides which side of an 11.4-point measured gap an application on a phone lives on, and it is the single highest value unknown in this field for anyone

building on a general-purpose handset.

- **Single-channel separation trained on real meeting audio.** Status: untried; every separation model submitted to CHiME-8 Task 2 was trained on simulated data only, and the organisers name real-data fine-tuning as the open path. Trigger: a single-channel system within a few points of the multi-channel track on NOTSOFAR-1 or its successor. Why it matters: it is the only route by which a device with one usable audio channel could close a gap that is currently a factor of two.
- **A public meeting corpus recorded on smartphones.** Status: does not exist; a recent survey's 36-row table of meeting corpora contains no smartphone entry, and the nearest neighbours are smart glasses, body-worn binaural rigs and a commercially licensed conversational telephone set. Trigger: a challenge or corpus release whose recording devices include phones on a table alongside an array. Why it matters: it is the only way the phone-versus-dedicated-recorder question becomes answerable by anyone without funding a recording programme.
- **A published named-entity word error rate on far-field meeting audio.** Status: none exists, in any language, on any corpus. Trigger: any benchmark reporting entity-level error separately from overall word error rate on distant conversational speech. Why it matters: the whole argument that a 25 to 30 percent transcript still summarises well is conditional on where the errors land, and this is the number that would test it.
- **Growth of the on-device language-model context window.** Status: 4,096 tokens on both platforms' on-device models, with a queryable context-size property on one of them that removes the hard-coded assumption, and a next-generation on-device model announced with expanded context. Trigger: a platform release raising the window past the length of an hour-long transcript. Why it matters: it is what stands between an on-device summary of a whole meeting and an unevaluated chunk-and-merge workaround.
- **A diarization module in a public platform speech framework.** Status: withheld by both platforms while one vendor's own recorder has labelled speakers since 2022. Trigger: speaker labels appearing in a public platform speech interface. Why it matters: it would remove a per-application cost and a differentiator at the same moment, and it is the clearest available signal of how the platforms intend to treat this category.
- **A benchmark for action-item extraction with owners.** Status: none; the largest labelled set in the world is on the order of 381 items in 101 meetings of one corpus, and that count is itself contested. Trigger: a released test set with annotated commitments and assignees, or a leaderboard scoring invented commitments. Why it matters: the third of this category's three outputs currently has no accuracy number anyone can cite, in any product or paper.
- **Streaming or online diarization and separation.** Status: challenge systems are offline, multi-pass and ensemble-based, and the most efficiency-oriented one still runs at a real-time factor above 2. Trigger: a challenge-competitive system reported at a real-time factor below 1 on commodity hardware. Why it matters: it converts a research-quality transcript from an overnight batch job into something deliverable within minutes of a meeting ending.
- **A published capture-survival rate for any shipping recorder application.** Status: none, on either platform, from any vendor or any academic study. Trigger: a vendor engineering post, a platform reliability report or an instrumented field study publishing a session-completion rate with a device and operating-system breakdown. Why it matters: the whole category is bought as insurance against a lost meeting, and nobody has ever said how often the insurance fails.
- **Dialogue-aware summary metrics that penalise misattribution.** Status: G-Eval consistency is the only measure that responds meaningfully to transcription error, at PCC -0.54, and it is a model judging a model. Trigger: a metric validated against human annotation of hallucinated and misattributed commitments. Why it matters: without it, no product in this category can substantiate a quality claim about its summary or its action items, and neither can anyone auditing one.