Recapp

Literature Index - far-field meeting transcription and meeting summarisation

This vault answers what a published error rate for meeting audio actually is, under which recording condition it was measured, and what can and ...

2026-09-01

Prepared by Copundit

AI for Advanced Ideation™

This vault answers what a published error rate for meeting audio actually is, under which recording condition it was measured, and what can and cannot be measured about a summary or an extracted action item. The papers cover the CHiME-6, CHiME-7, CHiME-8 DASR and CHiME-8 NOTSOFAR-1 distant-speech challenges and several of their submitted systems, the 2026 state of the art in single-microphone speaker-attributed transcription, six audio and annotation corpora with their recording conditions and licences, surveys and methods for meeting summarisation and action-item extraction, five studies of whether a summary can be scored at all, and one platform vendor's report on the language model that runs on a phone. A paper is in this vault only because its PDF is in `literature/` and was read end to end.

Papers indexed: 30

Contents

How to use this index

Every number in this index is written with the audio condition it was measured under and the named test set it was measured on, because a word error rate from a seven-microphone tabletop array and a word error rate from a single distant microphone stream are different numbers about different products, and the project's contract forbids quoting one for the other. Blocks headed "Quotable stats" are complete sentences ready to paste into a brief; blocks headed "Direct quotes" are verbatim ground truth from the PDF, page-referenced, and are what to fall back on when a downstream document has drifted. Before citing any paper here for a mechanism claim, read its "Mechanism family" tag and its "What this paper does NOT establish" block: several of these papers are routinely mis-cited for claims they explicitly did not test.

Papers are grouped into six thematic clusters:

- What far-field meeting transcription actually scores, and under which recording condition
- How the best-scoring systems are built, and what they require to work
- What meeting summarisation is, and what it has been built and scored on
- Whether a summary or an action item can be scored at all
- What the corpora actually contain, and which of the three outputs they can supervise
- What a phone can actually run, and what it was asked to summarise

Admission gate and evidence tiers

Adopted 2026-09-01 from the studio default template, with the domain venue list filled for this project. It is project-owned from here: edit or delete it and the index will follow the edited version.

- Classify each paper as METHOD (a model or system you would actually run), THEORY (a mathematical result with no artifact) or DOMAIN (empirical or benchmark grounding).
- Venue check, class-aware. METHOD and THEORY papers need a top venue for their field, or an arXiv preprint with verifiable acceptance at one: for this project's field that list is Interspeech, the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), Computer Speech and Language, the CHiME Workshop, and for the language side ACL, EMNLP, NAACL, the Transactions of the Association for Computational Linguistics (TACL) and their Findings volumes. DOMAIN papers need a top venue of the same list, since this project's grounding discipline is speech and language technology rather than a separate clinical or physical science. A canonical foundational reference is always admissible and is tagged `[ref] foundational`.
- Citation check, age-adjusted OR-gate. Admit if citations divided by years since publication is at least 5, or if the paper is under twelve months old, has released code, and passes the venue check.
- Implementation check, METHOD papers only. A public runnable implementation is required. A METHOD paper with no released code is flagged `[C]` rather than silently indexed.
- Evidence tier by strong signals: a top venue, a citation rate that clears the gate, and an independent reproduction or an official released artifact whose numbers match the paper. `[A] proven` is two or three signals, `[B] credible` is exactly one, `[C] watch` is none or a failed gate arm. DOMAIN and THEORY papers are tiered on the first two signals only. Nothing is deleted by the gate; a `[C]` is a decision for the reader, not an eviction.

Two papers in this vault carry a flag a reader needs to see before quoting them. Niu 2024 is a METHOD paper that names no repository for its own system, so it fails the implementation arm despite a strong venue and a strong citation rate. Kumar 2022 is an arXiv preprint with no conference or journal venue and a citation rate below the floor, and its central artifact is a leaderboard compiled from other papers' reported numbers rather than from its own measurements.

Mechanism-family taxonomy

The controlled vocabulary used in this index. Two papers with different tags must never be cited interchangeably for the same claim. The resolution here is set by the project's own exposure: the charter proposes capturing an in-person meeting with a phone's own microphones, so the tags separate what was measured on one microphone stream from what was measured on a microphone array, and separate scoring the words from scoring the summary.

- **Far-field single-channel transcription benchmark** - a recognition result measured on ONE distant microphone stream, with the microphone away from the speakers, typically on a table. This is the only family whose numbers are a defensible analogue for a phone lying on a meeting-room table.
- **Far-field multi-channel transcription benchmark** - a recognition result measured on two or more synchronised distant microphones, usually a purpose-built circular or linear array. Spatial information across microphones is available to the system by construction, so these numbers do not transfer to a single device stream.
- **Near-field close-talk reference condition** - audio from a headset, lapel or binaural microphone worn by the speaker. In these papers it is used for annotation, for training data and as an upper bound, and it is never the deployment condition being evaluated.
- **Joint speaker-attributed transcription metric** - a single number scoring words, speaker attribution and, for the time-constrained variants, segmentation together: time-constrained minimum-permutation word error rate (tcpWER), concatenated minimum-permutation word error rate (cpWER), diarization-attributed word error rate (DA-WER), speaker-dependent character error rate (SD-CER).
- **Speaker-agnostic transcription metric** - a number scoring words only, with speaker identity ignored or optimally reassigned: word error rate (WER), character error rate (CER), time-constrained optimal reference combination word error rate (tcorcWER).
- **Speaker attribution metric** - a number scoring who spoke when with no words involved: diarization error rate (DER), Jaccard error rate (JER), and their components missed speech, false alarm and speaker confusion.
- **Array front-end signal processing** - guided source separation (GSS), continuous speech separation (CSS), minimum variance distortionless response (MVDR) beamforming, weighted prediction error (WPE) dereverberation, channel selection. Everything in this family except CSS requires two or more microphones.
- **Abstractive summarisation method** - a system that reads a transcript and generates new sentences that paraphrase and condense it.
- **Extractive summarisation method** - a system that selects and concatenates sentences already present in the transcript, changing no words.
- **Decision and action-item extraction** - identifying decisions taken and tasks assigned, as a task distinct from producing a general summary of the meeting.
- **Summarisation evaluation metric** - a score assigned to a generated summary: ROUGE, BLEU, METEOR, BERTScore, perplexity, BLANC, LENS, QuestEval, UniEval, G-Eval.
- **Human error-taxonomy annotation** - trained human annotators labelling generated summaries against a defined list of error types, with an inter-annotator agreement figure.
- **Corpus resource** - the dataset itself: its recordings, its microphones, its speakers and its labels.
- **Overlapped speech detection** - a frame-level binary decision that more than one person is speaking, with no words recognised and no speaker named. Scored as detection precision, recall and F1, never as an error rate, and therefore never comparable with a transcription number.
- **Contextual biasing for rare words** - adapting a recogniser at inference time to a per-user list of names, places or device labels. Scored as a relative named-entity word error rate reduction against an un-personalised baseline, so its figures are relative improvements and not absolute accuracies.
- **Downstream-task robustness to transcription error** - a measurement of how a text task degrades as the transcript's word error rate rises. The measured object is the task score, not the transcript, and the result belongs to the task, the model and the way the errors were generated.
- **Synthetic training-data generation** - building simulated multi-talker conversations from single-speaker seed audio to train a recogniser or a diarizer. A result here is about training data, not about

a deployable system, and its evaluation numbers may be produced under conditions a deployed system never has.

- **On-device language model** - a general-purpose text model built to run on a phone, described with its own architecture, training recipe and quantisation. It touches no audio and no meeting.
- **Discourse segmentation annotation** - splitting a speech transcript into elementary discourse units. It labels structure, not importance, and supervises neither a summary nor an action item.

Common confusions in this domain, to read before pasting any of these papers into a pitch:

- `Far-field multi-channel transcription benchmark` is constantly quoted as if it were `Far-field single-channel transcription benchmark`, because both are described in the source papers as "distant" or "far-field" meeting transcription. The distinguishing parameter is the number of synchronised microphone signals the system was allowed to use, and on the same 170 meetings the same winning team scored 10.8 percent with seven microphones and 22.2 percent with one (Abramovski 2025, Niu 2024). Quoting the first for a single-device product roughly halves the stated error.
- `Speaker attribution metric` is confused with `Joint speaker-attributed transcription metric` because both are called "diarization performance". The distinguishing parameter is whether words are scored at all. A system with a better DER can produce a worse tcpWER, and the winning NOTSOFAR-1 system had the worst reported DER of the leading entries (Abramovski 2025).
- `Summarisation evaluation metric` scores are read as evidence that a summary is correct. They are not: no metric in Kirstein 2024 correlates strongly with any human-annotated error type, and two of them reward errors rather than penalise them.
- `Abstractive summarisation method` benchmark scores on AMI and ICSI are read as end-to-end evidence that a meeting can be summarised from audio. They are not: those scores were computed on gold transcripts that were human-produced or human-corrected, with no recognition error present (Rennard 2023), which is the opposite of a product that starts from a phone recording.
- `Far-field single-channel transcription benchmark` is confused with `Near-field close-talk reference condition` whenever a vendor cites a dictation or read-speech error rate for a meeting product. The distinguishing parameter is where the microphone was: on the talker, or across the table from several talkers at once.
- A `Joint speaker-attributed transcription metric` computed with GROUND-TRUTH diarization supplied is not the same measurement as the same metric computed by a system that had to diarize for itself, and the two appear in the literature under the same name, tcpWER. The distinguishing parameter is stated once in a methods paragraph and never in a table caption: Polok 2026 reports 16.3 percent on NOTSOFAR-1 single-channel with oracle diarization, and Abramovski 2025 reports 22.2 percent for the challenge winner that produced its own. The first does not beat the second; they are different tasks.
- `Overlapped speech detection` scores are read as transcription quality because both are percentages on meeting audio. They are not: an F1 of 82.76 percent (Sun 2025) says how often a detector correctly notices that two people are talking, and says nothing about any word.
- `Contextual biasing for rare words` figures are relative reductions against an un-personalised baseline whose absolute error rate the paper may never state, and they were measured on single-speaker voice assistant commands (Pandey 2023), not on meeting audio.
- The word "single-channel" carries two incompatible meanings in this literature. In the NOTSOFAR-1 family it means a commercial conference-room device's output stream AFTER that device's own echo cancellation, dereverberation, beamforming and noise suppression, and the dataset authors say so explicitly (Vinnikov 2024). In LOTUSDIS it means one literal microphone element with no processing (Tipaksorn 2025). A phone application sits closer to the second than to the first.

Quick-reference table

Short cite	Venue / Year	Mechanism family	N	Headline result
Cornell 2025	Computer Speech and Language, 2025	Far-field multi-channel transcription benchmark, Joint speaker-attributed transcription metric, Summarisation evaluation metric	32 systems, 9 teams, 4 scenarios	Best macro tcpWER 33.6 percent across four scenarios; CHiME-6 stays above 30 percent for every system; summary quality correlates with tcpWER at only PCC -0.51 (G-Eval overall)
Cornell 2024	CHiME 2024 Workshop	Far-field multi-channel transcription benchmark, Corpus resource	4 scenarios, 2 baselines	Baseline macro tcpWER 56.5 percent (NeMo) and 62.6 percent (ESPnet) on eval; speaker counting named as the dominant error source
Abramovski 2025	Computer Speech and Language, 2025	Far-field single-channel transcription benchmark, Far-field multi-channel transcription benchmark	315 meetings, 30 rooms, 14 submissions	Best single-channel tcpWER 22.2 percent vs best multi-channel 10.8 percent on the same 170-meeting eval set
Niu 2024	CHiME 2024 Workshop	Far-field single-channel transcription benchmark, Array front-end signal processing	NOTSOFAR-1 eval, 170 meetings	Won both tracks: tcpWER 22.2 percent single-channel, 10.8 percent multi-channel, using an ensemble of three modified Whisper models
Shi 2023	APSIPA ASC 2023	Far-field multi-channel transcription benchmark, Joint speaker-attributed transcription metric	AliMeeting, 104.75 h train / 4 h eval / 10 h test	Best average SD-CER 28.3 percent with an 8-channel array vs 34.4 percent from the beamformed single channel

Short cite	Venue / Year	Mechanism family	N	Headline result
Rennard 2023	TACL, 2023	Abstractive summarisation method, Corpus resource	3 English meeting corpora, about 280 h total	Only three English meeting-summarisation corpora exist, all with human-produced or human-corrected transcripts; best AMI ROUGE-1 55.27
Kumar 2022	arXiv preprint, 2022	Abstractive summarisation method, Extractive summarisation method	over 40 papers surveyed	Leaderboard compiled from published results: best reported AMI ROUGE-1 56.26, best reported ICSI ROUGE-1 60.7
Kirstein 2024	Findings of EMNLP 2024	Summarisation evaluation metric, Human error-taxonomy annotation	35 QMSum meetings, 175 annotated summaries, 4 annotators	No metric correlates strongly with any error type; about a third of metric-error pairs ignore or reward the error; perplexity rewards wrong speaker references at r = +0.44
Apple 2024	arXiv technical report, 2024 (v2 2026)	On-device language model, Abstractive summarisation method	2 models; ~3B on-device	The 4,096 is the core pre-training sequence length, not a runtime context cap; on-device summariser built for emails, messages and notifications, 71.3 to 74.9 percent good-result ratio
Chen 2016	LREC 2016	Corpus resource, Decision and action-item extraction	22 ICSI meetings, 21,035 utterances	Only 318 turns (1.5 percent) carry an actionable item; binary annotator agreement kappa 0.644, action-type agreement 1.000

Short cite	Venue / Year	Mechanism family	N	Headline result
Dai 2026	arXiv preprint, 2026	Far-field single-channel transcription benchmark, Joint speaker-attributed transcription metric	AMI-SDM, AliMeeting, AISHELL-4	23.32 percent cpWER on AMI single distant microphone; 25.43 on AliMeeting first array channel; speaker count accuracy 76.67 percent on AMI
Golia 2023	NLPIR 2023 (ACM)	Abstractive summarisation method, Decision and action-item extraction	AMI corpus, human transcripts	BERTScore 64.98 and ROUGE-1 36.27 on AMI gold transcripts; adding action items raises BERTScore and lowers ROUGE; action items themselves never scored on AMI
Gong 2024	arXiv preprint, 2024	Summarisation evaluation metric	QMSum (232 meetings) plus 139 internal Zoom examples	Language-model judges correlate with human meeting-summary scores at Pearson 0.5 on QMSum and 0.11 on real Zoom data, against 0.95 on short news summaries
Huo 2026	arXiv preprint, 2026	Far-field single-channel transcription benchmark, Joint speaker-attributed transcription metric, Speaker attribution metric	AMI-SDM, AliMeeting far	24.84 percent DER and 42.55 percent cpWER on AMI-SDM; in overlapping speech every system tested was above 33 percent DER
Jones 2022	LREC 2022	Corpus resource	202 speakers, ~2,359 calls, ~840 videos	Telephone and self-recorded video only; no meetings, no transcripts, no summaries; supervises speaker identity and nothing else

Short cite	Venue / Year	Mechanism family	N	Headline result
Kalda 2024	CHiME 2024 Workshop	Far-field single-channel transcription benchmark, Array front-end signal processing	NOTSOFAR-1 dev-set-2 and eval set	41.2 percent tcpWER on eval with continuous speech separation and NO diarization; distinct from the 41.4 percent published challenge baseline
Kirstein 2024b	arXiv preprint, 2024	Summarisation evaluation metric, Human error-taxonomy annotation	170 QMSum Mistake summaries, 4 annotators	ROUGE-LSum correlates POSITIVELY with repetition (+0.26), structure (+0.23), coreference and hallucination (+0.19) errors; best evaluator reaches Spearman -0.58
Li 2026	arXiv preprint, 2026	Far-field single-channel transcription benchmark, Near-field close-talk reference condition, Joint speaker-attributed transcription metric	AMI-IHM, AMI-SDM, AliMeeting, AISHELL-4, Fisher	21.26 percent cpWER on AMI single distant microphone against 16.40 on the worn-headset mix, using an external diarization prior
F. Liu 2008	ACL 2008 short papers	Summarisation evaluation metric, Extractive summarisation method	6 ICSI meetings, 36 human and 24 system summaries	ROUGE-SU4 correlates with human judgement of system meeting summaries at Spearman 0.08; ROUGE-1 at -0.07
J. Liu 2023	ICASSP 2023	Decision and action-item extraction, Corpus resource	AMC-A 424 Chinese meetings; AMI 101 meetings	AMI's 381 action items are derived from summary links, not annotated; ICSI's 18-meeting annotation is no longer public; best English positive F1 43.12

Short cite	Venue / Year	Mechanism family	N	Headline result
Y. Liu 2023	arXiv preprint, 2023	Human error-taxonomy annotation, Summarisation evaluation metric	22,000 summary-level annotations, 28 systems, 3 written-text datasets	The 150-hour annotation cost is for news and written chat summarisation, not meetings; reference-free human rating correlates 0.926 with input-blind preference
Pandey 2023	arXiv preprint, 2023	Contextual biasing for rare words	16k + 29k + 3,558 in-house far-field voice-assistant utterances	43.9 and 57.2 percent relative named-entity WER reduction over an un-personalised RNN-T; 37.1 and 50.0 for the prior text-only adapter
Polok 2026	arXiv preprint, 2026	Synthetic training-data generation, Joint speaker-attributed transcription metric, Speaker attribution metric	8 evaluation conditions, 2 model families	Best 16.3 percent tcpWER on NOTSOFAR-1 single-channel, measured with GROUND-TRUTH diarization supplied; synthetic plus real beats real-only everywhere
Prevot 2025	SIGDIAL 2025	Corpus resource, Discourse segmentation annotation	73 French meetings, ~24 h manually annotated	Human coders agree at Cohen's kappa 0.85-0.89 on discourse-unit boundaries; a fine-tuned model reaches F 0.86, plateauing at human agreement
Shapira 2025	ACL 2025 long papers	Downstream-task robustness to transcription error, Abstractive summarisation method	QMSum 281 instances, QAConv 2,083 questions, MRDA 1,200 utterances	Summary quality drops significantly at a noise-toleration point between 0.07 and 0.3 WER, about 0.2 on average; repairing named entities is the most effective correction

Short cite	Venue / Year	Mechanism family	N	Headline result
Shon 2023	arXiv preprint, 2023	Downstream-task robustness to transcription error, Corpus resource	4 datasets, >20 recognition systems	Word error rate and ROUGE-L correlate at Pearson -0.9 on single-speaker TED talks; the paper makes no claim about ICSI action-item sparsity
Sun 2025	Interspeech 2025	Overlapped speech detection	AMI test set; AliMeeting and LibriHeavyMix for training	F1 82.76 percent for overlapped speech detection on far-field AMI; overlap is 19 percent of AMI and 42.27 percent of AliMeeting
Tipaksorn 2025	arXiv preprint, 2025	Corpus resource, Far-field single-channel transcription benchmark, Near-field close-talk reference condition	114 h, 90 sessions, 86 speakers, 9 devices at 0.12-10 m	Zero-shot WER 36.4 percent at 15 cm, 44.2 at 2 m, 96.3 at 3 m, 104.2 at 10 m; far-field macro 81.6 zero-shot to 49.5 fine-tuned; CC BY-SA 4.0
Vinnikov 2024	Interspeech 2024	Corpus resource, Far-field single-channel transcription benchmark, Far-field multi-channel transcription benchmark	Table 1: 107 + 36 + 137 = 280 meetings	Dataset paper's own table gives 280 meetings and states NO licence; dev baseline 46.8 percent single-channel vs 32.4 multi-channel vs 12.0 close-talk
Watanabe 2020	CHiME-6 challenge, 2020	Corpus resource, Far-field multi-channel transcription benchmark, Speaker attribution metric	20 dinner parties, 4 participants each	Six Kinect 4-microphone linear arrays plus worn binaural mics in real homes; automatic diarization costs about 15 absolute WER points; DiPCo is a separate corpus

Note: no duplicate files. Cornell 2025 and Cornell 2024 are two different papers by an overlapping author group, the first a journal review of both challenges and the second the CHiME-8 DASR challenge description; they are indexed separately and must not be collapsed. Abramovski 2025 and Niu 2024 report the same two headline numbers (22.2 and 10.8 percent) because Niu 2024 is the system that produced them and Abramovski 2025 is the challenge summary that ranked it.

Short-cite disambiguation, because three surnames repeat in this vault. `Kirstein 2024` is "What's under the hood: Investigating Automatic Metrics on Meeting Summarization" (Findings of EMNLP 2024); `Kirstein 2024b` is "Is my Meeting Summary Good? Estimating Quality with a Multi-LLM Evaluator" (arXiv, November 2024), by an overlapping author group. `F. Liu 2008` is Feifan Liu and Yang Liu on ROUGE and human evaluation; `J. Liu 2023` is Jiaqing Liu and colleagues on action-item detection; `Y. Liu 2023` is Yixin Liu and colleagues on the RoSE summarisation-evaluation benchmark. The three Liu papers share no authors and no domain.

Note on disagreeing published counts for NOTSOFAR-1. The dataset paper's own Table 1 gives 107 + 36 + 137 = 280 meetings with 22 training and development speakers and 10 evaluation speakers (Vinnikov 2024). The challenge summary gives 110 + 35 + 170 = 315 meetings with 22 and 13 speakers (Abramovski 2025). A third paper's corpus-comparison table repeats 315 sessions and 35 speakers, and records the licence as CC BY-NC-ND 4.0 (Tipaksorn 2025), while the dataset paper states no licence at all. Each entry reports what its own paper says; no reconciliation is asserted here.

What far-field meeting transcription actually scores, and under which recording condition

Abramovski 2025 - the single-microphone penalty, measured on the same meetings (N=315 meetings, 30 rooms)

- **Citation:** Igor Abramovski, Alon Vinnikov, Shalev Shaer, Naoyuki Kanda, Xiaofei Wang, Amir Ivry, Eyal Krupka. "Summary of the NOTSOFAR-1 Challenge: Highlights and Learnings." Computer Speech and Language, 2025. DOI: 10.1016/j.csl.2025.101796. Preprint arXiv:2501.17304v2, 9 March 2025.
- **File:** `literature/arxiv-2501.17304.pdf`
- **Links:** arXiv:2501.17304 | DOI: 10.1016/j.csl.2025.101796 | Code: challenge baseline and data hosted at chimechallenge.org/challenges/chime8/task2 (named in the paper, not verified here) | Weights: not released by these authors | Project page: https://www.chimechallenge.org/challenges/chime8/task2 (named in the paper, not verified here) | Citations: 1 (OpenAlex, published record, as of 2026-09) | Reproduced: not applicable, this is a challenge summary of independently built systems
- **Evidence tier:** [B] credible DOMAIN, published in a top venue of the field (Computer Speech and Language) but with a citation rate below the gate's floor on the counts retrievable today.
- **Mechanism family:** `Far-field single-channel transcription benchmark`, `Far-field multi-channel transcription benchmark` (+ `Speaker attribution metric`, `Array front-end signal processing`)
- **Study type:** challenge summary and comparative analysis of independently submitted systems; not a controlled experiment by these authors.
- **Population:** 315 unique recorded English meetings across 30 different rooms, each averaging six minutes, each with 4 to 8 adult participants. 22 unique speakers appear in the training and development sets and a separate 13 speakers in the evaluation set. Split as reported in the paper's Table 1: 110 training meetings (20 rooms, 22 speakers), 35 development meetings (5 rooms, 11 speakers), 170 blind evaluation meetings (13 rooms, 13 speakers). No age, sex or ethnicity breakdown is reported. A separate simulated training set of about 1000 hours was synthesised using 15,000 real acoustic transfer functions.
- **Imaging / measurement:** each meeting captured simultaneously by 4 multi-channel devices and 5 or 6 single-channel streams. The multi-channel track device geometry is known and fixed: one central microphone and six surrounding microphones, seven in total. Total audio: 55 hours single-channel and 44 hours multi-channel in training, 17.5 and 14 hours in development, 102 and 68 hours in evaluation. Transcription used a multi-judge human annotation process with machine assistance deliberately avoided, to keep annotation bias out. Per-meeting metadata hashtags mark adversarial conditions such as transient noise, laughter, debate-style overlap and a speaker standing at a whiteboard.

- **Methods / model:** two ranking metrics. tcpWER, the time-constrained minimum-permutation word error rate, scores words and speaker attribution inside a time-constrained window, so it penalises bad segmentation as well as bad recognition. tcorcWER, the time-constrained optimal reference combination word error rate, scores only the words and ignores speaker identity. Both were computed per session and then averaged over sessions to give each system's score. Submitted systems are classified by the authors into two pipeline shapes: Dia-Sep-ASR, which diarizes first, then separates the audio using the diarization output, then recognises; and CSS-ASR-Dia, which separates continuously first, then recognises, then diarizes.
- **Statistical detail:**
  - Test: none. This paper reports challenge scores and correlational observations; it runs no significance test and reports no confidence intervals.
  - Per-arm N: 9 submissions ranked in the single-channel track, 5 NOTSOFAR-1 submissions plus 3 CHiME-8 DASR submissions in the multi-channel track.
  - Effect sizes: reported as relative improvements only, for example the 51 percent relative tcpWER improvement of the winning multi-channel system over the winning single-channel system.
  - p-values: none reported.
  - Power / pre-registration: not applicable and not reported.
- **Key findings:**
  - Single-channel track, NOTSOFAR-1 evaluation set (170 meetings, about 6 minutes each, 4 to 8 speakers, one distant microphone stream): winner USTC-NERCSLIP 22.2 percent tcpWER and 17.7 percent tcorcWER; second NPU-TEA 30.0 / 25.7; third NJU-AALab 33.5 / 30.4; the challenge baseline 41.4 / 35.5; the worst ranked submission 74.1 / 37.0.
  - Multi-channel track, same evaluation set, seven-microphone tabletop array with known geometry: winner USTC-NERCSLIP 10.8 percent tcpWER and 9.5 percent tcorcWER; second STCON 13.8 / 11.6; third NTT 15.9 / 12.6; the challenge baseline 28.3 / 24.6.
  - The winning multi-channel system beat the winning single-channel system, both from the same team, by 51 percent relative on tcpWER and 46 percent relative on tcorcWER. The authors state plainly that this gap is one "that single-channel systems are currently unable to bridge" (p. 3).
  - Pipeline shape matters and the loser is the shape a single-microphone product would most naturally use. All three top multi-channel systems were Dia-Sep-ASR. The best CSS-based system reached 18.7 percent tcpWER against the winner's 10.8 percent on the multi-channel track.
  - Lower diarization error rate did not predict lower tcpWER. The winning team reported the worst DER of the leaders: 14.19 percent on dev-set-2 against 7.9 percent (STCON), 9.72 percent (NTT) and 10.4 percent (BUT/JHU). The authors hypothesise the win came from the recognition component, not the diarization one.
  - Adversarial conditions hit both tracks, but unevenly. Transient noise and laughter degrade single-channel and multi-channel systems most. Debate-style overlapping speech is specifically a single-channel problem: multi-channel systems handle it at roughly their average accuracy, single-channel systems do not.
  - Adapting the recognition model to far-field separated audio was the single most repeated win. The winning team's own ablation cut tcpWER from 16.57 percent to 9.87 percent on dev-set-2, a 40 percent relative improvement, purely by adding real multi-channel training audio and simulated LibriSpeech audio processed with oracle guided source separation into the fine-tuning set.
  - Prompting the recognition model for verbatim output matters for the metric. One team cut tcorcWER from 37.6 to 31.0 percent on the development set by filtering empty segments, merging short segments and prompting Whisper to return verbatim text rather than let it silently clean up disfluencies.
  - Language-model rescoring gave very little. A fine-tuned Llama-2-7B rescorer produced relative tcpWER reductions of 1.42 to 2.94 percent across CHiME-8 datasets, 1.87 percent on the NOTSOFAR-1 development set (19.07 to 18.72 percent). The winning system used no language-model rescoring at all.

- **Author's framing of the contribution:** > "We compare the performance of single-channel and multi-channel systems and confirm a significant gap that > single-channel systems are currently unable to bridge." (p. 3)
- **What this paper does NOT establish:**
  - It does NOT measure a phone. Every single-channel number comes from a commercially available far-field array device's monaural output stream after that device's own internal acoustic front end, or from a tabletop device channel, not from a handset microphone at an arbitrary orientation on a table.
  - It does NOT test any live or streaming condition. All submissions processed archived files offline, with no latency budget, no constraint on compute, and several using ensembles of multiple recognition models.
  - It does NOT report any latency, real-time factor, model size or cost figure for any submitted system.
  - It does NOT test summarisation, action-item extraction, or any downstream text task. tcpWER and tcorcWER are transcription metrics; this paper draws no line from either to whether a summary would be usable.
  - It does NOT establish that the reported DER figures are comparable with each other. The authors explicitly disclaim them as self-reported and computed with different tools and different collar settings.
  - It does NOT cover any language other than English, and does not test code-switching.
  - It does NOT test any meeting longer than about six minutes, so it says nothing about a one-hour meeting.
- **Direct quotes (verbatim, source ground truth):**
  - "We compare the performance of single-channel and multi-channel systems and confirm a significant gap that single-channel systems are currently unable to bridge." (p. 3)
  - "Single-channel speaker diarization and source separation is more challenging due to the absence of spatial information, leaving only voice characteristics as the signal for the models to utilize." (p. 15)
  - "Transient noise and laughter have the most significant impact on the performance of both SC and MC systems." (p. 18)
  - "Debate-style overlapping speech is particularly challenging for SC systems, but MC systems can handle it effectively, achieving accuracy comparable to their average performance on all meetings" (p. 18)
  - "Disclaimer: DER numbers mentioned in this section are self-reported by participants and could be calculated using different tools with different parameters" (p. 12)
  - "Given the observation that lower DER does not necessarily translate to lower tcpWER, and the inferior diarization accuracy of the winning system, we hypothesize ASR improvements offered by USTC played a key role in their system achieving the best tcpWER in both single-channel and multi-channel tracks." (p. 11)
- **Limitations:**
  - Author-stated: the reported diarization error rates are self-reported by the competing teams with different tools and different collar parameters, which the authors say strictly makes them incomparable and their conclusions from them unreliable; they publish them anyway as guidance.
  - Author-stated: the hypotheses about why Dia-Sep-ASR beat CSS-ASR-Dia are offered as hypotheses, not validated, because no ablation across systems was possible.
  - Inferred: the paper's own Table 6 on page 18 has the single-channel and multi-channel rows transposed relative to Tables 2 and 3, labelling the 10.8 percent result single-channel and the 22.2 percent result multi-channel. Tables 2 and 3, and the winning team's own paper (Niu 2024), make the correct assignment clear. Anyone quoting Table 6 directly will state the finding backwards.
  - Inferred: the same section's sentence "The best tcpWER on the evaluation dataset achieved by a Dia-Sep-ASR system was 18.7% (NPU-TEA)" (p. 15) contradicts its own surrounding paragraph, which is about CSS-based systems and in which NPU-TEA is classified as CSS-based.
  - Inferred: the evaluation set has only 13 unique speakers across 170 meetings, so speaker-

attribution results rest on a narrow set of voices.
  - Inferred: the meetings were partly steered by a professional actor whose job was to start and guide the conversation, which is not how an unmediated business meeting behaves.
- **Cross-references in this index:**
  - See also: Niu 2024 (the system that produced both headline numbers, with its own ablations); Cornell 2024 (the twin challenge that used the same NOTSOFAR-1 data inside a four-scenario generalisation task); Cornell 2025 (the review that recomputes NOTSOFAR-1 scores by accumulating error statistics rather than averaging per session, and adds the summarisation evaluation this paper does not attempt).
  - Contrast with: Shi 2023 (also a multi-channel versus single-channel comparison, but the single-channel condition there is a beamformed mixdown of an 8-microphone array, not a genuine single device, and the corpus is Mandarin AliMeeting rather than English office meetings). Contrast with Rennard 2023 and Kumar 2022 (summarisation quality, a different family entirely; nothing in this paper licenses a claim about summary usefulness).
- **Relevance to platform:** This is the closest published measurement to the charter's `phone-is-the-microphone` constraint, and it is the single most important entry in the vault. On identical meetings, the best system in the world scored 22.2 percent tcpWER with one distant microphone stream and 10.8 percent with a seven-microphone tabletop array, and the off-the-shelf baseline scored 41.4 percent single-channel. That is the honest bound on what a phone on a table can be expected to deliver as a verbatim transcript with correct speaker labels, and it directly stresses the charter's working note that the capability comparison against the hardware is the load-bearing technical work. The paper also names the specific asymmetry the charter asked the research round to look for: spatial information across microphones, which one phone does not have, and which is exactly what lets the multi-channel systems survive debate-style overlapping speech.
- **Quotable stats (paste-ready):**
  - "On the 170-meeting NOTSOFAR-1 evaluation set of real six-minute office meetings with four to eight speakers, the best single-channel system scored 22.2 percent time-constrained minimum-permutation word error rate, against 10.8 percent for the best multi-channel system on the same meetings (Abramovski 2025)."
  - "The published NOTSOFAR-1 baseline system, using one distant microphone stream over 170 real office meetings, scored 41.4 percent time-constrained minimum-permutation word error rate (Abramovski 2025)."
  - "Across 170 real office meetings, moving from a single distant microphone to a seven-microphone tabletop array cut the speaker-attributed word error rate by 51 percent relative and the speaker-agnostic word error rate by 46 percent relative, for the same team's systems (Abramovski 2025)."
  - "Nine teams submitted to the NOTSOFAR-1 single-channel track and only one beat 30 percent word error rate; the remaining eight scored between 30.0 and 74.1 percent (Abramovski 2025)."
  - "Adapting the recognition model to real far-field meeting audio cut the winning team's error from 16.57 to 9.87 percent on the NOTSOFAR-1 development set, a 40 percent relative improvement from training data alone (Abramovski 2025)."

Cornell 2024 - the CHiME-8 DASR challenge, and why speaker counting drives the error (N=4 scenarios)

- **Citation:** Samuele Cornell, Taejin Park, Steve Huang, Christoph Boeddeker, Xuankai Chang, Matthew Maciejewski, Matthew Wiesner, Paola Garcia, Shinji Watanabe. "The CHiME-8 DASR Challenge for Generalizable and Array Agnostic Distant Automatic Speech Recognition and Diarization." 8th International Workshop on Speech Processing in Everyday Environments (CHiME 2024). DOI: 10.21437/chime.2024-1. Preprint arXiv:2407.16447v1, 23 July 2024.
- **File:** `literature/arxiv-2407.16447.pdf`
- **Links:** arXiv:2407.16447 | DOI: 10.21437/chime.2024-1 | Code: https://github.com/chimechallenge/chime-utils (named in the paper, not verified here) | Weights: baselines built on public ESPnet and NeMo models, no new weights released by this paper | Project page: https://www.chimechallenge.org/current/task1

(named in the paper, not verified here) | Citations: 19 (OpenAlex, published record, as of 2026-09) | Reproduced: yes, both baselines are released as runnable recipes and the challenge scored 5 independent submissions against them

- **Evidence tier:** [A] proven DOMAIN, top venue of the field plus a citation rate of about 9.5 per year, comfortably above the gate's floor.
- **Mechanism family:** `Far-field multi-channel transcription benchmark`, `Corpus resource` (+ `Speaker attribution metric`, `Array front-end signal processing`)
- **Study type:** challenge description paper with baseline system results; methods paper, not a study of human subjects.
- **Population:** four English conversational scenarios. CHiME-6, dinner parties with 4 participants in home environments, sessions of about 2 to 2.5 hours, 24 sessions in total. DiPCo, dinner parties with 4 participants in a single room, sessions of about 20 to 30 minutes. Mixer 6, one-to-one interviews with 2 speakers, about 15 minutes of interview per session. NOTSOFAR-1, office meetings with 4 to 8 participants, about 6 minutes each. Reported dataset statistics: CHiME-6 evaluation 5 hours 12 minutes, 11,028 utterances, 8 speakers, 2 sessions; DiPCo evaluation 2 hours 36 minutes, 16 speakers, 5 sessions; Mixer 6 evaluation 5 hours 45 minutes, 18 speakers, 23 sessions; NOTSOFAR-1 evaluation 16 hours 29 minutes, 38,662 utterances, 12 speakers, 160 sessions. Overlapped-speech ratio over total duration on the evaluation splits: CHiME-6 26.7 percent, DiPCo 24.9 percent, Mixer 6 13.9 percent, NOTSOFAR-1 29.6 percent.
- **Imaging / measurement:** deliberately heterogeneous recording setups, which is the point of the challenge. CHiME-6 uses 6 linear Kinect arrays of 4 microphones each, 24 microphones in total. DiPCo uses 5 circular arrays of 7 microphones each, 35 in total. Mixer 6 uses 10 heterogeneous far-field devices including an Acoustimagic array, a RODE NT6 single-microphone device and a Panasonic camcorder. NOTSOFAR-1 uses one circular array of 7 microphones per session. Close-talk microphones exist in every scenario for annotation and training but are never the evaluated condition. Systems are forbidden from using domain identification or any prior knowledge of the array geometry, the microphone count or the speaker count.
- **Methods / model:** the ranking metric changed from the previous edition's DA-WER to tcpWER computed with a generous 5 second collar, macro-averaged across the four scenarios. Both baselines follow the same shape: multi-channel diarization, then guided source separation guided by that diarization, then a monaural recognition model per separated utterance. The ESPnet baseline diarizes with a multi-channel extension of the Pyannote 2.1 pipeline using a local end-to-end neural diarization module with a 5 second context and ECAPA-TDNN speaker embeddings, and recognises with a hybrid CTC/attention transformer encoder-decoder over WavLM features. The NeMo baseline diarizes with a transformer multi-scale diarization decoder in the target-speaker voice activity detection family, over multi-scale TitaNet embeddings at 3, 1.5 and 0.5 seconds, and recognises with a NeMo Conformer transducer with an n-gram language model in beam search. Both use envelope-variance channel selection ahead of guided source separation.
- **Statistical detail:**
  - Test: none. Challenge scores only; no significance testing, no confidence intervals.
  - Per-arm N: 2 baseline systems scored on 4 scenarios, each on a development and an evaluation split.
  - Effect sizes: reported as absolute error rates per scenario and as macro averages.
  - p-values: none reported.
  - Power / pre-registration: not applicable.
- **Key findings:**
  - Baseline macro-averaged results on the evaluation splits: NeMo 52.9 percent cpWER and 56.5 percent tcpWER; ESPnet 58.8 percent cpWER and 62.6 percent tcpWER. These are the numbers a competent off-the-shelf multi-channel pipeline produced, before any competitor's work.
  - Per-scenario evaluation tcpWER, ESPnet baseline: CHiME-6 99.1 percent, DiPCo 56.6 percent, Mixer 6 43.8 percent, NOTSOFAR-1 50.7 percent. NeMo baseline: CHiME-6 73.8, DiPCo 57.1, Mixer 6 23.1, NOTSOFAR-1 72.0.
  - Neither baseline is better everywhere. ESPnet is far better on NOTSOFAR-1 (50.7 against 72.0) and far worse on CHiME-6 (99.1 against 73.8), and the authors attribute this directly to speaker

    counting: the two systems make opposite counting errors on scenarios with opposite session characteristics.
  - Diarization error rates on the evaluation splits, ESPnet baseline: CHiME-6 60.0 percent, DiPCo 20.5, Mixer 6 10.3, NOTSOFAR-1 12.8. NeMo baseline: CHiME-6 56.7, DiPCo 36.2, Mixer 6 13.4, NOTSOFAR-1 47.0.
  - DER on its own does not rank the systems correctly. On CHiME-6 evaluation the two baselines' DERs are close (60.0 versus 56.7 percent) yet the ESPnet baseline has much higher speaker confusion and much worse speaker counting, producing a far worse cpWER and tcpWER.
  - Adding the short, many-speaker NOTSOFAR-1 scenario measurably degraded the same baseline on the older scenarios, because its hyperparameters had to be retuned for a different speaker-count regime.
  - The NOTSOFAR-1 data is deliberately treated one device at a time, so each device's recording of a meeting is a separate session, on the stated grounds that having just one far-field device is the practically common case.
- **Author's framing of the contribution:** > "Results from the two baseline systems, one implemented in ESPnet and one in NeMo, suggest that one of the > most challenging aspects is accurate total meeting speaker counting, as this component is the one > responsible for most downstream recognition errors." (p. 5)
- **What this paper does NOT establish:**
  - It does NOT report a single-channel result. CHiME-8 DASR is a multi-channel task; the NOTSOFAR-1 single-channel recordings appear here only as extra training material.
  - It does NOT report any participant system's score. This is the challenge description with baselines; the submitted systems are analysed in Cornell 2025.
  - It does NOT measure latency, real-time factor, memory or cost for either baseline.
  - It does NOT test a mobile or consumer device of any kind; every recording device is a purpose-built array or a professional recorder.
  - It does NOT test any downstream task. No summary, no action item, no user judgement appears anywhere.
  - It does NOT cover any language other than English.
- **Direct quotes (verbatim, source ground truth):**
  - "Results from the two baseline systems, one implemented in ESPnet and one in NeMo, suggest that one of the most challenging aspects is accurate total meeting speaker counting, as this component is the one responsible for most downstream recognition errors." (p. 5)
  - "This choice was made due to the fact that having just one far-field device is a highly practical occurring situation, and it is thus of great interest for many application scenarios." (p. 3)
  - "It is also evident that by looking only at the DER value it is not always possible to assess which one of the two baselines is better on which scenario." (p. 5)
  - "The main rationale is to discourage automatic or manual domain identification or any use of a-priori information from the scenarios." (p. 4)
  - "In C8DASR, we manually re-annotated the development set of Mixer 6." (p. 2)
- **Limitations:**
  - Author-stated: the ranking metric DA-WER used in the previous edition was in practice equivalent to cpWER and insensitive to segmentation, which is why it was replaced; results across editions are therefore not directly comparable without rescoring.
  - Author-stated: adding NOTSOFAR-1 forced a retuning of the baseline that degraded it on the other scenarios, so the baseline numbers here are worse than the previous edition's on shared scenarios.
  - Inferred: with only 2 sessions in the CHiME-6 evaluation split and 5 in DiPCo, per-scenario numbers rest on very few recordings and single-session variance is large.
  - Inferred: the evaluation data was not fully blind for three of the four scenarios, since CHiME-6, DiPCo and Mixer 6 have been public for years; only NOTSOFAR-1 was genuinely unseen.
- **Cross-references in this index:**
  - See also: Cornell 2025 (the review that reports the submitted systems' results on exactly these four scenarios and this metric); Abramovski 2025 (the twin challenge on the NOTSOFAR-1 scenario alone, with known array geometry and a single-channel track).

    - Contrast with: Niu 2024 (a system tuned to one known scenario and one known device geometry, which is the opposite of this challenge's array-agnostic rule, and which is why its numbers are much lower).
- **Relevance to platform:** This paper sets the baseline that a product would start from rather than the ceiling it might reach: a competent multi-channel pipeline built by the challenge organisers scored between 43.8 and 99.1 percent tcpWER depending on the room, with a macro average around 56 to 63 percent. It also names the failure mode that a two-tap consumer app cannot ask the user to fix, which is counting how many people are in the room, and shows that a system tuned for short many-speaker meetings gets worse at long four-speaker ones. For the charter's `finished-before-the-walk-back` constraint, note the silence: this paper reports no latency figure at all.
- **Quotable stats (paste-ready):**
    - "The CHiME-8 DASR baseline systems, using purpose-built microphone arrays, scored macro-averaged time-constrained word error rates of 56.5 and 62.6 percent across four meeting scenarios (Cornell 2024)."
    - "On the CHiME-6 dinner-party evaluation set, one of the two published CHiME-8 DASR baselines scored 99.1 percent time-constrained word error rate, meaning its output carried roughly as many errors as words (Cornell 2024)."
    - "The organisers of CHiME-8 DASR identify total meeting speaker counting as the component responsible for most downstream recognition errors across four different meeting scenarios (Cornell 2024)."
    - "Overlapped speech accounts for 29.6 percent of total duration in the NOTSOFAR-1 evaluation set and 26.7 percent in the CHiME-6 evaluation set (Cornell 2024)."
    - "The CHiME-8 DASR organisers treat a single far-field device as the practically common case, and split NOTSOFAR-1 into one session per device for exactly that reason (Cornell 2024)."

Cornell 2025 - how weakly summary quality tracks transcription quality (N=32 systems, 9 teams)

- **Citation:** Samuele Cornell, Christoph Boeddeker, Taejin Park, He Huang, Desh Raj, Matthew Wiesner, Yoshiki Masuyama, Xuankai Chang, Zhong-Qiu Wang, Stefano Squartini, Paola Garcia, Shinji Watanabe. "Recent Trends in Distant Conversational Speech Recognition: A Review of CHiME-7 and 8 DASR Challenges." Computer Speech and Language, 2025. DOI: 10.1016/j.csl.2025.101901. Preprint arXiv:2507.18161v2, 1 November 2025.
- **File:** `literature/arxiv-2507.18161.pdf`
- **Links:** arXiv:2507.18161 | DOI: 10.1016/j.csl.2025.101901 | Code: https://github.com/chimechallenge/chime-utils, plus the ESPnet recipe at github.com/espnet/espnet/blob/master/egs2/chime8_task1 and https://github.com/chimechallenge/C8DASR-Baseline-NeMo (all named in the paper, not verified here) | Weights: baselines use public WavLM, Whisper and NeMo Conformer checkpoints; no new weights released | Project page: https://www.chimechallenge.org (named in the paper, not verified here) | Citations: 8 (Semantic Scholar, merged record, as of 2026-09) | Reproduced: yes, baselines released as runnable recipes and independently run by 9 participating teams
- **Evidence tier:** [A] proven DOMAIN, top venue of the field with a citation rate of about 10 per year and publicly released tooling that reproduces the reported baselines.
- **Mechanism family:** `Far-field multi-channel transcription benchmark`, `Joint speaker-attributed transcription metric`, `Summarisation evaluation metric` (+ `Speaker attribution metric`, `Array front-end signal processing`, `Corpus resource`)
- **Study type:** review and comparative meta-analysis of two challenge editions, with new experiments by the authors on downstream meeting summarisation.
- **Population:** 32 systems submitted by 9 teams across CHiME-7 DASR (27 systems, of which 14 in the main track and 13 in the optional oracle-diarization track) and CHiME-8 DASR (5 systems), plus the CHiME-8 NOTSOFAR-1 submissions from 5 further teams for the joint comparison. Underlying speech data as in Cornell 2024: four English scenarios, CHiME-6, DiPCo, Mixer 6 and NOTSOFAR-1. Mixer 6 itself consists of 1425 recording sessions among 594 native English speakers recorded in 2009 and 2010, of which 450 interview portions were annotated for these challenges. Turn-taking statistics computed across all splits show utterances averaging roughly 2 to 4 seconds in every

scenario.
- **Imaging / measurement:** as in Cornell 2024, plus a new signal-level characterisation. The authors compute signal-to-distortion ratio (SDR) per utterance from every far-field device against the speaker's close-talk microphone, over all splits, using a 4096-tap filter and taking the maximum over nine time offsets between -8192 and +8192 samples to work around synchronisation problems in DiPCo. The reported pattern: DiPCo and NOTSOFAR-1 give relatively consistent SDR across devices, while CHiME-6 shows a large spread between the best and worst microphone for the same utterance, and Mixer 6 shows a gap of roughly 10 dB between best and worst device.
- **Methods / model:** all results are rescored with the CHiME-8 text normalisation and reported as tcpWER for comparability across editions, including the CHiME-7 submissions that were originally ranked by DA-WER. Diarization is additionally reported as Jaccard error rate with a 250 ms collar. For the downstream experiment, the authors generate meeting summaries with Gemini 2.0 Flash from each submitted system's transcript on the NOTSOFAR-1 scenario, with a fixed prompt asking for about 200 words and explicitly requiring speaker attributions to be preserved. Hypothesis speaker tags are re-mapped onto reference speaker tags using the tcpWER-optimal permutation before summarisation, so the evaluation is well defined. Eight reference summaries per session are generated from the ground-truth transcript by varying the seed, and eight hypothesis summaries per system per session, giving 64 hypothesis-reference combinations per meeting for the reference-based metrics and 8 scores per meeting for the reference-free one. Summaries are scored with G-Eval (GPT-4o based, reference-free, scoring coherence, consistency, relevance and fluency), UniEval (T5-based, reference-based) and ROUGE-1, ROUGE-2 and ROUGE-L F1.
- **Statistical detail:**
  - Test: Pearson correlation coefficient (PCC) and Spearman rank correlation (SRC) between metrics. No hypothesis tests, no p-values.
  - Per-arm N: correlations between tcpWER and summarisation metrics computed over each submitted system crossed with each session, 1760 samples in total. Correlations between tcpWER and diarization metrics computed over 22 systems (14 CHiME-7, 5 CHiME-8, 3 baselines).
  - Effect sizes with CI when reported: correlation coefficients only; standard error bars are shown on the summarisation metric figures but no numeric confidence intervals are tabulated.
  - p-values: none reported anywhere in the paper.
  - Power / pre-registration: not applicable and not reported.
- **Key findings:**
  - Best macro-averaged tcpWER over the four CHiME-8 scenarios: STCON 33.6 percent, NTT 35.3 percent, NTTsmall 44.8 percent, against 56.5 percent for the NeMo baseline and 62.6 percent for the ESPnet baseline. An oracle-diarization reference system, guided source separation plus Whisper large-v3 with ground-truth speaker segmentation, scored 37.2 percent macro. The two best submitted systems beat it.
  - CHiME-6 is unsolved. Across every system and both editions, no system got below about 30 percent tcpWER on the dinner-party scenario, and the authors put it in words: nearly one word in three is wrong, despite ensembles.
  - Diarization quality predicts recognition quality across the field but stops predicting it at the top. Macro JER against macro tcpWER over the best system per team gives PCC 0.93 and SRC 0.89. Over all 22 systems and per scenario, JER against tcpWER gives PCC 0.92 and SRC 0.92 overall, but only PCC 0.58 on Mixer 6. Restricted to the top six teams, the per-scenario correlation collapses to PCC 0.16 on CHiME-6 and 0.21 on DiPCo.
  - Speaker counting errors compound catastrophically through the pipeline, which is why JER, which accounts for speaker counting, tracks tcpWER better than DER does.
  - Guided source separation was not displaced. All participating teams in both editions relied on it; every neural separation front end that was tried was used to initialise or refine it rather than replace it, and one team reported that a continuous-separation approach was too brittle under fast turn-taking and low signal-to-noise ratio.
  - Oracle diarization buys a lot on the hard scenario and little on the easy one. Macro tcpWER improvements from switching to oracle diarization ranged up to 25 percentage points for the best CHiME-7 team, with the CHiME-6 gains dominating and Mixer 6 gains small.

  - Large language model post-processing gave almost nothing. The one team that used the optional LLM sub-track with a fine-tuned Llama-2 rescorer improved macro tcpWER by 0.5 percentage points absolute.
  - Practicality is not solved. The most efficiency-oriented system in either challenge, NTTsmall, which used no ensembling and no diarization refinement, still had a real-time factor above 2, meaning it took more than two seconds of processing per second of audio, unoptimised.
  - The downstream summarisation result, which is the paper's most consequential finding for a product: G-Eval overall correlates with tcpWER at PCC -0.51 and SRC -0.50, G-Eval consistency at -0.54 and -0.55, G-Eval relevance at -0.46 and -0.45, G-Eval coherence at -0.27 and -0.28, and G-Eval fluency at only -0.22 and -0.13. UniEval overall reaches only -0.15 and -0.18, and UniEval fluency +0.01. ROUGE-1 F1 reaches -0.34 and -0.36, ROUGE-2 -0.28 and -0.32, ROUGE-L -0.33 and -0.35.
  - Concretely: on the NOTSOFAR-1 scenario, systems with tcpWER above 50 percent produced summaries that scored roughly on par with systems near 11 percent. Only artificially corrupted transcripts showed a clear drop, and those had to be severe: randomly deleting each word with 50 percent probability gave 49.83 percent tcpWER, randomly reassigning speaker labels gave 114 percent, and doing both gave 101 percent.
  - Fluency scores saturate. G-Eval assigned high fluency even to the randomly corrupted transcripts, because every summary was written by the same modern language model.
  - The authors found it necessary to put explicit speaker-attribution instructions in the summarisation prompt; without them the summaries were too generic and no correlation with transcription quality appeared at all, on any metric.
- **Author's framing of the contribution:** > "Downstream evaluation via LLM-based meeting summarization in the NOTSOFAR-1 scenario (Section 5.3.1) > revealed a weak correlation with transcription quality." (p. 9)
- **What this paper does NOT establish:**
  - It does NOT establish that a bad transcript produces a good summary in a user's judgement. Every summarisation score here is automatic, from GPT-4o, a T5-based model or n-gram overlap; no human read any summary and rated it, and the authors say the evaluation is inherently noisy and the metrics themselves are unreliable.
  - It does NOT establish that action items survive transcription error. The prompt asked for key points, decisions, action items and significant exchanges in one 200-word summary, and no metric in the paper scores action items separately from the rest of the summary.
  - It does NOT report a single-channel result of its own. Its NOTSOFAR-1 figures are multi-channel-track systems; it says explicitly that single-channel results were excluded from that comparison.
  - It does NOT measure any latency for the summarisation stage, or any end-to-end time from end of meeting to finished output.
  - It does NOT test a phone, a handset microphone or any consumer device.
  - It does NOT cover any language other than English, a limitation the authors name.
  - It does NOT use fully blind evaluation domains: the authors state that all evaluation scenarios were known to participants in advance and that this could have biased system development.
- **Direct quotes (verbatim, source ground truth):**
  - "On the NOTSOFAR-1 scenario, even systems with over 50% time-constrained minimum permutation WER can perform roughly on par with the most effective ones (around 11%)." (p. 2)
  - "Even for the best performing systems, the tcpWER for CHiME-6 remains above 30%. This means that nearly one in three words is incorrectly transcribed, despite advances in ASR and diarization technology and the fact that almost every participant used ASR systems ensembles." (p. 45)
  - "Accurate speaker counting in the initial diarization pass is crucial, as errors propagate catastrophically through the subsequent separation and recognition stages" (p. 9)
  - "traditional ASR and diarization metrics derived from WER (e.g. tcpWER, cpWER) or DER/JER may inadequately reflect actual user experience and over-estimate the amounts of errors." (p. 61)
  - "When an application requires reliable or even acceptable transcription quality, meeting sum-

marization may not be a good proxy evaluation task due to its high robustness to transcription errors." (p. 67)
  - "we found it crucial to include explicit speaker attribution instructions in the LLM summarization prompt; without these, the produced summaries were too generic, and no appreciable correlation was observed across all summarization metrics, including G-Eval." (p. 69)
- **Limitations:**
  - Author-stated: the datasets cover English only and do not test code-switching.
  - Author-stated: no evaluation scenario was hidden from participants, so generalisation is not truly measured.
  - Author-stated: the speaker alignment step in the summarisation experiment uses the tcpWER-optimal permutation at transcript level rather than the theoretically preferable per-summary permutation, which would need 720 metric evaluations for a six-speaker meeting.
  - Author-stated: better summarisation evaluation metrics for multi-speaker dialogue do not yet exist, and building them would need new data collection with human-annotated reference summaries that explicitly ground key facts, decisions and action items.
  - Author-stated: most submitted systems rely on ensembling and are far from practical.
  - Inferred: the summarisation correlations are computed over systems whose transcripts all come from the same four-scenario data with the same annotation, so the range of transcription quality being correlated is the range these challenge systems happen to span, roughly 11 to 70 percent tcpWER, and nothing outside it is measured.
  - Inferred: using one language model (Gemini 2.0 Flash) to write all summaries and another (GPT-4o) to score them leaves an unquantified shared-model bias, which the authors acknowledge as possible for fluency.
- **Cross-references in this index:**
  - See also: Cornell 2024 (the challenge this reviews, same four scenarios, same metric); Abramovski 2025 (the twin challenge, whose systems are compared here on the NOTSOFAR-1 scenario); Niu 2024 (one of the systems analysed here); Kirstein 2024 (independent evidence, from human annotation rather than LLM judging, that summarisation metrics do not track summary errors).
  - Contrast with: Rennard 2023 and Kumar 2022 (both report ROUGE leaderboards on AMI and ICSI computed on clean human transcripts; this paper is the only entry in the vault that measures summarisation quality downstream of real recognition errors). Contrast with Shi 2023 (different corpus, different language, different metric family).
- **Relevance to platform:** Two findings here cut directly at the charter. First, the transcription bound: on a multi-room, noisy, four-speaker conversation nothing published gets below about 30 percent tcpWER, so the charter's kill criterion "the phone cannot actually do it" needs to be tested against the acoustic condition and not just against operating-system limits. Second, and more usefully, the summarisation result is the best evidence in the vault that the charter's `three-outputs` deliverable may be more robust than the transcript underneath it: a system at 50 percent tcpWER produced summaries scoring roughly on par with one at 11 percent, which means the product's promise may survive a transcript that would look bad quoted as a number. That is a genuine positive for the thesis, and it is also the exact place where the project must not overclaim, because the finding is about automatic summary scores, not about a person reading the summary. The real-time factor above 2 for the most practical system is the vault's only direct evidence bearing on `finished-before-the-walk-back`, and it points the wrong way for a research-grade pipeline.
- **Quotable stats (paste-ready):**
  - "On real office meetings, systems whose transcripts carried more than 50 percent speaker-attributed word error produced automatic summaries scoring roughly on par with systems near 11 percent error (Cornell 2025, 1760 system-session samples)."
  - "Across 1760 system-session pairs, the correlation between speaker-attributed word error rate and the best-performing automatic summary score, G-Eval overall, was only PCC -0.51; ROUGE-1 reached -0.34 and UniEval overall -0.15 (Cornell 2025)."
  - "No system in the CHiME-7 or CHiME-8 distant speech challenges scored below about 30 percent time-constrained word error rate on four-speaker dinner-party audio, despite ensembles of multiple recognition models (Cornell 2025, 32 systems from 9 teams)."

- "The best generalist meeting transcription systems reached 33.6 and 35.3 percent macro-averaged time-constrained word error rate across four different meeting scenarios, against 56.5 and 62.6 percent for the published baselines (Cornell 2025)."
- "The most efficiency-oriented system submitted to either CHiME-7 or CHiME-8 DASR still ran at a real-time factor above 2, taking more than two seconds of computation per second of meeting audio (Cornell 2025)."

How the best-scoring systems are built, and what they require to work

Niu 2024 - the system that won both NOTSOFAR-1 tracks, and what it cost to build (N=NOTSOFAR-1 dev and eval sets)

- **Citation:** Shutong Niu, Ruoyu Wang, Jun Du, Gaobin Yang, Yanhui Tu, Siyuan Wu, Shuangqing Qian, Huaxin Wu, Haitao Xu, Xueyang Zhang, Guolong Zhong, Xindi Yu, Jieru Chen, Mengzhi Wang, Di Cai, Tian Gao, Genshun Wan, Feng Ma, Jia Pan, Jianqing Gao. "The USTC-NERCSLIP Systems for the CHiME-8 NOTSOFAR-1 Challenge." 8th International Workshop on Speech Processing in Everyday Environments (CHiME 2024), pp. 31-36. DOI: 10.21437/CHiME.2024-7. Preprint arXiv:2409.02041v2, 24 October 2024.
- **File:** `literature/arxiv-2409.02041.pdf`
- **Links:** arXiv:2409.02041 | DOI: 10.21437/CHiME.2024-7 | Code: no repository for this system is named in the paper; it cites the official challenge baseline at chimechallenge.org/current/task2/baseline and the JSALT 2020 simulation scripts at github.com/jsalt2020-asrdiar/jsalt2020_simulate | Weights: not released; the system builds on the public Whisper large-v2 and large-v3 checkpoints and WavLM | Project page: none | Citations: 16 (Semantic Scholar, merged record, as of 2026-09) | Reproduced: unknown, no independent reproduction identified from the paper itself
- **Evidence tier:** [C] watch, gate-flagged METHOD. It passes the venue arm (CHiME Workshop) and the citation arm comfortably (about 8 per year), but fails the implementation arm: the paper names no public runnable implementation of its own system, only of the baseline it modified.
- **Mechanism family:** `Far-field single-channel transcription benchmark`, `Far-field multi-channel transcription benchmark`, `Array front-end signal processing` (+ `Speaker attribution metric`)
- **Study type:** challenge system description with ablation studies; technical report, not a controlled study.
- **Population:** not human subjects. Evaluated on the NOTSOFAR-1 development sets (Dev-set-1, Dev-set-2) and the blind evaluation set, that is, real office meetings of about six minutes with 4 to 8 participants recorded across many rooms. Training material as reported: the official simulated training set, NOTSOFAR-1 Train-set-1, Train-set-2 and Dev-set-1, near-field recordings from those sets used to simulate diarization and separation training data, plus 960 hours of LibriSpeech, plus MUSAN noise. The recognition training table lists thirteen data sources totalling roughly 1080 hours before augmentation, dominated by LibriSpeech at 960 hours; the real meeting material contributes 14, 16 and 10 hours per processing variant.
- **Imaging / measurement:** multi-channel track uses the NOTSOFAR-1 seven-microphone circular tabletop array; single-channel track uses one stream. Window lengths are stated throughout: the overlap detector and the joint diarization-separation model use 800 frames, which is 12.8 seconds at a 16 ms frame; the neural diarization module uses 800 frames at a 10 ms frame, which is 8 seconds; the complex angular central Gaussian mixture model rectification step uses a 120 second window with a 60 second shift; continuous speech separation uses 3 second and 8 second windows.
- **Methods / model:** a Dia-Sep-ASR pipeline. Front end: weighted prediction error dereverberation, then overlap detection, then continuous speech separation on overlapping segments and minimum variance distortionless response beamforming on non-overlapping ones, then clustering-based speaker diarization using spectral clustering over ResNet-221 speaker embeddings trained on VoxCeleb and LibriSpeech, then neural speaker diarization (NSD-MS2S, memory-aware multi-speaker embedding with a sequence-to-sequence architecture), then complex angular central Gaussian mixture model rectification, then guided source separation, then re-clustering, producing four al-

ternative diarization outputs. Separation has three variants: guided source separation initialised from neural diarization boundaries (V1), guided source separation initialised in the time-frequency domain from masks predicted by a jointly trained diarization-and-separation model (V2), and beamforming driven directly by those masks (V3). Back end: "Enhanced Whisper", the public Whisper encoder-decoder modified with WavLM features injected at an intermediate encoder layer, a ConvNeXt branch parallel to the original 1D convolutions, bias relative positional encoding instead of absolute, a sigmoid gating mechanism in the transformer block, a depthwise convolution after multi-head attention, and a mixture-of-experts component in the final encoder layer. Training adds a word-level timestamp prediction task, previous-utterance text conditioning, and a consistency regulariser the authors call Noise KLD, which feeds original and augmented audio through the model and penalises the Kullback-Leibler divergence between the two predictions. The submitted result fuses posterior probabilities from three Whisper variants across three diarization priors, nine systems in total.

- **Statistical detail:**
  - Test: none. Ablation tables of single runs; no repeated seeds, no variance, no significance testing.
  - Per-arm N: results are reported per dataset split, not per subject or per session with dispersion.
  - Effect sizes: absolute and relative error-rate reductions only.
  - p-values: none reported.
  - Power / pre-registration: not applicable.
- **Key findings:**
  - Final challenge result, NOTSOFAR-1 blind evaluation set: 22.2 percent tcpWER on the single-channel track and 10.8 percent on the multi-channel track, first place in both.
  - Development results, Dev-set-2: 14.265 percent tcpWER multi-channel and 22.989 percent single-channel for the submitted fusion systems.
  - The single-channel path is the same pipeline minus the spatial components. Removing multi-channel input roughly doubles the error on the same data: 14.265 against 22.989 percent on Dev-set-2, 10.8 against 22.2 on the evaluation set.
  - Real meeting training data is worth more than architecture. With a fixed Whisper large-v3 and oracle separation, training on the original datasets gave 16.57 percent tcpWER on Dev-set-2; adding real multi-channel meeting audio processed with guided source separation gave 12.07 percent; adding all sets plus simulated LibriSpeech gave 9.87 percent. Architecture changes on the same data moved the recognition number from 8.46 to 7.50 percent on Dev-set-1.
  - The same holds for diarization: adding real NOTSOFAR training data to the simulated data cut DER on Dev-set-1 from 21.51 to 16.52 percent.
  - Neural separation improved on classical continuous separation but did not displace guided source separation. On Train-set-1 with a fixed recognition model, continuous separation with beamforming gave 26.68 percent tcpWER, adding overlap detection gave 25.14, the joint diarization-separation model gave 20.62, adding better speaker boundaries gave 20.29, adding matched diarization segmentation gave 19.95, and extending the window from 3 to 8 seconds gave 17.47.
  - Better diarization did not reliably mean better recognition. Between the reported stages, DER improved from 23.43 to 21.07 percent while tcpWER got worse, from 12.87 to 14.13 percent.
  - Compute, as reported by the authors: diarization training about 88 hours; the joint diarization-separation model about 4 days; recognition training about 20 hours; and decoding all of Dev-set-2 takes about 1 hour for diarization, about 1 hour for separation and about 6 hours for recognition, on A100 GPUs for training and V100 or A40 GPUs for testing.
- **Author's framing of the contribution:** > "In the NOTSOFAR-1 challenge, our system achieved the tcpWERs of 22.2% and 10.8% in the single-channel and > multi-channel tracks of the evaluation set, respectively, winning first place in both tracks." (p. 5)
- **What this paper does NOT establish:**
  - It does NOT establish that this accuracy is reachable in real time, on a device, or at consumer cost. The submitted system is a nine-way fusion of three modified Whisper models over three diarization priors, and the authors report roughly eight GPU-hours to decode one development set.

  - It does NOT establish that the single-channel result generalises off the NOTSOFAR-1 device. The single-channel audio here comes from the challenge's own device streams, not from an arbitrary handset.
  - It does NOT release the system. No repository, no weights and no inference recipe for the submitted system are named anywhere in the paper.
  - It does NOT test any language other than English, nor any meeting longer than about six minutes.
  - It does NOT evaluate transcription usefulness downstream: no summary, no action item, no user study.
  - It does NOT report variance. Every ablation number is a single run.
- **Direct quotes (verbatim, source ground truth):**
  - "Our system attained a Time-Constrained minimum Permutation Word Error Rate (tcpWER) of 14.265% and 22.989% on the CHiME-8 NOTSOFAR-1 Dev-set-2 multi-channel and single-channel tracks, respectively." (p. 1)
  - "In the NOTSOFAR-1 challenge, our system achieved the tcpWERs of 22.2% and 10.8% in the single-channel and multi-channel tracks of the evaluation set, respectively, winning first place in both tracks." (p. 5)
  - "As shown in the table, adding real training data in NOTSOFAR [18] leads to a substantial improvement (DER from 21.51% to 16.52%)." (p. 4)
  - "For diarization, the training requires approximately 88 hours, and testing all sentences in Dev-set-2 takes about 1 hour." (p. 4)
  - "For ASR, training requires about 20 hours, and testing all sentences in Dev-set-2 consumes about 6 hours." (p. 4)
  - "At the same time, stage 2 shows a relatively effective improvement in DER compared to stage 1, but the recognition performance actually become worsens." (p. 4)
- **Limitations:**
  - Author-stated: multi-stage optimisation of the diarization priors was needed, and improvements in diarization error rate did not consistently translate into recognition improvements.
  - Inferred: the submitted results are ensembles of nine systems, so none of the headline numbers is attributable to a single deployable model.
  - Inferred: no code or weights are released, so nothing here is directly reusable or independently checkable.
  - Inferred: the ablations are on different splits (Train-set-1, Dev-set-1, Dev-set-2) with different fixed components, so the individual contributions do not add up to the final result and cannot be recombined.
  - Inferred: single-run numbers with no variance make small differences, such as 14.265 against 14.286 percent between fusion variants, uninterpretable.
- **Cross-references in this index:**
  - See also: Abramovski 2025 (the challenge summary that ranked this system and cross-checks its ablations); Cornell 2025 (which analyses this system against the generalist array-agnostic entries and notes it dropped the forced-alignment trick its predecessor used).
  - Contrast with: Shi 2023 (also a speaker-attributed multichannel system, but on Mandarin AliMeeting with a character error rate metric and an 8-channel array, and not comparable number for number). Contrast with Cornell 2024 (whose rules forbid exactly the device-specific tuning this system relies on).
- **Relevance to platform:** This is what the 22.2 percent single-channel number actually costs: a nine-model ensemble, roughly eight GPU-hours to decode one development set, no released code, and a front end that was still built around array processing even in the single-channel path. For the charter's `finished-before-the-walk-back` constraint, this paper is the clearest evidence that best-published accuracy and minutes-after-the-meeting delivery are, today, two different products. It also gives the project its most actionable lever: the largest single win reported anywhere here came from fine-tuning on real meeting audio matched to the deployment condition, 16.57 to 9.87 percent, which is a data problem a small team can attack without inventing anything.
- **Quotable stats (paste-ready):**
  - "The system that won both CHiME-8 NOTSOFAR-1 tracks scored 22.2 percent time-

constrained word error rate on one distant microphone stream and 10.8 percent on a seven-microphone array, over the same 170 real office meetings (Niu 2024)."
- "The winning NOTSOFAR-1 system is a fusion of nine configurations, three modified Whisper models across three diarization priors, and its authors report about six hours of GPU time to decode a single development set (Niu 2024)."
- "Fine-tuning the recognition model on real meeting audio matched to the recording condition cut error from 16.57 to 9.87 percent on the NOTSOFAR-1 development set, a larger gain than any architecture change the same team reported (Niu 2024)."
- "Adding real meeting recordings to simulated training data cut the diarization error rate from 21.51 to 16.52 percent on the NOTSOFAR-1 development set (Niu 2024)."
- "In the winning NOTSOFAR-1 system's own ablation, a diarization stage that improved diarization error rate from 23.43 to 21.07 percent made the transcription worse, from 12.87 to 14.13 percent (Niu 2024)."

Shi 2023 - what the extra microphones bought on AliMeeting (N=104.75 h train, 4 h eval, 10 h test)

- **Citation:** Mohan Shi, Jie Zhang, Zhihao Du, Fan Yu, Qian Chen, Shiliang Zhang, Li-Rong Dai. "A Comparative Study on Multichannel Speaker-Attributed Automatic Speech Recognition in Multi-party Meetings." Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2023. DOI: 10.1109/APSIPAASC58517.2023.10317185. Preprint arXiv:2211.00511v3, 2 March 2023.
- **File:** `literature/arxiv-2211.00511.pdf`
- **Links:** arXiv:2211.00511 | DOI: 10.1109/APSIPAASC58517.2023.10317185 | Code: not released; no repository named in the paper | Weights: not released | Project page: none | Citations: 4 (OpenAlex, published record, as of 2026-09) | Reproduced: unknown, no independent reproduction identified from the paper itself
- **Evidence tier:** [C] watch, gate-flagged METHOD. APSIPA ASC is not on this project's top-venue list, the citation rate is about 1.3 per year against a floor of 5, and no public implementation is named. Admitted because it is the only entry in the vault that reports a same-corpus microphone-count comparison with a speaker-attributed metric, but every number from it should be treated as single-source.
- **Mechanism family:** `Far-field multi-channel transcription benchmark`, `Joint speaker-attributed transcription metric` (+ `Array front-end signal processing`, `Corpus resource`)
- **Study type:** methods paper with a controlled comparison of three architectures against three single-channel predecessors.
- **Population:** the AliMeeting corpus: 104.75 hours for training, 4 hours for evaluation and 10 hours for testing, in Mandarin Chinese. Each session is a 15 to 30 minute discussion among 2 to 4 participants. No age or sex breakdown is reported. A further 50 hours of simulated mixed training data was generated from the near-field recordings.
- **Imaging / measurement:** 8-channel far-field audio recorded by a microphone array, referred to as Ali-far, plus single-channel near-field audio from each participant's headset, referred to as Ali-near. The single-channel comparison condition is Ali-far-bf, the 8-channel far-field audio reduced to one stream by a constrained differential beamformer, so it is a beamformed mixdown of an array and not an independent single-microphone recording. Input features are 80-dimensional log Mel filterbank coefficients computed with the ESPnet toolkit; the modelling units are 4950 common Chinese characters. Oracle voice activity detection and oracle sentence segmentation are used, and speaker profiles are given.
- **Methods / model:** three multichannel speaker-attributed designs, each a multichannel version of a published single-channel one. MC-FD-SOT combines frame-level diarization by target-speaker voice activity detection with serialized output training recognition, fusing channels with multi-frame cross-channel attention in the recogniser and channel-level cross-channel attention in the diarizer. MC-WD-SOT does word-level diarization, fusing channels with frame-level cross-channel attention before the speech encoder. MC-TS-ASR jointly trains a neural beamformer (embedding and beamforming network, EaBNet) for target-speaker separation with the recognition model. Model

sizes are matched: 45 M parameters for the multichannel serialized-output recogniser and 44 M for the comparison conformer recogniser. Training uses Adam, an initial learning rate of 1.5e-4 halved on plateau for the front end, and 1e-3 with 10000 warmup steps for joint fine-tuning; the joint loss is an equal-weighted (lambda = 0.5) combination of a negative scale-invariant signal-to-noise ratio separation loss and a joint attention-CTC recognition loss.

- **Statistical detail:**
  - Test: none. Single-run error rates per split, plus averages of the two splits.
  - Per-arm N: Eval and Test splits only, 4 hours and 10 hours respectively.
  - Effect sizes: relative error-rate reductions, for example 18.8 percent, 16.6 percent and 17.7 percent relative for the three designs against their single-channel counterparts.
  - p-values: none reported.
  - Power / pre-registration: not applicable.
- **Key findings:**
  - Speaker-dependent character error rate (SD-CER), averaged over Eval and Test. Single-channel beamformed baselines: frame-level diarization 41.2 percent, word-level diarization 36.8 percent, target-speaker separation 34.4 percent. Multichannel versions: 33.5, 30.7 and 28.3 percent respectively. Relative reductions 18.8, 16.6 and 17.7 percent.
  - The best system, MC-TS-ASR, scored 30.4 percent on Eval and 27.5 percent on Test with an 8-microphone array, against 32.5 and 35.1 percent for the beamformed single channel.
  - Speaker-agnostic recognition is far better than speaker-attributed recognition on the same audio. The multichannel serialized-output recogniser reached 17.3 percent character error rate on Eval and 18.4 percent on Test, against 30.7 percent average SD-CER for the best word-level-diarization system built on it. Roughly half the speaker-attributed error is attribution, not transcription.
  - Better separated audio does not mean better recognition. Joint fine-tuning cut SD-CER by 34.2 percent relative on Eval (46.2 to 30.4 percent) and 37.2 percent on Test (43.8 to 27.5 percent), while the scale-invariant signal-to-noise ratio of the separated audio fell from 17.24 dB to -1.62 dB on Eval and 16.95 to -1.99 dB on Test.
  - Estimated diarization can beat oracle speaker labels for downstream recognition. Using the diarizer's output rather than oracle labels gave 30.4 against 33.2 percent SD-CER on Eval, and discarding diarized utterances shorter than 0.9 seconds improved it further to 28.2 percent.
  - The diarization component itself is strong on this corpus: 2.26 percent DER on Eval and 2.98 percent on Test for the cited multichannel target-speaker voice activity detection system, which is an order of magnitude better than the English meeting scenarios in the CHiME papers.
- **Author's framing of the contribution:** > "To our knowledge, this work is the first attempt to the multichannel SA-ASR problem in real meeting > scenarios." (p. 4)
- **What this paper does NOT establish:**
  - It does NOT compare an array against a genuine single microphone. Its single-channel condition is a beamformed mixdown of the same 8-microphone array, which already carries spatial processing, so the reported 16 to 19 percent relative gains understate the array-versus-one-microphone gap rather than measure it.
  - It does NOT test English. AliMeeting is Mandarin and the metric is a character error rate, which is not numerically comparable to a word error rate.
  - It does NOT run without oracle information. Oracle voice activity detection, oracle sentence segmentation and enrolled speaker profiles are supplied, none of which a two-tap consumer recording has.
  - It does NOT test more than 4 speakers, and does not test the 5-to-8-speaker regime that the NOTSOFAR-1 papers found hardest.
  - It does NOT report latency, streaming operation, model footprint or inference cost.
  - It does NOT release code or weights.
- **Direct quotes (verbatim, source ground truth):**
  - "Experimental results on the AliMeeting corpus reveal that our proposed multichannel SA-ASR models can consistently outperform the corresponding single-channel counterparts in terms of the speaker-dependent character error rate (SD-CER)." (p. 1)

  - "The AliMeeting corpus contains 104.75 hours data for training (Train), 4 hours for evaluation (Eval) and 10 hours for testing (Test)." (p. 3)
  - "the proposed MC-TS-ASR method, which calculates the filter weights of a neural beamformer and then performs filter-and-sum to obtain the fused spectrum, results in the best performance among all MC-SA-ASR systems and an average relative SD-CER reduction of 17.7% compared to SC-TS-ASR (from 34.4% to 28.3%)." (p. 4)
  - "This also shows that a better front-end signal quality (e.g., in SI-SNR) does not necessarily mean a better back-end ASR performance in practice." (p. 4)
  - "The obtained CERs on Eval and Test sets of AliMeeting corpus are 17.3% and 18.4%, respectively." (p. 3)
- **Limitations:**
  - Author-stated: joint fine-tuning improves recognition at the cost of separation quality, implying the recognition model is not actually consuming separated speech in the way the design intends.
  - Inferred: the single-channel comparison condition is a beamformed array mixdown, so this paper cannot answer the question a single-device product needs answered.
  - Inferred: oracle segmentation and enrolled speaker profiles make the setting substantially easier than a real recording, and the paper does not report a no-oracle condition.
  - Inferred: single-run results with no repeated seeds and no variance, on a 4-hour evaluation split.
  - Inferred: the strong 2.26 and 2.98 percent diarization error rates are quoted from a cited prior system, not measured in this paper.
- **Cross-references in this index:**
  - See also: Niu 2024 (a later system in the same family, target-speaker diarization plus guided separation plus a large pre-trained recogniser, on English office meetings); Abramovski 2025 (which reaches the same conclusion that spatial information is what single-channel systems lack).
  - Contrast with: Abramovski 2025 (whose single-channel track is a genuine single device stream, so its array-versus-single-stream gap of 51 percent relative is the number to quote, not this paper's 16 to 19 percent against a beamformed mixdown). Contrast with Cornell 2024 (which forbids the array-specific tuning this paper depends on).
- **Relevance to platform:** The useful contribution to this project is not the array comparison but the decomposition: on the same far-field audio, speaker-agnostic character error was 17.3 to 18.4 percent while speaker-attributed character error was about 30 percent, which says that roughly half the error a user would see in an attributed transcript is attribution error rather than misheard words. For a product whose deliverable is a summary and action items with names attached, that split matters more than the headline rate. The paper also shows, twice, that improving an intermediate metric can make the end result worse, which is a caution against optimising diarization or separation quality as a proxy for product quality.
- **Quotable stats (paste-ready):**
  - "On the AliMeeting corpus of 15-to-30-minute meetings with 2 to 4 participants, the best multichannel speaker-attributed system scored 28.3 percent speaker-dependent character error rate with an 8-microphone array, against 34.4 percent for the beamformed single-channel version (Shi 2023)."
  - "On the same far-field meeting audio, speaker-agnostic character error was 17.3 and 18.4 percent while speaker-attributed character error was around 30 percent, so roughly half the attributed error came from getting the speaker wrong rather than the words (Shi 2023, AliMeeting, 14 hours of test audio)."
  - "Joint fine-tuning cut speaker-attributed character error by 34 to 37 percent relative while the signal-to-noise ratio of the separated audio fell from about 17 dB to below zero, showing that cleaner audio and better recognition are not the same objective (Shi 2023)."
  - "Discarding diarized segments shorter than 0.9 seconds improved speaker-attributed character error from 30.4 to 28.2 percent on the AliMeeting evaluation set (Shi 2023)."

Dai 2026 - hierarchical speaker classification bolted onto a speech language model (N=3 meeting corpora)

- **Citation:** Yuhang Dai, Haopeng Lin, Jiale Qian, Ruiqi Yan, Hao Meng, Hanke Xie, Hanlin Wen, Shunshun Yin, Ming Tao, Xie Chen, Lei Xie, Xinsheng Wang. "Joint Learning Global-Local Speaker Classification to Enhance End-to-End Speaker Diarization and Recognition." arXiv:2603.25377v2, 27 March 2026. DOI: not provided. Affiliations as printed: Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science, Northwestern Polytechnical University; Soul AI Lab, China; X-LANCE Lab, Shanghai Jiao Tong University, China.
- **File:** `literature/arxiv-2603.25377.pdf`
- **Links:** arXiv:2603.25377 | DOI: not provided | Code: not released; the paper names no repository for GLSC-SDR | Weights: not released | Project page: none | Citations: not retrieved | Reproduced: unknown
- **Evidence tier:** [C] watch METHOD. The paper states no conference or journal acceptance, names no public implementation of its own system, and its citation count is unverified, so it fails the implementation arm of this index's gate and clears none of the three strong signals.
- **Mechanism family:** `Far-field single-channel transcription benchmark`, `Joint speaker-attributed transcription metric` (+ `Speaker-agnostic transcription metric`)
- **Study type:** methods paper with an ablation study; three benchmark corpora, no human subjects of the authors' own.
- **Population:** three existing meeting corpora, used as published. AMI (English), AliMeeting (Mandarin) and AISHELL-4 (Mandarin). No speaker demographics, ages or recruitment details are reported, because the authors recorded nothing. The clustering stage that builds the training labels used HDBSCAN over speaker embeddings with clusters merged when centroid cosine similarity exceeded 0.75, and the paper's own sweep found 200 clusters optimal on AliMeeting.
- **Imaging / measurement:** the audio condition is stated explicitly and matters. "For AMI, we adopt the Single Distant Microphone (SDM) subset. For AliMeeting and AISHELL-4, we use the first channel from the 8-channel far-field microphone array recordings." (p. 3). So every number in this paper is a far-field SINGLE-CHANNEL result: one microphone at a distance, no array processing, no spatial information. Audio was cut into turn groups, where consecutive speaker turns with no temporal gap are merged into one segment. Training-label construction discarded any single-speaker segment whose word error rate exceeded 30 percent or that had more than two insertion errors, on the assumption that such segments contain crosstalk.
- **Methods / model:** backbone is Qwen2.5-Omni-7B, a large audio-language model, adapted with Low-Rank Adaptation (LoRA) at rank 8 applied to the AudioEncoder, Aligner and Thinker modules, learning rate 1e-4, 30 epochs for the main results and 3 for the ablations. Speaker embeddings come from ERes2Net. The training objective is a weighted sum of a speaker-classification loss and a diarization-and-recognition loss. The classification labels are hierarchical: a global label from unsupervised clustering of speaker embeddings, and a local label distinguishing individuals inside each cluster.
- **Statistical detail:**
  - Test: none. No significance test, no confidence interval and no repeated-seed variance is reported.
  - Per-arm N: one run per configuration per corpus, on the corpora's own published test splits.
  - Effect sizes: reported as absolute and relative differences in error rate only.
  - p-values: none reported.
  - Power / pre-registration: not applicable and not reported.
- **Key findings:**
  - On AMI Single Distant Microphone, one far-field microphone: GLSC-SDR scored 17.49 percent word error rate and 23.32 percent concatenated minimum-permutation word error rate (cpWER), against 20.23 / 27.16 for the same backbone with ordinary supervised fine-tuning and 32.25 / 49.86 for the un-tuned Qwen2.5-Omni.
  - On AliMeeting, first channel of the 8-microphone far-field array: 20.09 percent word error rate and 25.43 percent cpWER, against 20.22 / 26.77 for supervised fine-tuning and 29.21 / 43.64 un-tuned.

- On AISHELL-4, same single-channel far-field condition: 21.36 percent word error rate and 23.49 percent cpWER, against 23.83 / 26.34 fine-tuned and 31.46 / 46.28 un-tuned.
- The gap between cpWER and word error rate, which the authors call delta-cp and use as a diarization proxy, was 5.83 on AMI-SDM, 5.34 on AliMeeting and 2.13 on AISHELL-4. On AliMeeting this is an 18.47 percent relative reduction in attribution error against their fine-tuned baseline.
- Speaker count accuracy was 76.67 percent on AMI-SDM, 83.26 on AliMeeting and 90.96 on AISHELL-4, so the system got the number of people in the room wrong in roughly one AMI segment in four.
- The reported comparison systems on AMI-SDM cpWER are TagSpeech 42.55, Gemini-2.5-pro 34.78, Gemini-3-pro 26.91, VibeVoice-ASR 28.82, all above GLSC-SDR's 23.32.
- Ablation: global-only classification scored 30.81 cpWER on AliMeeting and direct flat speaker-label classification 30.80, against 29.93 for the hierarchical version, at 3 training epochs.

- **Author's framing of the contribution:** > "We propose GLSC-SDR, a fully end-to-end joint training paradigm that tightly integrates speaker > classification with the SDR task." (p. 1)
- **What this paper does NOT establish:**
  - It does NOT measure a phone. AMI-SDM is a fixed tabletop microphone in an instrumented meeting room, and the AliMeeting and AISHELL-4 channels are one element of a purpose-built circular array.
  - It does NOT establish a latency, a real-time factor, a memory footprint or a cost. A 7-billion-parameter audio-language model with LoRA adapters is described, and no inference-time figure of any kind is given.
  - It does NOT report timestamps or a diarization error rate. delta-cp is a derived proxy, not a measured DER, and the paper never scores "when" anything was said.
  - It does NOT test on NOTSOFAR-1, so its numbers cannot be lined up against this index's other single-channel results (Abramovski 2025, Niu 2024) without changing the test set.
  - It does NOT test summarisation or action-item extraction; nothing here says a summary would improve.
  - It does NOT release code or weights, so the numbers cannot be checked.
- **Direct quotes (verbatim, source ground truth):**
  - "For AMI, we adopt the Single Distant Microphone (SDM) subset. For AliMeeting and AISHELL-4, we use the first channel from the 8-channel far-field microphone array recordings." (p. 3)
  - "Real-world conversations are characterized by rapid turn-taking, speaker overlaps, and acoustic variability, making robust speaker attribution a critical challenge" (p. 1)
  - "the vanilla Qwen2.5-Omni model exhibits poor performance on AliMeeting, confirming that unoptimized LALMs lack the necessary discriminative power for multi-speaker environments." (p. 4)
  - "while task-specific fine-tuning (Qwen2.5-Omni-SFT) improves results, its speaker attribution remains inferior to our proposed method. This indicates that simple fine-tuning is insufficient for robust speaker classification" (p. 4)
  - "cpWER evaluates the joint correctness of speech recognition and speaker attribution. It is calculated by first concatenating all utterances belonging to the same speaker and then finding the optimal permutation between the predicted and reference speakers that minimizes the WER" (p. 3)
- **Limitations:**
  - Author-stated: cluster granularity is a sensitive hyper-parameter; too few clusters crowd the groups and too many create redundancy, and the optimum was found empirically at 200 on one corpus.
  - Inferred: no venue, no released code, no seeds and no confidence intervals, and several of the headline margins over the authors' own fine-tuned baseline are under one absolute point.
  - Inferred: the training-label pipeline discards any segment above 30 percent word error rate as presumed crosstalk, which also discards the hardest genuine speech, so the speaker classifier is trained on an easier distribution than the one it is tested on.
  - Inferred: comparison numbers for Gemini-2.5-pro and Gemini-3-pro are marked in the paper's own Table 1 as converted from another paper, not measured by these authors.

- **Cross-references in this index:**
  - See also: Huo 2026 (the same AMI-SDM and AliMeeting far-field single-channel condition, same audio-language-model family, and the system this paper positions itself against); Li 2026 (same corpora, same cpWER metric, better AMI-SDM score with a diarization prior); Niu 2024 (the far stronger single-channel result, on a different corpus and with a purpose-built pipeline rather than a language model).
  - Contrast with: Abramovski 2025 (single-channel numbers on NOTSOFAR-1, a different corpus and a time-constrained metric, so 23.32 here and 22.2 there are not the same measurement); Shi 2023 (multi- channel, where the array does the work this paper asks a classifier to do).
- **Relevance to platform:** This is one of three 2026 papers in the vault that a later document could quote as "the state of the art for single-microphone meeting transcription". Its best far-field single-channel attributed error rate is 23.32 percent on AMI and 25.43 percent on AliMeeting, both far from any figure under 20, and its speaker-count accuracy on English far-field audio is 76.67 percent. For the charter's `phone-is-the-microphone` constraint, the reading is that the 2026 frontier for one distant microphone is still roughly a fifth to a quarter of words wrong once speaker labels are scored, and that the frontier is reached with a 7-billion-parameter model whose runtime cost the paper never states.
- **Quotable stats (paste-ready):**
  - "On the AMI Single Distant Microphone condition, one far-field tabletop microphone, the best system in Dai 2026 scored 23.32 percent concatenated minimum-permutation word error rate and 17.49 percent speaker-agnostic word error rate."
  - "On AliMeeting, using only the first channel of the 8-microphone far-field array, Dai 2026 reports 25.43 percent concatenated minimum-permutation word error rate."
  - "An un-adapted general-purpose audio-language model scored 49.86 percent attributed word error rate on AMI's single distant microphone, against 23.32 percent after task-specific training (Dai 2026)."
  - "Speaker count accuracy on English far-field meeting audio was 76.67 percent, so the system named the wrong number of participants in about one segment in four (Dai 2026, AMI-SDM)."

Huo 2026 - explicit timestamps inside a language model, and what that costs in words (N=2 meeting corpora)

- **Citation:** Mingyue Huo, Yiwen Shao, Yuheng Zhang. "TagSpeech: End-to-End Multi-Speaker ASR and Diarization with Fine-Grained Temporal Grounding." arXiv:2601.06896v2, 13 July 2026. DOI: not provided. Affiliations as printed: University of Illinois Urbana-Champaign; Johns Hopkins University.
- **File:** `literature/arxiv-2601.06896.pdf`
- **Links:** arXiv:2601.06896 | DOI: not provided | Code: https://github.com/AudenAI/Auden/tree/main/examples/tag stated in the paper to hold code, model and demo (named in the paper, not verified here) | Weights: stated as available at the same repository (named in the paper, not verified here) | Project page: none | Citations: not retrieved | Reproduced: unknown
- **Evidence tier:** [C] watch METHOD. A public implementation is named, which clears the implementation arm, but the paper states no conference or journal acceptance and its citation count is unverified, so it clears none of the three strong signals.
- **Mechanism family:** `Far-field single-channel transcription benchmark`, `Joint speaker-attributed transcription metric`, `Speaker attribution metric`
- **Study type:** methods paper with ablations; two benchmark corpora, no data collection by the authors.
- **Population:** AMI (English) and AliMeeting (Mandarin), used as published. No participant demographics are reported. Training used at most 100 hours of audio per dataset, and models were trained and evaluated separately within each dataset.
- **Imaging / measurement:** "We adopt the most challenging far-field settings, using the Single Distant Microphone (SDM) subset for AMI and the far-field subset for AliMeeting, taking the first channel from the 8-channel recordings." (p. 5). One microphone, at a distance, in both cases. Audio was segmented into utterance groups, where adjacent utterances with no duration gap are treated

as one sample. Diarization error rate is reported with a strict 0-second collar and overlap regions included, which is a harsher setting than the 0.25-second collar most challenge papers use.

- **Methods / model:** a frozen Qwen2.5-Instruct-7B language model receives two separate audio streams, a semantic stream from an encoder fine-tuned with Serialized Output Training and a speaker stream from a speaker encoder, both wrapped in XML-style tags. Only two projector modules, each a two-layer perceptron with temporal downsampling, are trained. Interleaved numeric time-anchor tokens are inserted into the input at a fixed frame interval to give the model an explicit timeline. Metrics: diarization error rate (DER), concatenated minimum-permutation word error rate (cpWER), global word error rate (gWER, all utterances concatenated chronologically with speaker identity ignored), speaker count accuracy (SCA), and a fail rate counting samples whose output could not be parsed.
- **Statistical detail:**
  - Test: none. No significance test, no confidence interval, no repeated seeds.
  - Per-arm N: one run per configuration, on each corpus's published test split.
  - Effect sizes: relative improvements only, for example "approximately 28% relative improvement on AMI and 36% on AliMeeting" in diarization error rate over end-to-end baselines.
  - p-values: none reported.
  - Power / pre-registration: not applicable and not reported.
- **Key findings:**
  - AMI Single Distant Microphone: TagSpeech scored 24.84 percent DER, 42.55 percent cpWER, 31.62 percent gWER, 70.01 percent speaker count accuracy, 1.27 percent fail rate.
  - AliMeeting far-field, first channel: 22.13 percent DER, 33.84 percent concatenated character error rate, 25.42 percent global character error rate, 81.63 percent speaker count accuracy, 3.04 percent fail rate.
  - The cascaded baseline of Pyannote 3.1 plus Whisper-large-v3 beat TagSpeech on AMI diarization, 23.05 against 24.84 percent DER, but lost on AliMeeting, 26.13 against 22.13.
  - The general-purpose audio-language baselines were far worse at diarization on this data: Gemini-2.0-flash 45.16 percent DER on AMI-SDM with a 31.90 percent fail rate on AliMeeting; Qwen2.5-Omni-7B 34.71; Qwen3-Omni-30B-A3B-Instruct 39.06.
  - Splitting DER by region shows where the win is. In non-overlapping speech the cascade was much better, 5.77 against 15.93 percent on AMI. In overlapping speech TagSpeech was better, 46.20 against 54.80 on AMI and 33.68 against 53.11 on AliMeeting. Every system was above 33 percent DER in overlap.
  - Serialized Output Training of the semantic encoder was the single largest ablation effect on AliMeeting: DER 16.12 against 18.02 pretrained, concatenated character error 25.99 against 30.82, fail rate 3.04 against 5.76.
  - Cross-lingual zero-shot transfer between AMI and AliMeeting collapsed transcription entirely, with concatenated word error rate above 100 percent, while diarization stayed near 20 percent DER.
- **Author's framing of the contribution:** > "In contrast, we define the task as the explicit prediction of transcription (what), speaker label (who), > and precise timestamps (when), enabling a truly unified formulation of multi-speaker ASR and diarization." > (p. 2)
- **What this paper does NOT establish:**
  - It does NOT measure a phone, and it does not measure any consumer device. AMI-SDM and the AliMeeting array channel are fixed room microphones.
  - It does NOT produce a competitive transcript. Its own AMI-SDM cpWER of 42.55 percent is worse than the cascaded Pyannote-plus-Whisper baseline on the same audio in the companion table of Li 2026, and the authors say plainly that content recognition "is not the best".
  - It does NOT report latency, throughput, memory or real-time factor for a frozen 7-billion-parameter model with two audio encoders.
  - It does NOT test overlapping speech beyond four concurrent speakers, and its separation capacity is bounded by the design.
  - It does NOT evaluate summarisation or action items.
  - It does NOT report significance or variance, so the 1.8-point DER gap it loses to the cascade on AMI and the 4.0-point gap it wins on AliMeeting are single measurements.

- **Direct quotes (verbatim, source ground truth):**
  - “We adopt the most challenging far-field settings, using the Single Distant Microphone (SDM) subset for AMI and the far-field subset for AliMeeting, taking the first channel from the 8-channel recordings.” (p. 5)
  - “While our content recognition performance (cpWER/gWER) is not the best, it remains consistently strong and reliable across both languages” (p. 6)
  - “To evaluate ‘who spoke when’, we report DER with a strict 0 s collar, capturing precise timestamp accuracy, with overlap region included.” (p. 5)
  - “Fail Rate denotes the proportion of samples where the model generates invalid or unparseable output, often due to formatting violations or hallucination loops. High fail rate implies lower reliability of other metrics due to survivor bias.” (p. 5)
  - “A clear trend emerges: missed speech is the dominant source of diarization error across all settings” (p. 7)
- **Limitations:**
  - Author-stated: the cascade baseline remains better in non-overlapping regions, which the authors attribute to end-to-end multi-task modelling introducing time-boundary error against a specialised voice-activity detector.
  - Author-stated: transcription collapses under cross-lingual shift.
  - Inferred: the fail-rate column is itself a finding. Gemini-2.0-flash failed on 31.90 percent of AliMeeting samples, and the paper computes the other metrics on survivors only, so several baseline numbers in the comparison table describe an easier subset than the full test set.
  - Inferred: no venue, no seeds, no confidence intervals.
- **Cross-references in this index:**
  - See also: Dai 2026 (same two corpora, same far-field single-channel condition, builds directly on this paper and beats its cpWER); Li 2026 (same AMI-SDM condition, reaches 21.26 percent cpWER by feeding in an external diarization prior); Niu 2024 (the far-field single-channel result from a purpose-built pipeline rather than a language model).
  - Contrast with: Abramovski 2025 and Cornell 2025 (time-constrained tcpWER on different corpora; cpWER here carries no temporal constraint, so the numbers are not interchangeable); Sun 2025 (scores overlap DETECTION on AMI, not overlap transcription).
- **Relevance to platform:** This paper is the clearest published statement of what happens when one model is asked for the transcript, the speaker and the time from one distant microphone: diarization at 24.84 percent error and an attributed transcript at 42.55 percent error on English meeting audio. For the charter, the regional breakdown is the useful part. In non-overlapping speech a conventional cascade was three times better at diarization; in overlapping speech every system tested was above 33 percent DER. Overlap, which is exactly what a phone on a table faces in a four-person meeting, remains the unsolved region regardless of architecture.
- **Quotable stats (paste-ready):**
  - “On AMI’s single distant microphone, TagSpeech scored 24.84 percent diarization error rate and 42.55 percent concatenated minimum-permutation word error rate (Huo 2026).”
  - “In overlapping speech on AMI’s single distant microphone, the best diarization error rate reported was 46.20 percent, against 5.77 percent for the same cascade in non-overlapping speech (Huo 2026).”
  - “A general-purpose commercial audio-language model failed to produce parseable output on 31.90 percent of AliMeeting far-field samples (Huo 2026, Gemini-2.0-flash).”
  - “Fine-tuning the semantic encoder with Serialized Output Training cut concatenated character error on AliMeeting far-field audio from 30.82 to 25.99 percent (Huo 2026).”

Kalda 2024 - the NOTSOFAR-1 entry that skipped diarization entirely (N=NOTSOFAR-1 dev-set-2 and eval set)

- **Citation:** Joonas Kalda, Tanel Alumae, Severin Baroudi, Martin Lebourdais, Herve Bredin, Ricard Marxer. “ToTaTo System Descriptions for the NOTSOFAR1 Challenge.” Proceedings of the 8th International Workshop on Speech Processing in Everyday Environments (CHiME 2024), 6 September 2024, Kos, Greece, pages 23-25. DOI: 10.21437/CHiME.2024-5. Affiliations as printed: Tallinn

University of Technology, Estonia; IRIT, Universite de Toulouse, CNRS; Universite de Toulon, Aix Marseille Univ, CNRS, LIS.

- **File:** `literature/isca-kalda24-totato.pdf`
- **Links:** DOI: 10.21437/CHiME.2024-5 | Code: not released for this system; the paper names third-party components only (faster-whisper, audiomentations, SpeechBrain ECAPA-TDNN) | Weights: not released | Project page: none | Citations: not retrieved | Reproduced: unknown
- **Evidence tier:** [C] watch METHOD. Published at the CHiME Workshop, which is on this index's venue list, but it is a two-page technical report with no released implementation of the submitted system, so it fails the implementation arm.
- **Mechanism family:** `Far-field single-channel transcription benchmark`, `Array front-end signal processing` (continuous speech separation only, which is the one member of that family that runs on a single microphone), `Joint speaker-attributed transcription metric`
- **Study type:** challenge system description; four submitted systems compared on one corpus.
- **Population:** the NOTSOFAR-1 development set 2 and the blind evaluation set. See Vinnikov 2024 and Abramovski 2025 in this index for the corpus itself. No population of the authors' own.
- **Imaging / measurement:** single-channel track only. Each meeting is one internally processed audio stream from a commercially available conference-room device. The authors did not enter the multi-channel track.
- **Methods / model:** Whisper large-v3 fine-tuned for one epoch on the challenge's in-domain single-channel training data, speed-perturbed at factors 0.9 and 1.1, plus voice-converted data generated by converting the close-talk microphone audio of each meeting to random LibriSpeech speakers with k-nearest-neighbour voice conversion and a HiFi-GAN vocoder, then remixed, reverberated with real room impulse responses from OpenSLR 28, noise-augmented, and distorted with air absorption, MP3 encode-decode and bit crushing. Effective batch size 64 segments, learning rate 1e-5, 50 warm-up steps, AdamW, SpecAugment. Inference used faster-whisper with voice activity detection to strip silence. Four submissions differ only in how speakers are separated and attributed: PixIT continuous speech separation with NO diarization; NVIDIA NeMo word-based diarization; PixIT diarization; and a pyannote 3.1 segmentation model with w2v-BERT 2.0 fine-tuned by LoRA plus ECAPA-TDNN clustering, called SseRiouSs.
- **Statistical detail:**
  - Test: none. Challenge scores only.
  - Per-arm N: four submitted systems, scored on dev-set-2 and the official evaluation set.
  - Effect sizes: absolute error-rate differences only.
  - p-values: none reported.
  - Power / pre-registration: not applicable; the evaluation set was blind at submission time.
- **Key findings:**
  - The best of the team's systems on the evaluation set was pixit-whisper: continuous speech separation by PixIT and NO speaker diarization at all, scoring 41.2 percent tcpWER and 29.2 percent tcorcWER.
  - This is a DIFFERENT system from the challenge's own baseline, which scored 41.4 percent tcpWER and 35.5 percent tcorcWER on the same evaluation set using the baseline separation front end and NeMo diarization. The two numbers are within 0.2 points of each other by coincidence, not by construction.
  - Development-set and evaluation-set rankings disagreed sharply. On dev-set-2 the best system was whisper-nemo at 37.6 percent tcpWER, which fell to 42.4 on eval; pixit-whisper was 39.8 on dev and 41.2 on eval. The worst reversal was whisper-sseriouss, 44.2 on dev and 70.4 on eval.
  - Voice-conversion augmentation helped on dev and hurt on eval. Whisper large-v3 alone: 39.8 dev / 42.6 eval tcpWER. Plus in-domain fine-tuning: 38.7 / 41.5. Plus voice conversion: 37.6 / 42.1. The authors state flatly that "voice conversion turned out to be actually not helpful on evaluation data" (p. 1).
  - Diarization error rate for the SseRiouSs system on dev-set-2 ranged 21.0 to 23.5 percent depending on which subset the clustering thresholds were tuned on, with speaker confusion the largest component in every case (10.4 to 11.8 percent).
  - The authors report an unexplained discrepancy between their own tcorcWER scores and the

organisers' official scores for what they say is the same system.

- **Author's framing of the contribution:** > "It performs CSS through the recently proposed PixIT framework which allows to skip speaker diarization > altogether." (p. 1)
- **What this paper does NOT establish:**
  - It does NOT show that skipping diarization is better. The no-diarization system won on the evaluation set and lost on the development set to a system that used NeMo diarization.
  - It does NOT measure a phone. The single-channel condition is a commercial conference-room device's own processed output stream.
  - It does NOT report latency, model size at inference, or compute cost, although it does note that all systems are one-pass with no ensembling.
  - It does NOT test the multi-channel track, so it says nothing about the array-versus-single-device gap.
  - It does NOT evaluate summarisation or action items.
  - It does NOT establish that voice conversion is a useful augmentation; on the blind set it was harmful.
- **Direct quotes (verbatim, source ground truth):**
  - "Our best-performing system utilizes a Whisper model fine-tuned on the challenge dataset and voice-converted data. It performs CSS through the recently proposed PixIT framework which allows to skip speaker diarization altogether. It achieves a tcpWER score of 41.2% on the challenge evaluation set." (p. 1)
  - "However, voice conversion turned out to be actually not helpful on evaluation data." (p. 1)
  - "There is a relatively large discprepancy between tcorcWER scores between the last line in Table 2 and the whisper-nemo scores in Table 1, which correspond to the same system. Currently we don't know the reason behind this" (p. 1)
  - "PixIT can be used for CSS with the added benefit that each file-level separated source corresponds to a single speaker. Thus speaker attribution requires no additional diarization." (p. 2)
- **Limitations:**
  - Author-stated: the discrepancy between self-computed and organiser-computed scores for the same system is unexplained.
  - Inferred: a two-page technical report with no ablation isolating PixIT from the fine-tuned recogniser, so the 41.2 percent cannot be attributed to the separation method alone.
  - Inferred: all four systems share one recogniser, so the comparison is only of the attribution stage.
  - Inferred: the dev-to-eval reversals are large enough (up to 26 points) that the ranking on 35 development meetings carries very little information about the 170 evaluation meetings.
- **Cross-references in this index:**
  - See also: Abramovski 2025 (the challenge summary that reports the 41.4 percent baseline this system nearly ties, and ranks all nine single-channel submissions); Vinnikov 2024 (the corpus and the baseline system this entry is measured against); Niu 2024 (the winning single-channel system at 22.2 percent).
  - Contrast with: Polok 2026 (also reports NOTSOFAR-1 single-channel numbers, but conditioned on ground-truth diarization, which this paper does not do); Shi 2023 (multi-channel).
- **Relevance to platform:** This entry settles a distinction the project needs. The 41.2 percent figure is a research team's best submitted single-channel system with no diarization stage, and the 41.4 percent figure is the challenge organisers' own published baseline with diarization; they are two different systems that happen to land 0.2 points apart. Neither is a floor a product could ship against. The other useful lesson for the charter is the dev-to-eval instability: a system that looked 5 points better on 35 meetings was 1 point worse on 170, and one system degraded by 26 points, so any internal benchmark this project builds on a handful of meetings will not predict field behaviour.
- **Quotable stats (paste-ready):**
  - "A NOTSOFAR-1 single-channel submission that performed continuous speech separation and no speaker diarization at all scored 41.2 percent time-constrained speaker-attributed word error rate on the evaluation set (Kalda 2024)."
  - "The NOTSOFAR-1 published single-channel baseline scored 41.4 percent time-constrained speaker-attributed word error rate on the same evaluation set; it is a different system from the

41.2 percent result, not the same one (Kalda 2024, Abramovski 2025)."
- "Fine-tuning Whisper large-v3 on in-domain far-field meeting audio moved evaluation-set attributed word error rate from 42.6 to 41.5 percent, while an additional voice-conversion augmentation that helped on the development set pushed it back up to 42.1 (Kalda 2024)."
- "Across four systems sharing one recogniser, development-set and evaluation-set rankings disagreed by up to 26 absolute points on the same corpus (Kalda 2024)."

Li 2026 - handing the language model a diarization prior instead of asking it to diarize (N=5 test sets)

- **Citation:** Li Li, Ming Cheng, Weixin Zhu, Yannan Wang, Juan Liu, Ming Li. "DM-ASR: Diarization-aware Multi-speaker ASR with Large Language Models." arXiv:2604.22467v1, 24 April 2026. DOI: not provided. Affiliations as printed: School of Artificial Intelligence, Wuhan University; School of Computer Science, Wuhan University; Tencent Ethereal Audio Lab, Tencent, Shenzhen; The Chinese University of Hong Kong, Shenzhen.
- **File:** `literature/arxiv-2604.22467.pdf`
- **Links:** arXiv:2604.22467 | DOI: not provided | Code: not released; the paper names third-party components (Whisper-large-v3-turbo, Qwen3, Gemma3, MeetEval) but no repository of its own | Weights: not released | Project page: none | Citations: not retrieved | Reproduced: unknown
- **Evidence tier:** [C] watch METHOD. No stated venue, no released implementation, citation count unverified.
- **Mechanism family:** `Far-field single-channel transcription benchmark`, `Near-field close-talk reference condition` (the AMI-IHM-Mix results), `Joint speaker-attributed transcription metric`, `Speaker attribution metric`
- **Study type:** methods paper with ablations across model size, training-data volume and chunk length; five published test sets.
- **Population:** Mandarin training used AliMeeting (2-4 speakers, 105 hours), AISHELL-4 (3-7 speakers, 107 hours), MISP2025, MagicData-RAMC and HKUST. English training used AMI, ICSI and one quarter of Fisher. Test sets: AliMeeting, AISHELL-4, AMI-IHM, AMI-SDM and Fisher. No demographics of the authors' own.
- **Imaging / measurement:** four distinct audio conditions appear and the paper keeps them separate. AMI-SDM is a single distant microphone in a meeting room. AMI-IHM-Mix is the mix of the participants' own individual headset microphones, a NEAR-FIELD condition and an upper bound, not a deployment condition. AliMeeting and AISHELL-4 are far-field circular-array corpora. Fisher is telephone speech. Diarization error rate is reported with a 0-second collar, with 0.5-second-collar figures also given in parentheses for comparison with papers that use the looser setting.
- **Methods / model:** Whisper-large-v3-turbo encodes the mixed multi-speaker audio; a two-layer perceptron projector with GELU maps the features into a small language model's embedding space (Gemma3-270M, Qwen3-0.6B or Qwen3-1.7B). An external diarization system, either DiariZen or S2SND, supplies speaker labels and segment boundaries as an explicit prior. Transcription is reformulated as multi-turn dialogue generation, one query per speaker per time segment. Word-level timestamps are optionally interleaved with word tokens. Metrics: DER, cpWER or cpCER, and tcpWER or tcpCER.
- **Statistical detail:**
  - Test: none. No significance test, no confidence interval, no repeated seeds.
  - Per-arm N: one run per configuration, on published test splits.
  - Effect sizes: absolute error-rate differences only.
  - p-values: none reported.
  - Power / pre-registration: not applicable and not reported.
- **Key findings:**
  - AMI Single Distant Microphone, far-field single channel: best configuration 13.72 percent DER, 21.26 percent cpWER, 22.07 percent tcpWER, with a 1.7-billion-parameter language model and S2SND diarization.
  - AMI-IHM, the near-field headset mix: 11.48 percent DER, 16.40 percent cpWER, 16.93 percent tcpWER. The 5-point cpWER gap between AMI-IHM and AMI-SDM in the same system is the

near-field-to-far-field penalty measured inside one paper.
  - AliMeeting far-field: 10.09 percent DER, 19.15 percent cpCER, 19.45 percent tcpCER at 1.7 billion parameters trained on 1300 hours of Mandarin. AISHELL-4: 10.56 / 17.66 / 18.10.
  - Fisher telephone speech: 11.16 percent DER, 15.91 percent cpWER, 16.10 percent tcpWER.
  - The cascade of DiariZen plus Whisper-large-v3 on AMI-SDM scored 14.61 percent DER and 43.91 percent cpWER, so the language-model back end roughly halved attributed word error on identical diarization input.
  - Reported comparison numbers on AMI-SDM cpWER: TagSpeech not reported for SDM; Gemini-2.5-Pro 34.78; Gemini-3.0-Pro 26.91; VibeVoice-ASR 28.82; Pyannote plus Whisper-large-v3 43.57.
  - Adding word-level timestamp prediction improved text accuracy, not just structure: on AliMeeting with a 270-million-parameter model, cpCER fell from 31.07 to 28.24 percent when word timestamps were added.
- **Author's framing of the contribution:** > "At the current stage, leveraging reliable speaker diarization as an explicit structural prior provides a > practical and efficient way to simplify this multi-speaker ASR task." (p. 2)
- **What this paper does NOT establish:**
  - It does NOT measure a phone. AMI-SDM is a room microphone, AMI-IHM is worn headsets, AliMeeting and AISHELL-4 are circular arrays, Fisher is a telephone line.
  - It does NOT report latency, real-time factor, memory or on-device feasibility, despite the small language models being the paper's efficiency argument.
  - It does NOT show an end-to-end system. Every headline number depends on an external diarization system that is itself a separate trained model, so the pipeline is two models deep, not one.
  - It does NOT test NOTSOFAR-1, so its results cannot be compared with this index's NOTSOFAR-1 numbers without changing the corpus.
  - It does NOT test summarisation or action-item extraction.
  - It does NOT provide code, seeds or confidence intervals.
- **Direct quotes (verbatim, source ground truth):**
  - "At the current stage, leveraging reliable speaker diarization as an explicit structural prior provides a practical and efficient way to simplify this multi-speaker ASR task." (p. 2)
  - "unlike traditional cascaded systems, it does not use diarization to split audio into speaker-wise utterances for a single-speaker ASR backend. Instead, it uses diarization as an explicit structural prior while allowing an LLM-based backend to recognize mixed multi-speaker speech directly." (p. 2)
  - "Our analysis shows that diarization systems provide more reliable speaker identities and segment-level boundaries, while LLMs excel at modeling linguistic content and long-range dependencies, demonstrating their complementary strengths." (p. 1)
  - "many existing methods mainly target who said what, while explicit modeling of when is still weakly represented in a large portion of the literature" (p. 1)
- **Limitations:**
  - Author-stated: the approach depends on the quality of the upstream diarization, and the paper's own evaluation settings exist to probe when the model can and cannot correct an imperfect prior.
  - Inferred: no venue, no code, no variance estimates.
  - Inferred: the AMI-SDM DER of 13.72 percent is inherited from the external diarization system, not produced by the proposed method, so it is not evidence about the method.
  - Inferred: comparison figures for Gemini and several baselines are marked as taken from other papers rather than measured here, and the paper flags that one cited DER used a 0.5-second collar while its own use none, which makes those columns not directly comparable.
- **Cross-references in this index:**
  - See also: Huo 2026 (same AMI-SDM and AliMeeting far-field single-channel conditions, weaker cpWER, no external diarization); Dai 2026 (same corpora, 23.32 percent cpWER on AMI-SDM against 21.26 here); Niu 2024 (the NOTSOFAR-1 single-channel winner, a different corpus).

  - Contrast with: Abramovski 2025 (tcpWER on NOTSOFAR-1; different corpus and different metric collar); Watanabe 2020 (cpWER on CHiME-6, a dinner-party rather than a meeting condition, where the same metric sat at 77.9 percent).
- **Relevance to platform:** This is the strongest 2026 far-field single-channel meeting number in the vault: 21.26 percent attributed word error rate on AMI's single distant microphone. It is also the clearest demonstration that the number is not produced by one model. It required a separate trained diarization system supplying who and when, then a language model supplying what, with the diarization stage alone contributing a 13.72 percent DER. For the charter's two-taps and finished-before-the-walk-back constraints, the honest reading is that the best published single-microphone attributed transcript in 2026 still costs roughly one word in five and comes from a two-stage pipeline whose runtime nobody has published.
- **Quotable stats (paste-ready):**
  - "On AMI's single distant microphone, the best system in Li 2026 scored 21.26 percent concatenated minimum-permutation word error rate and 22.07 percent time-constrained speaker-attributed word error rate."
  - "Measured inside one system on one corpus, moving from worn headset microphones to a single distant tabletop microphone raised attributed word error rate from 16.40 to 21.26 percent (Li 2026, AMI)."
  - "A conventional cascade of a diarization system and Whisper-large-v3 scored 43.91 percent attributed word error rate on AMI's single distant microphone; replacing only the recognition back end with a diarization-conditioned language model halved it to 21.26 percent (Li 2026)."
  - "Adding word-level timestamp prediction to the output lowered concatenated character error on far-field Mandarin meeting audio from 31.07 to 28.24 percent (Li 2026, AliMeeting, 270-million-parameter model)."

Pandey 2023 - pronunciation-aware biasing for personal names, measured on a voice assistant (N=3 test sets)

- **Citation:** Rahul Pandey, Roger Ren, Qi Luo, Jing Liu, Ariya Rastrow, Ankur Gandhe, Denis Filimonov, Grant Strimel, Andreas Stolcke, Ivan Bulyko. "PROCTER: PROnunCiation-aware conTextual adaptER for Personalized Speech Recognition in Neural Transducers." arXiv:2303.17131v1, 30 March 2023. DOI: not provided in the PDF. Affiliations as printed: Amazon Alexa AI, USA; George Mason University, USA.
- **File:** `literature/arxiv-2303.17131.pdf`
- **Links:** arXiv:2303.17131 | DOI: not provided | Code: not released; the system is trained on in-house de-identified data and no repository is named | Weights: not released | Project page: none | Citations: not retrieved | Reproduced: unknown
- **Evidence tier:** [C] watch METHOD. The PDF states no venue, names no public implementation, and its data are in-house and unreleasable, so it fails both the implementation arm and any possibility of independent reproduction.
- **Mechanism family:** `Contextual biasing for rare words` (+ `Speaker-agnostic transcription metric`)
- **Study type:** methods paper with a four-way ablation; industrial in-house evaluation.
- **Population:** in-house de-identified far-field utterances from interactions with a commercial virtual voice assistant, sampled from more than 20 domains including Communications, Weather, SmartHome and Music. No speaker demographics, no age, no sex breakdown, no recruitment description; the data are proprietary. The baseline recogniser was trained on 114,000 hours; the contextual adapter was trained on about 290 hours. Three test sets: 16,000 far-field English utterances with no personalised context; 29,000 far-field English utterances containing mentions of personalised entities; 3,558 utterances containing personalised device names.
- **Imaging / measurement:** FAR-FIELD, but the far field of a smart-speaker voice assistant, not of a meeting. Utterances are short commands and queries addressed to a device, one speaker at a time, not multi-party conversation. Input features are 64-dimensional log filterbank energies every 10 milliseconds with a 25-millisecond window, three frames stacked to 192 features per frame.
- **Methods / model:** a recurrent neural network transducer (RNN-T) with an 8-layer LSTM audio

encoder at 1280 units per layer, a 2-layer LSTM prediction network, a 512-unit joint network and a 4000-word-piece vocabulary, decoded with beam size 8. The proposed adapter adds a phoneme encoder and a grapheme encoder, each a BiLSTM (128 and 64 units), and biases intermediate audio-encoder outputs by scaled dot-product cross-attention over concatenated grapheme-phoneme embeddings. Queries come from a learned weighted sum of the last, third-last and fifth-last LSTM layer outputs. All pronunciations of an entity are kept as separate entries. The core recogniser is frozen; only 1.5 million adapter parameters are trained, about 1 percent of the model. Maximum 300 contextual entities, 600 grapheme-phoneme pairs. Adam, learning rate 5e-4, early stopping.

- **Statistical detail:**
  - Test: none. No significance test, no confidence interval, no seed variance.
  - Per-arm N: five model configurations on three test sets.
  - Effect sizes: reported only as relative word error rate reduction (WERR) and relative named-entity word error rate reduction (NE-WERR) against the un-personalised RNN-T baseline.
  - p-values: none reported.
  - Power / pre-registration: not applicable and not reported.
- **Key findings:**
  - The paper's headline "44% and 57%" are relative NAMED-ENTITY word error rate reductions against the vanilla RNN-T baseline: 43.9 percent on all personalised entities and 57.2 percent on rare personalised entities that appear only once in the test set.
  - The baseline those figures are measured against is a non-personalised RNN-T, not a competing personalised system. Against the previous state of the art, a text-only contextual adapter, the improvement is much smaller: 43.9 against 37.1 percent on all entities, and 57.2 against 50.0 percent on rare entities.
  - Overall word error rate reduction on the personalised-entity test set was 38.2 percent for PROCTER against 32.5 percent for the text-only adapter.
  - On the zero-shot personalised-device-name set, PROCTER gave 6.9 percent NE-WERR against 0.9 percent for the text-only adapter, the largest relative gap in the paper.
  - On the general test set with no personalised context, all adapter variants were within 1.0 percent WERR of the baseline, so biasing did not damage general recognition.
  - Removing the intermediate-layer queries cost almost nothing on the personalised-entity set (43.6 against 43.9 NE-WERR) but most of the zero-shot gain (3.9 against 6.9), which the authors read as the intermediate layers mattering only when context is sparse.
- **Author's framing of the contribution:** > "We propose a PROnunCiation-aware conTextual adaptER (PROCTER) that dynamically injects lexicon knowledge > into an RNN-T model by adding a phonemic embedding along with a textual embedding." (p. 1)
- **What this paper does NOT establish:**
  - It does NOT measure meeting audio. Every test set is short single-speaker utterances directed at a voice assistant, with no overlapping speech, no multi-party turn-taking and no conversational context.
  - It does NOT measure a phone. The condition is a far-field voice-assistant device.
  - It does NOT report an absolute word error rate anywhere. Every number is relative to an unstated baseline error rate, so the reader cannot know whether 43.9 percent relative means moving from 20 percent to 11 percent or from 4 percent to 2.2 percent.
  - It does NOT test any language other than English.
  - It does NOT show a gain for general speech; the general-set improvement is 0.2 to 1.0 percent relative, which the authors present as a non-degradation result, not a gain.
  - It does NOT release code, weights or data, and the evaluation data cannot be obtained.
- **Direct quotes (verbatim, source ground truth):**
  - "The experimental results show that the proposed PROCTER architecture outperforms the baseline RNN-T model by improving the word error rate (WER) by 44% and 57% when measured on personalized entities and personalized rare entities, respectively, while increasing the model size (number of trainable parameters) by only 1%." (p. 1)
  - "We use in-house de-identified far-field datasets coming from interactions with a virtual voice assistant." (p. 3)

  - "End-to-End (E2E) automatic speech recognition (ASR) systems used in voice assistants often have difficulties recognizing infrequent words personalized to the user, such as names and places." (p. 1)
  - "Given the general test set with no personalized context, we observe that the previous text-only adapter, proposed PROCTER, and all ablation experiments have no degradation over the baseline RNN-T model." (p. 4)
- **Limitations:**
  - Inferred: the entire evaluation is on proprietary data that no third party can obtain, so nothing here is independently checkable.
  - Inferred: absolute error rates are withheld, which is standard practice for this lab but means the practical size of the improvement cannot be judged.
  - Inferred: the comparison that matters, PROCTER against the prior text-only adapter, is 6.8 points of relative NE-WERR on entities and 7.2 on rare entities, considerably less than the 44 and 57 the abstract leads with.
  - Inferred: the method requires a pronunciation lexicon for every biased entity, which is a per-language engineering asset the paper does not cost.
- **Cross-references in this index:**
  - See also: Shapira 2025 (independently finds that named entities are the word class whose correction most improves a downstream summary, which is the reason a project would want this method at all).
  - Contrast with: every other recognition paper in this index. Abramovski 2025, Niu 2024, Huo 2026, Dai 2026 and Li 2026 all measure multi-party far-field MEETING audio; this paper measures single-speaker voice assistant commands, and its family tag is different for that reason. Its numbers must never be quoted as a meeting-transcription result.
- **Relevance to platform:** The charter's deliverable is a summary and action items, and both hang on people's names, project names and dates surviving recognition. This paper is the vault's only entry on how to protect exactly those words, and it reports that adding pronunciation information to a per-user entity list cuts named-entity errors by 43.9 percent relative on entities and 57.2 percent on rare ones, for about 1 percent extra trainable parameters. The condition it was measured under is a voice assistant, not a meeting, so this is a mechanism the project could investigate, not a result the project can claim.
- **Quotable stats (paste-ready):**
  - "Adding phonemic representations to a contextual biasing adapter cut named-entity word error rate by 43.9 percent relative on personalised entities and 57.2 percent on rare personalised entities, against a non-personalised recogniser, on far-field voice-assistant utterances (Pandey 2023, 29,000 utterances)."
  - "Against the prior text-only contextual adapter rather than the un-personalised baseline, the same method improved named-entity word error rate reduction from 37.1 to 43.9 percent on entities and from 50.0 to 57.2 percent on rare entities (Pandey 2023)."
  - "The biasing adapter added 1.5 million trainable parameters, about 1 percent of the recognition model, and left general-condition word error rate unchanged within 1.0 percent relative (Pandey 2023)."
  - "On previously unseen personalised device names, pronunciation-aware biasing gave a 6.9 percent relative named-entity word error rate reduction against 0.9 percent for text-only biasing (Pandey 2023, 3,558 utterances)."

Polok 2026 - what simulated conversations buy, and the oracle-diarization caveat under the numbers (N=2 tasks)

- **Citation:** Alexander Polok, Ivan Medennikov, Jan Cernocky, Shinji Watanabe, Lukas Burget, Samuele Cornell. "Mind the Gap: Impact of Synthetic Conversational Data on Multi-Talker ASR and Speaker Diarization." arXiv:2605.15442v1, 14 May 2026. DOI: not provided. Affiliations as printed: Brno University of Technology, Czechia; Carnegie Mellon University, USA; NVIDIA, USA.
- **File:** `literature/arxiv-2605.15442.pdf`
- **Links:** arXiv:2605.15442 | DOI: not provided | Code: https://github.com/popcornell/FastMSS,

stated in the paper as a fully open-source simulation toolkit (named in the paper, not verified here) | Weights: reference models named as https://huggingface.co/nvidia/diar_sortformer_4spk-v1 and https://huggingface.co/BUT-FIT/ DiCoW_v3_3 (named in the paper, not verified here) | Project page: none | Citations: not retrieved | Reproduced: unknown

- **Evidence tier:** [C] watch METHOD. A public simulator and public reference weights are named, clearing the implementation arm, but the paper states no venue acceptance and its citation count is unverified.
- **Mechanism family:** `Synthetic training-data generation` (+ `Far-field single-channel transcription benchmark`, `Joint speaker-attributed transcription metric`, `Speaker attribution metric`)
- **Study type:** controlled ablation study across four factors (turn-taking dynamics, source domain, acoustic augmentation, data mixing) on two model families.
- **Population:** no human subjects. Seed corpora for simulation: LibriSpeech read speech (960 hours), VoxPopuli semi-spontaneous parliamentary speech (543 hours), otoSpeech full-duplex conversational speech (141 hours), and the close-talk channels of AMI and NOTSOFAR-1 (the NOTSOFAR-1 close-talk seed is about 7.5 hours). All were re-aligned with the Montreal Forced Aligner. Noise from MUSAN with the speech class excluded. The real-data comparison set is about 314 hours drawn from NOTSOFAR-1, AMI, AliMeeting, DIHARD-III development and VoxConverse v0.3.
- **Imaging / measurement:** evaluation conditions are named per test set and they are mixed. For recognition: AMI Single Distant Microphone (far-field single channel), NOTSOFAR-1 single-channel, LibriSpeechMix (simulated 1-3 speaker mixtures), and Mixer6 channel 4. For diarization, additionally NOTSOFAR-1 and AMI Mix of Headset Mics (NEAR-FIELD), AliMeeting near and far, DIHARD-III evaluation 1-4 speaker subset, and MSDWild few-talker split. AMI and AliMeeting were cut to 180-second chunks because the offline diarizer cannot take full recordings. Ground-truth labels for AMI, AliMeeting and NOTSOFAR-1 were regenerated by forced alignment rather than taken from the corpora's own segment-level annotations, which the authors say are over-segmented for diarization.
- **Methods / model:** recognition is Diarization-Conditioned Whisper (DiCoW) on a Whisper-large-v3-turbo backbone. Diarization is Sortformer, an encoder-only end-to-end model on a 109-million-parameter NEST-FastConformer backbone, trained on 60-second crops from 90-second simulated segments with 1-4 speaker sessions balanced 1:3:6:10, averaged over three random seeds. The simulator, FastMSS, extends a two-speaker hidden-Markov turn-taking model to arbitrary speaker counts with four transition types (turn hold, turn switch, interruption, backchannel) whose probabilities can be fitted to any annotated corpus by maximum likelihood. THE CRITICAL MEASUREMENT DETAIL: "Consistent with the DiCoW/SE-DiCoW evaluation protocol, we report throughout this paper tcpWER conditioned on ground truth diarization, and we utilize greedy attention-only decoding." (p. 2). Every recognition number in the paper therefore assumes perfect knowledge of who spoke when. Recognition tcpWER uses a 5-second collar; diarization error rate uses a 0-second collar.
- **Statistical detail:**
  - Test: none. Diarization results are averaged over three random seeds; no variance, confidence interval or significance test is reported for any number.
  - Per-arm N: one recognition run and three diarization seeds per configuration.
  - Effect sizes: absolute error-rate differences only.
  - p-values: none reported.
  - Power / pre-registration: not applicable and not reported.
- **Key findings:**
  - The best recognition result on NOTSOFAR-1 single-channel was 16.3 percent tcpWER, from a model trained on real data plus simulated data. This is measured with GROUND-TRUTH diarization supplied, so it is not comparable with the NOTSOFAR-1 challenge's 22.2 percent winning score, which had to produce its own diarization. The paper does not state which NOTSOFAR-1 split "NSF-1 SC" refers to.
  - Real data alone on the same oracle-diarization protocol gave 17.7 percent on NOTSOFAR-1 and 15.5 percent on AMI-SDM. Synthetic data alone gave 20.1 and 16.0. Synthetic pre-training followed by real fine-tuning gave 16.3 and 14.9, the paper's best macro average at 8.7 percent.

  - Source diversity beat domain matching. A combined mixture of all five seed corpora gave a macro average of 10.0 percent against 10.9 for training on real AMI plus NOTSOFAR-1 recordings.
  - Turn-taking statistics matter and the optimum differs by task. Boosting overlap beyond natural levels improved recognition (24.8 to 22.1 percent on NOTSOFAR-1) and degraded diarization (macro DER 26.1 to 27.6 percent). The authors state plainly that “a single simulation recipe cannot optimally serve both ASR and diarization” (p. 3).
  - Acoustic augmentation is decisive for diarization and marginal for recognition. Adding reverberation cut AliMeeting-Far DER from 36.8 to 25.7 percent, an 11-point absolute reduction; noise plus reverberation reached 22.2 percent macro DER against 26.1 clean. For recognition the same augmentations moved the macro average by 0.2 to 0.3 points.
  - Best diarization overall was synthetic pre-training then real fine-tuning at 15.5 percent macro DER against 17.4 for real-only training and 22.8 for the public reference model.
  - The simulator itself generated 1,000 hours of annotated multi-talker audio in under five minutes on a 256-core machine, against roughly 85 simulated hours per minute for the next-fastest tool at 32 workers.
- **Author’s framing of the contribution:** > “Ultimately, synthetic-only training approaches real-data baselines, and combining simulated data with > real recordings yields substantial gains over real-only training across both tasks.” (p. 1)
- **What this paper does NOT establish:**
  - It does NOT report an end-to-end recognition result. Every tcpWER in the paper is conditioned on ground-truth diarization, which no deployed system has. Quoting 16.3 percent against a challenge score is a category error.
  - It does NOT state which NOTSOFAR-1 split its single-channel numbers were measured on, so a development-set-versus-evaluation-set comparison with the challenge results cannot be made from this paper alone.
  - It does NOT measure a phone, and does not include any handset or consumer-device condition.
  - It does NOT report latency, real-time factor or model footprint at inference.
  - It does NOT test summarisation or action-item extraction.
  - It does NOT report confidence intervals or significance for any comparison, including differences of 0.1 to 0.3 absolute points on which several of its conclusions rest.
- **Direct quotes (verbatim, source ground truth):**
  - “Consistent with the DiCoW/SE-DiCoW evaluation protocol, we report throughout this paper tcpWER conditioned on ground truth diarization, and we utilize greedy attention-only decoding.” (p. 2)
  - “our findings reveal that optimal simulation recipes are highly task-dependent: increasing speech overlap benefits ASR but degrades diarization.” (p. 1)
  - “This finding has a practical implication: a single simulation recipe cannot optimally serve both ASR and diarization.” (p. 3)
  - “the synthetic-only Combined setup already outperforms training on real data alone in macro average (10.0% vs. 10.9%), confirming that source diversity outweighs exact domain matching” (p. 3)
  - “Real meeting corpora, such as AMI, NOTSOFAR-1, MCoREC, and CHiME-5/6 provide spontaneous conversational data but are limited in scale, typically comprising only tens to a few hundreds of hours.” (p. 1)
  - “Notably, we use forced-alignment based ground-truth labels for AMI, AliMeeting and NSF-1, because default segment-level annotations are not well-suited for diarization due to severe over-segmentation” (p. 2)
- **Limitations:**
  - Author-stated: concatenative simulation lacks inter-turn semantic coherence, which the authors mitigate by freezing the recognition decoder during training rather than by fixing the simulation.
  - Inferred: the oracle-diarization protocol is stated once, in a single sentence at the end of a methods paragraph, and never repeated in any table caption. A reader skimming the tables

will take 16.3 percent for an end-to-end number.
  - Inferred: re-deriving ground-truth labels by forced alignment means the diarization numbers are not comparable with any published result on the same corpora that used the corpora's own references.
  - Inferred: no venue and no significance testing.
- **Cross-references in this index:**
  - See also: Kalda 2024 (also fine-tunes Whisper on NOTSOFAR-1 with heavy augmentation, and also finds an augmentation that helps on development data and not on evaluation data); Niu 2024 (independently reports that adding real and simulated matched training audio was the single largest win available); Vinnikov 2024 (the 1000-hour simulated training set this line of work reacts to).
  - Contrast with: Abramovski 2025 (the 22.2 percent NOTSOFAR-1 single-channel figure there is produced by a system that had to diarize for itself; the 16.3 percent here is not the same measurement and does not beat it).
- **Relevance to platform:** This paper is the vault's main evidence on whether a team without a proprietary meeting corpus can build a competitive system, and the answer is a qualified yes: synthetic conversations built from public read speech and public close-talk channels came within about one point of real in-domain data, and combining the two beat real data on every benchmark. It is also the vault's clearest example of a number that will be mis-quoted. The 16.3 percent single-channel figure would appear to beat the NOTSOFAR-1 challenge winner's 22.2 percent, and it does not, because it was measured with the answer to "who spoke when" handed to the model in advance.
- **Quotable stats (paste-ready):**
  - "Training a multi-talker recognition model on simulated conversations alone reached 20.1 percent time-constrained speaker-attributed word error rate on NOTSOFAR-1 single-channel audio against 17.7 percent for real in-domain recordings, both measured with ground-truth diarization supplied (Polok 2026)."
  - "Combining simulated and real conversational data beat real-only training on every benchmark tested, improving the macro average from 10.9 to 8.7 percent (Polok 2026, 8 test conditions)."
  - "Adding reverberation to simulated training audio cut diarization error rate on far-field Mandarin meeting audio from 36.8 to 25.7 percent (Polok 2026, AliMeeting far)."
  - "A simulation recipe tuned for recognition degrades diarization: boosting speech overlap improved attributed word error from 24.8 to 22.1 percent while worsening macro diarization error from 26.1 to 27.6 percent (Polok 2026)."

Sun 2025 - detecting when two people talk at once, on far-field meeting audio (N=3 corpora, AMI test set)

- **Citation:** Zhaokai Sun, Li Zhang, Qing Wang, Pan Zhou, Lei Xie. "Towards Robust Overlapping Speech Detection: A Speaker-Aware Progressive Approach Using WavLM." arXiv:2505.23207v1, 29 May 2025. DOI: not provided. The PDF's page-4 footer names the Interspeech 2025 organisers, which is the venue this camera-ready targets. Affiliations as printed: Audio, Speech and Language Processing Group (ASLP@NPU), School of Software, Northwestern Polytechnical University, China; Space AI, Li Auto.
- **File:** `literature/arxiv-2505.23207.pdf`
- **Links:** arXiv:2505.23207 | DOI: not provided | Code: not released; the paper names third-party components (WavLM at https://github.com/microsoft/unilm/tree/master/wavlm, CampPlus on ModelScope) but no repository of its own | Weights: not released | Project page: none | Citations: not retrieved | Reproduced: unknown
- **Evidence tier:** [B] credible METHOD. Interspeech is on this index's venue list, which is one strong signal, but no implementation of the proposed system is released and the citation rate is unverified.
- **Mechanism family:** `Overlapped speech detection` (+ `Far-field single-channel transcription benchmark` only in the sense of the audio condition; no words are scored anywhere in this paper)
- **Study type:** methods paper with three ablation studies; three corpora.
- **Population:** three corpora, described in the paper's Table 1 as: AliMeeting, style "Meeting/Far", 104.75 hours, 42.27 percent overlap ratio; AMI, "Meeting/Far", 75 hours, 19 percent overlap ratio;

LibriHeavyMix, “Multi/Near”, 240 hours, 42.56 percent overlap ratio, an open-source simulated conversation dataset. No participant demographics of the authors’ own.

- **Imaging / measurement:** AMI and AliMeeting are used in their FAR-FIELD condition; the paper does not name which specific far-field channel of AMI. LibriHeavyMix is near-field simulated. Audio is cut into 5-second segments, 25-millisecond frames with a 20-millisecond shift. Training data was manually rebalanced to a 1:1:1 ratio of silence, single-speaker and overlapping segments, and overlap labels were softened from 1 to 0 over 10 frames near each boundary.
- **Methods / model:** WavLM-Large supplies frame-level acoustic features; CampPlus supplies frame-level speaker embeddings; a cross-attention module fuses the two with the acoustic features as query. A voice-activity-detection decoder runs first and its logits mask the fused representation before an overlapped-speech-detection decoder runs on the masked features. Both decoders are stacks of Conformer blocks. Pre-training on LibriHeavyMix for five epochs, then fine-tuning on real data. Adam, learning rate 1e-4, weight decay 1e-4, cosine schedule.
- **Statistical detail:**
  - Test: none. No significance test, no confidence interval, no seeds.
  - Per-arm N: one run per configuration, on the AMI test set.
  - Effect sizes: absolute and relative F1 differences.
  - p-values: none reported.
  - Power / pre-registration: not applicable and not reported.
- **Key findings:**
  - Best system on the AMI test set: F1 82.76 percent, recall 81.48, precision 84.08. The authors call this a 4.4 percent relative improvement over the previous best, an XLSR-Conformer at F1 79.21.
  - Reported comparison systems on the same AMI test set: CNN F1 56.1 (recall 44.6, precision 75.8); x-vectors 61.0 (47.7 / 85.5); pyannote 75.3 (80.7 / 70.5); Conformer 64.73 (65.03 / 64.43); XLSR-Conformer 79.21 (79.38 / 79.04).
  - The speaker-attention module is worth about 3.3 F1 points: 82.76 with cross-attention, 81.62 with a mean-squared-error alignment, 79.47 with no speaker information at all.
  - The self-supervised front end is worth roughly 13 points: WavLM plus speaker attention 82.76, versus a Conformer with no self-supervised front end at 65.95.
  - Progressive training, where voice-activity logits mask the features before overlap detection, gave 0.6 percent relative over the unified three-class formulation (82.76 against 82.20).
  - Adding reverberation augmentation during training did not help and was dropped from the final pipeline.
- **Author’s framing of the contribution:** > “This work proposes a speaker-aware progressive OSD model that leverages a progressive training strategy > to enhance the correlation between subtasks such as voice activity detection (VAD) and overlap detection.” > (p. 1)
- **What this paper does NOT establish:**
  - It does NOT transcribe anything. No word error rate, character error rate or attributed error rate appears in the paper. Overlap detection is a frame-level binary decision, and an F1 of 82.76 percent is not a transcription accuracy.
  - It does NOT diarize. It does not say who is speaking, only that more than one person is.
  - It does NOT measure a phone. AMI and AliMeeting far-field are fixed room microphones and arrays.
  - It does NOT name the specific AMI far-field channel used, so its number cannot be lined up precisely with AMI-SDM results elsewhere in this index.
  - It does NOT report latency or streaming capability, and the architecture (WavLM-Large plus CampPlus plus two Conformer decoders) is not costed at inference.
  - It does NOT show that reverberation augmentation helps, and reports the opposite.
- **Direct quotes (verbatim, source ground truth):**
  - “Experimental results show that the proposed method achieves state-of-the-art performance, with an F1 score of 82.76% on the AMI test set, demonstrating its robustness and effectiveness in OSD.” (p. 1)
  - “Due to the scarcity of real-world dialogue datasets with sufficient overlap durations, most OSD research relies on simulated data or selectively filtered open-source datasets.” (p. 1)

- "One major challenge in OSD training is the inherent class imbalance, as overlapping speech occurs far less frequently than non-overlapping speech." (p. 3)
- "we investigate the impact of reverberation augmentation during training. However, our experiments indicate that adding reverberation does not yield significant performance improvements; therefore, it is not incorporated into our final training pipeline." (p. 4)

- **Limitations:**
  - Author-stated: overlap is rare relative to single-speaker speech, and the training distribution had to be manually rebalanced, so the training distribution does not match the test distribution.
  - Inferred: the AMI far-field channel is unnamed, which makes the headline number hard to reproduce exactly.
  - Inferred: no released code, no seeds, no confidence intervals, and the progressive-training gain the paper is named for is 0.56 absolute F1 points.
  - Inferred: an F1 of 82.76 percent still means roughly one frame in five is misclassified in a task that is upstream of every attribution decision.
- **Cross-references in this index:**
  - See also: Huo 2026 (independently shows that overlapping regions are where every meeting system fails, with diarization error above 33 percent in overlap on both AMI and AliMeeting); Abramovski 2025 (debate-style overlapping speech named as the specifically single-channel failure mode); Polok 2026 (finds that simulating MORE overlap helps recognition and hurts diarization).
  - Contrast with: every transcription entry in this index. This paper scores a detection F1, not an error rate, and the two must not be presented as comparable quality figures.
- **Relevance to platform:** Overlapping speech is the specific asymmetry the charter's working notes asked the research round to find: it is what a microphone array handles and a single device does not. This paper is the vault's measurement of how well the overlap can even be DETECTED from far-field meeting audio, before anyone tries to transcribe it, and the answer is F1 82.76 percent on AMI with a large self-supervised front end. AliMeeting's overlap ratio of 42.27 percent and AMI's 19 percent also bound how much of a real meeting is at stake.
- **Quotable stats (paste-ready):**
  - "The best published overlapped-speech detection on far-field AMI meeting audio scored an F1 of 82.76 percent, with recall 81.48 and precision 84.08 (Sun 2025)."
  - "Overlapping speech accounts for 19 percent of the AMI meeting corpus and 42.27 percent of the AliMeeting corpus (Sun 2025, Table 1)."
  - "Removing frame-level speaker information from an overlap detector cost 3.3 F1 points, from 82.76 to 79.47 percent, on far-field AMI audio (Sun 2025)."
  - "Replacing a self-supervised speech front end with a plain Conformer cost about 17 F1 points on far-field AMI overlap detection, from 82.76 to 65.95 percent (Sun 2025)."

What meeting summarisation is, and what it has been built and scored on

Kumar 2022 - a survey and leaderboard of meeting summarisation systems (N=over 40 papers surveyed)

- **Citation:** Lakshmi Prasanna Kumar, Arman Kabiri. "Meeting Summarization: A Survey of the State of the Art." arXiv:2212.08206v1, 16 December 2022. IMRSV Data Labs, Ottawa, Canada. DOI: not provided in the PDF beyond the arXiv identifier.
- **File:** `literature/arxiv-2212.08206.pdf`
- **Links:** arXiv:2212.08206 | DOI: 10.48550/arXiv.2212.08206 | Code: not released, this is a survey | Weights: not applicable | Project page: none | Citations: 10 (Semantic Scholar, as of 2026-09) | Reproduced: not applicable
- **Evidence tier:** [C] watch DOMAIN. No conference or journal venue, and a citation rate of about 2.5 per year against a floor of 5. Useful as a map of the field and as a compiled leaderboard, not as a source of measurements.

- **Mechanism family:** `Abstractive summarisation method`, `Extractive summarisation method` (+ `Summarisation evaluation metric`, `Corpus resource`)
- **Study type:** narrative survey with a compiled results table; not a systematic review, no stated inclusion protocol, no PRISMA checklist.
- **Population:** not human subjects. The authors state they studied over forty papers covering meeting summarisation techniques. The corpora discussed are AMI (137 meetings), ICSI (59 meetings), QMSum (1,808 query-summary pairs over 232 meetings, being 137 AMI, 59 ICSI and 36 parliamentary committee meetings) and ConvoSumm.
- **Imaging / measurement:** none of its own. The reported corpus statistics are quoted from other papers: AMI and ICSI transcripts have 289 and 464 turns and 4757 and 10,189 words on average, and their summaries 322 and 534 words on average.
- **Methods / model:** taxonomy plus leaderboard. Extractive approaches are divided into maximal marginal relevance based, graph based, optimisation based and supervised. Abstractive approaches are divided into graph based, template based and deep learning based, with the deep learning branch split into query based, multi-modal, long-meeting and supplementary-information based. All leaderboard entries are ROUGE-1, ROUGE-2 and ROUGE-L F1 scores adopted from published literature rather than recomputed.
- **Statistical detail:**
  - Test: none. No statistics are computed in this paper.
  - Per-arm N: not applicable.
  - Effect sizes: not applicable.
  - p-values: none.
  - Power / pre-registration: not applicable.
- **Key findings:**
  - Best reported ROUGE-1 on AMI in the compiled leaderboard: RetrievalSum 56.26, then DialogLM 54.49, then Longformer-BART with argument mining 54.47. Extractive baselines are far behind: TextRank 35.19, SummaRunner 30.98.
  - Best reported ROUGE-1 on ICSI: a domain-terminology method at 60.7, then DialogLM-sparse at 49.56 and DialogLM at 49.25, against HMNet at 46.28.
  - ROUGE-2 scores are much lower than ROUGE-1 across the board, from 3.7 to 34.9 on AMI and 3.7 to 37.1 on ICSI, which is what a bigram overlap metric looks like on abstractive summaries that reword everything.
  - The leaderboard is incomplete by construction: many cells are empty because the results were never reported, and the authors state that in those cases the code was not made public either, so the missing cells cannot be filled in.
  - The authors name factual inconsistency, that is, hallucination, as a live problem for meeting summarisers, together with transcript length exceeding transformer context and a shortage of annotated meeting data because industry meetings are proprietary.
- **Author's framing of the contribution:** > "In this survey, we aim to cover recent meeting summarization techniques." (p. 1)
- **What this paper does NOT establish:**
  - It does NOT measure anything. Every number in it is adopted from another paper's reported results, and the authors say so.
  - It does NOT establish that the leaderboard entries are comparable with each other. Different rows use different ROUGE-L conventions, different input lengths and different training regimes, and the paper applies no normalisation.
  - It does NOT establish that any of these scores were obtained from recognised speech. Its own claim that meeting transcripts "are obtained using Automatic Speech Recognition" is contradicted by Rennard 2023, which states that AMI, ICSI and ELITR all ship human-produced or human-corrected transcripts, and the AMI and ICSI leaderboard results are computed on those clean transcripts.
  - It does NOT evaluate action-item extraction. Action items appear only in passing, in the context of email to-do generation, and no meeting action-item benchmark or score is reported.
  - It does NOT report any human evaluation, any inter-annotator agreement, or any error analysis.
  - It does NOT cover work after 2022, so it predates the entire instruction-tuned large language

model era in this task.

- **Direct quotes (verbatim, source ground truth):**
  - "An abstractive text summarizer often suffers from a factual inconsistency problem. These problems are also called as hallucinations." (p. 8)
  - "In addition to this, there is not enough annotated datasets for meetings as most of the meetings performed in industry are proprietary in nature." (p. 8)
  - "AMI (Carletta et al., 2005) and ICSI (Janin et al., 2003) are the two main corpora used for meeting summarization task. The meeting transcripts are obtained using Automatic Speech Recognition (ASR)." (p. 7)
  - "Missing values in the leaderboard indicate that the results are not available for the corresponding dataset. Besides, in all of these cases, the code is not publicly made available to reproduce the experiments on the other datasets for which results are not reported." (p. 7)
  - "The number of tokens in a meeting transcript is typically more than what can be handled by a transformer architecture." (p. 8)
- **Limitations:**
  - Author-stated: none. The paper has no limitations section.
  - Inferred: no stated search protocol, no inclusion or exclusion criteria and no coverage claim, so the survey's completeness cannot be assessed.
  - Inferred: the leaderboard mixes ROUGE-L computed with and without sentence splitting, which the table footnote flags for one row but which plainly affects several, since ROUGE-L values in the table range from 12.97 to 52.51 for comparable systems.
  - Inferred: the claim that AMI and ICSI transcripts come from automatic speech recognition is wrong as stated, and it is the kind of error that would license exactly the mis-citation this vault exists to prevent.
- **Cross-references in this index:**
  - See also: Rennard 2023 (a peer-reviewed survey of the same task with an overlapping leaderboard and a stated taxonomy; prefer it wherever the two disagree).
  - Contrast with: Kirstein 2024 (which measures what the ROUGE numbers in this leaderboard actually track, and finds it is not summary correctness). Contrast with Cornell 2025 (the only entry here that measures summarisation quality downstream of real recognition error rather than on clean transcripts).
- **Relevance to platform:** Useful to the project as a map of what has been tried and as a warning label. The headline ROUGE scores it compiles, up to 56.26 on AMI and 60.7 on ICSI, are the numbers a competitor or an investor is most likely to quote as evidence that meeting summarisation is solved, and this entry records what they are: overlap scores against a single human reference, computed on clean human transcripts of role-played or academic meetings, with no recognition error anywhere in the chain. The project's `three-outputs` constraint promises a summary and action items, and this survey shows the second of those three has essentially no benchmark behind it.
- **Quotable stats (paste-ready):**
  - "The best reported ROUGE-1 scores for meeting summarisation are 56.26 on the AMI corpus and 60.7 on the ICSI corpus, both computed against human reference summaries on clean transcripts (Kumar 2022, compiling published results)."
  - "Extractive meeting summarisers score far below abstractive ones on the same corpora: TextRank reaches 35.19 ROUGE-1 on AMI against 56.26 for the best abstractive system (Kumar 2022)."
  - "AMI meeting transcripts average 4,757 words with 322-word summaries, and ICSI transcripts average 10,189 words with 534-word summaries (Kumar 2022, quoting Zhu et al. 2020)."
  - "A survey of over forty meeting summarisation papers names factual inconsistency, transcript length beyond transformer context, and a shortage of annotated meeting data as the field's three standing challenges (Kumar 2022)."

Rennard 2023 - the survey, and the corpus scarcity behind every number in it (N=3 corpora, about 280 h)

- **Citation:** Virgile Rennard, Guokan Shang, Julie Hunter, Michalis Vazirgiannis. "Abstractive Meeting Summarization: A Survey." Transactions of the Association for Computational Linguistics, 2023. DOI: 10.1162/tacl_a_00578. Preprint arXiv:2208.04163v2, 25 April 2023.
- **File:** `literature/arxiv-2208.04163.pdf`
- **Links:** arXiv:2208.04163 | DOI: 10.1162/tacl_a_00578 | Code: preprocessed corpora released at https://github.com/guokan-shang/ami-and-icsi-corpora and https://github.com/guokan-shang/elitr-minuting-corpus (named in the paper, not verified here) | Weights: not applicable | Project page: none | Citations: 40 (Semantic Scholar, merged record, as of 2026-09) | Reproduced: not applicable to a survey; the released corpora are the reusable artifact
- **Evidence tier:** [A] proven DOMAIN, top venue (TACL) with a citation rate of about 11 per year, well above the gate's floor, plus a released data artifact.
- **Mechanism family:** `Abstractive summarisation method`, `Corpus resource` (+ `Extractive summarisation method`, `Decision and action-item extraction`, `Summarisation evaluation metric`)
- **Study type:** systematic survey organised around a three-stage taxonomy, with a comparative benchmark table adopted from published results.
- **Population:** not human subjects. Three English meeting corpora, about 280 hours in total. AMI: 137 scenario-driven meetings, about 65 hours, 15 to 45 minutes each, four participants playing fixed roles in a fictitious electronics company designing a television remote control. ICSI: 75 naturally occurring weekly research meetings, about 72 hours, roughly one hour each, six participants on average, real colleagues discussing technical topics. ELITR: 113 English and 53 Czech technical project meetings, over 160 hours in total, of which the Czech portion is roughly 50 hours. Cited workplace context: American employees and managers average 6 and 23 hours per week in meetings respectively.
- **Imaging / measurement:** the survey's central measurement point about the data is that all three corpora provide gold transcripts that were either fully human-produced or human-corrected from recogniser output, precisely to avoid compounding recognition errors into the summarisation task. AMI and ICSI both carry topic segmentation, dialogue act labels, extractive summaries, abstractive summaries and abstractive communities, the last being the mapping from each abstractive sentence back to the set of extractive sentences supporting it. AMI and ICSI abstractive summaries follow a fixed four-part structure: Abstract, Decisions, Problems and Actions, each capped at 200 words and each optional. ELITR annotators were given no structure at all.
- **Methods / model:** the survey organises systems by which stage of Jones's summarisation pipeline they address. Interpretation systems enrich the transcript before summarising, using discourse structure (Rhetorical Structure Theory, Segmented Discourse Representation Theory), dialogue act labels, or multi-modal signals such as visual focus of attention estimated from head orientation and eye gaze. Transformation systems build an intermediate representation, by topic segmentation, abstractive community detection, template filling or query-related clustering. Generation systems attack the length and style problem directly, with sliding windows, multi-stage split-then-summarise frameworks, long-sequence transformers such as Longformer, hierarchical transformers such as HMNet and HAT, and domain-adaptive pre-training such as DialogLM. Comparison is by ROUGE-1, ROUGE-2 and ROUGE-L on AMI and ICSI, with the authors stating explicitly why they use ROUGE despite its inadequacy.
- **Statistical detail:**
  - Test: none. The paper computes no statistics of its own.
  - Per-arm N: 16 systems tabulated on AMI, 10 of them also on ICSI.
  - Effect sizes: ROUGE F1 scores only.
  - p-values: none.
  - Power / pre-registration: not applicable.
- **Key findings:**
  - Only three English meeting corpora with summaries exist, totalling around 280 hours, and only ELITR contains a second language. This is the hard constraint under everything else in the field.

  - Every published meeting-summarisation score rests on clean transcripts. The corpora deliberately supply human-produced or human-corrected text so that recognition errors do not compound into the summarisation task.
  - Meeting transcripts are an order of magnitude longer than the documents most summarisers were built for: one AMI transcript averages 4,757 tokens with a 322-token summary, against 781 tokens and a 56-token summary for a CNN/DailyMail article.
  - Benchmark table, best reported scores. On AMI: Longformer-BART with argument mining 55.27 ROUGE-1 and 20.89 ROUGE-2; DialogLM 54.49 and 20.03; SUMM-N 53.44 and 20.30; HMNet 53.02 and 18.57; the interpretation-focused DDAMS 53.15 and 22.32. On ICSI: DialogLM 49.25 ROUGE-1 and 12.31 ROUGE-2; HMNet 46.28 and 10.60; SUMM-N 45.57 and 11.49.
  - Generation-focused systems, which are large pre-trained models adapted to the domain, currently outscore interpretation-focused and transformation-focused ones, and the authors read this as a shift in the field over the preceding two years rather than a settled verdict.
  - Transformation-focused systems, the ones that segment a meeting before summarising it, score worst, which the authors attribute partly to age and partly to the difficulty of segmenting a meeting where participants hold side conversations, forget things and come back, and get interrupted.
  - ROUGE is inadequate for this task and the authors demonstrate it with a worked example: a summary that substitutes a false noun scores higher than a correct paraphrase, because ROUGE matches surface strings.
  - Humans prefer abstractive summaries for conversation, unlike for documents, and the cited reason is that extractive summaries copy the noise and grammatical mistakes of spoken language and lose coherence.
  - Factual consistency is a standing failure: the authors cite a finding that nearly 30 percent of summaries generated by neural sequence-to-sequence models suffer from fact fabrication.
  - Decisions and action items are treated as a distinct sub-task with its own literature, based on identifying decision-related dialogue acts, not as a by-product of general summarisation.
- **Author's framing of the contribution:** > "In this paper, we provide an overview of the challenges raised by the task of abstractive meeting > summarization and of the data sets, models and evaluation metrics that have been used to tackle the > problems." (p. 1)
- **What this paper does NOT establish:**
  - It does NOT report any result computed from recognised speech. Every ROUGE score in it comes from a system reading a clean human transcript.
  - It does NOT report any end-to-end audio-to-summary system, and no system in its tables takes audio as input.
  - It does NOT provide a benchmark for action-item extraction. Actions are one of four sections of an AMI or ICSI reference summary, and the tabulated ROUGE scores are computed over whole summaries, not per section.
  - It does NOT evaluate any instruction-tuned large language model on meeting summarisation. The prompting paradigm is discussed as a future direction and the authors state explicitly that they know of no attempt to apply it to meeting summarisation.
  - It does NOT report human evaluations of the tabulated systems, nor any inter-annotator agreement of its own.
  - It does NOT establish that ROUGE differences between the top systems are meaningful, and the authors cite work showing that metrics disagree about rankings within any narrow scoring range.
  - It does NOT test speaker attribution error, or any condition where the summariser is given the wrong speaker labels.
- **Direct quotes (verbatim, source ground truth):**
  - "We note that all three corpora provide gold transcripts that have been either fully human-produced or human-corrected based on ASR-output to avoid compounded errors from ASR transcripts." (p. 4)
  - "On average, one AMI transcript contains 4,757 tokens and its summary has 322, while an article from the CNN/DailyMail dataset (Hermann et al., 2015) has an average of 781 tokens

and its summary, 56." (p. 2)
  - "Unfortunately, the ROUGE metric (Lin, 2004), which remains the standard for both meeting and general text summarization, scores system-produced summaries based purely on surface lexicographic matches with a (usually single) gold summary, making it unideal for assessing abstractive summaries." (p. 5)
  - "it is reported that nearly 30% of summaries generated by neural seq2seq models suffer from fact fabrication (Cao et al., 2018)." (p. 11)
  - "Automatic speech recognition (ASR) systems can produce transcription errors, for example, that are compounded as we progress through the summarization pipeline, making it risky to skip the laborious task of manual correction." (p. 4)
  - "In the absence of a clear winner for summary evaluation metrics, none of the alternatives has yet to be widely adopted." (p. 5)
- **Limitations:**
  - Author-stated: evaluation is itself a very challenging task, no metric correlates better with human judgement than the others across datasets, and none of the ROUGE alternatives has been widely adopted, so the survey's own comparison table rests on a metric its authors describe as unideal.
  - Author-stated: existing reference-based and reference-free metrics cannot reliably evaluate the zero-shot summaries that prompted models produce, giving falsely low scores to more abstractive output.
  - Inferred: the benchmark table pools results from papers that used different preprocessing, different input truncation and, for ROUGE-L, different sentence-splitting conventions, which the table marks for some rows with an asterisk but which makes cross-row comparison unsafe.
  - Inferred: AMI, the most used corpus, is role-played by participants who did not know each other, following a designed scenario, which the authors themselves call arguably overly well-behaved.
  - Inferred: the survey closes in 2022 and so predates the instruction-tuned model era it anticipates.
- **Cross-references in this index:**
  - See also: Kumar 2022 (an overlapping survey of the same task; this one is peer reviewed, has a stated taxonomy and corrects Kumar 2022's claim about how AMI and ICSI transcripts were produced); Kirstein 2024 (which supplies the empirical test of the metric inadequacy this survey argues for from first principles).
  - Contrast with: Cornell 2025 (which measures summary quality downstream of real recognition error; every number in this survey is upstream of it). Contrast with Abramovski 2025 and Niu 2024 (transcription accuracy, a different family; a system's tcpWER says nothing about the ROUGE score of a summary written from its output, and vice versa).
- **Relevance to platform:** This is the entry that shows where the seam is in the charter's `three-outputs` promise. The whole published record for meeting summarisation is built on around 280 hours of English meetings whose transcripts were cleaned by hand, and none of it measures a summary written from a recogniser's output. So a competitor's summarisation benchmark claim and this project's actual product are not the same measurement, and the project should not cite one for the other. The survey also gives the best available definition of what the third output is: decisions and actions are a separate sub-task with its own dialogue-act-based literature, not something a general summariser produces for free, and the AMI and ICSI reference summaries treat Decisions and Actions as their own capped sections. Finally, the 30 percent fact-fabrication figure it cites is the honest counterweight to the Cornell 2025 finding that summaries survive transcription error: they survive the errors, but they invent things on their own.
- **Quotable stats (paste-ready):**
  - "There are only three English meeting corpora with summaries in existence, together offering about 280 hours, and all three ship human-produced or human-corrected transcripts rather than recogniser output (Rennard 2023)."
  - "An AMI meeting transcript averages 4,757 tokens against 781 for a news article, and its reference summary 322 tokens against 56, so meeting summarisation is roughly a six-times-longer input problem than the task most summarisers were built for (Rennard 2023)."

- "The best reported abstractive meeting summarisation scores are 55.27 ROUGE-1 on AMI and 49.25 on ICSI, both measured on clean human transcripts with no recognition error in the pipeline (Rennard 2023)."
- "A peer-reviewed survey of abstractive meeting summarisation reports that nearly 30 percent of summaries generated by neural sequence-to-sequence models suffer from fact fabrication (Rennard 2023, citing Cao et al. 2018)."
- "The AMI and ICSI reference summaries treat Decisions and Actions as their own separately annotated sections, capped at 200 words each, which is evidence that action-item extraction is a distinct task rather than a by-product of summarising (Rennard 2023)."

Golia 2023 - action items folded into the summary itself, scored on gold AMI transcripts (N=AMI corpus)

- **Citation:** Logan Golia, Jugal Kalita. "Action-Item-Driven Summarization of Long Meeting Transcripts." In 2023 7th International Conference on Natural Language Processing and Information Retrieval (NLPIR 2023), December 15-17, 2023, Seoul, Republic of Korea. ACM, New York. DOI: 10.1145/3639233.3639253. Preprint arXiv:2312.17581v2, 6 January 2024. Affiliations as printed: Rice University, USA; University of Colorado, Colorado Springs, USA.
- **File:** `literature/arxiv-2312.17581.pdf`
- **Links:** arXiv:2312.17581 | DOI: 10.1145/3639233.3639253 | Code: https://github.com/logangolia/meeting-summarization (named in the paper, not verified here) | Weights: not released; the paper uses a publicly available BART checkpoint fine-tuned on XSUM and SAMSUM | Project page: none | Citations: not retrieved | Reproduced: unknown
- **Evidence tier:** [C] watch METHOD. Code is released, which clears the implementation arm, but NLPIR is not on this index's venue list and the citation count is unverified, so no strong signal is cleared.
- **Mechanism family:** `Abstractive summarisation method`, `Decision and action-item extraction` (+ `Summarisation evaluation metric`)
- **Study type:** methods paper with a four-way segmentation comparison; one corpus, no human evaluation.
- **Population:** the AMI meeting corpus, described in this paper as 137 scenario-driven meetings with corresponding summaries. No participant demographics. The action-item classifier was trained on a separate public dataset from a GitHub repository containing 2,750 dialogue statements labelled for whether they contain an action item.
- **Imaging / measurement:** NO AUDIO IS PROCESSED ANYWHERE IN THIS PAPER. Every experiment runs on the AMI corpus's human transcripts and is scored against its human reference summaries. There is no recogniser, no word error rate and no recording condition; the input is clean text.
- **Methods / model:** a recursive divide-and-conquer pipeline. Long transcripts are split into topical chunks by one of four methods: linear segmentation by token count (the baseline), chunked linear segmentation that never cuts a speaker turn, simple cosine segmentation that starts a new chunk when the cosine similarity between consecutive turns' MPNet sentence embeddings falls to zero or below, and complex cosine segmentation that additionally suppresses short meaningless turns. Each chunk is summarised in parallel by a BART model fine-tuned on the XSUM and SAMSUM datasets, with a 1024-token input limit. Action items are extracted per chunk by a BertForSequenceClassification model fine-tuned on the 2,750-statement dataset, and appended to the chunk before summarisation. The concatenated chunk summaries are then fed back into the summariser recursively until they fit. Scoring: BERTScore plus ROUGE-1, ROUGE-2 and ROUGE-L.
- **Statistical detail:**
  - Test: none. No significance test, no confidence interval, no seed variance.
  - Per-arm N: eight configurations (four segmentation methods, with and without action items), each scored once across the AMI corpus.
  - Effect sizes: absolute metric differences and relative percentages.
  - p-values: none reported.
  - Power / pre-registration: not applicable and not reported.
- **Key findings:**

  - Best configuration, action-item-driven summaries with chunked linear segmentation: BERTScore 64.98, ROUGE-1 36.27, ROUGE-2 8.31, ROUGE-L 19.62.
  - The comparison state of the art, Shinde et al. 2022, is reported in the same table as BERTScore 60, ROUGE-1 45.2, ROUGE-2 13.3. So this paper's claimed 4.98 percent improvement is on BERTScore only, and on the ROUGE metrics it is roughly 9 ROUGE-1 points and 5 ROUGE-2 points WORSE than the system it says it improves on.
  - Adding action items lowered ROUGE and raised BERTScore. General summaries with chunked linear segmentation: BERTScore 64.77, ROUGE-1 38.93. Action-item-driven with the same segmentation: BERTScore 64.98, ROUGE-1 36.27. The authors attribute the ROUGE drop to action items adding words the human reference summaries do not contain.
  - The best segmentation method beat plain linear segmentation by 1.36 BERTScore points (64.77 against 63.41 for general summaries), and simply not cutting a speaker turn mid-sentence accounted for that entire gain; both cosine-similarity methods did worse than chunked linear.
  - The action-item classifier reached 95.4 percent classification accuracy on the held-out test split of the 2,750-statement dataset it was trained on.
  - Raising the cosine-similarity threshold from 0 to 0.2 cost more than 1 percent on both BERTScore and ROUGE-L, so the segmentation method is sensitive to a hyper-parameter chosen by manual inspection.
- **Author's framing of the contribution:** > "Our novel parallel and recursive meeting summarization algorithm properly generates action-item-driven > summaries and improves upon the performance of current state-of-the-art models by approximately 4.98% in > terms of the BERTScore metric." (p. 2)
- **What this paper does NOT establish:**
  - It does NOT process audio. Every number comes from AMI's human transcripts, so nothing here says anything about a summary produced from recognised speech.
  - It does NOT score the action items. The 95.4 percent accuracy belongs to a sentence classifier on an unrelated 2,750-statement dataset, not to the action items appearing in the AMI summaries. No precision, recall or human judgement of the extracted action items on AMI is reported anywhere.
  - It does NOT beat the prior state of the art on ROUGE, and the paper's abstract does not say so.
  - It does NOT include any human evaluation. Nobody read a summary.
  - It does NOT report latency, cost or the number of model calls the recursive algorithm makes, although parallelism is claimed as a contribution.
  - It does NOT test any corpus other than AMI, and does not test ICSI.
- **Direct quotes (verbatim, source ground truth):**
  - "Our pipeline achieved a BERTScore of 64.98 across the AMI corpus, which is an approximately 4.98% increase from the current state-of-the-art result produced by a fine-tuned BART (Bidirectional and Auto-Regressive Transformers) model." (p. 1)
  - "Another very important component of any good meeting summary is what each participant has accomplished and what they need to accomplish before the next meeting; so for each chunk of text, we need to extract the action items. Although recording action items is an important part of many meeting summaries, the issue has been ignored in prior work." (p. 3)
  - "This training method proved effective with a classification accuracy of 95.4% on the test" (p. 3)
  - "unlike a dialogue, useful meeting minutes have additional features that are often not included in the automated summary of the meeting: action items, main topics, tension levels, decisions made, etc." (p. 1)
  - "As seen in Table 1, our action-item-driven summaries achieve slightly higher BERTScores than our general summaries (without action items)" (p. 6)
- **Limitations:**
  - Author-stated: ROUGE scores fall when action items are added, because action items are absent from the human reference summaries.
  - Inferred: the headline improvement is metric-selective. On the two ROUGE metrics reported

    for the prior system, this pipeline is substantially worse, and the paper's abstract quotes only BERTScore.
  - Inferred: the action-item component, which the title is built on, is never evaluated on AMI at all.
  - Inferred: the whole result depends on BERTScore being the better proxy for human judgement, which the paper asserts by citation rather than by measuring anything.
  - Inferred: the segmentation improvement of 1.36 BERTScore points comes from not splitting speaker turns, which is an implementation detail rather than a topic-segmentation method.
- **Cross-references in this index:**
  - See also: Rennard 2023 (the survey establishing that AMI and ICSI summarisation scores are all computed on human transcripts); J. Liu 2023 (the other action-item paper in this vault, which does score action items directly and reports 43.12 positive F1 on AMI); Kirstein 2024 and Kirstein 2024b (why a BERTScore of 64.98 is not evidence that the summary is correct).
  - Contrast with: Cornell 2025 (the one experiment in the vault that summarises RECOGNISED speech rather than a gold transcript); Shapira 2025 (measures what happens to summary quality as transcript error rises, which this paper does not do).
- **Relevance to platform:** The charter names action items as one of three deliverables, and this is the only paper in the vault that builds them into the summarisation pipeline itself rather than treating them as a separate classification task. The result the project should take is negative and useful: the standard metrics punish a summary for containing action items, because the human reference summaries were not written to include them. That means the project cannot use ROUGE to tell whether its action items are good, and this paper's own approach of switching to the metric that moved in the right direction is exactly the failure mode the vault's evaluation cluster documents.
- **Quotable stats (paste-ready):**
  - "The best action-item-driven summarisation pipeline scored BERTScore 64.98 and ROUGE-1 36.27 on the AMI corpus, computed entirely on human transcripts with no recognition error present (Golia 2023)."
  - "Adding extracted action items to a meeting summary raised BERTScore from 64.77 to 64.98 and lowered ROUGE-1 from 38.93 to 36.27, because the human reference summaries do not contain action items (Golia 2023, AMI)."
  - "The prior state of the art on the same corpus scored BERTScore 60 but ROUGE-1 45.2 and ROUGE-2 13.3, against 36.27 and 8.31 here, so the reported improvement holds on one metric and reverses on two (Golia 2023)."
  - "Ensuring that transcript chunks never split a speaker's turn improved BERTScore by 1.36 points over splitting purely by token count (Golia 2023, AMI)."

J. Liu 2023 - the action-item corpus that exists, and the two that do not (N=424 Chinese meetings, 101 AMI)

- **Citation:** Jiaqing Liu, Chong Deng, Qinglin Zhang, Qian Chen, Wen Wang. "Meeting Action Item Detection with Regularized Context Modeling." Accepted paper, IEEE, 2023 (ICASSP 2023). DOI: not provided in the PDF. Preprint arXiv:2303.16763v1, 27 March 2023. Affiliation as printed: Speech Lab of DAMO Academy, Alibaba Group.
- **File:** `literature/arxiv-2303.16763.pdf`
- **Links:** arXiv:2303.16763 | DOI: not provided | Code: https://github.com/alibaba-damo-academy/SpokenNLP/tree/main/action-item-detection (named in the paper, not verified here) | Weights / data: the AMC-A corpus is stated to be released at https://www.modelscope.cn/datasets/modelscope/A (named in the paper, not verified here) | Project page: none | Citations: not retrieved | Reproduced: unknown
- **Evidence tier:** [B] credible METHOD. The PDF carries an IEEE accepted-paper notice, and ICASSP is on this index's venue list, which is one strong signal; a public implementation and a public corpus are both named, which clears the implementation arm; the citation rate is unverified.
- **Mechanism family:** `Decision and action-item extraction`, `Corpus resource`
- **Study type:** corpus construction plus a methods paper with an ablation; two corpora, five random seeds per configuration.

- **Population:** two corpora. AMC-A (AliMeeting-Action Corpus), the corpus this paper builds: 424 Chinese meetings, 306,846 utterances, 1,506 action items, split 295 / 65 / 64 meetings into train / dev / test at a 70:15:15 ratio. Each session is a 15-to-30-minute discussion by 2 to 4 participants on topics biased towards work meetings in various industries. It extends 224 meetings previously published as AliMeeting with 200 additional meetings. Mean 3.55 action items per meeting, standard deviation 3.97. AMI, used for comparison: 101 annotated meetings with 381 action items, 80,298 utterances, mean 3.77 action items per meeting.
- **Imaging / measurement:** AMC-A annotation is on MANUAL TRANSCRIPTS of meeting recordings, with punctuation inserted by hand; no audio is processed in the experiments. Sentences are defined as semantic units ending in a manually labelled period, question mark or exclamation. Inter-annotator agreement: average Cohen's kappa of 0.47 between pairs of annotators on AMC-A. Three annotators labelled each candidate sentence independently; an expert reviewed the majority vote and modified 5 to 10 percent of the majority-voting labels. Annotation cost was reduced by pre-selecting candidate sentences containing both a temporal expression and an action-related verb and highlighting them.
- **Methods / model:** action-item detection is framed as sentence-level binary classification. Pre-trained language models compared: BERT, RoBERTa (Chinese RoBERTa-wwm-ext), StructBERT and Longformer (Erlangshen-Longformer-110M, used for a sequence-labelling formulation with a 4096 sliding window). All are BERT-base size; input truncated to 128 tokens for the classification formulation. Two contributions: Context-Drop, a contrastive regularisation forcing the prediction distribution for a bare focus sentence and for the same sentence plus its context to agree via bidirectional Kullback-Leibler divergence; and Lightweight Model Ensemble, initialising the encoder from one pre-trained model and the pooler layer from another. Local context is the preceding and following sentence; global context is the top-2 most similar sentences by n-gram cosine similarity. Batch size 32, dropout 0.3, five runs per experiment with a grid search over learning rate {1e-5, 2e-5} and epochs {2, 3}. Metric: positive-class F1.
- **Statistical detail:**
  - Test: none. No significance test is reported.
  - Per-arm N: five runs with different random seeds per configuration; mean and standard deviation reported.
  - Effect sizes: absolute F1 differences with standard deviations, for example 70.82 +/- 1.33 against 67.84 +/- 1.20.
  - p-values: none reported.
  - Power / pre-registration: not applicable and not reported.
- **Key findings:**
  - AMI's 101 action-item-annotated meetings are DERIVED, not annotated. The paper states there are no direct action-item annotations for AMI, and that indirect labels are generated by treating dialogue acts linked to the action-related abstractive summary as positive samples, following prior work. This is how the 381 action items arise.
  - ICSI has no publicly available action-item annotation. The paper says ICSI "comprises only 75 meetings without publicly available action item annotations" (p. 1) and separately that ICSI "has action item annotations for 18 meetings" from Purver et al. 2007 which "are no longer publicly available" (p. 2).
  - Action-item annotation has low agreement even with quality control. AMC-A reached an average pairwise Cohen's kappa of 0.47 (per split: 0.46 train, 0.49 dev, 0.50 test). The paper cites a kappa of 0.36 on the ICSI corpus from prior work as the state of the field before this.
  - Best result on AMC-A: 70.82 +/- 1.33 positive F1, with the focus sentence plus local and global context and dynamic Context-Drop, against 67.84 +/- 1.20 for the bare sentence baseline, a gain of 2.98 absolute.
  - Best result on AMI: 43.12 +/- 0.74 positive F1, with the focus sentence plus local context and fixed Context-Drop, against 38.67 +/- 1.25 for the bare sentence baseline. The absolute level on English AMI is roughly 27 F1 points below the Chinese corpus.
  - Removing the Kullback-Leibler regularisation term degraded both corpora, so the gain is from contrastive regularisation rather than from data augmentation.
  - Long-sequence modelling gave almost nothing. Longformer as a sequence labeller beat BERT

by 0.59 F1, while switching from BERT to StructBERT as a sentence classifier gained 3.08.

- **Author's framing of the contribution:** > "We construct and make available a Chinese meeting corpus with action item annotations, to alleviate > scarcity of resources and prompt related research. To the best of our knowledge, this is so far the largest > meeting action item detection corpus." (p. 1)
- **What this paper does NOT establish:**
  - It does NOT process audio. All annotation and all experiments are on manual transcripts with hand-inserted punctuation, so nothing here says how action-item detection behaves on a recognised transcript.
  - It does NOT establish that AMI has 381 human-annotated action items. Those labels are derived from dialogue acts linked to the action-related section of the abstractive summary.
  - It does NOT extract the content of an action item. The task is binary sentence classification: does this sentence contain action-item information, yes or no. Owner, deadline and task description are not predicted.
  - It does NOT produce a summary or evaluate one.
  - It does NOT test English beyond the derived AMI labels, and its own corpus is Chinese only.
  - It does NOT report inference latency or model size beyond "BERT base size".
- **Direct quotes (verbatim, source ground truth):**
  - "We obtain 101 annotated AMI meetings with 381 action items following previous works. The ICSI meeting corpus comprises only 75 meetings without publicly available action item annotations." (p. 1)
  - "Although there are no direct annotations for action items for this corpus, indirect annotations can be generated based on annotations of the summary. Following previous works, we consider dialogue acts linked to the action-related abstractive summary as positive samples for action item detection and otherwise negative samples. In this way, we obtain 101 annotated meetings with 381 action items." (p. 2)
  - "Another public meeting corpus, the ICSI meeting corpus, has action item annotations for 18 meetings and is much smaller for action item detection research. Also, these annotations are no longer publicly available." (p. 2)
  - "As found in previous research and our experience, annotations of action items have high subjectivity and low consistency, e.g., only a Kappa coefficient of 0.36 on the ICSI corpus." (p. 2)
  - "With these quality control methods, the average Kappa coefficient on AMC-A between pairs of annotators is 0.47. For inconsistent labels from three annotators, an expert reviews the majority voting results and decides on final labels." (p. 2)
  - "the expert only modifies 5%-10% of the majority voting labels" (p. 4)
- **Limitations:**
  - Author-stated: action-item annotation is highly subjective with low consistency, which is why the guidelines, the candidate pre-selection and the expert adjudication were necessary.
  - Inferred: a pairwise kappa of 0.47 is moderate agreement at best, and it caps how good any automatic system on this corpus can be judged to be.
  - Inferred: candidate sentences were pre-selected by a heuristic requiring both a temporal expression and an action verb, so action items phrased without either are absent from the corpus by construction, and the reported F1 is on an easier distribution than a raw transcript.
  - Inferred: the 43.12 F1 on AMI means that on the only English data in the paper, more than half the positive predictions or positives are wrong.
- **Cross-references in this index:**
  - See also: Chen 2016 (the other action-item resource in this vault, 22 ICSI meetings annotated with 10 actionable-item intents, which is a different annotation scheme on a different subset from the 18-meeting set this paper says is unavailable); Golia 2023 (folds action items into the summary and never scores them); Rennard 2023 (the survey that treats action items as a named sub-task with no benchmark).
  - Contrast with: every recognition entry in this index, because none of them supply the transcript this task would run on in production.
- **Relevance to platform:** This paper settles what the project can and cannot know about one of

its three named deliverables. There is exactly one purpose-built action-item corpus with manual annotation, it is Chinese, it is 424 meetings, and its inter-annotator kappa is 0.47. The English resource everyone cites, 101 AMI meetings with 381 action items, is derived from summary links rather than annotated, and ICSI's action-item labels are not publicly available. The best English positive F1 reported here is 43.12. For the charter's `three-outputs` constraint, that means the project cannot benchmark its action items against anything English and human-labelled, and any internal claim about action-item quality will need its own annotation effort at an agreement level around 0.47 kappa.

- **Quotable stats (paste-ready):**
  - "The AMI corpus has no direct action-item annotation; the widely cited 101 meetings with 381 action items are derived by treating dialogue acts linked to the action-related abstractive summary as positive samples (J. Liu 2023)."
  - "The ICSI corpus has action-item annotations for 18 meetings that are no longer publicly available, so there is no public English meeting corpus with human-annotated action items (J. Liu 2023)."
  - "The largest purpose-built action-item corpus is AMC-A: 424 Chinese meetings, 306,846 utterances and 1,506 action items, with an average pairwise Cohen's kappa of 0.47 between annotators (J. Liu 2023)."
  - "The best reported positive F1 for action-item detection is 70.82 on Chinese meeting transcripts and 43.12 on English AMI, both on manual transcripts with no recognition error (J. Liu 2023)."
  - "Prior work reports a Cohen's kappa of only 0.36 for action-item annotation on the ICSI corpus (cited in J. Liu 2023)."

Shapira 2025 - how much transcription error a downstream task survives, and which errors matter (N=3 tasks)

- **Citation:** Ori Shapira, Shlomo E. Chazan, Amir DN Cohen. "Measuring the Effect of Transcription Noise on Downstream Language Understanding Tasks." Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 29978-30004, July 27 - August 1, 2025. Association for Computational Linguistics. DOI: not provided in the PDF. Affiliation as printed: OriginAI.
- **File:** `literature/acl-2025.acl-long.1449.pdf`
- **Links:** Code: https://github.com/OriShapira/ENDow (named in the paper, not verified here) | DOI: not provided | Weights: not applicable; the paper uses public checkpoints | Project page: none | Citations: not retrieved | Reproduced: unknown
- **Evidence tier:** [B] credible METHOD. ACL main conference long paper, a top venue on this index's list, which is one strong signal, and code is released; the citation rate is unverified.
- **Mechanism family:** `Downstream-task robustness to transcription error` (+ `Abstractive summarisation method`, `Summarisation evaluation metric`)
- **Study type:** methods paper introducing an evaluation framework, applied across three tasks, four models, seven noise levels and seven transcript-cleaning techniques.
- **Population:** three test sets, none of them re-annotated by these authors. QMSum, for summarisation: 35 transcripts, 281 query instances, averaging 592 utterances per instance (range 131 to 1368), drawn from AMI product meetings, ICSI academic meetings and parliamentary committee meetings. QAConv, for question answering: 505 transcripts, 2,083 questions, from court cases and interviews. MRDA, for dialogue-act classification: 12 transcripts, 1,200 utterances, from ICSI and research-oriented group meetings; the first and last 50 utterances of each transcript were used.
- **Imaging / measurement:** THE AUDIO IS SYNTHETIC AND SO IS THE DEGRADATION. The datasets' reference transcripts were converted to speech with the tortoise-tts text-to-speech library, then each audio file was reverberated with the rir-generator library and mixed with a clipped recording of background office sounds at one of five signal-to-noise ratios, giving six increasingly damaged audio sets plus the clean one. Those were transcribed with openai/whisper-small.en. The resulting word error rates span roughly 0 to 0.9. There is no real meeting recording anywhere in the pipeline,

no microphone, no overlapping speakers and no multi-party turn-taking in the acoustics.

- **Methods / model:** the framework defines a noise-toleration point (NTP), the lowest word error rate at which a task score becomes statistically significantly lower than the score on clean reference transcripts, and an area under the score-versus-error curve (AUC) for comparing models. Seven cleaning techniques repair one word class each in the noisy transcript by aligning it to the reference with jiwer: nouns, verbs, adjectives, adverbs, all content words, all non-content words, and named entities. These are an upper-bound oracle, since they use the reference transcript. A cleaning-effectiveness score (CES) rewards task-score gain per unit of word-error-rate repair. Four instruction-tuned language models run every task zero-shot: Mistral-7B-Instruct-v0.3, Meta-Llama-3-8B-Instruct, Llama-3.1-8B-Instruct and gpt-4o-mini. For Mistral and Llama-3 the transcripts do not fit the context window, so summarisation and question answering were done in segments. SUMMARY QUALITY WAS JUDGED BY A LANGUAGE MODEL, NOT BY PEOPLE: pairwise comparison ranking with gpt-4o-mini as the judge, alongside ROUGE-1, ROUGE-2 and ROUGE-L.
- **Statistical detail:**
  - Test: significance at $p < 0.05$ is used to define the noise-toleration point, comparing the lower confidence bound of the clean score against the upper confidence bound of the noisy score.
  - Per-arm N: 281 summarisation instances, 2,083 questions, 1,200 utterances, each run at seven noise levels with four models and eight cleaning conditions.
  - Effect sizes: AUC reported with a margin of error at the 95 percent confidence level, for example "5.82 +/- 0.32".
  - p-values: not reported individually; the $p < 0.05$ threshold is applied to the confidence bands.
  - Power / pre-registration: not reported.
- **Key findings:**
  - THE THRESHOLD CLAIM, EXACTLY AS THE PAPER STATES IT: "models tend to tolerate a noise level of about 0.2 WER (NTP between 0.07 and 0.3), i.e., task scores are not significantly lower ($p < 0.05$) until that level of noise" (p. 7). The per-model noise-toleration points on QMSum under pairwise ranking are Mistral 0.195, Llama3 0.070, Llama3.1 0.249 and gpt-4o-mini 0.297. Only two of the four models reach 0.25, and one degrades significantly at 7 percent word error rate.
  - The conditional half is measured, and it is the paper's real contribution. Repairing only named entities was the most cost-effective cleaning technique for summarisation with gpt-4o-mini, cleaning-effectiveness 0.499, just above repairing all content words at 0.479 and well above nouns 0.305, adjectives 0.135, verbs 0.073 and adverbs -0.023. A negative score means the repair made the task result worse on average.
  - The same ordering holds for question answering across all four models: named entities first, then nouns, then content words, with verbs and adverbs last.
  - Dialogue-act classification behaves oppositely: repairing NON-content words was effective, consistent with function words carrying the dialogue act.
  - Repairing content words in a transcript at roughly 0.9 word error rate brought it to about 0.4, and the summaries from that repaired transcript were much preferred over summaries from a differently-damaged transcript at the same 0.4 word error rate. The type of error, not just the amount, changes the outcome.
  - Model ranking flips with noise level. gpt-4o-mini produced the most preferred summaries at low word error rate and fell to the bottom at high word error rate, while Llama-3.1 overtook it.
  - Noise-toleration points for dialogue-act classification were high (0.439 to 0.709), which the authors attribute not to robustness but to the models being poor at that task in the first place, citing lower macro-F1 than published results.
- **Author's framing of the contribution:** > "Our first-of-its-kind framework examines task model behavior under varying noise intensities and types, > providing quantitative metrics and facilitating qualitative analyses." (p. 2)
- **What this paper does NOT establish:**
  - It does NOT establish a 25-to-30-percent word error rate threshold for meeting summarisation. Its own stated range is a noise-toleration point between 0.07 and 0.3 across four models, centred at about 0.2, and the two models reaching 0.25 or above are two of four.
  - It does NOT use real meeting audio. The audio is text-to-speech synthesis of reference

transcripts, degraded with a room-impulse-response simulator and an office-noise recording. There is no overlapping speech, no genuine far-field microphone and no real speaker.
  - It does NOT have humans judge any summary. Summary quality is pairwise ranking by gpt-4o-mini plus ROUGE.
  - It does NOT provide a usable cleaning method. All seven cleaning techniques repair the noisy transcript by consulting the reference transcript, which is an oracle and is stated as such.
  - It does NOT test speaker attribution. The pipeline synthesises and re-transcribes text; who said what is never at risk and never scored.
  - It does NOT test non-English data, other acoustic settings, or a real recognition system other than whisper-small.en.
- **Direct quotes (verbatim, source ground truth):**
  - “models tend to tolerate a noise level of about 0.2 WER (NTP between 0.07 and 0.3), i.e., task scores are not significantly lower ($p < 0.05$) until that level of noise.” (p. 7)
  - “In all three tasks, fixing only adverbs or verbs is less effective. On the other end, nouns, and named-entities particularly, are more effective for the transcript-level tasks (summarization and QA).” (p. 8)
  - “This further stresses the importance of analyzing the types of errors in transcripts and not just the amount, which is a known limitation of the WER metric.” (p. 8)
  - “as the analysis will show, prioritizing named-entity accuracy may be more effective than generally minimizing the WER of the ASR system.” (p. 8)
  - “In our experiments, the pipelines are initiated with relatively clean and clear audio files, and the subsequent acoustic deterioration is done in a specific manner (reverberation and background sounds). Other acoustic settings are indeed possible for initiating the SLU pipeline, e.g., with a low-resourced lingual dialect, different speaker voices per turn, overlapping speech, microphone settings, and many other parameters.” (p. 9)
  - “The cleaning techniques we used depend on the reference transcript in order to identify the word/phrase types that we want to include in our analysis.” (p. 9)
- **Limitations:**
  - Author-stated: the pipeline starts from clean synthetic audio and the degradation is a specific reverberation-plus-noise recipe; overlapping speech, different speaker voices per turn and microphone settings are named as untested parameters.
  - Author-stated: the cleaning techniques are reference-dependent and therefore an upper bound.
  - Author-stated: the conclusions depend on the task metric chosen, and can differ under a different metric for the same task and data.
  - Inferred: the summarisation judgement is a language model judging language-model output, exactly the protocol that Kirstein 2024 and Gong 2024 in this index show to be weakly correlated with human judgement on meeting summaries.
  - Inferred: two of four models needed segmentation to fit the transcripts in context, so their scores conflate context-window effects with noise effects.
- **Cross-references in this index:**
  - See also: Pandey 2023 (the recognition-side method for protecting exactly the named entities this paper identifies as the highest-value words); Cornell 2025 (the only other measurement in the vault of summary quality against transcription error, on real challenge audio, finding a weak correlation); Shon 2023 (finds a strong negative correlation of -0.9 between word error rate and ROUGE-L on TED talk summarisation).
  - Contrast with: Kirstein 2024 and Kirstein 2024b (score summary ERRORS against human annotation, not summary preference against transcript noise); Golia 2023 and Rennard 2023 (summarisation on clean gold transcripts, no noise axis at all).
- **Relevance to platform:** This is the paper the project’s whole summarisation argument rests on, and read precisely it supports a weaker and more conditional claim than the one in circulation. Four instruction-tuned models tolerated roughly 20 percent word error rate before summary quality dropped significantly, with the range across models running from 7 to 30 percent; and which words the errors fall on changed the outcome as much as how many there were, with named entities the single most valuable class to protect. For the charter’s `three-outputs` constraint that is genuinely

encouraging, because the far-field single-channel error rates elsewhere in this vault sit at 21 to 42 percent and the top of that range is not obviously fatal to a summary. But the audio here is synthesised, undegraded by overlap, and the judge is a language model, so the project cannot present this as evidence that its own recordings will summarise well.
- **Quotable stats (paste-ready):**
  - "Across four instruction-tuned language models summarising meeting transcripts, summary quality did not drop significantly until roughly 20 percent word error rate, with per-model thresholds ranging from 7 to 30 percent (Shapira 2025, QMSum, 281 instances)."
  - "Repairing only the named entities in a noisy transcript was the most cost-effective correction for summarisation, scoring 0.499 on the paper's cleaning-effectiveness measure against 0.073 for verbs and -0.023 for adverbs (Shapira 2025, gpt-4o-mini on QMSum)."
  - "The authors state that prioritising named-entity accuracy may be more effective than minimising overall word error rate (Shapira 2025)."
  - "All of these results come from text-to-speech audio artificially reverberated and noised, transcribed by whisper-small.en, with summaries ranked by gpt-4o-mini rather than by human readers (Shapira 2025)."

Shon 2023 - a spoken-language benchmark, and what it does and does not say about meeting corpora (N=4 tasks)

- **Citation:** Suwon Shon, Siddhant Arora, Chyi-Jiunn Lin, Ankita Pasad, Felix Wu, Roshan Sharma, Wei-Lun Wu, Hung-Yi Lee, Karen Livescu, Shinji Watanabe. "SLUE Phase-2: A Benchmark Suite of Diverse Spoken Language Understanding Tasks." arXiv:2212.10525v2, 15 June 2023. DOI: not provided in the PDF. Affiliations as printed: ASAPP; Carnegie Mellon University; National Taiwan University; Toyota Technological Institute at Chicago.
- **File:** `literature/arxiv-2212.10525.pdf`
- **Links:** arXiv:2212.10525 | DOI: not provided | Code: stated in the paper as "To be released"; no repository URL is given in the PDF | Weights: not released in the PDF | Project page: none | Citations: not retrieved | Reproduced: unknown
- **Evidence tier:** [C] watch METHOD. The PDF states no venue acceptance, its own code footnote reads "To be released", and the citation count is unverified.
- **Mechanism family:** `Downstream-task robustness to transcription error`, `Corpus resource` (+ `Abstractive summarisation method`, `Summarisation evaluation metric`)
- **Study type:** benchmark construction and baseline evaluation across four tasks, with a sensitivity analysis over more than 20 recognition models.
- **Population:** four new or extended datasets, none of them meeting corpora. SLUE-HVB for dialogue-act classification: the Harper Valley Bank scripted consumer-banking corpus, 23 hours of audio, 1,446 conversations simulated by 59 speakers, re-annotated by professional annotators with a new 18-class dialogue-act set; splits of 11,344 / 1,690 / 6,121 utterances. SLUE-SQA-5 for question answering: 46,186 fine-tune, 1,939 dev and 2,382 test questions, spoken by 931 / 41 / 51 crowdworkers on Amazon Mechanical Turk, paired with 40-second spoken Wikipedia documents from about 400 real speakers. SLUE-TED for summarisation: 3,384 / 425 / 424 TED talks, 664 / 81 / 84 hours. SLUE-VoxPopuli for named entity localization: 1,750 dev and 1,838 test utterances, 5.0 and 4.9 hours.
- **Imaging / measurement:** every audio condition here is single-speaker and near-field or broadcast-quality. Harper Valley Bank is scripted simulated telephone banking; SLUE-SQA-5 questions are crowdworker recordings and documents are Spoken Wikipedia readings; SLUE-TED is stage-microphone conference talks; SLUE-VoxPopuli is European Parliament recordings. There is no far-field meeting audio and no overlapping speech in the benchmark.
- **Methods / model:** for each task, an end-to-end baseline built on wav2vec2, and pipeline baselines pairing a recogniser (wav2vec2 fine-tuned in-domain, NVIDIA NeMo models, or Whisper) with a text model (DeBERTa for classification and question answering, a Longformer Encoder-Decoder initialised from BART-large for summarisation). An oracle pipeline using ground-truth transcripts is run for every task as an upper bound. Summarisation is scored with ROUGE, METEOR and BERTScore.

- **Statistical detail:**
  - Test: Pearson correlation between recognition word error rate and task score, with significance reported.
  - Per-arm N: baseline systems plus all available NeMo models, more than 20 recognisers in total.
  - Effect sizes: Pearson correlation coefficients.
  - p-values: reported as $p < 0.01$ for the question-answering and summarisation correlations.
  - Power / pre-registration: not reported.
- **Key findings:**
  - RC-21 CHECK. This paper does NOT say that ICSI Actions annotations are sparse, and it makes no claim about action items at all. Its only statement about ICSI is a single clause in the related-work section: “Other corpora, such as the ICSI and AMI meeting summarization corpora, contain relatively less annotated data” (p. 2). ICSI appears nowhere else in the paper except in the reference list. The word “action” appears in this paper only in the phrase “action prediction” describing the unrelated SLURP benchmark, and in the body of a quoted transcript.
  - Recognition quality and downstream task quality are strongly coupled across all four tasks. Dialogue-act classification against word error rate: Pearson -0.9. Question answering, document word error rate against frame-F1: Pearson -0.89, $p < 0.01$. Summarisation, word error rate against ROUGE-L: Pearson -0.9, $p < 0.01$.
  - The oracle-transcript ceiling is far above every real system. Dialogue-act classification: 72.3 macro-F1 with ground-truth transcripts against 70.7 with the best pipeline and 57.9 end-to-end. Question answering: 62.3 frame-F1 oracle against 43.3 best pipeline and 21.8 end-to-end. Summarisation: ROUGE-L 19.3 oracle against 18.6 best pipeline and 16.3 end-to-end.
  - The summarisation ceiling is low in absolute terms. The best system reached ROUGE-1 30.1, ROUGE-2 7.7 and ROUGE-L 19.3 even with a perfect transcript.
  - Named entities are where summarisation fails. In the end-to-end summarisation model, “only 6.6% of entities in the reference summary were found in the predicted summary” (p. 8), and the authors attribute the majority of summarisation errors to proper nouns.
  - On SLUE-TED, 66 percent of the words in a talk’s title and 57.4 percent of the words in its abstract already appear in the transcript.
- **Author’s framing of the contribution:** > “We introduce SLUE Phase-2, a set of SLU tasks that complement the existing SLU datasets or benchmarks.” > (p. 1)
- **What this paper does NOT establish:**
  - It does NOT say ICSI’s action-item or “Actions” annotations are sparse, scarce or inadequate. It says the ICSI and AMI SUMMARIZATION corpora contain relatively less annotated data, in one clause, with no numbers.
  - It does NOT evaluate any meeting corpus. AMI and ICSI are mentioned in related work and never used.
  - It does NOT test far-field, multi-speaker or overlapping audio anywhere.
  - It does NOT test action-item extraction, decision detection or any meeting-specific task.
  - It does NOT release code in the version indexed here; the footnote reads “To be released”.
  - It does NOT establish that its strong word-error-rate-to-summary-quality correlation transfers to meetings. It is measured on single-speaker TED talks scored by ROUGE-L, and the vault’s only meeting-audio measurement of the same relationship (Cornell 2025) found a much weaker coupling.
- **Direct quotes (verbatim, source ground truth):**
  - “Other corpora, such as the ICSI (Janin et al., 2003) and AMI (McCowan et al., 2005) meeting summarization corpora, contain relatively less annotated data.” (p. 2)
  - “We observe a strong correlation (Pearson correlation coefficient=-0.9, p-value<0.01) between WER and ROUGE-L scores, suggesting that we can boost SUMM performance using a stronger ASR model.” (p. 7)
  - “A similar analysis on the percentage of exact matches for named entities shows that only 6.6% of entities in the reference summary were found in the predicted summary.” (p. 8)
  - “Based on this analysis, we infer that the current speech summarization models struggle to correctly extract entities for the summary.” (p. 8)

  - "All the baseline models perform worse than the pipeline-oracle model suggesting room for improvement." (p. 7)
- **Limitations:**
  - Inferred: none of the four datasets contains multi-party conversational audio, so the benchmark cannot speak to meeting conditions despite being widely cited in meeting contexts.
  - Inferred: the summarisation task is TED-talk abstract generation, a single-speaker prepared monologue, which is a much easier discourse structure than a four-person meeting.
  - Inferred: the code was unreleased at the time of this version, so the baselines are not independently checkable from the PDF alone.
  - Inferred: the strong word-error-rate correlations are computed across recogniser families whose word error rates span 12 to 34 percent, which is a narrower and cleaner range than far-field meeting audio produces.
- **Cross-references in this index:**
  - See also: Shapira 2025 (the same question, transcription noise against downstream task quality, with a controlled noise axis rather than a set of recognisers); Cornell 2025 (the same question on real meeting audio, with a much weaker correlation of PCC -0.51).
  - Contrast with: Rennard 2023 and Kumar 2022 (meeting summarisation proper); J. Liu 2023 (the paper that does report ICSI's action-item annotation status, and reports it as 18 meetings no longer publicly available, not as sparse).
- **Relevance to platform:** This entry exists in the vault mainly to close a citation. A research response used this paper as the authority for a claim about ICSI's action-item annotations being sparse; the paper says nothing of the kind, and its only sentence about ICSI concerns summarisation annotation volume. What the paper does contribute is a well-powered measurement that word error rate and downstream text quality are strongly coupled on single-speaker audio, Pearson -0.9 for summarisation, which is worth holding beside the much weaker coupling Cornell 2025 measured on real meeting audio. The two disagree, and the difference is the audio condition.
- **Quotable stats (paste-ready):**
  - "On single-speaker TED talk summarisation, word error rate and ROUGE-L correlate at Pearson -0.9, $p < 0.01$, across more than 20 recognition systems (Shon 2023)."
  - "Even with a perfect ground-truth transcript, the best speech-summarisation baseline reached only ROUGE-L 19.3 and ROUGE-1 30.1 (Shon 2023, SLUE-TED, 424 test talks)."
  - "Only 6.6 percent of the named entities in the reference summary appeared in the end-to-end model's generated summary (Shon 2023, SLUE-TED)."
  - "Across four spoken-language understanding tasks, pipelines using ground-truth transcripts beat every speech-driven system, by 19 frame-F1 points on question answering and 1.6 macro-F1 points on dialogue-act classification (Shon 2023)."

Whether a summary or an action item can be scored at all

Kirstein 2024 - human annotators against nine automatic metrics (N=35 meetings, 175 annotated summaries)

- **Citation:** Frederic Kirstein, Jan Philip Wahle, Terry Ruas, Bela Gipp. "What's under the hood: Investigating Automatic Metrics on Meeting Summarization." Findings of the Association for Computational Linguistics: EMNLP 2024, November 2024. DOI: 10.18653/v1/2024.findings-emnlp.393. Preprint arXiv:2404.11124v2, 18 October 2024. University of Goettingen, Germany.
- **File:** `literature/arxiv-2404.11124.pdf`
- **Links:** arXiv:2404.11124 | DOI: 10.18653/v1/2024.findings-emnlp.393 | Code: https://github.com/FKIRSTE/emnlp2 Meeting-Sum-Metrics, stated in the paper to hold the codebase, annotations and annotator guidelines (named in the paper, not verified here) | Weights: not applicable | Project page: none | Citations: 2 (OpenAlex, published record, as of 2026-09) | Reproduced: unknown
- **Evidence tier:** [B] credible DOMAIN. Top venue for the language side (Findings of EMNLP) and a released artifact of code, annotations and guidelines, but a citation rate of about 1 per year, below the gate's floor.

- **Mechanism family:** `Summarisation evaluation metric`, `Human error-taxonomy annotation` (+ `Abstractive summarisation method`)
- **Study type:** literature review followed by a human-annotation study and a correlation analysis between human annotations and automatic metrics.
- **Population:** two populations. The papers: an initial pool of 300 papers from Google Scholar published between 2019 and 2024, narrowed under a PRISMA checklist to 18 core articles addressing challenges directly and 40 more addressing them implicitly. The annotators: four graduate students, two male and two female, aged 24 to 28, two in computer science and one each in psychology and communication science, all native or C1-C2 certified English speakers, employed on standard contracts as interns or doctoral candidates, onboarded over one week and annotating over five weeks at about 43 minutes per sample. The data: 35 general-summary samples from the QMSum test set, each summarised by five models, giving 175 annotated samples. QMSum itself comprises 232 meetings, 137 from AMI, 59 from ICSI and 36 Welsh and Canadian parliamentary committee meetings, averaging 556.8 turns, 9.2 speakers and 9,069.8 words per meeting with 69.6-word summaries.
- **Imaging / measurement:** no audio. Every input is a written transcript. Annotators answered a binary yes/no question per error type, for example “Does the given summary omit crucial information or provide insufficient detail about salient points?”, then rated frequency of challenges and impact of errors on 1-to-5 Likert scales for the primary model’s outputs, and gave written reasoning for each decision. Inter-annotator agreement by Krippendorff’s alpha: 0.79 for challenge detection and 0.82 for challenge frequency; 0.78 for error detection on encoder-decoder models, 0.76 on decoder-only models, and 0.83 for error impact; 0.81 on average.
- **Methods / model:** five summarisers, grouped into encoder-decoder models (Longformer Encoder Decoder, DialogLED, PEGASUS-X, all fine-tuned on the QMSum training subset for three epochs) and decoder-only models (GPT-3.5 turbo and Zephyr-7B-alpha, both zero-shot with a chunk-based prompt). Nine metrics across three families: count-based ROUGE-1, ROUGE-2, ROUGE-L, BLEU and METEOR; model-based BERTScore (rescaled F), perplexity computed with GPT-2, BLANC and LENS; and question-answering-based QuestEval. Challenges and error types were linked with point-biserial correlation, metrics were linked to annotated errors with point-biserial correlation, and metrics were linked to annotated error impact with Spearman correlation.
- **Statistical detail:**
  - Test: point-biserial correlation for metric-against-error-presence and challenge-against-error; Spearman rank correlation for metric-against-error-severity; Krippendorff’s alpha for agreement.
  - Per-arm N: 35 samples per model, 175 annotated samples in total; correlations between metrics and errors are computed on the 35 summaries generated by the Longformer Encoder Decoder model.
  - Effect sizes with CI when reported: correlation coefficients, no confidence intervals reported.
  - p-values: reported only as thresholds, one asterisk for $p <= 0.05$ and two for $p <= 0.01$. Exact p-values are not given.
  - Power / pre-registration: no power analysis and no pre-registration reported. The authors state that 35 samples provide statistical significance while acknowledging the sample covers only a fraction of meeting types.
- **Key findings:**
  - No metric correlates highly with all error types, and most correlations are weak to moderate. The largest absolute correlations in the whole table are 0.46 and under.
  - Metrics that behave as intended: ROUGE-1 against missing information, -0.40 ($p <= 0.05$); ROUGE-L against structural disorganisation, -0.46 ($p <= 0.01$); ROUGE-1 against structural disorganisation, -0.41; METEOR against structural disorganisation, -0.38; BLANC against incoherence, -0.42.
  - Metrics that reward errors, which is the finding that matters most for a product: perplexity against wrong references, +0.44 ($p <= 0.01$), meaning a summary that attributes a statement to the wrong participant scores better as long as it reads fluently; LENS against structural disorganisation, +0.45 ($p <= 0.01$).
  - In the severity analysis, BLEU against hallucination is +0.35 and QuestEval against hallucina-

tion is +0.34, both significant at p <= 0.05, so two metrics reward hallucinated content more the worse it gets.

  - About a third of the metric-error combinations either ignore the error, with a near-zero correlation, or reward it, with a positive one.
  - Hallucination is essentially invisible to every metric tested: its correlations run from -0.13 to +0.32 with no negative significance anywhere.
  - Error rates in the summaries themselves, out of 35 samples per model. Missing information appears in 89 to 97 percent of summaries across all five models. Structural disorganisation appears in 57 to 80 percent. Wrong references appear in 9 to 11 percent for four of the five models but 40 percent for PEGASUS-X. Hallucination appears in 14 to 37 percent. Incorrect reasoning appears in 3 percent for the two decoder-only models and 20 to 46 percent for the encoder-decoder ones.
  - The two model families fail differently. Encoder-decoder models tie incoherence, structural disorganisation and redundancy to spoken language and speaker dynamics. Decoder-only models tie wrong references strongly to spoken language and low information density, and tend to list topics without context.
  - The challenges are almost universal in the source data. In the 35 QMSum transcripts, speaker dynamics, coreference and low information density were detected in 100 percent, discourse structure in 97 percent, contextual turn-taking in 94 percent and spoken language in 91 percent. Only implicit context was rare, at 14 percent.
- **Author's framing of the contribution:** > "Current established metrics struggle to capture the observable errors, showing weak to moderate > correlations, with a third of the correlations indicating error masking." (p. 1)
- **What this paper does NOT establish:**
  - It does NOT test any audio or any recogniser output. Every transcript is a clean human transcript from QMSum, so this paper says nothing about how errors compound from a microphone onward.
  - It does NOT test action-item extraction as a scored task. Its error taxonomy covers missing information, redundancy, wrong references, incorrect reasoning, hallucination, incoherence, linguistic inaccuracy and structural disorganisation, and none of these is action-item accuracy.
  - It does NOT test any large-language-model-as-judge metric such as G-Eval, which is precisely the metric that Cornell 2025 found correlated best with transcription quality. Its nine metrics are all count-based, model-based or question-answering-based, and the authors name LLM-based metrics as future work.
  - It does NOT test the current generation of models. Its decoder-only models are GPT-3.5 turbo and Zephyr-7B-alpha as of 2024.
  - It does NOT test any language other than English, a limitation the authors state.
  - It does NOT establish that the metric-error correlations generalise beyond the Longformer Encoder Decoder outputs, since the metric correlation tables are computed on that model's 35 summaries.
  - It does NOT test hierarchical summarisation architectures, which the authors say were excluded for accessibility reasons.
- **Direct quotes (verbatim, source ground truth):**
  - "Current established metrics struggle to capture the observable errors, showing weak to moderate correlations, with a third of the correlations indicating error masking." (p. 1)
  - "Structural disorganization errors are often penalized, matching human judgments, but hallucination errors are sometimes rewarded." (p. 1)
  - "Our analysis reveals that no metric consistently correlates highly with all error types, underscoring the complexity of meeting summarization and the absence of a universal metric that captures human judgment well" (p. 7)
  - "Perplexity favors incorrect references if they preserve linguistic coherence and fluency, aligning with the language model's expectations and yielding a lower (better) score." (p. 8)
  - "BLEU and QuestEval display significant positive correlations with hallucination errors, underscoring their vulnerability to hallucinated content." (p. 8)
  - "Metrics that work well for other summarization tasks either did not react to errors or cannot

reflect the impact on quality in their scores." (p. 9)

- **Limitations:**
  - Author-stated: English only; QMSum's 35 samples represent a fraction of possible meeting types; challenges are inferred from human annotation rather than directly tested; the model selection misses hierarchical architectures; LENS and QuestEval may have biased scores because they take the meeting transcript as input and are not domain-trained; the holistic challenge-to-error linking may dilute strong connections because challenges were not isolated.
  - Author-stated: reliance on Google Scholar as the sole search database, which favours citation counts and may include less reputable sources.
  - Inferred: four annotators, each of whom annotated all 35 samples of one model per week, means each model-sample pair carries one annotator's judgement, with agreement measured on the overlap rather than on every item.
  - Inferred: no exact p-values are reported anywhere, only threshold asterisks, so effect precision cannot be assessed.
  - Inferred: with 35 samples, a correlation of 0.44 is a wide-interval estimate; the paper reports no confidence intervals to bound it.
- **Cross-references in this index:**
  - See also: Rennard 2023 (which argues from first principles that ROUGE is unideal for this task; this paper is the measurement behind that argument); Cornell 2025 (independent evidence from the other direction, using LLM-based judges on recogniser output rather than human annotation on clean transcripts, and reaching a compatible conclusion that the metrics do not track what matters).
  - Contrast with: Kumar 2022 (whose ROUGE leaderboard is exactly the kind of ranking this paper shows cannot distinguish a hallucinating summariser from a faithful one). Contrast with Abramovski 2025, Niu 2024 and Shi 2023 (transcription metrics, a different family; a word error rate is a well-defined measurement against a reference and does not suffer the failure this paper documents).
- **Relevance to platform:** This entry is the direct answer to the vault's question about what is measurable in a summary or an action item, and the answer is: less than the field's leaderboards imply. Two findings bear on the charter's `three-outputs` constraint specifically. First, wrong references, that is, attributing a statement or an action to the wrong participant, is exactly the failure a product that promises action items must not make, and perplexity rewards it while no metric tested penalises it significantly. Second, missing information appeared in 89 to 97 percent of generated summaries across all five models on clean transcripts, which sets a floor on what "the summary" can be promised to contain. For the project's own evaluation planning, the implication is that any quality claim about Recapp's summary or action items will have to be defended by a rubric and human raters, not by a metric, because the metrics available do not separate a good summary from a fluent wrong one.
- **Quotable stats (paste-ready):**
  - "In a study of nine automatic summarisation metrics against expert human error annotations on 35 meetings, no metric correlated strongly with any error type and about a third of metric-error pairs either ignored or rewarded the error (Kirstein 2024)."
  - "Perplexity rewards summaries that attribute statements to the wrong meeting participant, correlating at +0.44 with that error (p <= 0.01), because a fluent wrong attribution reads well to a language model (Kirstein 2024, 35 meetings)."
  - "BLEU and QuestEval both correlate positively with hallucination severity, at +0.35 and +0.34 respectively, meaning they score invented content higher rather than lower (Kirstein 2024)."
  - "Human annotators found missing information in 89 to 97 percent of machine-generated meeting summaries across five different models, working from clean transcripts with no recognition error (Kirstein 2024, 175 annotated summaries)."
  - "Across 35 real meeting transcripts, expert annotators detected speaker dynamics, coreference and low information density as active challenges in 100 percent of them (Kirstein 2024)."

Gong 2024 - language-model judges fail on meeting summaries, measured against human scores (N=2 datasets)

- **Citation:** Ziwei Gong, Lin Ai, Harshsaiprasad Deshpande, Alexander Johnson, Emmy Phung, Zehui Wu, Ahmad Emami, Julia Hirschberg. "CREAM: Comparison-Based Reference-Free ELO-Ranked Automatic Evaluation for Meeting Summarization." arXiv:2409.10883v1, 17 September 2024. DOI: not provided. Affiliations as printed: Machine Learning Center of Excellence, JPMorgan Chase & Co.; Department of Computer Science, Columbia University.
- **File:** `literature/arxiv-2409.10883.pdf`
- **Links:** arXiv:2409.10883 | DOI: not provided | Code: not released; no repository is named in the paper | Weights: not applicable; the paper evaluates commercial OpenAI models | Project page: none | Citations: not retrieved | Reproduced: unknown
- **Evidence tier:** [C] watch METHOD. No stated venue, no released implementation, and one of its two evaluation datasets is internal and unreleasable, so the central result cannot be reproduced.
- **Mechanism family:** `Summarisation evaluation metric` (+ `Human error-taxonomy annotation`)
- **Study type:** diagnostic study of existing evaluators followed by a proposed replacement metric; four datasets, no new human annotation of the authors' own beyond the internal data.
- **Population:** four datasets. FRANK: 2,246 summaries with sentence-level faithfulness annotation, built on CNN/DailyMail and XSum. REALSumm: 2,500 summaries with key-fact-level alignment, also news. QMSum: 1,808 query-summary pairs from 232 meetings, averaging 556.8 speaker turns, derived from AMI, ICSI and parliamentary committee meetings. IZMS (Internal Zoom Meeting Summarization): 139 examples from internal Zoom meetings, comprising transcripts from 10 product meetings plus an additional 40 meetings. This dataset is proprietary and unreleased.
- **Imaging / measurement:** no audio is processed. QMSum transcripts are human-produced. The IZMS transcripts come from Zoom recordings and the paper notes that the factual errors it found in them "mostly stem from ASR errors in transcripts" (p. 5), which is the only place in the paper where recognition error enters, and it is an observation rather than a measurement.
- **Methods / model:** three evaluators are compared: GPT-4-omni (gpt-4o-2024-05-13), GPT-4-turbo (gpt-4-turbo-2024-04-09) and GPT-3.5 (gpt-35-turbo-16k-0613), scoring completeness, conciseness and faithfulness by extracting key facts and aligning them to the summary. The proposed replacement, CREAM, extracts key facts from the concatenation of two candidate summaries, compares each fact against each summary with chain-of-thought prompting, and ranks systems by Elo rating over the resulting pairwise wins, using no reference summary and no source transcript.
- **Statistical detail:**
  - Test: Pearson and Spearman correlations against human annotations; balanced accuracy for factuality detection.
  - Per-arm N: 2,246 FRANK summaries, 2,500 REALSumm summaries, QMSum and IZMS as above.
  - Effect sizes: correlation coefficients, reported without confidence intervals.
  - p-values: none reported anywhere in the paper.
  - Power / pre-registration: not reported.
- **Key findings:**
  - On SHORT news summaries, a language-model evaluator tracks humans well. With human-annotated key facts as the reference, GPT-4o reached a rank correlation of 0.95 (Pearson) with human completeness scores on REALSumm.
  - On MEETING summaries the same evaluators fail. "all versions of GPT have a weak correlation (Pearson's r = 0.5) with gold standard human scores with a strong self-bias" (p. 4) on QMSum, whose transcripts average 556.8 turns.
  - The self-bias is severe and quantified. Using machine-generated summaries as the key-fact reference, GPT evaluators assigned completeness scores of 100.0 percent to summaries from GPT-3.5, GPT-4 and GPT-4o alike, against human-summary-referenced scores of 54.8 to 77.9 percent for the same summaries.
  - Factuality detection is weak on designed benchmarks and near-useless on real meetings. On FRANK, balanced accuracy 61.5 percent at sentence level, summary-level Pearson 0.38 and

Spearman 0.35, system-level Spearman 1.0. On the internal Zoom meeting data, summary-level Pearson fell to 0.11 and Spearman to 0.14.
- Factual errors are rare in real meeting summaries. Error ratios were 16.3 percent for short summaries and 15.2 percent for long ones on the internal data, with no significant difference, and the authors report that most flagged errors were false alarms.
- The proposed comparison-and-Elo method raised the system-level ranking correlation with human preference from 0.5 to 1.0 (Pearson) on both completeness and conciseness. This is a ranking of three systems, so a perfect correlation means getting three items in the right order.

- **Author's framing of the contribution:** > "Do current LLM-based automatic evaluators work effectively for meeting summarization? (Our research shows > that they do not.)" (p. 1)
- **What this paper does NOT establish:**
  - It does NOT show that CREAM measures summary quality. It shows that CREAM ranks three GPT models in the same order humans did. A perfect rank correlation over three items is a weak claim.
  - It does NOT report a single p-value or confidence interval for any correlation.
  - It does NOT process audio or measure recognition error, despite naming recognition errors as the source of the factual errors it found.
  - It does NOT evaluate action-item extraction.
  - It does NOT release the internal Zoom dataset on which its most striking negative result (Pearson 0.11) rests, so that result cannot be checked.
  - It does NOT compare against human evaluation at the summary level for CREAM; the validation is a three-system ranking.
- **Direct quotes (verbatim, source ground truth):**
  - "Do current LLM-based automatic evaluators work effectively for meeting summarization? (Our research shows that they do not.)" (p. 1)
  - "our evaluation revealed that all versions of GPT have a weak correlation (Pearson's r = 0.5,) with gold standard human scores with a strong self-bias. Both GPT-4 and GPT-3.5 tend to give near-perfect scores for summaries generated by LLMs, with GPT-4o performing only slightly better but still showing a bias towards its own generated content." (p. 4)
  - "However, when applying the same method to our IZMS data, the results are less promising. The summary-level correlations were low (Pearson = 0.11 and Spearman = 0.14), indicating the model's ability to find factual error across different contexts is limited." (p. 5)
  - "Analysis of the IZMS dataset indicates that factuality errors are rarely observed in human annotation feedback and mostly stem from ASR errors in transcripts. in human annotations or out-of-context information from GPT models. Yet GPT models often report out-of-context error, which are false alarms when compared to transcripts." (p. 5)
  - "For both completeness and conciseness, LLM evaluators show low correlation with human preference and high self-bias in long-context dialogues like meeting summaries, although perform well in evaluating shorter summaries." (p. 4)
- **Limitations:**
  - Author-stated: models that produce more complete summaries do so at the cost of conciseness, and all three GPT models struggle to balance the two, so a single quality score conflates a trade-off.
  - Inferred: no significance testing anywhere, and the headline validation is a rank correlation over three systems.
  - Inferred: the internal dataset is unreleasable, so the paper's strongest evidence about real meetings is unverifiable.
  - Inferred: GPT-4o is both an evaluated summariser and the evaluator in the proposed method, which the paper itself identifies as the source of self-bias in the systems it criticises.
- **Cross-references in this index:**
  - See also: Kirstein 2024 (nine automatic metrics against expert human error annotation on 175 meeting summaries, reaching the same conclusion by a different route); Kirstein 2024b (the follow-on framework that reports Point-Biserial and rank correlations with significance, which this paper does not); F. Liu 2008 (the same finding for ROUGE on extractive meeting summaries, seventeen years earlier).

  - Contrast with: Y. Liu 2023 (the same metric-versus-human question on NEWS and written dialogue summarisation, where the metrics behave far better); Shapira 2025 and Cornell 2025 (both use a language model as the summary judge, which is precisely the protocol this paper measures and rejects for meetings).
- **Relevance to platform:** The charter promises a summary and action items, and this paper measures whether the cheap way to check them works. On meeting transcripts it does not: language-model evaluators correlated with human scores at Pearson 0.5 on QMSum and 0.11 on real Zoom meeting data, and they assigned perfect completeness scores to machine summaries they were shown as their own reference. Any internal quality dashboard this project builds on a model-as-judge score will report near-perfect numbers regardless of the summary, which is the specific way this measurement fails.
- **Quotable stats (paste-ready):**
  - "Language-model evaluators correlated with human summary scores at Pearson 0.5 on meeting transcripts averaging 557 speaker turns, against 0.95 on short news summaries (Gong 2024, QMSum and REALSumm)."
  - "On real internal Zoom meeting data, a language-model factuality evaluator correlated with human judgement at Pearson 0.11 and Spearman 0.14 at the summary level (Gong 2024, 139 examples)."
  - "When shown a machine-generated summary as the reference, GPT evaluators assigned completeness scores of 100.0 percent to every model's summaries, against 54.8 to 77.9 percent when the human summary was the reference (Gong 2024)."
  - "A language-model evaluator detected sentence-level factual errors in summaries at 61.5 percent balanced accuracy on a purpose-built factuality benchmark (Gong 2024, FRANK)."

Kirstein 2024b - a multi-agent evaluator scored against human error annotation (N=170 meeting summaries)

- **Citation:** Frederic Kirstein, Terry Ruas, Bela Gipp. "Is my Meeting Summary Good? Estimating Quality with a Multi-LLM Evaluator." arXiv:2411.18444v1, 27 November 2024. DOI: not provided. Affiliation as printed: University of Goettingen, Germany. Short-cite note: this is a different paper from Kirstein 2024 elsewhere in this index, which is "What's under the hood: Investigating Automatic Metrics on Meeting Summarization" (Findings of EMNLP 2024).
- **File:** `literature/arxiv-2411.18444.pdf`
- **Links:** arXiv:2411.18444 | DOI: not provided | Code: not released; no repository for MESA is named in the PDF | Weights: not applicable; the framework runs on commercial GPT-4o | Project page: none | Citations: not retrieved | Reproduced: unknown
- **Evidence tier:** [C] watch METHOD. No stated venue in this version, no released implementation, citation count unverified.
- **Mechanism family:** `Summarisation evaluation metric`, `Human error-taxonomy annotation`
- **Study type:** methods paper with component ablations; one annotated dataset, re-annotated by the authors.
- **Population:** the QMSum Mistake dataset, 170 samples from academic meetings (ICSI), business meetings (AMI) and parliamentary meetings, summarised by five models: LED, DialogLED, Pegasus-X, GPT-3.5 and Phi-3. Four annotators updated the existing human Likert scores (0 to 5) and reasoning traces to match the paper's modified error definitions, reaching an inter-annotator agreement of 0.793 Krippendorff's alpha.
- **Imaging / measurement:** no audio. The transcripts are QMSum's human transcripts, so every summary evaluated here was written from an error-free transcript.
- **Methods / model:** MESA evaluates eight error types separately: omission (OM), repetition (REP), incoherence (INC), coreference (COR), hallucination (HAL), language (LAN), structure (STR) and irrelevance (IRR). For each error type it runs a three-step pipeline (identify candidate error instances, rate each instance's severity, assign a 0-to-5 Likert score) with chain-of-thought prompting and verbose confidence scores from 0 to 10. Optionally a multi-agent discussion protocol has one agent draft and others challenge and refine. A self-training loop compares MESA's output with available human annotations, uses a second model as a judge of reasoning quality, and appends a consol-

idated feedback report to the evaluation prompt. Configurations are named Single-n and Multi-n for n rounds of self-training. Backbone: GPT-4o. Baselines: ROUGE-1, ROUGE-2, ROUGE-LSum, BERTScore and a modified G-Eval-4 given the same eight criteria and access to the transcript.

- **Statistical detail:**
  - Test: Point-Biserial correlation between metric score and binary human error presence; Kendall's tau and Spearman's rho between metric score and human-rated error impact. Significance is marked at $p <= 0.05$ (*) and $p <= 0.01$ (**).
  - Per-arm N: 170 annotated summaries, eight error types, nine metric configurations.
  - Effect sizes: correlation coefficients per error type, plus the absolute gap between mean model-assigned and mean human-assigned Likert scores.
  - p-values: reported as thresholds only, marked per cell.
  - Power / pre-registration: not reported.
- **Key findings:**
  - Traditional metrics have essentially no relationship with human error judgements, and several have the WRONG SIGN. Point-Biserial correlations for ROUGE-LSum were POSITIVE for repetition (+0.26, $p <= 0.01$), coreference (+0.19, $p <= 0.05$), hallucination (+0.19, $p <= 0.05$) and structure (+0.23, $p <= 0.05$), meaning the metric rewarded the presence of those errors. Only irrelevance had the desired negative sign (-0.20).
  - BERTScore correlated with human error presence at between -0.32 and +0.08 across the eight error types, with only two cells significant.
  - The best language-model baseline, G-Eval-4, reached Point-Biserial correlations of -0.13 to -0.49, with only repetition passing $p <= 0.01$.
  - The proposed Multi-1 configuration reached -0.21 to -0.69, an average of about 0.13 better than G-Eval-4 across all eight error types, and its strongest cells are repetition -0.69 and incoherence -0.63.
  - For error IMPACT rather than presence, Multi-1 reached Spearman -0.22 to -0.58 and Kendall -0.16 to -0.46. The two hardest error types for every method were omission and hallucination, which are the two that matter most for a meeting summary a person will act on.
  - Self-training closes an overestimation gap. The mean absolute distance between model-assigned and human-assigned Likert scores for the Multi configuration fell from 0.92-3.24 across error types before self-training to 0.22-2.46 after one round.
  - Even the best configuration's rank correlation with human error impact is mid-range, not high: no cell in the paper exceeds 0.58 Spearman in absolute value.
- **Author's framing of the contribution:** > "Using GPT-4o as its backbone, MESA achieves mid to high Point-Biserial correlation with human judgment in > error detection and mid Spearman and Kendall correlation in reflecting error impact on summary quality, on > average 0.25 higher than previous methods." (p. 1)
- **What this paper does NOT establish:**
  - It does NOT produce a metric that agrees with humans on the errors that matter most. Omission and hallucination remain the weakest cells for every method including its own.
  - It does NOT process audio, and every summary it scores was written from a human transcript.
  - It does NOT evaluate action-item extraction; none of the eight error types concerns actions or tasks.
  - It does NOT release code, so the framework cannot be run by anyone else from this paper.
  - It does NOT report cost or latency, although it acknowledges the framework is more expensive than count-based metrics.
  - It does NOT test any backbone other than GPT-4o, so the result is entangled with one commercial model.
- **Direct quotes (verbatim, source ground truth):**
  - "Established metrics such as ROUGE and BERTScore have a relatively low correlation with human judgments and fail to capture nuanced errors." (p. 1)
  - "current LLM-based evaluators risk masking errors and can only serve as a weak proxy, leaving human evaluation the gold standard despite being costly and hard to compare across studies." (p. 1)
  - "Traditional count- and model-based metrics (ROUGE, BERTScore) perform poorly across

most dimensions as expected. LLM-based methods show higher, desired negative correlations with human judgment, suggesting them as a preferred choice." (p. 5)
  - "INC, LANG, and IRR benefit most, while OM and HAL remain challenging, aligning with recent findings on LLMs' struggle with contextualization" (p. 5)
  - "These metrics correlate relatively poorly with human judgment, potentially masking or rewarding certain error types (e.g., QuestEval (Scialom et al., 2021) favors missing information)." (p. 6)
- **Limitations:**
  - Author-stated: the framework is more computationally costly than count-based and model-based metrics, and a lighter variant is included for that reason.
  - Inferred: no released code and a single commercial backbone.
  - Inferred: the annotated set is 170 summaries from one dataset, and the eight error types are evaluated on subsets of those where the error occurs, so several correlation cells rest on small counts.
  - Inferred: the reported improvement over G-Eval-4 is an average of about 0.13 in Point-Biserial correlation, and the paper's abstract quotes 0.25, which is the gap on the error-impact measure rather than the error-detection one.
- **Cross-references in this index:**
  - See also: Kirstein 2024 (the same group's prior study, nine metrics against expert annotation on 175 meeting summaries, which supplies the error taxonomy this paper extends); Gong 2024 (independently finds language-model evaluators unreliable on meeting summaries); F. Liu 2008 (the same result for ROUGE two decades earlier).
  - Contrast with: Y. Liu 2023 (news and written-dialogue summarisation, where automatic metrics reach much higher system-level correlations with human annotation); Golia 2023 (uses BERTScore as its headline metric on meeting summaries, which this paper shows correlates between -0.32 and +0.08 with human error judgements).
- **Relevance to platform:** This is the second independent confirmation in the vault that there is no off-the-shelf number the project can use to say its meeting summary is good. ROUGE actively rewards repetition, hallucination, coreference errors and structural errors on meeting summaries; BERTScore is close to uncorrelated; the best purpose-built language-model evaluator reaches Spearman -0.58 at best and is weakest on omission and hallucination, which are the two failure modes that would make a summary dangerous to act on. Any quality claim in the charter's `three-outputs` deliverable will have to be carried by a rubric and human raters.
- **Quotable stats (paste-ready):**
  - "On meeting summaries, ROUGE-LSum correlated POSITIVELY with the presence of repetition (+0.26), structure (+0.23), coreference (+0.19) and hallucination (+0.19) errors, meaning it rewarded those errors rather than penalising them (Kirstein 2024b, 170 annotated summaries)."
  - "BERTScore's correlation with human error judgements on meeting summaries ranged from -0.32 to +0.08 across eight error types, with only two of eight significant (Kirstein 2024b)."
  - "The best purpose-built language-model evaluator reached Point-Biserial correlations of -0.21 to -0.69 with human error presence, and was weakest on omission and hallucination (Kirstein 2024b)."
  - "Four annotators re-scoring 170 meeting summaries against a fixed eight-type error taxonomy reached Krippendorff's alpha of 0.793 (Kirstein 2024b)."

F. Liu 2008 - ROUGE against human judgement on meeting summaries, and how to make it less bad (N=60 summaries)

- **Citation:** Feifan Liu, Yang Liu. "Correlation between ROUGE and Human Evaluation of Extractive Meeting Summaries." Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 201-204, Columbus, Ohio, USA, June 2008. Association for Computational Linguistics. DOI: not provided in the PDF. Affiliation as printed: The University of Texas at Dallas.
- **File:** `literature/acl-P08-2051-rouge-human-meeting.pdf`
- **Links:** DOI: not provided | Code: not released | Weights: not applicable | Project page: none | Citations: not retrieved | Reproduced: unknown

- **Evidence tier:** [B] credible DOMAIN, and `[ref] foundational` for this index. ACL is a top venue on this index's list, which is one strong signal. It is also the origin of the finding that ROUGE does not track human judgement on meeting summaries, which every later paper in this cluster restates.
- **Mechanism family:** `Summarisation evaluation metric`, `Extractive summarisation method`, `Human error-taxonomy annotation`
- **Study type:** correlation study between an automatic metric and human ratings; within-corpus, no model training.
- **Population:** six ICSI test meetings, the same six used in the prior meeting-summarisation literature. For each meeting: 6 human-written extractive summaries (3 from the corpus's original annotators plus 3 collected by these authors) and 4 system summaries, giving 36 human and 24 system summaries in total. Human summaries select about 6.5 percent of dialogue acts and 13.5 percent of words; system summaries select about 5 percent of dialogue acts and 16 percent of words. Inter-annotator Cohen's kappa across the six annotators varied from 0.11 to 0.35 by meeting. Five human raters, all undergraduate computer-science students, each rated 24 summaries.
- **Imaging / measurement:** ICSI is a naturally occurring research-meeting corpus. All experiments here use the HUMAN TRANSCRIPTIONS, with the authors stating explicitly that this is to isolate meeting characteristics from recognition and automatic-segmentation error. Speaker information was available because the corpus provides separate channels per speaker.
- **Methods / model:** ROUGE with the DUC evaluation options, reporting the F-measure for ROUGE-1 (unigram) and ROUGE-SU4 (skip-bigram with a maximum gap of 4). Two manipulations of the ROUGE input were tested: removing hand-annotated disfluencies from the summaries, and attaching the speaker identifier to every word so that the same word from different speakers cannot match. Human raters agreed or disagreed with nine statements on a 1-to-5 scale, grouped into four categories: informative structure (IS), informative coverage (IC), informative relevance (IRV) and informative redundancy (IRD). Correlation is Spearman's rank coefficient.
- **Statistical detail:**
  - Test: Spearman's rank correlation coefficient (rho).
  - Per-arm N: 36 human summaries and 24 system summaries, correlated separately to avoid the inherent difference between them.
  - Effect sizes: rho values reported per category.
  - p-values: NONE reported anywhere in the paper. The authors describe differences as "significant" in prose without a significance test.
  - Power / pre-registration: not reported.
- **Key findings:**
  - Baseline correlation between ROUGE and human judgement is very low. On human summaries, ROUGE-1 reached rho 0.09 overall and ROUGE-SU4 0.18. On system summaries, ROUGE-1 reached -0.07 and ROUGE-SU4 0.08.
  - Per category on human summaries with ROUGE-SU4: informative structure 0.33, informative coverage 0.38, informative relevance 0.04, informative redundancy -0.30. ROUGE tracks coverage weakly and tracks redundancy in the wrong direction.
  - Per category on system summaries with ROUGE-SU4 before any adjustment: 0.05, 0.01, -0.15, 0.14. The metric carries essentially no information about the summaries that matter, the machine-generated ones.
  - Removing disfluencies helped system summaries most, raising informative structure from 0.05 to 0.22 and informative coverage from 0.01 to 0.19, while leaving the overall correlation unchanged at 0.08.
  - Adding speaker identity to every word raised the overall system-summary correlation from 0.08 to 0.14 and informative redundancy from -0.07 to 0.21, and slightly degraded correlations on human summaries.
  - ROUGE-SU4 correlates better with human judgement than ROUGE-1 throughout, and the authors report that other ROUGE variants correlate above rho 0.9 with one of these two, so the two reported variants cover the metric family.
- **Author's framing of the contribution:** > "In this paper, we have made a first attempt to systematically investigate the correlation of automatic > ROUGE scores with human evaluation for meeting summarization." (p. 4)

- **What this paper does NOT establish:**
  - It does NOT test abstractive summarisation. Every summary here is extractive, a selection of existing dialogue acts, so no generated sentence is ever judged.
  - It does NOT process audio; all experiments use human transcriptions, deliberately.
  - It does NOT report a single p-value, so its comparisons between correlation values are descriptive.
  - It does NOT test action items, decisions or any structured output.
  - It does NOT establish that speaker-aware ROUGE is a usable metric. The best system-summary correlation it achieves is rho 0.14 overall, which is still near zero.
  - It does NOT generalise beyond six meetings from one corpus.
- **Direct quotes (verbatim, source ground truth):**
  - "Our experiments show that generally the correlation is rather low, but a significantly better correlation can be obtained by accounting for several unique meeting characteristics, such as disfluencies and speaker information, especially when evaluating system-generated summaries." (p. 1)
  - "We can see that R-SU4 obtains a higher correlation with human evaluation than R-1 on the whole, but still very low, which is consistent with the previous conclusion from (Murray et al., 2005)." (p. 3)
  - "we found low correlation on system generated summaries, suggesting it is more challenging to evaluate those summaries both by humans and the automatic metrics." (p. 3)
  - "All the experiments in this paper are based on human transcriptions, with a central interest on whether some characteristics of the meeting recordings affect the correlation between ROUGE and human evaluations, without the effect from speech recognition or automatic sentence segmentation errors." (p. 2)
  - "The Kappa statistics for those 6 different annotators varies from 0.11 to 0.35 for different meetings." (p. 2)
- **Limitations:**
  - Author-stated: the analysis uses the average over all nine statements rather than examining each statement's relationship to ROUGE, and a larger dataset is needed.
  - Author-stated: evaluation on recognition output is left to future work.
  - Inferred: six meetings, 60 summaries and five undergraduate raters is a small study, and no significance testing is reported.
  - Inferred: the human summaries themselves have Cohen's kappa as low as 0.11, so the reference against which ROUGE is computed is itself unstable.
- **Cross-references in this index:**
  - See also: Kirstein 2024 (the same finding with nine metrics and an expert error taxonomy, sixteen years later); Kirstein 2024b (ROUGE positively correlated with several error types on meeting summaries); Gong 2024 (the same finding for language-model judges).
  - Contrast with: Y. Liu 2023 (news and written-dialogue summarisation, where ROUGE reaches system-level Kendall correlations of 0.84 against a human protocol); Golia 2023 and Rennard 2023 (report ROUGE on AMI and ICSI as a quality figure without this caveat attached).
- **Relevance to platform:** This is the oldest and most citable statement of the vault's central negative finding: on meeting summaries, ROUGE's rank correlation with human judgement is 0.08 to 0.18, and on machine-generated meeting summaries it is 0.08 or worse. It has been the state of affairs since 2008, three further papers in this vault re-confirm it with modern metrics and modern models, and it means the charter's summary deliverable has no cheap quality measure available at any point in the project's life.
- **Quotable stats (paste-ready):**
  - "Spearman rank correlation between ROUGE and human judgement of machine-generated extractive meeting summaries was 0.08 for ROUGE-SU4 and -0.07 for ROUGE-1 (F. Liu 2008, 24 system summaries over 6 ICSI meetings)."
  - "On human-written meeting summaries the same correlation reached only 0.18 for ROUGE-SU4 and 0.09 for ROUGE-1 (F. Liu 2008, 36 summaries)."
  - "Attaching the speaker identity to every word before scoring raised ROUGE's correlation with human judgement of system summaries from 0.08 to 0.14, and its correlation with judged

redundancy from -0.07 to 0.21 (F. Liu 2008)."
- "Inter-annotator agreement among six people writing extractive summaries of the same meetings ranged from Cohen's kappa 0.11 to 0.35 (F. Liu 2008, ICSI)."

Y. Liu 2023 - what a rigorous human summarisation protocol costs, and what it was applied to (N=22,000 annotations)

- **Citation:** Yixin Liu, Alexander R. Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, Dragomir Radev. "Revisiting the Gold Standard: Grounding Summarization Evaluation with Robust Human Evaluation." arXiv:2212.07981v2, 6 June 2023. DOI: not provided in the PDF. Affiliations as printed: Yale University; Salesforce AI; Carnegie Mellon University.
- **File:** `literature/arxiv-2212.07981.pdf`
- **Links:** arXiv:2212.07981 | DOI: not provided | Code / data: https://github.com/Yale-LILY/ROSE (named in the paper, not verified here) | Weights: not applicable | Project page: none | Citations: not retrieved | Reproduced: unknown
- **Evidence tier:** [B] credible DOMAIN. The benchmark and all system outputs are released, which is one strong signal; the PDF states no venue acceptance and the citation rate is unverified.
- **Mechanism family:** `Human error-taxonomy annotation`, `Summarisation evaluation metric`
- **Study type:** annotation-protocol design, benchmark construction, and a comparative study of four human evaluation protocols and 50 automatic metrics.
- **Population:** THE DOMAIN IS TEXT SUMMARISATION, NOT MEETINGS. Three datasets: CNN/DailyMail news articles, XSum news articles, and SamSum, a corpus of short written messenger-style chats. Annotations cover 28 top-performing summarisation systems. About 21,800 unit-level annotations and about 22,000 summary-level annotations, aggregated over about 50,000 individual summary-level judgements. Annotation was by a mix of in-house annotators and Amazon Mechanical Turk crowdworkers.
- **Imaging / measurement:** no audio, no speech, no meetings and no speakers. The inputs are written documents and written chat logs.
- **Methods / model:** the Atomic Content Unit (ACU) protocol, modified from the Pyramid and LitePyramid protocols: a reference summary is decomposed into fine-grained content units and a candidate summary is scored by how many of those units it contains, which makes the judgement close to objective. Three other protocols are compared: Prior (rate the summary without seeing the input), Ref-free (rate against the source document) and Ref-based (rate against the reference summary). Fifty automatic metrics are then correlated against the ACU scores, including ROUGE, METEOR, CHRF, BERTScore, BARTScore, SummaQA, QAEval, Lite3Pyramid, GPTScore and G-Eval on GPT-3.5 and GPT-4.
- **Statistical detail:**
  - Test: Krippendorff's alpha for inter-annotator agreement; Pearson, Spearman and Kendall correlations for metric and protocol comparisons; bootstrapped confidence intervals; explicit statistical power analysis.
  - Per-arm N: 500 annotated CNN/DailyMail examples for the power analysis; system pairs bucketed into six equal-sized groups by score difference.
  - Effect sizes: correlation coefficients and statistical power curves against sample size.
  - p-values: significance is handled through power analysis and bootstrapped intervals rather than reported per comparison.
  - Power / pre-registration: power analysis is a central contribution; a power of 0.80 is treated as the threshold for a sufficiently powered experiment.
- **Key findings:**
  - THE COST FIGURE, IN ITS ACTUAL CONTEXT: "the RoSE benchmark, consisting of 22000 summary-level annotations and requiring over 150 hours of in-house annotation, across three summarization datasets" (p. 2). The 150 hours covers in-house annotation across CNN/DailyMail, XSum and SamSum, none of which is a meeting corpus, and the 22,000 summary-level judgements were collected through a mix of in-house work and crowdsourcing.

    - The ACU protocol reaches high agreement where earlier protocols did not: Krippendorff's alpha 0.7571 at summary level and 0.7528 at unit level, against 0.66 for RealSumm and 0.49 for SummEval crowdworkers.
    - The three unconstrained protocols reach much lower agreement on the same summaries: 0.3455 for Prior, 0.2201 for Ref-free and 0.2741 for Ref-based.
    - Reference-free human evaluation measures annotator preference more than summary quality. Prior scores, collected without showing the annotator the input at all, correlate with reference-free scores at system-level Pearson 0.926, and both correlate with summary length at 0.833 and 0.875.
    - Reference-free and reference-based human evaluation have a near-zero system-level correlation, -0.061.
    - Language-model-based metrics did not beat traditional ones. On this benchmark, ROUGE, BARTScore and QAEval reached higher system-level Kendall correlations than GPTScore and G-Eval, and G-Eval-3.5's correlations were negative in the buckets where systems were closest.
    - Automatic metrics degrade sharply when the systems being compared are close. In the closest-performing bucket, ROUGE-1's Kendall correlation with human scores was 0.091 and G-Eval-3.5's was -0.091.
    - Statistical power is the binding constraint. The 50-to-100-sample human evaluations typical of recent papers reach power 0.80 only when systems differ by more than 5 ROUGE-1 recall points.
- **Author's framing of the contribution:** > "we introduce the Atomic Content Unit (ACU) protocol for summary salience evaluation, which is modified > from the Pyramid and LitePyramid protocols." (p. 1)
- **What this paper does NOT establish:**
    - It does NOT annotate meetings. CNN/DailyMail and XSum are news articles; SamSum is short written messenger-style dialogue. No spoken conversation, no multi-party meeting and no transcript appears.
    - The 150-hour figure is NOT a meeting-annotation cost. It is in-house annotation time for three written-text summarisation datasets under the ACU protocol.
    - It does NOT process audio or involve any recognition system.
    - It does NOT evaluate action items, decisions or any structured extraction.
    - It does NOT test summary faithfulness or hallucination directly; the ACU protocol scores salience, meaning coverage of reference content, not correctness of invented content.
    - It does NOT include any language other than English.
- **Direct quotes (verbatim, source ground truth):**
    - "We curate the Robust Summarization Evaluation (RoSE) benchmark, consisting of 22000 summary-level annotations over 28 top-performing systems on three datasets." (p. 1)
    - "We curate the RoSE benchmark, consisting of 22000 summary-level annotations and requiring over 150 hours of in-house annotation, across three summarization datasets, which can lay a solid foundation for training and evaluating automatic metrics." (p. 2)
    - "Reference-free and reference-based human evaluation results have a near-zero correlation." (p. 2)
    - "Reference-free human evaluation strongly correlates with input-agnostic, annotator preference." (p. 2)
    - "Despite their successes in other human evaluation benchmarks such as SummEval, LLM-based automatic evaluation cannot outperform traditional methods such as ROUGE on RoSE." (p. 8)
    - "A high statistical power is difficult to reach when the system performance is similar." (p. 5)
- **Limitations:**
    - Author-stated: annotator bias and pre-training data bias may be present; the benchmark is English only; noise is inevitable in ACU writing and matching, and high agreement does not guarantee correctness.
    - Inferred: everything here is written text, so its transfer to spoken multi-party meeting summaries is an assumption, and the meeting-specific papers in this cluster report far worse met-

ric behaviour.
  - Inferred: the protocol's high agreement comes from decomposing the reference summary, which requires a reference summary to exist; for a meeting product there is none.
- **Cross-references in this index:**
  - See also: Kirstein 2024, Kirstein 2024b and Gong 2024 (the same metric-versus-human question applied to meetings, where the metrics behave much worse); F. Liu 2008 (the meeting-domain original).
  - Contrast with: every meeting paper in this index. On this written-text benchmark ROUGE reaches system-level Kendall 0.84 against the reference-based human protocol; on meeting summaries the same metric reaches 0.08 (F. Liu 2008) or the wrong sign (Kirstein 2024b). The difference is the domain, and it is the reason this paper's numbers must not be quoted as evidence about meeting summaries.
- **Relevance to platform:** This paper is in the vault to stop a specific mistake. Its "over 150 hours for 22,000 annotations" figure has been read as the cost of annotating meetings, and it is not: it is in-house annotation of news articles and written chat summaries. What it does contribute to this project is the protocol design and the power analysis. If the project ever runs its own human evaluation of summaries, this paper says a reference-free rating scale will mostly measure the rater's preference for longer text (correlating 0.875 with summary length and 0.926 with input-blind prior preference), and that a 50-to-100 sample study cannot detect anything but a large difference.
- **Quotable stats (paste-ready):**
  - "A fine-grained content-unit protocol reached Krippendorff's alpha of 0.7571 among crowdworkers on summarisation, against 0.66 and 0.49 for two earlier benchmarks (Y. Liu 2023, 22,000 annotations across three written-text datasets)."
  - "Reference-free human ratings of summaries correlate with input-blind annotator preference at system-level Pearson 0.926 and with summary length at 0.875, and with reference-based ratings at only -0.061 (Y. Liu 2023)."
  - "Human evaluations of 50 to 100 samples, the typical size in recent summarisation papers, reach a statistical power of 0.80 only when systems differ by more than 5 ROUGE-1 recall points (Y. Liu 2023)."
  - "The RoSE benchmark's 22,000 summary-level annotations required over 150 hours of in-house annotation on news articles and written chat logs, not on meeting transcripts (Y. Liu 2023)."

What the corpora actually contain, and which of the three outputs they can supervise

Chen 2016 - ten assistant actions annotated onto 22 ICSI meetings, and how rare they are (N=21,035 utterances)

- **Citation:** Yun-Nung Chen, Dilek Hakkani-Tur. "AIMU: Actionable Items for Meeting Understanding." Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pages 739-743. DOI: not provided in the PDF. Affiliation as printed: Microsoft Research, Redmond, WA, USA.
- **File:** `literature/acl-L16-1117-aimu.pdf`
- **Links:** DOI: not provided | Data: the paper gives http://research.microsoft.com/projects/meetingunderstanding/ as the distribution point (named in the paper, not verified here) | Code: not released | Weights: the paper states that convolutional deep structured semantic model embeddings are shipped with the data | Project page: as above | Citations: not retrieved | Reproduced: unknown
- **Evidence tier:** [B] credible DOMAIN. LREC is the field's language-resources venue and the annotation is released with the paper, which is one strong signal; the citation rate is unverified.
- **Mechanism family:** `Corpus resource`, `Decision and action-item extraction`
- **Study type:** annotation resource paper with a baseline detection experiment.
- **Population:** 22 meetings from the ICSI meeting corpus, specifically the meetings used as test and development data in prior work, comprising 21,035 utterances segmented by speaker turn.

Three ICSI meeting types: Bed (Even Deeper Understanding, 4 meetings, 4,544 utterances), Bmr (Meeting Recorder, 9 meetings, 9,227 utterances) and Bro (Robustness, 8 meetings, 7,264 utterances). No participant demographics are reported; the underlying recordings are ICSI's own research-group meetings.

- **Imaging / measurement:** no audio is processed. Annotation and experiments run on ICSI's transcripts. The underlying ICSI audio is distributed by the Linguistic Data Consortium under LDC2004S02 and LDC2004T04, which the paper cites as its language resources; no licence for the AIMU annotations themselves is stated.
- **Methods / model:** the annotation adapts a semantic intent schema designed for a commercial virtual personal assistant to the human-human meeting genre. Ten intents across five domains: find calendar entry, create calendar entry, open agenda, add agenda item (Calendar); create reminder (Reminders); send email, find email, make call (Communication); open setting (OnDevice); search (Search). Each intent carries typed arguments, for example contact name, start date, start time, entry type, title, absolute location, implicit location, reminder text, email subject, email content, from contact name, setting type and query term. The baseline detector is a support vector machine with a radial basis function kernel over convolutional deep structured semantic model embeddings, scored by area under the precision-recall curve.
- **Statistical detail:**
  - Test: Cohen's kappa for annotation agreement, computed on two randomly selected meetings annotated twice.
  - Per-arm N: two doubly-annotated meetings for agreement; 22 meetings and 21,035 utterances for the statistics; one detection experiment per feature set.
  - Effect sizes: kappa values and area-under-curve percentages.
  - p-values: none reported.
  - Power / pre-registration: not reported.
- **Key findings:**
  - Actionable items are extremely sparse in real meetings: 318 turns out of 21,035 carry one, which is 1.5 percent. By meeting type the rate is 4.2 percent for Bed (192 of 4,544), 1.3 percent for Bmr (116 of 9,227) and 1.5 percent for Bro (110 of 7,264).
  - Deciding WHETHER a turn contains an actionable item is the hard part; deciding WHICH action it is, is not. Average binary actionable-utterance agreement was Cohen's kappa 0.644 (0.699 and 0.642 on the two doubly-annotated meetings), while agreement on the action TYPE among turns both annotators marked positive was 1.000. Overall eleven-way agreement was 0.673.
  - Detection is hard from text alone. Area under the precision-recall curve: 52.84 with n-gram features (n = 1, 2, 3), 59.79 with paragraph vectors from doc2vec, 64.33 with predictive CDSSM embeddings, 65.58 with generative and 69.27 with bidirectional CDSSM embeddings.
  - The distribution of actions differs by meeting type, with create-reminder and find-calendar-entry frequent across all types and open-setting and make-call rare in all of them.
- **Author's framing of the contribution:** > "This paper presents an extended set of annotations for the ICSI meeting corpus with a goal of deeply > understanding meeting conversations, where participant turns are annotated by actionable items that could > be performed by an automated meeting assistant." (p. 1)
- **What this paper does NOT establish:**
  - It does NOT process audio, so it says nothing about how these labels survive a recognised transcript.
  - It does NOT annotate action items in the sense the meeting-summarisation literature uses. These are assistant intents (open a calendar, find an email, launch a device setting), not commitments a participant made to do work after the meeting.
  - It does NOT cover the whole ICSI corpus. 22 of 75 meetings are annotated, chosen because prior work used them as development and test data.
  - It does NOT state a licence for the annotations, and the distribution URL is a Microsoft Research project page from 2016.
  - It does NOT evaluate argument extraction, only intent detection; the argument annotations are released but no baseline scores them.

  - It does NOT report any inter-annotator agreement beyond two meetings.
- **Direct quotes (verbatim, source ground truth):**
  - "There are total 318 turns annotated with actionable items, which account for about 1.5% of all of the turns." (p. 2)
  - "The average agreement about whether an utterance includes an actionable item is 0.644." (p. 2)
  - "The average agreement is 1.000, indicating that both annotators always decide on the same action if they both agree that there is an action in this utterance. It also suggests that most actions may not be ambiguous." (p. 2)
  - "A total of 10 defined intents/actions are considered as actionable items in meetings. Turns that include actionable intents were annotated for 22 public ICSI meetings, that include a total of 21K utterances, segmented by speaker turns." (p. 1)
  - "It suggests that the number of actionable utterances may depend on the meeting type." (p. 2)
- **Limitations:**
  - Inferred: agreement is computed on two meetings only, and the binary kappa of 0.644 is moderate.
  - Inferred: the intent schema is imported from a virtual-assistant product and then applied to human-human speech, so it captures what an assistant could do rather than what participants agreed to do.
  - Inferred: the released embeddings are from a 2016 model family, so the baseline number is of historical interest only.
  - Inferred: the distribution URL is a decade-old research project page and the paper states no licence.
- **Cross-references in this index:**
  - See also: J. Liu 2023 (the other action-item resource paper in this vault, which reports separately that ICSI has action-item annotations for 18 meetings that are no longer publicly available; those are a different annotation on a different subset from these 22 meetings); Golia 2023 (extracts action items and never scores them); Rennard 2023 (treats action items as a named sub-task with no benchmark).
  - Contrast with: Vinnikov 2024, Watanabe 2020, Tipaksorn 2025 and Jones 2022 (audio corpora; this one adds text labels to an existing corpus and contains no new recordings).
- **Relevance to platform:** For the charter's `three-outputs` constraint, the single most useful number here is 1.5 percent. Across 21,035 turns of real research meetings only 318 carried an actionable item, and human annotators agreed on whether a turn was one of them at kappa 0.644. That sets the difficulty of the action-items deliverable: it is a needle-in-a-haystack detection problem where a false positive rate of even a few percent would bury the true items, and where two trained people disagree on a third of the calls.
- **Quotable stats (paste-ready):**
  - "Only 318 of 21,035 speaker turns in 22 real research meetings contained an actionable item, 1.5 percent of all turns (Chen 2016, ICSI)."
  - "Two annotators agreed on whether a turn contained an actionable item at Cohen's kappa 0.644, but agreed on which of ten actions it was at kappa 1.000 whenever they both marked it positive (Chen 2016)."
  - "The rate of actionable turns varied by meeting type from 1.3 to 4.2 percent, so how many action items a meeting yields depends on what kind of meeting it is (Chen 2016, ICSI)."
  - "The best actionable-item detector reported reached 69.27 percent area under the precision-recall curve on human transcripts (Chen 2016)."

Jones 2022 - a telephone and video corpus for speaker recognition, with no meetings in it (N=202 speakers)

- **Citation:** Karen Jones, Kevin Walker, Christopher Caruso, Jonathan Wright, Stephanie Strassel. "WeCanTalk: A New Multi-language, Multi-modal Resource for Speaker Recognition." Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3451-3456, Marseille, 20-25 June 2022. European Language Resources Association (ELRA). The proceedings volume is licensed CC-BY-NC-4.0. DOI: not provided in the PDF. Affiliation as printed: Linguistic Data Consortium, University of Pennsylvania, Philadelphia, USA.

- **File:** `literature/acl-2022.lrec-1.369-wecantalk.pdf`
- **Links:** DOI: not provided | Data: the paper states the corpus "will be published in the LDC catalog"; no catalogue number or licence for the corpus itself is given in the PDF | Code: not released | Weights: not applicable | Project page: none | Citations: not retrieved | Reproduced: unknown
- **Evidence tier:** [B] credible DOMAIN. LREC is the field's language-resources venue and the corpus was delivered to NIST for the 2021 Speaker Recognition Evaluation, which is one strong signal; the citation rate is unverified and the corpus is behind an LDC catalogue licence.
- **Mechanism family:** `Corpus resource` (speaker identity only)
- **Study type:** data-collection resource paper. No models are trained and no benchmark result is reported.
- **Population:** 202 distinct speakers, all native Cantonese speakers recruited in Hong Kong who were also fluent in at least one other language, all adults, recruited by word of mouth via Hong Kong Polytechnic University. 315 participants were recruited to yield 202 complete speakers. Sex: 154 female, 48 male. Year of birth: 1 in 1960-69, 4 in 1970-79, 4 in 1980-89, 131 in 1990-99, 61 in 2000-2009, one not reported. Collection ran June to October 2020. Institutional Review Board approval at the University of Pennsylvania and at Hong Kong Polytechnic University; all speakers gave informed consent and were compensated.
- **Imaging / measurement:** TELEPHONE AND SELF-RECORDED VIDEO. Each speaker contributed at least 10 telephone calls of 8 to 10 minutes through a custom platform physically located in Hong Kong, plus at least 3 videos of 3 to 10 minutes in which they were visible and audible, plus one selfie image. Calls were delivered to NIST as full-length narrowband single-channel 8 kHz a-law files. 2,241 calls came from mobile phones and 118 from landlines; device modes were internal microphone 934, speakerphone 558, headset 867. At least 25 percent of calls were required to be made in noisy environments; auditors judged 803 calls noisy and 1,555 not noisy, and 77 videos noisy and 763 not noisy. Language distribution across calls: Cantonese 1,227, Mandarin 700, English 367, mixed 56, other 9. Across videos: Cantonese 503, English 162, Mandarin 98, mixed 70, other 8. Every call and video was manually audited for recording quality, language, amount of speech and speaker identity, by trained annotators who were native Cantonese speakers fluent in Mandarin and English.
- **Methods / model:** none. This is a collection-and-auditing paper. Automatic validation used file duration checks, the LDC HMM Speech Activity Detector v1.0.5 and md5 uniqueness; manual auditing followed a fixed question set for calls, videos and speaker consistency.
- **Statistical detail:**
  - Test: none. No inferential statistics are reported anywhere.
  - Per-arm N: 202 complete speakers, roughly 2,359 audited calls and 840 audited videos.
  - Effect sizes: not applicable.
  - p-values: not applicable.
  - Power / pre-registration: not applicable.
- **Key findings:**
  - The corpus is 202 multilingual Cantonese speakers, with at least 10 telephone calls and 3 videos each, collected specifically to support the NIST 2021 Speaker Recognition Evaluation.
  - It is the first LDC speaker-recognition corpus to include both video and telephony from every speaker, and the first to use multilingual Cantonese speakers rather than separate speakers per language.
  - Code-switching is present but rare: about 2 percent of calls and 8 percent of videos contain a mixture of languages, despite one call and one video per speaker being designated "Freestyle" with any language mix permitted.
  - The speaker population is heavily skewed: 76 percent female, and 95 percent born in 1990 or later.
  - Recruitment was disrupted by the 2019-2020 Hong Kong protests and by COVID-19 campus closures, which the authors say reduced the number of multi-party video conversations below what was originally planned.
- **Author's framing of the contribution:** > "The WeCanTalk (WCT) Corpus is a new multi-language,

multi-modal resource for speaker recognition." (p. 1)

- **What this paper does NOT establish:**
  - It contains NO MEETINGS. Every recording is a two-party telephone call or a self-recorded video, not a multi-party in-person discussion.
  - It has NO TRANSCRIPTS. Nothing in the corpus is transcribed; auditing recorded language, speaker identity, noise and quality judgements only.
  - It therefore cannot supervise or evaluate ANY of this project's three outputs. It has no verbatim text, no summaries and no action items. Its only label is speaker identity.
  - It is NOT far-field. Telephone handsets, speakerphones and headsets at conversational distance, plus phone-camera video.
  - It reports NO benchmark result of any kind; no speaker-recognition system is run on it in this paper.
  - It is NOT freely available. The paper states the corpus will be published in the LDC catalog and gives no open licence.
- **Direct quotes (verbatim, source ground truth):**
  - "The corpus contains Cantonese, Mandarin and English telephony and video speech data from over 200 multilingual speakers located in Hong Kong." (p. 1)
  - "The WeCanTalk Corpus has been used to support the NIST 2021 Speaker Recognition Evaluation and will be published in the LDC catalog." (p. 1)
  - "LDC delivered the complete set of WCT call recordings to NIST as full-length narrowband 1-channel 8-kHz a-law files." (p. 5)
  - "Approximately 2% of calls consisted of a mix of languages compared to 8% of videos containing a mixture." (p. 6)
  - "WCT data collection took place in June-October 2020, a time characterized by significant political and social turmoil in Hong Kong including widespread political protests and increasing impact of the COVID-19 pandemic" (p. 1)
- **Limitations:**
  - Author-stated: recruitment and video collection were curtailed by the protests and the pandemic, producing fewer multi-party video conversations than planned.
  - Inferred: the demographic skew (76 percent female, 95 percent born 1990 or later) limits generalisation.
  - Inferred: callees were anonymous, unenrolled and had no demographic data collected, so only one side of each conversation is usable as labelled speaker data.
  - Inferred: no transcripts of any kind means the corpus cannot be repurposed for recognition or summarisation work without new annotation.
- **Cross-references in this index:**
  - See also: nothing in this index shares its family closely; it is the only speaker-recognition-only corpus in the vault.
  - Contrast with: Vinnikov 2024, Watanabe 2020 and Tipaksorn 2025 (multi-party far-field audio with verbatim transcripts and speaker labels, which is what a meeting product would train and test on); Chen 2016 and J. Liu 2023 (action-item labels).
- **Relevance to platform:** This corpus appears in the project's data landscape and its entry here exists to bound what it can be used for. It supervises exactly one thing, who is speaking, and it does so on telephone and self-recorded video audio from a two-party or single-speaker setting. It has no transcripts, no summaries and no action items, so it can test none of the charter's three outputs, and its recording condition is not the in-person meeting the charter describes. Its one point of contact with the project is that 2,241 of its 2,359 calls were made from mobile phones, with 934 on the internal microphone, 558 on speakerphone and 867 on a headset, which is a rare published record of how people actually hold a phone while talking.
- **Quotable stats (paste-ready):**
  - "The WeCanTalk corpus holds 202 multilingual speakers contributing at least 10 telephone calls and 3 videos each, with no transcripts, no summaries and no action items (Jones 2022)."
  - "Of 2,359 audited calls, 2,241 were made from mobile phones and 118 from landlines, split by mode into 934 internal microphone, 558 speakerphone and 867 headset (Jones 2022)."
  - "Auditors judged 803 of the calls noisy and 1,555 not noisy, against a collection requirement

of at least 25 percent noisy (Jones 2022)."
- "Only about 2 percent of calls and 8 percent of videos contained a mixture of languages, despite the collection deliberately targeting code-switching in a bilingual population (Jones 2022)."

Prevot 2025 - a French meeting corpus segmented into discourse units, and what that costs (N=73 meetings)

- **Citation:** Laurent Prevot, Roxane Bertrand, Julie Hunter. "Segmenting a French Meeting Corpus into Elementary Discourse Units." Proceedings of the 26th Annual Meeting of SIGDIAL, pages 183-191, Aug 25-27,

2025. Association for Computational Linguistics. DOI: not provided in the PDF. Affiliations as printed: Aix Marseille Univ. & CNRS, LPL, and CNRS & MEAE, CEFC; CNRS & Aix Marseille Univ., LPL; LINAGORA Labs, Toulouse, France.

- **File:** `literature/acl-2025.sigdial-1.14.pdf`
- **Links:** DOI: not provided | Data: the underlying SUMM-RE corpus is at https://huggingface.co/datasets/linagora/S RE (named in the paper, not verified here) | Code: the paper names https://github.com/phimit/jiant-discut as the fine-tuning framework and states the segmentation model and its application to the full dataset will be released alongside the code repository | Weights: stated as forthcoming | Project page: none | Citations: not retrieved | Reproduced: unknown
- **Evidence tier:** [B] credible DOMAIN. SIGDIAL is an ACL-affiliated venue and the underlying corpus is publicly hosted, which is one strong signal; the segmentation model was unreleased at publication and the citation rate is unverified.
- **Mechanism family:** `Corpus resource`, `Discourse segmentation annotation`
- **Study type:** annotation campaign plus a fine-tuning baseline; one corpus, four annotators.
- **Population:** the SUMM-RE corpus: approximately 100 sessions, each made of three 20-minute meetings, with 2 to 4 participants, on event planning. Each group takes part in three sessions with different tasks: discussing ideas for the event (more monologue-oriented), deciding what to do (dialogic, with divergent opinions), and planning the practical organisation including task delegation. The dev and test portions, 73 meetings and about 24 hours of recording, have been manually corrected and discourse-annotated. The paper describes roughly 20 hours of manually transcribed and segmented conversation plus 80 hours automatically transcribed and segmented. Annotators were four undergraduate students from the School of Humanities, described by the authors as naive coders, trained over three sessions.
- **Imaging / measurement:** NEAR-FIELD. "Most sessions were recorded face-to-face using head-mounted microphones, with a few additional sessions conducted via Zoom." (p. 4). The full corpus has been automatically transcribed; the dev and test portions were then manually corrected. Segmentation was performed speaker by speaker in Praat, with the other participants' contributions available as context, so a four-participant session was analysed four times.
- **Methods / model:** annotation is into elementary discourse units (EDUs), defined by the authors as a span of contiguous tokens whose combined meaning corresponds to a semantic proposition or a single speech act, with preparatory disfluent material included in the unit. A two-step process was used: a small manually segmented subset trained a rough model, that model pre-segmented the whole corpus, and coders then edited the pre-segmented files. The automatic segmenter is xlm-roberta-large fine-tuned in the DISRPT framework with a binary per-token label for discourse-unit start, learning rate 1e-5, batch size 1, gradient accumulation 4, up to 30 epochs with patience 10. Scoring is F-score, precision and recall computed on discourse boundaries only, excluding true negatives.
- **Statistical detail:**
  - Test: Cohen's kappa for inter-coder agreement during training; F-score comparisons between coders and against the model.
  - Per-arm N: 73 manually annotated meetings; four coders; models trained at a range of training-set sizes.
  - Effect sizes: kappa and F-score values.
  - p-values: none reported. The authors report "no significant differences" between files anno-

tated from scratch and files annotated over a pre-segmentation without naming a test.
  - Power / pre-registration: not reported.
- **Key findings:**
  - Semi-naive coders reached mean Cohen's kappa in the 0.85 to 0.89 range after training; a few expert annotations by the paper's authors reached about 0.9.
  - The automatic segmenter reaches an F-score of approximately 0.86 on spontaneous French meeting speech, which the authors say is below what the same task achieves on written documents and which plateaus at the level of inter-coder agreement.
  - Model performance plateaus early with respect to training-set size, and the plateau appears to be the maximum theoretically achievable given human disagreement, so the authors conclude they annotated more than was strictly necessary for the model.
  - Corpus statistics: a dialogue hosts on average 534 elementary discourse units, and the mean length of an EDU is 8 tokens with a standard deviation of 6.
  - Pre-segmenting files with a rough model before human editing did not measurably bias the result; the authors compared inter-coder agreement on files annotated from scratch against pre-segmented files and found the difference smaller than the variability between coders.
- **Author's framing of the contribution:** > "we present the creation of a large, discourse-segmented corpus of French meetings, comprising > approximately 20 hours of manually transcribed and segmented conversations, along with 80 hours of > automatically transcribed and segmented data." (p. 1)
- **What this paper does NOT establish:**
  - It contains NO SUMMARIES and NO ACTION ITEMS in the annotation described here. The labels are discourse unit boundaries, so of this project's three outputs it can supervise only the transcript, and then only its segmentation.
  - It is NOT far-field. The recordings use head-mounted microphones, which is a near-field reference condition, not a tabletop device.
  - It is NOT English. The corpus is French, and the authors note that discourse segmentation performance declines for languages other than English.
  - It does NOT report segmentation performance on the automatically transcribed portion; the authors name that as future work.
  - It does NOT report any significance test, despite claiming no significant difference between two annotation procedures.
  - It does NOT release the trained segmentation model in this paper; the release is stated as forthcoming.
- **Direct quotes (verbatim, source ground truth):**
  - "Most sessions were recorded face-to-face using head-mounted microphones, with a few additional sessions conducted via Zoom." (p. 4)
  - "Overall, semi-naive coder reached mean kappa-scores within the 0.85-0.89 range, while a few expert annotations conduced by the authors of the paper have shown higher reliability, with kappa-scores around 0.9." (p. 4)
  - "While an F-score of approximately 0.86 makes the method more promising than typical proxies, it still falls short of the scores typically achieved on written documents for this task." (p. 5)
  - "Although discourse segmentation into elementary discourse units (EDUs) is considered to be nearly solved for canonical written texts, conversational spontaneous speech transcripts present different challenges." (p. 1)
  - "the dialogue hosted in average 534 EDUs and mean length of an EDUs is 8 +/- 6" (p. 3)
- **Limitations:**
  - Author-stated: performance falls short of written-text discourse segmentation, and applying the method to the automatic-transcription version of the corpus is untested.
  - Author-stated: the annotators were naive rather than expert, and collecting expert annotations is named as a next step.
  - Inferred: the annotation was performed on manually corrected transcripts, so the F-score of 0.86 is an upper bound that a recognised transcript would not reach.
  - Inferred: no significance testing, and the reported human agreement kappa range of 0.85 to 0.89 is the ceiling on what the model can be credited with.

- **Cross-references in this index:**
  - See also: Vinnikov 2024, Watanabe 2020 and Tipaksorn 2025 (the other corpus papers in this cluster); Rennard 2023 (the survey establishing how few meeting corpora with summaries exist).
  - Contrast with: J. Liu 2023 and Chen 2016 (annotate actionable content rather than discourse structure, and report kappa of 0.47 and 0.644 respectively, far below the 0.85-0.89 reached here; segmentation is a much more objective judgement than deciding what counts as an action item).
- **Relevance to platform:** This entry is in the vault for one comparison. Human annotators segmenting the same French meeting transcripts into discourse units agreed at Cohen's kappa 0.85 to 0.89, and a fine-tuned model reached an F-score of 0.86 against them. That is what a well-defined structural annotation task on meeting speech looks like, and it sits directly against the 0.47 kappa reported for action-item annotation and the 0.644 for actionable-item detection. The difference is not annotator quality, it is that structure is decidable and importance is not. For the charter's summary and action-item deliverables, that is the relevant asymmetry.
- **Quotable stats (paste-ready):**
  - "Human coders segmenting French meeting transcripts into elementary discourse units agreed at Cohen's kappa 0.85 to 0.89, and experts at about 0.9 (Prevot 2025, 73 meetings)."
  - "A fine-tuned cross-lingual language model reached an F-score of about 0.86 on discourse segmentation of spontaneous French meeting speech, plateauing at the level of human agreement (Prevot 2025)."
  - "The SUMM-RE meeting corpus comprises about 100 sessions of three 20-minute meetings with 2 to 4 participants, of which 73 meetings and about 24 hours are manually corrected and annotated (Prevot 2025)."
  - "A French meeting dialogue contains on average 534 elementary discourse units, with a mean unit length of 8 tokens (Prevot 2025)."

Tipaksorn 2025 - the same conversation on nine devices at measured distances, 0.12 m to 10 m (N=114 hours)

- **Citation:** Pattara Tipaksorn, Sumonmas Thatphithakkul, Vataya Chunwijitra, Kwanchiva Thangthai. "LOTUSDIS: A Thai Far-Field Meeting Corpus for Robust Conversational ASR." arXiv:2509.18722v1, 23 September

2025. DOI: not provided. Affiliation as printed: Speech and Text Understanding Research Team, NECTEC.

- **File:** `literature/arxiv-2509.18722.pdf`
- **Links:** arXiv:2509.18722 | DOI: not provided | Code / data: https://github.com/kwanchiva/LOTUSDIS, stated in the paper to hold the corpus and a reproducible baseline system | Weights: the paper's baselines fine-tune the public Pathumma-whisper-th-large-v3 and Whisper-large-v3 checkpoints; no fine-tuned weights are named | Project page: as above | Citations: not retrieved | Reproduced: unknown
- **Evidence tier:** [B] credible DOMAIN. The corpus is released under a permissive licence with a reproducible baseline, which is one strong signal; the PDF states no venue acceptance and the citation rate is unverified.
- **Mechanism family:** `Corpus resource`, `Far-field single-channel transcription benchmark`, `Near-field close-talk reference condition`
- **Study type:** corpus resource paper with a benchmark experiment across zero-shot and fine-tuned recognition, front-end processing and augmentation.
- **Population:** 86 unique participants aged 19 to 48, mean age 27; 47 male and 24 female across the splits as tabulated (train 35 male / 39 female participants, dev 5 / 7, test 7 / 5). 90 sessions of about 15 minutes each, three participants per session, spontaneous unscripted conversation on a pre-assigned topic. Splits: 69 train sessions with 40 topics, 10 dev with 10 topics, 11 test with 11 topics. Utterance counts 120,245 / 13,090 / 27,580. Average overlap ratio 31.5 percent train, 40.3 percent dev, 29.0 percent test. Language: Thai.
- **Imaging / measurement:** THIS IS THE CLOSEST PUBLISHED DESIGN IN THE VAULT TO THIS

PROJECT'S OWN QUESTION. Nine independent single-channel microphones of six device types recorded the same conversations simultaneously, at measured distances from 0.12 m to 10 m, with no microphone array anywhere. Near-field, at or under 0.5 m: three lavalier microphones and three table-mounted condensers, positioned 12 to 15 cm from each speaker's mouth, described by the authors as high direct-to-reverberant-ratio reference recordings. Far-field, at or beyond 2 m: a tabletop JBL loudspeaker unit at 2 m, and two Bluetooth speakerphones at 3 m (BT3m) and 10 m (BT10m). Room: a furnished office 16 by 9.5 by 2.7 metres. The heating, ventilation and air-conditioning system and common appliances including a water cooler and a refrigerator ran normally throughout, deliberately, to create a natural noise floor. SYNCHRONISATION: "Synchronization across channels was achieved with slate pulses and verified by cross-correlation, ensuring sub-sample alignment and minimal drift throughout each session." (p. 2). All microphones were fixed at measured distances with line of sight maintained. Total: 114 hours across all microphones, from about 20 hours of unique meeting sessions; 88 hours train, 12.8 dev, 13.3 eval. LICENCE: "The corpus is available under CC-BY-SA 4.0" (p. 1), repeated as "a permissive CC-BY-SA 4.0 license" (p. 1).

- **Methods / model:** annotation is two-stage: three trained annotators segment and transcribe under a unified guideline covering tokenisation, code-switching and non-lexical events, then a senior annotator reviews every session to resolve discrepancies, correct boundaries and validate transcripts. The released corpus provides utterance-level transcripts with speaker labels and an explicit overlap mask. A Thai tag set marks noise, silences over 300 ms, unintelligible spans and dialectal uncertainty; overlaps are indicated by concatenating speaker identifiers with an ampersand. Baselines fine-tune Whisper variants for five epochs on a single NVIDIA H200; word tokenisation for scoring uses the newmm segmenter from PyThaiNLP. Front ends tested: weighted prediction error (WPE) dereverberation via NaraWPE, and minimum mean-square-error log-spectral-amplitude spectral subtraction.
- **Statistical detail:**
  - Test: none. No significance test, no confidence interval, no seeds. Utterance-level word error rate distributions are shown as half-violin plots with quartiles for single-speaker versus overlapped speech.
  - Per-arm N: the test split of 11 sessions and 27,580 utterances, scored per microphone.
  - Effect sizes: absolute word error rate differences.
  - p-values: none reported.
  - Power / pre-registration: not applicable and not reported.
- **Key findings:**
  - THE TWO WHISPER FIGURES, WITH THEIR CONDITION: for Pathumma-whisper-th-large-v3, a Thai-specific Whisper fine-tune, zero-shot far-field macro-average word error rate was 81.57 percent and near-field 36.99, for an overall 64.32. After fine-tuning on LOTUSDIS, far-field fell to 49.54 and near-field to 21.59, overall 38.33. So 81.6 and 49.5 are the FAR-FIELD MACRO AVERAGES across the JBL at 2 m, BT3m and BT10m, in Thai, before and after in-domain fine-tuning.
  - Distance dominates. Zero-shot per-microphone word error rates for the same model: lavalier 37.55, condenser 36.43, JBL at 2 m 44.22, BT3m 96.27, BT10m 104.22. A word error rate above 100 percent means insertions exceed correct words.
  - Fine-tuning helps most where it is worst: BT3m from 96.27 to 58.15 and BT10m from 104.22 to 64.04, while the condenser improved from 36.43 to 20.40.
  - Off-the-shelf multilingual Whisper-large-v3 was worse than the Thai fine-tune throughout: zero-shot 51.95 lavalier, 117.03 BT3m, 125.52 BT10m, overall 79.84.
  - Classical single-channel front ends HURT. Weighted prediction error dereverberation raised near-field word error from 21.59 to 35.92 and far-field from 49.54 to 56.12; spectral subtraction raised them to 24.92 and 54.55. The authors conclude that a uniform front-end strategy is suboptimal.
  - Training on one microphone type overfits catastrophically. A model fine-tuned only on the near-field condenser reached 19.26 percent on that device, better than the all-microphone baseline, but collapsed to 97.95 percent on BT3m. Adding reverberation augmentation to the single-microphone training recovered far-field to 65.39 percent, the best single-microphone

result.
- Overlap and distance compound. The utterance-level word error distribution shifts upward under overlap for every microphone, most severely on BT3m and BT10m, with an increase in deletions and short-token drops and more substitutions on Thai tone-bearing syllables.
- THE CORPUS LICENCE TABLE, VERBATIM FROM TABLE 1 (p. 2). AMI Meeting Corpus: English, 100 hours, 137 sessions, 4 speakers per session, 0.3-3 m close-talk plus mic arrays, CC BY 4.0. ICSI Meeting Corpus: English, 72 hours, 75 sessions, 3-10 speakers, close-talk worn plus table top, CC BY 4.0. AliMeeting: Chinese, 120 hours, 220 sessions, 2-4 speakers, 481 speakers, 0.3-5 m headset plus circular array, CC BY-SA 4.0. LibriCSS: English, 10 hours, 10 sessions, 8 speakers, 40 speakers, 0.3-4 m circular array (audio play back), CC BY 4.0. CHiME-6 (CHiME-5 data): English, 50 hours, 20 sessions, 4 speakers, 48 speakers, 0.2-4 m close-talk plus linear arrays, CC BY-SA 4.0. NOTSOFAR-1: English, 28 hours (260 all mics), 315 sessions, 4-8 speakers, 35 speakers, close talk plus far-field, CC BY-NC-ND 4.0. DiPCo: English, 5 hours, 10 sessions, 4 speakers, 32 speakers, 1-4 m close-talk plus circular arrays, CDLA-Permissive. AISHELL-4: Chinese, 120 hours, 211 sessions, 4-8 speakers, 61 speakers, 0.6-6 m headset plus circular array, CC BY-SA 4.0. LOTUSDIS: Thai, 20 hours (114 all mics), 90 sessions, 3 speakers, 86 speakers, 0.12-10 m lavalier plus table top, CC BY-SA 4.0.

- **Author's framing of the contribution:** > "LOTUSDIS addresses this gap by offering a non-array, single-channel evaluation framework across diverse > microphone types and distances." (p. 2)
- **What this paper does NOT establish:**
  - It is NOT English. Every number is Thai, scored with a Thai word segmenter, and the paper attributes part of the error to Thai tone-bearing syllables under masking. The 81.6 and 49.5 figures cannot be quoted as English far-field results.
  - It does NOT measure a phone. The far-field devices are a tabletop loudspeaker unit and two Bluetooth speakerphones; no handset is in the rig.
  - It does NOT report a speaker-attributed error rate, a diarization error rate or any joint metric. Every figure is plain word error rate; speaker labels exist in the corpus but no attribution is scored.
  - It does NOT test the NOTSOFAR-1, AMI or CHiME licences it tabulates; Table 1's licence column is the authors' own summary of other corpora and disagrees with at least one of those corpora's own distributions.
  - It does NOT test summarisation or action items; the corpus has transcripts, speaker labels and an overlap mask, and no summaries.
  - It does NOT include a mobile or moving talker condition; all microphones and all distances are fixed with line of sight.
- **Direct quotes (verbatim, source ground truth):**
  - "Speech was recorded simultaneously by nine independent single-channel devices spanning six microphone types at distances from 0.12 m to 10 m, preserving the authentic effects of reverberation, noise, and device coloration without relying on microphone arrays." (p. 1)
  - "Fine-tuning on LOTUSDIS dramatically improved robustness: a Thai Whisper baseline reduced overall WER from 64.3% to 38.3% and far-field WER from 81.6% to 49.5%, with especially large gains on the most distant microphones." (p. 1)
  - "All microphones were fixed at measured distances with line-of-sight maintained. Synchronization across channels was achieved with slate pulses and verified by cross-correlation, ensuring sub-sample alignment and minimal drift throughout each session." (p. 2)
  - "The corpus is available under CC-BY-SA 4.0" (p. 1)
  - "These results confirm that single-channel training leads to severe overfitting to device-specific acoustics, undermining generalization to new conditions." (p. 4)
  - "a uniform front-end strategy is suboptimal. In this setting, fine-tuned models trained directly on distance-diverse data achieved greater robustness than either dereverberation or denoising." (p. 4)
  - "their reliance on microphone arrays presents a different set of research problems and may not reflect more common deployment scenarios where only single-channel, off-the-shelf devices are available." (p. 2)
- **Limitations:**

  - Inferred: Thai is a low-resource language for speech recognition and the absolute error rates are much higher than English equivalents, so the SHAPE of the distance curve transfers better than its level.
  - Inferred: three participants per session and 29 to 40 percent overlap is a smaller and more overlapped conversation than a four-to-eight-person business meeting.
  - Inferred: no significance testing, no seeds, one test split of 11 sessions.
  - Inferred: the corpus-comparison table's licence and size figures for other corpora are secondary reporting; its entry for NOTSOFAR-1 gives 315 sessions and 35 speakers, which disagrees with the NOTSOFAR-1 dataset paper's own Table 1 (280 meetings, 32 speakers), and its CC BY-NC-ND 4.0 licence for NOTSOFAR-1 is not a licence the NOTSOFAR-1 dataset paper itself states.
- **Cross-references in this index:**
  - See also: Vinnikov 2024 (the NOTSOFAR-1 dataset paper whose own counts and licence this paper's Table 1 reports differently); Abramovski 2025 (the single-channel-versus-array gap on English office meetings); Polok 2026 (independently finds that reverberation augmentation is what rescues far-field performance).
  - Contrast with: Shi 2023 and Cornell 2024 (array corpora, where the extra microphones are the point); Sun 2025 (measures overlap detection rather than transcription on far-field meeting audio).
- **Relevance to platform:** This is the vault's most direct answer to the charter's `phone-is-the-microphone` question, because it is the only paper that puts several single-channel devices at measured distances on the SAME conversation and reports what each one scored. The distance curve is steep: on identical speech, word error rate ran 36 to 38 percent at 12 to 15 cm, 44 percent at 2 m, 96 percent at 3 m and 104 percent at 10 m with an off-the-shelf model, and in-domain fine-tuning brought the 3 m and 10 m figures to 58 and 64 percent while leaving them roughly triple the near-field rate. It also settles two engineering questions the project would otherwise have to test: classical dereverberation and denoising front ends made things worse, and training on one device type collapsed on every other device type unless reverberation augmentation was added. The language is Thai, so the levels do not transfer, but the shape of the curve and both negative results are about acoustics, not about Thai.
- **Quotable stats (paste-ready):**
  - "On the same conversations captured simultaneously by nine single-channel devices, off-the-shelf recognition word error rate ran 36.4 percent on a table condenser at 12 to 15 cm, 44.2 percent on a tabletop unit at 2 m, 96.3 percent on a speakerphone at 3 m and 104.2 percent at 10 m (Tipaksorn 2025, Thai)."
  - "Fine-tuning on distance-diverse in-domain data cut far-field macro-average word error rate from 81.6 to 49.5 percent and overall word error rate from 64.3 to 38.3 percent (Tipaksorn 2025, LOTUSDIS test set)."
  - "Classical single-channel front ends degraded recognition: weighted prediction error dereverberation raised far-field word error rate from 49.5 to 56.1 percent and near-field from 21.6 to 35.9 (Tipaksorn 2025)."
  - "A model fine-tuned on one near-field microphone reached 19.3 percent word error rate on that device and 98.0 percent on a speakerphone 3 m away; adding reverberation augmentation recovered the far-field average to 65.4 percent (Tipaksorn 2025)."
  - "LOTUSDIS is released under CC-BY-SA 4.0, and its corpus-comparison table records NOTSOFAR-1 as CC BY-NC-ND 4.0 (Tipaksorn 2025, Table 1)."

Vinnikov 2024 - the NOTSOFAR-1 datasets as the dataset paper itself describes them (N=280 meetings)

- **Citation:** Alon Vinnikov, Amir Ivry, Aviv Hurvitz, Igor Abramovski, Sharon Koubi, Ilya Gurvich, Shai Pe'er, Xiong Xiao, Benjamin Martinez Elizalde, Naoyuki Kanda, Xiaofei Wang, Shalev Shaer, Stav Yagev, Yossi Asher, Sunit Sivasankaran, Yifan Gong, Min Tang, Huaming Wang, Eyal Krupka. "NOTSOFAR-1 Challenge: New Datasets, Baseline, and Tasks for Distant Meeting Transcription." Proc. Interspeech 2024, 1-5 September 2024, Kos, Greece, pages 5003-5007. DOI:

10.21437/Interspeech.2024-1788. Affiliation as printed: Microsoft.
- **File:** `literature/isca-vinnikov24-notsofar1-dataset.pdf`
- **Links:** DOI: 10.21437/Interspeech.2024-1788 | Code: github.com/microsoft/NOTSOFAR1-CHALLENGE, given in the paper as the baseline system with inference, training, data handling and evaluation code (named in the paper, not verified here) | Data / project page: chimechallenge.org/current/task2/index (named in the paper, not verified here) | Weights: not named | Citations: not retrieved | Reproduced: yes, in the sense that fourteen independent teams built systems on this data and their results are reported in Abramovski 2025
- **Evidence tier:** [A] proven DOMAIN. Interspeech is a top venue on this index's list, and the dataset and an open-source baseline were released and then used by fourteen independent teams, which is a second strong signal; the citation rate is unverified.
- **Mechanism family:** `Corpus resource`, `Far-field single-channel transcription benchmark`, `Far-field multi-channel transcription benchmark`, `Near-field close-talk reference condition`
- **Study type:** dataset and challenge description with a baseline system.
- **Population:** THE DATASET PAPER'S OWN TABLE 1, VERBATIM IN CONTENT. Meetings: 107 training, 36 development, 137 evaluation. That is 280 meetings in total, and the body text agrees: "a benchmark dataset of 280 English meetings" (p. 1) and "featuring 280 unique meetings with authentic multi-participant English conversations" (p. 3). Average duration about 6 minutes throughout. Devices per meeting, given as multi-channel / single-channel: 4 / 5 in training, about 3 / 5 in development, about 4 / 6 in evaluation. Sessions, where a session is one device's recording of one meeting: 428 / 535 training, 106 / 177 development, 548 / 822 evaluation. Session hours: 42 / 53 training, 10 / 17 development, 54 / 82 evaluation. Rooms: "20 rooms in total" spanning training and development, and 10 for evaluation. Speakers: "22 speakers in total" for training and development, and 10 for evaluation. The body text adds that the meetings span 30 different rooms overall, that attendees typically number 4 to 8 adults, that gender representation was balanced, and that the dataset contains around 150 hours of single-channel and 110 hours of multi-channel audio across all sessions. The evaluation set is entirely disjoint from training and development, with no overlap in speakers or rooms. LICENCE: THE DATASET PAPER STATES NO LICENCE. There is no licence statement anywhere in the PDF, for the recorded meeting dataset, for the simulated dataset or for the baseline code.
- **Imaging / measurement:** each meeting was captured simultaneously by several devices, each positioned differently: around 5 single-channel devices each producing one internally processed stream, and 4 multi-channel devices each producing 7 raw streams from a geometry of one central and six surrounding microphones. The paper is explicit that "single-channel" here means a commercial conference-room device's processed output, not a raw single microphone: those devices "are typically equipped with microphone arrays and employ proprietary on-device processing involving echo cancellation, de-reverberation, beamforming, and noise suppression, yielding a single stream output that we refer to in this paper as 'single-channel'" (p. 3). Attendees wore close-talk microphones for annotation. The recordings include speakers at varying distances and volumes, in-seat movement, walking, standing or sitting, speaking near a whiteboard, and entering or leaving the room; conversational dynamics include overlapping speech, interruptions, rapid speaker change, laughter, coughing and fillers. Recordings are exclusively in English by native or near-native speakers, deliberately, to keep language and accent out of the far-field question.
- **Methods / model:** annotation is bias-free by construction: the development and evaluation transcription process relies on close-talk recordings and is conducted entirely without human access to machine-generated transcriptions, because a machine-then-human pipeline would penalise algorithms that deviate from the transcription algorithm used. Spectral subtraction between close-talk channels pre-processed an automatic segmenter tuned never to miss the target speaker; human annotators corrected segmentation errors; throat microphones were used for quality control; most meetings were transcribed independently by two professional transcribers with a third judge resolving non-consensus, while about 50 training meetings used two transcribers with the second deciding. The simulated training set is about 1000 hours built from 15,000 real acoustic transfer functions measured in real conference rooms with a mouth simulator emitting chirps, convolved with the upper mean-opinion-score quartile of the LibriVox corpus (about 500 hours), augmented with inserted silences to create realistic overlap, mixed up to three speakers per mixture into seg-

ments averaging 50 seconds, with real recorded transient and stationary noise added. Room impulse responses are split at a 50-millisecond cutoff into a direct-plus-early-reflections component and a late-reverberation component, and the baseline uses the former as its separation target. The baseline system is continuous speech separation with a Conformer separator trained with permutation invariant training, then Whisper large-v3 with word-level timestamps, then NeMo multi-scale speaker embeddings clustered by normalised maximum eigengap spectral clustering. The only difference between the single-channel and multi-channel variants is whether the separator takes 1 or 7 channels and whether masks are applied by multiplication or by minimum variance distortionless response beamforming.

- **Statistical detail:**
  - Test: none, but the paper argues that treating each meeting as an approximately independent and identically distributed sample makes confidence intervals computable, and states that its baseline code includes a simple method for calculating them in absolute terms and relative to the baseline algorithm.
  - Per-arm N: 106 multi-channel and 177 single-channel development sessions for the baseline table.
  - Effect sizes: absolute error rates per condition.
  - p-values: none reported.
  - Power / pre-registration: the argument for prioritising many short meetings over long ones is explicitly a statistical-power argument: “the narrowness of these confidence intervals is determined by the number of meetings rather than the total recorded hours” (p. 3).
- **Key findings:**
  - Baseline results on the DEVELOPMENT set, reported as multi-channel / single-channel. All sessions: tcpWER 32.4 / 46.8 percent, tcORC WER 26.7 / 38.5 percent. Meetings tagged NaturalMeeting: 32.3 / 47.6 and 26.2 / 40.2. DebateOverlaps: 38.0 / 54.9 and 31.4 / 44.7. TurnsNoOverlap: 21.2 / 32.4 and 18.8 / 29.7. TransientNoise high: 33.6 / 51.0 and 29.1 / 43.7. TalkNearWhiteboard: 39.9 / 55.4 and 31.2 / 43.9.
  - The close-talk upper bound on the same 36 development meetings, processed through the same recognition and diarization modules with no separation front end: 12.0 percent tcpWER and 9.3 percent tcORC WER. This is the near-field reference against which every far-field number in the challenge should be read.
  - The single-channel penalty in the baseline is roughly 14 absolute points of tcpWER across every condition, and it is largest on debate-style overlap (38.0 against 54.9) and on talking near a whiteboard (39.9 against 55.4).
  - The paper’s stated motivation for the corpus is the inadequacy of the alternatives: AMI is “confined to three rooms” with development and evaluation sets of only 18 and 16 sessions and at most four attendees; AliMeeting’s test set is 20 sessions with at most four attendees; neither has a matching simulated training set.
  - Most meetings were semi-professional role-play, with participants acting as professionals discussing a work-related issue, and some were non-work topics such as favourite television shows.
- **Author’s framing of the contribution:** > “We launch two new datasets: First, a benchmark dataset of 280 English meetings, averaging 6 minutes each, > capturing a broad spectrum of acoustic and conversational patterns across 30 rooms with 4-8 attendees. > Second, a 1000-hour simulated training dataset, synthesized for real-world generalization, incorporating > 15,000 real acoustic transfer functions.” (p. 1)
- **What this paper does NOT establish:**
  - It does NOT state a licence for any of its data or code.
  - Its “single-channel” condition is NOT a bare microphone and is NOT a phone. It is a commercial conference-room device’s output after that device’s own echo cancellation, dereverberation, beamforming and noise suppression, and the paper says so explicitly and calls the distinction out as important.
  - It does NOT report evaluation-set results. Every number in the paper is on the development set.
  - It does NOT test any language other than English, deliberately.

  - It does NOT test summarisation or action-item extraction, although it names meeting summary and note taking as the applications motivating the work.
  - It does NOT contain time-varying acoustic transfer functions in the simulated training set, which the authors name as its notable limitation.
  - It does NOT report latency, real-time factor or compute cost for the baseline.
- **Direct quotes (verbatim, source ground truth):**
  - "We launch two new datasets: First, a benchmark dataset of 280 English meetings, averaging 6 minutes each, capturing a broad spectrum of acoustic and conversational patterns across 30 rooms with 4-8 attendees." (p. 1)
  - "Our dataset consists of approximately 280 relatively short meetings, each lasting roughly six minutes." (p. 3)
  - "Distinguishing single-channel from single-microphone: Speech recognition systems often deal with audio from conference room devices. These devices are typically equipped with microphone arrays and employ proprietary on-device processing involving echo cancellation, de-reverberation, beamforming, and noise suppression, yielding a single stream output that we refer to in this paper as 'single-channel'. This significantly differs from the single-microphone audio commonly used for evaluation." (p. 3)
  - "our transcription process for the development and evaluation sets relies on close-talk recordings and is conducted entirely without human access to machine-generated transcriptions" (p. 3)
  - "the utility of a benchmark is determined by the distribution it spans and the sample size it provides" (p. 3)
  - "A notable limitation of the simulated training set is the static nature of the ATFs, which, despite being recorded in real rooms, remain fixed throughout each simulated utterance." (p. 4)
- **Limitations:**
  - Author-stated: the simulated training set has static acoustic transfer functions, so it cannot train robustness to a moving speaker or a changing environment.
  - Author-stated by implication: about 50 training meetings were transcribed by only two people rather than three with adjudication.
  - Inferred: the paper states no licence, which is a material gap for any commercial use of the data.
  - Inferred: meetings are role-played on assigned topics and average six minutes, so nothing here bears on a one-hour unscripted meeting.
  - Inferred: 22 speakers across the training and development sets and 10 in evaluation is a narrow set of voices for a corpus of 280 meetings.
- **Cross-references in this index:**
  - See also: Abramovski 2025 (the challenge summary built on this data, which reports the results of fourteen submissions and gives a different split table); Kalda 2024 and Niu 2024 (two submitted systems); Cornell 2024 (the twin CHiME-8 challenge that used this data as one of four scenarios); Polok 2026 (uses the NOTSOFAR-1 close-talk channels as a simulation seed).
  - Contrast with: Tipaksorn 2025 (single-channel in the literal sense of one microphone element, which this paper is careful to say its own single-channel track is NOT); Watanabe 2020 (a dinner-party rather than an office-meeting condition).
- **Relevance to platform:** Two things in this paper bear directly on the charter. First, its warning that "single-channel" in this literature means a conference-room device's beamformed, echo-cancelled, noise-suppressed output stream, not a raw microphone. Every single-channel number this project might quote from the NOTSOFAR-1 family, including the 22.2 percent winning score, was measured downstream of hardware signal processing that a phone application does not necessarily get. Second, the close-talk reference of 12.0 percent tcpWER against 46.8 percent for the single-channel baseline on the same 36 meetings, which is the cleanest statement in the vault of what the microphone position alone costs. That gap is the charter's `phone-is-the-microphone` bet stated as a number. DATA-COUNT NOTE FOR DOWNSTREAM DOCUMENTS: this dataset paper's own Table 1 gives 107 + 36 + 137 = 280 meetings, 20 rooms and 22 speakers for training plus development, and 10 rooms and 10 speakers for evaluation. The challenge summary (Abramovski 2025) reports 110 + 35 + 170 = 315 meetings with 13 evaluation rooms and 13 evaluation speakers, and

Tipaksorn 2025 repeats 315 sessions and 35 speakers. The three published counts disagree; this entry reports what the dataset paper's own table says.

- **Quotable stats (paste-ready):**
  - "The NOTSOFAR-1 dataset paper's own Table 1 gives 107 training, 36 development and 137 evaluation meetings, 280 in total, across 30 rooms with 4 to 8 attendees each (Vinnikov 2024)."
  - "The NOTSOFAR-1 dataset paper states no licence for its recorded meeting dataset, its 1000-hour simulated training set or its baseline code (Vinnikov 2024)."
  - "On the same 36 development meetings, the baseline scored 12.0 percent time-constrained speaker-attributed word error rate from participants' close-talk microphones against 46.8 percent from a single distant device stream (Vinnikov 2024)."
  - "The baseline single-channel penalty was worst on debate-style overlapping speech, 54.9 percent against 38.0 for the seven-microphone version of the same system (Vinnikov 2024, development set)."
  - "In this literature 'single-channel' means a conference-room device's output after its own echo cancellation, dereverberation, beamforming and noise suppression, which the dataset authors distinguish explicitly from single-microphone audio (Vinnikov 2024)."

Watanabe 2020 - twenty real dinner parties on six Kinect arrays, and what diarization costs (N=20 parties)

- **Citation:** Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, David Snyder, Aswin Shanmugam Subramanian, Jan Trmal, Bar Ben Yair, Christoph Boeddeker, Zhaoheng Ni, Yusuke Fujita, Shota Horiguchi, Naoyuki Kanda, Takuya Yoshioka, Neville Ryant. "CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings." arXiv:2004.09249v2, 2 May 2020. DOI: not provided in the PDF. Affiliations as printed: Johns Hopkins University; The City University of New York; University of Sheffield; Inria; Paderborn University; Hitachi; Microsoft; Linguistic Data Consortium.
- **File:** `literature/arxiv-2004.09249.pdf`
- **Links:** arXiv:2004.09249 | DOI: not provided | Code: Kaldi recipes at github.com/kaldi-asr/kaldi/tree/master/egs/chime6/s5_track1 and .../s5_track2, plus the array synchronisation tool at github.com/chimechallenge/chime6-synchronisation and the scoring tool github.com/nryant/dscore (all named in the paper, not verified here) | Weights: pretrained speech-activity and diarization models at kaldi-asr.org/models/m12 (named in the paper, not verified here) | Project page: https://chimechallenge.github.io/chime6/ | Citations: not retrieved | Reproduced: yes, in the sense that the challenge received independent submissions
- **Evidence tier:** [A] proven DOMAIN. Published in the CHiME challenge series on this index's venue list, with complete open-source baselines and pretrained models released, and used by independent teams; the citation rate is unverified.
- **Mechanism family:** `Corpus resource`, `Far-field multi-channel transcription benchmark`, `Near-field close-talk reference condition`, `Speaker attribution metric`, `Joint speaker-attributed transcription metric`, `Array front-end signal processing`
- **Study type:** challenge description with baseline systems and results; no new recordings, since the audio is the CHiME-5 material with corrected synchronisation.
- **Population:** twenty separate dinner parties in real homes. Each party has four participants, two acting as hosts and two as guests; all are friends who know each other well and were instructed to behave naturally. Each party lasted a minimum of two hours and was composed of three phases in three locations: kitchen (preparing the meal), dining (eating) and living (a post-dinner period in a separate room), each phase at least 30 minutes. Participants moved freely between locations. Topics were unconstrained and there was no scenario. Background television and commercial music were disallowed to avoid recording copyrighted content. Some personally identifying material was redacted post-recording as part of the consent process. Training, development-test and evaluation-test sets feature non-overlapping homes and speakers. A range of room acoustics from 20 different homes, each with two or three separate recording areas, with real domestic

background noise from kitchen appliances, air conditioning and movement.

- **Imaging / measurement:** THE CHiME-6 RIG, EXACTLY AS DESCRIBED. Each party was recorded with a set of SIX MICROSOFT KINECT DEVICES, placed so that at least two capture the activity in each location. Each Kinect has a LINEAR ARRAY OF 4 SAMPLE-SYNCHRONISED MICROPHONES and a camera; raw microphone signals and video were recorded, each Kinect onto a separate laptop. In addition, to facilitate transcription, each participant wore Soundman OKM II Classic Studio BINAURAL microphones recorded through a Soundman A3 adapter onto Tascam DR-05 stereo recorders worn by the participants. So CHiME-6 is 24 far-field microphones in 6 four-element linear arrays, plus 4 worn binaural pairs, in a real home, with the talkers moving between three rooms. THE DINNER PARTY CORPUS IS A DIFFERENT CORPUS ON DIFFERENT HARDWARE. This paper describes DiPCo only in its related-work section, as an external corpus: "The DIPCO corpus, inspired by the CHiME-5 challenge but of shorter duration, provides recordings of dinner table interactions between four participants recorded simultaneously on several commercially available microphone arrays." (p. 1). DiPCo is not part of CHiME-6, was not recorded on Kinects, and none of the results in this paper are DiPCo results. The array synchronisation contribution is specific: frame-dropping, which affects the Kinect devices only, is compensated by inserting zeros at locations detected by comparing Kinect audio to an uncorrupted stereo signal recovered from the recorded video files; clock drift, which affects all devices, is estimated by cross-correlating each device against the session's reference binaural recording at regular intervals and corrected with a linear fit, an adjustment typically under 100 ms over a 2.5-hour session. Two devices required a piecewise linear fit because they temporarily changed speed.
- **Methods / model:** two tracks. Track 1 is recognition given ground-truth diarization; Track 2 is diarization plus recognition from raw recordings. Both are multi-array. The baseline is a Kaldi recipe: a 127,712-word lexicon, a maximum-entropy 3-gram language model, point-source noise augmentation from the corpus's own noise regions, 13-dimensional MFCC features, GMM-HMM training and data cleaning, then a factorised time-delay neural network adapted from the Switchboard 7q recipe. Two enhancement front ends: weighted prediction error dereverberation plus guided source separation plus beamforming using only the first and last microphone of each array, and weighted prediction error plus weighted delay-and-sum BeamformIt on the reference array. Track 2 adds a TDNN-plus-LSTM speech activity detector and an x-vector diarization system with a 5-layer TDNN trained on VoxCeleb, probabilistic linear discriminant analysis trained on CHiME-6 data, and agglomerative hierarchical clustering that is told there are exactly four speakers. The Track 2 baseline performs speech activity detection, diarization and recognition on the U06 array only. Scoring for Track 2 is concatenated minimum-permutation word error rate (cpWER) over 24 speaker permutations; diarization is scored as diarization error rate and Jaccard error rate with dscore. A refined reference for diarization was produced by forced-aligning the transcripts to the cleaner binaural recordings, merging words separated by at most 300 ms of silence.
- **Statistical detail:**
  - Test: none. Baseline scores only; no significance tests, confidence intervals or seeds.
  - Per-arm N: development and evaluation splits with non-overlapping homes and speakers.
  - Effect sizes: absolute word error rate and diarization error rate differences.
  - p-values: none reported.
  - Power / pre-registration: not applicable; the evaluation set was blind.
- **Key findings:**
  - Track 1, recognition with ORACLE segmentation: 51.8 percent word error rate on development and 51.3 percent on evaluation with guided source separation. With BeamformIt instead: 69.8 and 61.2. The CHiME-5 baseline on the equivalent task was 81.1 and 73.3, and the best CHiME-5 system was 45.6 and 46.6.
  - Track 2, full pipeline with system diarization: 84.3 percent word error rate on development and 77.9 on evaluation, using BeamformIt.
  - The cost of diarization is stated as a subtraction: comparing Track 1 and Track 2 with the same BeamformIt front end and the same acoustic and language models, "the main degradation (around 15% absolute) comes from speaker diarization" (p. 5).
  - Diarization itself is very poor on this data. Against the human-annotation reference, diarization error rate was 61.6 percent on development and 62.0 on evaluation, with Jaccard error rates

of 69.8 and 71.4. Against the forced-alignment reference used as the official one, 63.4 / 68.2 and 70.8 / 72.5. The authors say plainly that "both metrics show over 60% error rates and improving the diarization performance is one of the main challenges in Track 2" (p. 5).
  - Speech activity detection is comparatively easy: total error 3.3 percent on development and 5.9 on evaluation against the annotation reference, 2.6 and 5.8 against the alignment reference.
  - The paper notes that the evaluation set is harder than the development set for speech activity detection, and that the two reference formats differ significantly on development and marginally on evaluation.
- **Author's framing of the contribution:** > "Of note, Track 2 is the first challenge activity in the community to tackle an unsegmented multispeaker > speech recognition scenario with a complete set of reproducible open source baselines providing speech > enhancement, speaker diarization, and speech recognition modules." (p. 1)
- **What this paper does NOT establish:**
  - It is NOT a meeting corpus. It is dinner parties in homes with people moving between a kitchen, a dining area and a living room; the acoustics, the movement and the conversational register are all different from a seated conference-room meeting.
  - It does NOT contain the Dinner Party Corpus. DiPCo is cited in related work as a separate corpus recorded on commercially available microphone arrays; no DiPCo result appears here.
  - It is NOT single-channel. Every baseline uses multiple synchronised microphones, and Track 1's better front end uses guided source separation across arrays.
  - It does NOT measure a phone. Six Kinect four-element linear arrays and worn binaural microphones.
  - It does NOT test summarisation or action-item extraction.
  - It does NOT report latency, real-time factor or compute cost, and the pipeline includes two-stage decoding with i-vector refinement.
- **Direct quotes (verbatim, source ground truth):**
  - "Each party has been recorded with a set of six Microsoft Kinect devices. The devices have been strategically placed such that there are always at least two capturing the activity in each location. Each Kinect device has a linear array of 4 sample-synchronised microphones and a camera." (p. 2)
  - "In addition to the Kinects, to facilitate transcription, each participant is wearing a set of Soundman OKM II Classic Studio binaural microphones. The audio from these is recorded via a Soundman A3 adapter onto Tascam DR-05 stereo recorders being worn by the participants." (p. 2)
  - "The DIPCO corpus [33], inspired by the CHiME-5 challenge [34] but of shorter duration, provides recordings of dinner table interactions between four participants recorded simultaneously on several commercially available microphone arrays." (p. 1)
  - "The dataset is made up of the recording of twenty separate dinner parties taking place in real homes. Each dinner party has four participants - two acting as hosts and two as guests." (p. 2)
  - "In spite of using a state-of-the-art diarization technique, both metrics show over 60% error rates and improving the diarization performance is one of the main challenges in Track 2." (p. 5)
  - "by comparing Tracks 1 and 2 with BeamformIt, we can observe that the main degradation (around 15% absolute) comes from speaker diarization." (p. 5)
- **Limitations:**
  - Author-stated: the clock-drift correction failed for two devices that temporarily changed speed and needed a piecewise linear fit.
  - Author-stated by implication: the Track 2 baseline runs speech activity detection, diarization and recognition on one array only, for simplicity, and the authors name multi-array fusion as an integral part of the challenge rather than something they solved.
  - Inferred: word error rates above 50 percent even with oracle segmentation mean this corpus is far harder than office meetings, and its numbers should not be read as a bound on meeting transcription.

- Inferred: the diarization system is told there are exactly four speakers, which no deployed system knows.
- **Cross-references in this index:**
  - See also: Cornell 2024 and Cornell 2025 (the CHiME-7 and CHiME-8 successors, where CHiME-6 remains the hardest scenario and no system got below 30 percent tcpWER on it); Vinnikov 2024 (the office-meeting counterpart, with far lower error rates).
  - Contrast with: every office-meeting entry in this index. CHiME-6's 51.3 percent oracle-segmentation word error rate and 77.9 percent full-pipeline cpWER are dinner-party-in-a-home numbers, and quoting them for a seated meeting overstates the difficulty by a large margin.
- **Relevance to platform:** This paper is the authority for what CHiME-6 actually is, and the project needed it because a research response had placed the Dinner Party Corpus on CHiME-6's recording rig. It does not belong there: CHiME-6 is twenty dinner parties captured on six four-microphone Kinect arrays plus worn binaural microphones, and DiPCo is a separate, shorter corpus on commercially available arrays that this paper cites only in related work. The number worth carrying forward is the 15-absolute-point cost of automatic diarization measured inside one system with everything else held constant, which is the cleanest isolation of that cost anywhere in the vault.
- **Quotable stats (paste-ready):**
  - "CHiME-6 recorded twenty two-hour dinner parties in real homes on six Microsoft Kinect devices, each a linear array of four sample-synchronised microphones, plus worn binaural microphones for transcription (Watanabe 2020)."
  - "With ground-truth segmentation and guided source separation across arrays, the CHiME-6 baseline scored 51.3 percent word error rate on the evaluation set; with automatic diarization it scored 77.9 percent (Watanabe 2020)."
  - "Holding the front end, acoustic model and language model constant, automatic speaker diarization cost about 15 absolute points of word error rate (Watanabe 2020, CHiME-6)."
  - "A state-of-the-art x-vector diarization system reached 62.0 percent diarization error rate and 71.4 percent Jaccard error rate on the CHiME-6 evaluation set (Watanabe 2020)."
  - "The Dinner Party Corpus is a separate resource recorded on commercially available microphone arrays, not part of CHiME-6 and not recorded on its Kinect rig (Watanabe 2020, related work)."

What a phone can actually run, and what it was asked to summarise

Apple 2024 - the on-device foundation model, its 4,096, and the summaries it was built for (N=2 models)

- **Citation:** Apple. "Apple Intelligence Foundation Language Models." arXiv:2407.21075v2, 27 May 2026 (v1 July 2024). DOI: not provided. Corporate author; the PDF names no individual authors and no affiliations beyond Apple.
- **File:** `literature/arxiv-2407.21075.pdf`
- **Links:** arXiv:2407.21075 | DOI: not provided | Code: not released; the report names third-party tools (JAX, AXLearn) but no repository for the models | Weights: not released | Project page: none | Citations: not retrieved | Reproduced: no; nothing in this report is independently reproducible
- **Evidence tier:** [C] watch METHOD. An industrial technical report with no stated venue, no released code or weights, and evaluation on internal benchmarks with unnamed human graders, so it fails the implementation arm and clears no strong signal. It is indexed because it is the primary source for what an on-device model from a major platform vendor is, and because the project's claims about on-device summarisation rest on it.
- **Mechanism family:** `On-device language model`, `Abstractive summarisation method`
- **Study type:** industrial technical report describing architecture, training data, training recipe, inference optimisation and evaluation.
- **Population:** no human subjects for the models themselves. Human graders were used for the summarisation evaluation; the report states no number of graders, no demographics, no recruitment

and no agreement statistic.

- **Imaging / measurement:** NO AUDIO ANYWHERE. This is a text-only language model report. There is no speech recognition, no microphone and no meeting in it.
- **Methods / model:** two models are described. AFM-on-device is a roughly 3-billion-parameter model with 2.58 billion non-embedding parameters and 0.15 billion embedding parameters, model dimension 3072, head dimension 128, 24 query heads, 8 key-value heads (grouped-query attention), 26 layers, SwiGLU activation, RMSNorm, query and key normalisation, and rotary positional embeddings with a base frequency of 500,000 for long-context support. Tokeniser: byte-pair encoding with byte fallback, 49,000 tokens for the on-device model and 100,000 for the server model. THE 4,096, EXACTLY AS THE REPORT USES IT. Section 3.2.1, Core pre-training: "We train AFM-server from scratch for 6.3T tokens on 8192 TPUv4 chips, using a SEQUENCE LENGTH of 4096 and a batch-size of 4096 sequences." (p. 6). The batch size of 4096 was chosen from a scaling-law fit, with the predicted optimum around 3072 and 4096 selected for chip utilisation. For the on-device model the report states: "All training hyper-parameters except for batch-size are kept the same as AFM-server." (p. 7). So 4,096 in this report is a TRAINING SEQUENCE LENGTH in the core pre-training stage, and separately a batch size in sequences. It is not stated anywhere as a runtime context cap. Training continues past that stage. Section 3.2.2: "For both models we perform continued pre-training at a sequence length of 8192, with another 1T tokens". Section 3.2.3: "Finally, we conduct a further 100B tokens of continued pre-training at a sequence length of 32768 tokens", raising the rotary base frequency from 500,000 to 6,315,089; that section does not say explicitly whether it covers both models or the server model alone, and the only long-context evaluation reported is for AFM-server. AFM-on-device was initialised from a pruned 6.4-billion-parameter model, pruning only the feed-forward hidden dimension using Soft-Top-K masks learned over 188 billion tokens, then trained for a full 6.3 trillion tokens with a distillation loss weighting the teacher's top-1 predictions at 0.9. The report says pruning improved final benchmark results by 0 to 2 percent and distillation added about 5 percent on MMLU and 3 percent on GSM8K. Training used 2048 TPUv5p chips for the on-device model. Quantisation for deployment: mixed-precision palletisation with 4-bit default and some layers pushed to 2-bit, giving about 3.5 bits per weight on average without significant quality loss, with 3.7 bits per weight used in production because it already met the memory requirement. The embedding layer, shared between input and output, is quantised per channel to 8-bit integers. Accuracy-recovery LoRA adapters are trained on top of the quantised model, and feature-specific adapters are initialised from them.
- **Statistical detail:**
  - Test: none. No significance test, no confidence interval, no inter-rater agreement is reported for any human evaluation.
  - Per-arm N: not reported for the summarisation evaluation.
  - Effect sizes: reported as percentage ratios of good and poor results.
  - p-values: none reported.
  - Power / pre-registration: not reported.
- **Key findings:**
  - The summarisation feature the on-device model was built for is EMAILS, MESSAGES AND NOTIFICATIONS. Not meetings, not transcripts and not conversation. "We worked with our design teams to create specifications for summaries of Emails, Messages, and Notifications." (p. 18).
  - The base on-device model could not follow the summary specification, and a LoRA adapter was fine-tuned on top of the quantised model specifically to make it conform: "While AFM-on-device is good at general summarization, we find it difficult to elicit summaries that strictly conform to the specification." (p. 18).
  - The adapter's training summaries were SYNTHETIC, generated by the server model according to the product's requirements, then filtered by rule-based heuristics for length, formatting, point of view and voice, and by model-based filters for entailment.
  - Prompt injection is an acknowledged failure mode of the on-device summariser: "We find that AFM-on-device is prone to following instructions or answering questions that are present in the input content instead of summarizing it." (p. 18). The mitigation was to generate more synthetic training data with the server model, which the report says does not exhibit the be-

haviour.
  - Human satisfaction with the summarisation feature, judged along five dimensions where a result is good only if all dimensions are good and poor if any is poor. AFM-on-device plus adapter: Email 71.3 percent good and 7.2 percent poor; Message 63.0 good and 15.9 poor; Notification 74.9 good and 10.0 poor. Comparison models: Gemma-7B 70.9 / 4.5 on Email, 51.1 / 18.3 on Message, 60.9 / 12.9 on Notification; Phi-3-mini 44.7 / 12.0, 32.8 / 31.4, 56.6 / 18.3; Llama-3-8B 36.1 / 12.8, 21.3 / 35.3, 27.7 / 48.3.
  - On the Email summarisation task, Gemma-7B had a LOWER poor-result ratio than the on-device model plus adapter, 4.5 against 7.2 percent, while the good-result ratios were within 0.4 points.
  - Long-context capability is reported for the SERVER model only, and it degrades with length: a RULER-style evaluation score of 91.7 at a 4096 context against 43.3 at 32768.
- **Author's framing of the contribution:** > "These models are designed to perform a wide range of tasks efficiently, accurately, and responsibly. This > report describes the model architecture, the data used to train the model, the training process, how the > models are optimized for inference, and the evaluation results." (p. 1)
- **What this paper does NOT establish:**
  - It does NOT state a runtime context length for AFM-on-device. The 4,096 figure is the core pre-training sequence length (and, separately, the batch size in sequences); continued pre-training then ran at 8,192 and context lengthening at 32,768.
  - It does NOT summarise a meeting, a transcript or any conversation. The summarisation feature covers emails, messages and notifications, which are short written texts with one author each.
  - It does NOT process audio, speech or a microphone anywhere.
  - It does NOT report latency, tokens per second, time to first token, memory footprint or battery cost on any device.
  - It does NOT report how many human graders judged the summaries, who they were, or what agreement they reached.
  - It does NOT release the model, its weights, its code or its evaluation harness, so no number in it is independently checkable.
- **Direct quotes (verbatim, source ground truth):**
  - "We train AFM-server from scratch for 6.3T tokens on 8192 TPUv4 chips, using a sequence length of 4096 and a batch-size of 4096 sequences." (p. 6)
  - "All training hyper-parameters except for batch-size are kept the same as AFM-server." (p. 7)
  - "For both models we perform continued pre-training at a sequence length of 8192, with another 1T tokens from a mixture that upweights math and code, and down-weights the bulk web-crawl." (p. 6)
  - "Finally, we conduct a further 100B tokens of continued pre-training at a sequence length of 32768 tokens, using the data mixture from the continued pre-training stage, augmented with synthetic long-context Q&A data." (p. 6)
  - "We use the AFM-on-device model to power summarization features. We worked with our design teams to create specifications for summaries of Emails, Messages, and Notifications." (p. 18)
  - "While AFM-on-device is good at general summarization, we find it difficult to elicit summaries that strictly conform to the specification. Therefore, we fine tune a LoRA adapter on top of the quantized AFM-on-device for summarization." (p. 18)
  - "We find that AFM-on-device is prone to following instructions or answering questions that are present in the input content instead of summarizing it." (p. 18)
  - "On average, AFM-on-device can be compressed to only about 3.5 bits per weight (bpw) without significant quality loss. We choose to use 3.7 bpw in production as it already meets the memory requirements." (p. 17)
- **Limitations:**
  - Author-stated: the base on-device model does not conform to summary specifications without an adapter, and is prone to prompt injection from the content it is summarising.
  - Author-stated: long-context support was not the focus for this version of AFM, and the long-

context evaluation covers the server model.
  - Inferred: no released artifacts, no reported grader counts and no agreement statistics, so the human satisfaction figures are not verifiable and their uncertainty is unknown.
  - Inferred: the summarisation adapter was trained on synthetic summaries generated by a larger model, so the quality ceiling is the server model's behaviour on emails and messages.
  - Inferred: this is a vendor report on a shipped product, and it reports no failure case of its own beyond prompt injection.
- **Cross-references in this index:**
  - See also: nothing in this index shares its family; it is the only on-device-model entry in the vault.
  - Contrast with: Kirstein 2024, Kirstein 2024b, Gong 2024 and F. Liu 2008 (all of which measure how badly summary quality can be judged, on meeting transcripts, where this report's good-result ratios were never tested); Rennard 2023 and Cornell 2025 (meeting summarisation proper, a longer and structurally different input than an email).
- **Relevance to platform:** This is the vault's only evidence about what a phone can actually run, and it bounds the charter's `finished-before-the-walk-back` and `three-outputs` constraints from the platform side. A roughly 3-billion-parameter model quantised to about 3.7 bits per weight ships on a phone and summarises emails, messages and notifications at a 71.3 to 74.9 percent good-result ratio judged by human graders. It does not summarise transcripts, it needed a fine-tuned adapter to follow a summary specification at all, and it is prone to obeying instructions found inside the content it is summarising, which is a real hazard for a meeting transcript in which someone says "ignore that and do this instead". The 4,096 figure the project has been carrying is a core pre-training sequence length, not a runtime context cap, and the same report describes later training stages at 8,192 and 32,768 tokens.
- **Quotable stats (paste-ready):**
  - "Apple's on-device foundation model is about 3 billion parameters, 2.58 billion of them non-embedding, with 26 layers and a model dimension of 3072, quantised to about 3.7 bits per weight in production (Apple 2024)."
  - "The 4,096 in Apple's foundation model report is the core pre-training sequence length and the pre-training batch size in sequences, not a runtime context limit; continued pre-training ran at 8,192 tokens and context lengthening at 32,768 (Apple 2024)."
  - "The on-device model's summarisation feature was built and evaluated for emails, messages and notifications, and required a fine-tuned adapter because the base model would not conform to the summary specification (Apple 2024)."
  - "Human graders rated the on-device summariser's output good on 71.3 percent of emails, 63.0 percent of messages and 74.9 percent of notifications, and poor on 7.2, 15.9 and 10.0 percent respectively (Apple 2024)."
  - "The report identifies prompt injection as a known failure of the on-device summariser: it is prone to following instructions present in the content it is meant to be summarising (Apple 2024)."

Summary Notes for Platform Work

The single-microphone penalty is measured, and it is roughly a factor of two. On the same 170 real office meetings, the same team's system scored 22.2 percent time-constrained speaker-attributed word error rate from one distant microphone stream and 10.8 percent from a seven-microphone tabletop array (Abramovski 2025, Niu 2024). The published baseline for the same single-channel condition was 41.4 percent. The mechanism behind the gap is named in both papers and is not fixable with a better model on one device: spatial information across microphones is what lets a system separate two people talking at once, and one device does not have it. This number is the vault's most direct evidence bearing on the charter's `phone-is-the-microphone` constraint, and it is also the number a competitor selling a hardware recorder would quote back. Note that all these measurements come from the far-field array device family and none from a handset, so the phone's actual position on this scale is unmeasured, not favourable.

Half the error a user would see in an attributed transcript is attribution, not misheard words. On the same far-field meeting audio, speaker-agnostic character error was 17.3 and 18.4 percent while speaker-attributed character error was around 30 percent (Shi 2023, `Far-field multi-channel transcription benchmark`), and the NOTSOFAR-1 winner's speaker-agnostic score was 17.7 percent single-channel against a speaker-attributed 22.2 percent (Abramovski 2025, same family). For a product whose promise is a summary and action items with names attached, the attribution half is the half that shows.

Improving the intermediate metric can make the product worse, and this is reported three separate times. A diarization stage that improved diarization error rate from 23.43 to 21.07 percent made transcription worse, from 12.87 to 14.13 percent (Niu 2024). Joint fine-tuning cut speaker-attributed character error by 34 to 37 percent relative while the separated audio's signal-to-noise ratio fell by roughly 19 dB (Shi 2023). Across 22 challenge systems, diarization error rate predicts transcription error well overall but the correlation collapses to about 0.16 to 0.21 among the top systems (Cornell 2025). The practical reading for this project is that a dashboard of component metrics will mislead, and only an end-to-end measurement of the three delivered outputs is worth trusting.

A summary appears to survive a transcript that would look bad quoted as a number, but the evidence is automatic scores, not readers. On the NOTSOFAR-1 scenario, systems above 50 percent tcpWER produced summaries scoring roughly on par with a system near 11 percent, and the correlation between transcription error and the best summary metric was only PCC -0.51 (Cornell 2025). This is the strongest positive signal in the vault for the charter's thesis, because it means the deliverable the customer sees may be more robust than the raw recognition quality suggests. It is also the vault's most citable-in-the-wrong-direction finding, because every one of those summary scores came from a language model judging another language model's output, no human read any summary, and the authors themselves say the metrics are unreliable. This bullet crosses two mechanism families deliberately, `Joint speaker-attributed transcription metric` and `Summarisation evaluation metric`, and the crossing is the finding: they do not move together.

Nothing published measures a summary written from recognised speech, except the one experiment above. Every meeting-summarisation benchmark number in the vault, up to 56.26 ROUGE-1 on AMI and 60.7 on ICSI, was computed on transcripts that were human-produced or human-corrected precisely so that recognition errors would not compound (Rennard 2023, correcting a contrary claim in Kumar 2022). So the field's headline summarisation scores and this project's actual product are two different measurements, and the project must never cite the first as evidence for the second. Only three English meeting corpora with summaries exist at all, together about 280 hours.

The metrics that score a summary do not separate a good one from a fluent wrong one. Across nine metrics against expert human error annotations, none correlated strongly with any error type, about a third of the metric-error pairs ignored or rewarded the error, perplexity rewarded wrong speaker attribution at +0.44, and BLEU and QuestEval rewarded hallucination severity at +0.35 and +0.34 (Kirstein 2024). Independently, a peer-reviewed survey cites nearly 30 percent of neural sequence-to-sequence summaries suffering fact fabrication (Rennard 2023). Any quality claim this project makes about its summary or its action items will have to rest on a rubric and human raters; there is no number available to buy it with.

Action-item extraction is close to unmeasured. It is a named sub-task with its own dialogue-act-based literature and its own separately annotated section in the AMI and ICSI reference summaries (Rennard 2023), but no paper in this vault reports a benchmark score for it, no error taxonomy in Kirstein 2024 scores it, and the one downstream summarisation experiment folded action items into a single 200-word summary prompt and never scored them separately (Cornell 2025). The charter names action items as one of three deliverables; the literature offers no way to say how good they are.

Nothing in this vault demonstrates minutes-after-the-meeting delivery, and the one figure that exists points the wrong way. The most efficiency-oriented system across both CHiME challenges, built specifically to be practical and using no ensembling and no diarization refinement, still ran at a real-time factor above 2 (Cornell 2025). The NOTSOFAR-1 winner reports about six hours of GPU time to decode one development set (Niu 2024). Neither is a product engineering effort and both are unoptimised research pipelines, so this is not evidence that the charter's `finished-before-the-walk-back` constraint is unreachable; it is evidence that published accuracy and published speed have never been demonstrated together, and that the project cannot cite a research word error rate and a product latency in the same breath.

Real matched training data beats architecture, and it is the lever a small team actually has. The largest single improvement reported anywhere in the vault came from fine-tuning the recognition model on real meeting audio processed the same way the deployment audio would be: 16.57 to 9.87 percent tcpWER, a 40 percent relative cut, against 8.46 to 7.50 percent from every architecture modification the same team made (Niu 2024, confirmed in Abramovski 2025). The same effect appears in diarization, 21.51 to 16.52 percent DER. For a project whose capture condition is a specific device in a specific pose, this says the defensible asset is condition-matched data, not a better model.

The 2026 frontier for one distant microphone is roughly one word in five, and it takes two models to get there. Three independent 2026 systems evaluated on the AMI Single Distant Microphone condition, one tabletop microphone in a meeting room, report attributed word error rates of 42.55 percent (Huo 2026), 23.32 percent (Dai 2026) and 21.26 percent (Li 2026). The best of them reaches that figure only by taking speaker labels and segment boundaries from a separate trained diarization system and giving them to a language model as a prior; the diarization stage alone contributes 13.72 percent diarization error. Measured inside that same system, moving from worn headset microphones to the single tabletop microphone costs 4.9 absolute points of attributed word error, 16.40 to 21.26 (Li 2026). None of the three reports latency, memory or real-time factor, so the charter's `finished-before-the-walk-back` constraint is untested against any of them. A fourth 2026 paper reports 16.3 percent on NOTSOFAR-1 single-channel (Polok 2026), which looks like a further improvement and is not: that number is computed with ground-truth diarization supplied, and the paper says so in one sentence that appears in no table caption.

Distance, measured directly, is worse than the meeting-corpus numbers suggest. The only study in the vault that puts several single-channel devices at measured distances on the same conversation reports word error rates of 36.4 percent on a table condenser at 12 to 15 cm, 44.2 percent on a tabletop unit at 2 m, 96.3 percent at 3 m and 104.2 percent at 10 m from an off-the-shelf model, falling to 20.4, 26.4, 58.2 and 64.0 after in-domain fine-tuning (Tipaksorn 2025). The language is Thai, so the levels do not transfer to English, but the shape does, and it is corroborated in English by the NOTSOFAR-1 baseline's 12.0 percent from close-talk microphones against 46.8 percent from a single device stream on the same 36 meetings (Vinnikov 2024). Two engineering results from the same study transfer without qualification because they are about acoustics rather than language: classical dereverberation and denoising front ends made recognition WORSE at every distance, and a model fine-tuned on one microphone type collapsed on every other type unless reverberation augmentation was added. For a product whose capture device and pose vary across users, the second of those is the binding constraint.

"Single-channel" in this literature is not what a phone application gets. The NOTSOFAR-1 dataset paper states plainly that its single-channel condition is a commercial conference-room device's output stream after that device's own echo cancellation, dereverberation, beamforming and noise suppression, and it flags the distinction from single-microphone audio as important enough to give its own subsection (Vinnikov 2024). Every NOTSOFAR-1 single-channel figure this project might quote, including the 22.2 percent winning score (Abramovski 2025), was therefore measured downstream of hardware signal processing. The only vault entry whose single-channel condition is one literal microphone element is LOTUSDIS (Tipaksorn 2025), and its far-field numbers are far higher. This distinction moves in the wrong direction for the charter's `phone-is-the-microphone` bet and it is invisible in every number as normally quoted.

A summary survives noise better than a transcript deserves, but only two of four models survived 25 percent, and the audio was synthetic. The threshold result the project has been carrying reads, in the authors' own words, "models tend to tolerate a noise level of about 0.2 WER (NTP between 0.07 and 0.3)" (Shapira 2025), with per-model points of 0.070, 0.195, 0.249 and 0.297. The conditional half of the claim is the better-supported half: repairing only the named entities in a damaged transcript was the single most cost-effective correction for summarisation, and a transcript at 0.9 word error rate repaired down to 0.4 by fixing content words produced summaries much preferred over a differently-damaged transcript at the same 0.4. Two things bound it. The audio was text-to-speech synthesis reverberated and noised in software, with no overlapping speech and no real microphone, and the summaries were ranked by a language model rather than read by people. The one measurement in the vault that used real meeting audio found a much weaker coupling, PCC -0.51 (Cornell 2025), while the one that used single-speaker TED talks found a much stronger one, Pearson -0.9 (Shon 2023). The three do not agree, and the axis they differ on is the audio condition.

Action items are the least measurable of the three deliverables, and the resources say why. There is no public English meeting corpus with human-annotated action items: AMI's widely cited 101 meetings and 381 action items are DERIVED by treating dialogue acts linked to the action-related section of the abstractive summary as positive, and ICSI's 18-meeting annotation is no longer publicly available (J. Liu 2023). The largest purpose-built corpus is Chinese, 424 meetings, and its inter-annotator agreement is Cohen's kappa 0.47. The best reported English positive F1 is 43.12, on human transcripts. Independently, the one ICSI annotation that is public covers 22 meetings and finds that only 318 of 21,035 turns carry an actionable item, 1.5 percent, with binary annotator agreement of kappa 0.644 (Chen 2016). Set that against kappa 0.85 to 0.89 for annotating discourse-unit boundaries on the same kind of speech (Prevot 2025): structure is decidable and importance is not, and the gap is a property of the task, not of the annotators.

Nothing has changed since 2008 about whether a meeting summary can be scored automatically, and four papers now say so. ROUGE's rank correlation with human judgement of machine-generated meeting summaries was 0.08 in 2008 (F. Liu 2008). In 2024 ROUGE-LSum correlated POSITIVELY with the presence of repetition, structure, coreference and hallucination errors on meeting summaries, meaning it rewards them, and BERTScore ran between -0.32 and +0.08 (Kirstein 2024b). Language-model judges, the modern replacement, correlate with human scores at Pearson 0.5 on QMSum and 0.11 on real Zoom meeting data, and assign perfect completeness scores to machine summaries shown to them as their own reference (Gong 2024). On written news and chat summarisation the same metrics behave far better, reaching system-level Kendall 0.84 (Y. Liu 2023), which is why the failure is specific to meetings and why borrowing a written-text evaluation result would hide it. Note also that Y. Liu 2023's "over 150 hours for 22,000 annotations" is the cost of annotating news articles and written chat logs, not meetings.

The platform vendor's on-device model has never been asked to do this job. The only evidence in the vault about what a phone can run is a roughly 3-billion-parameter model quantised to about 3.7 bits per weight, whose summarisation feature was specified, trained and evaluated for emails, messages and notifications (Apple 2024). It needed a fine-tuned adapter before it would follow a summary specification at all, its adapter training summaries were synthetic output from a larger server model, and the report names prompt injection as a live failure: the model follows instructions found inside the content it is meant to summarise, which a meeting transcript will contain by accident. The 4,096 figure this project has been carrying is the core pre-training sequence length and, separately, the pre-training batch size in sequences; the same report describes later training stages at 8,192 and 32,768 tokens, and states no runtime context cap anywhere. No latency, throughput or memory figure appears in the report.