

Keeping the Ums: Conversation-Analysis-Informed Verbatim ASR for Web Speech

Steven Coats

English, Faculty of Humanities, University of Oulu
90014 University of Oulu, Finland
steven.coats@oulu.fi

Abstract

Web video makes massive spoken-language corpora possible, but automatic speech recognition (ASR) routinely normalizes away fillers, repetitions, repairs, and hesitations. Because Conversation Analysis (CA) transcripts meticulously document these interactional phenomena, they offer an ideal yet underutilized resource for ASR adaptation. We show that fine-tuning Whisper large-v3 on just 4.47 hours of CA-annotated Australian and US conversation substantially improves verbatim transcription on heterogeneous web speech without sacrificing lexical accuracy. On a 57-clip benchmark from 39 Australian and New Zealand local-government channels, fine-tuning reduces verbatim word error rate from 16.06% to 11.39% (a 29% relative reduction) and raises filler recovery (F_1) from .117 to .819. These gains transfer directly to New Zealand speech absent from training (reducing verbatim error from 18.24% to 11.71%), demonstrating that small CA datasets teach the model a generalizable transcription policy rather than narrow acoustic memorization. We argue that targeted verbatim adaptation should become standard practice in web-as-corpus pipelines to ensure large-scale resources like CoANZSE preserve the interactional evidence needed for linguistic research.

1 Introduction

The web has made it possible to build speech corpora at a scale that was once impractical. CoANZSE Audio, for example,¹ provides searchable access to more than 195 million words from videos uploaded to YouTube by Australian and New Zealand councils and other local-government entities (Coats, 2024). Such resources can support work on regional variation, morphosyntax, phonetics, discourse, and interaction. Their linguistic value, however, depends on what the transcript represents.

¹<https://coanzse.org>

Modern ASR systems are often optimized for readable text, a design philosophy which has been characterized as emphasizing "clarity of intent" (Lea et al., 2023). Systems such as Whisper (Radford et al., 2023) often omit *uh*, *um*, repeated words, and false starts, and normalize non-standard inflectional forms, making transcripts less useful for researchers studying dialectal speech or interactional features. While disfluencies, repetitions, or filler particles carry relatively little semantic content alone, they can play a crucial role in discourse. From an interactional perspective, filled pauses can signal production delay and upcoming complexity (Clark and Fox Tree, 2002), prompting listeners to anticipate new information (Arnold et al., 2003), or can be associated with dispreferred responses (Kendrick and Torreira, 2015) such as disagreement or rejection. From a sociolinguistic perspective, these items can exhibit systematic regional variation (Wieling et al., 2016) and evolve over time in the same manner as content words (Fruehwald, 2016; Tottie, 2011). For speech corpora, transcriptions which preserve filler particles, repetitions, and repairs (i.e. verbatim transcriptions) can therefore provide important evidence for the analysis of discourse content, interaction, and language variation and change. At web-corpus scale, transcription normalization can hinder dialectological, sociolinguistic, and interactional research, necessitating the development of models specifically designed to recover these features.

Recent work has addressed verbatim ASR directly. As might be expected, fine-tuning on in-domain transcripts in which fillers are carefully preserved improves recovery of these items (Hori et al., 2022; Lin et al., 2025), but many gold standard transcripts used as training data do not carefully preserve such features, and manual annotation is time-consuming and expensive. The importance of domain adaptation as part of the corpus-design pipeline is also increasingly recognized in corpus

linguistics and research on varieties of English. Fine-tuning an ASR model with conversational Singapore English speech, for example, makes locally salient forms such as *lah*, *lor*, *sia*, and *kena* available for corpus analysis, whereas they are largely absent from the output of default ASR (Coats et al., forthcoming). Re-engineering parts of the ASR system such as the tokenizer and the procedures used to derive timing information can improve verbatim performance (Zusag et al., 2024). More extensive interventions allow controllable decoding: by prepending a prompt token (e.g., <verbatim> or <intended>) to the decoder, a single model can be instructed either to transcribe literal hesitations and repairs or to filter them out for readable text (Wagner et al., 2026). From the practical perspective of corpus linguists, however, finding an appropriate balance between training data selection, manual annotation, and system redesign is crucial. The practical question for a web corpus of national varieties of English remains open: can a small amount of existing, carefully transcribed conversation improve verbatim ASR on heterogeneous Australian and New Zealand web video without sacrificing lexical accuracy?

We address this question by adapting conversation-analytic and related interactionally detailed data available for research use. Conversation Analysis (CA) is an approach to recordings of talk-in-interaction in which detailed transcripts are used to analyze how participants organize social action through turn-taking, sequence organization, repair, overlap, and timing, typically through qualitative sequential analysis. These transcripts preserve precisely the hesitations, repetitions, cut-offs, and other features that general-purpose ASR frequently normalizes away, making them a potentially valuable but underused source of supervision for corpus construction.

We contribute (i) a context-assisted manual benchmark from 100 clips in a candidate source-recording pool for a future CoANZSE release, with a prespecified 57-clip quality-controlled primary subset; (ii) an evaluation of two Whisper large-v3 models adapted on this CA data, namely a verbatim-only full fine-tuned model and a controllable dual-policy model using Low-Rank Adaptation (LoRA; Hu et al., 2022), benchmarked against base Whisper and CrisperWhisper2 for lexical accuracy and the preservation of interactionally relevant phenomena; and (iii) evidence that only 4.47 hours of detailed conversational supervision transfers across

recording domains and to New Zealand speech absent from training. Supplementary source-corpus experiments examine adaptation of CrisperWhisper2 and the effect of adding a third Australian corpus. Code, evaluation manifests, sanitized predictions, and scoring scripts are available in an anonymous repository.²

2 Data and method

2.1 Interactionally detailed training data

The starting point of our experiments comprised two corpora with interactionally detailed transcripts and the corresponding audio: the La Trobe Corpus of Spoken Australian English (LTCSAusE; Mullan, 2002) and the Santa Barbara Corpus of Spoken American English (SBCSAE; Du Bois et al., 2000–2005). LTCSAusE comprises conversational interaction recorded in Melbourne; SBCSAE includes diverse US conversational interactions. Both corpora preserve substantial conversational detail, including precise timing information and annotations for overlapping speech, pauses, pitch and prosody, hesitations and disfluencies, repetitions, and filler particles.

From the source transcripts, we normalized corpus-specific notation into a common lexical convention for fillers (e.g. *uh/er*, *um/erm*, *mm*), fragments (e.g. *d-didn't*), repetitions, and non-speech events (e.g. music and noise). Training examples were cut to exact acoustic boundaries and filtered to avoid substantial overlap and mixed-speaker target segments. The resulting split contains 798 training clips (4.47 h), 167 validation clips (1.10 h), and 210 held-out clips (1.24 h). Only the training and validation splits were used during model development.

For an additional data-ablation experiment, we prepared the Griffith Corpus of Spoken Australian English (GCSAusE; Haugh and Chang, 2013), a registered-access TalkBank corpus with CA-style transcripts and 40 recordings. The same normalization and boundary criteria yielded 192 non-overlapping, single-speaker clips (0.96 h): 124 training, 30 validation, and 38 test clips. Adding these items produced a three-source dataset with 922 training, 197 validation, and 248 test clips. Recording-level splits prevented material from the same source recording from crossing partitions.

Although limited in size by the amount of suitable data available, this dataset provides consistent

²<https://github.com/stcoats/CAWhisperWac13>

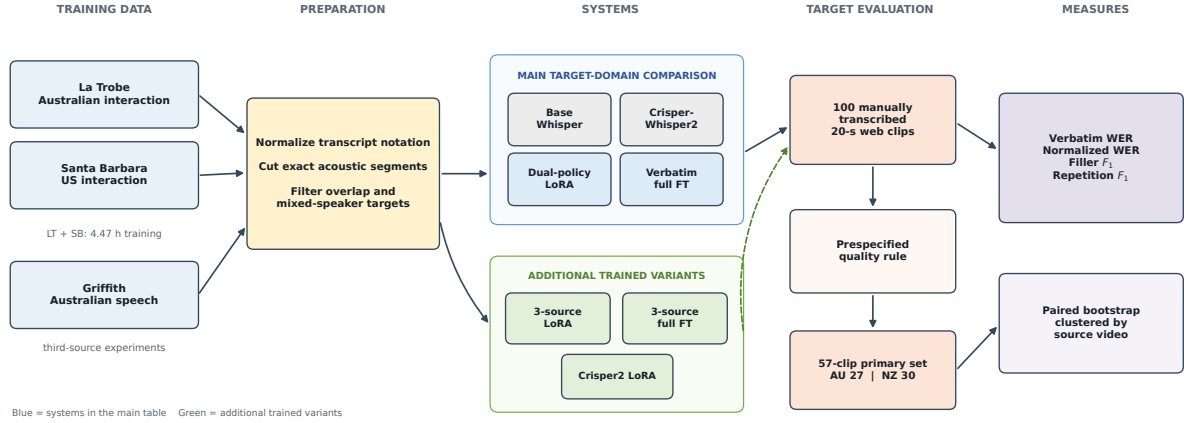


Figure 1: Experimental workflow for verbatim ASR adaptation and evaluation on candidate CoANZSE web speech.

positive supervision for features commonly omitted by general-purpose ASR and weak web captions. While CA transcripts also document fine-grained prosody, overlap, and temporal latching, the present study focuses specifically on recovering lexicalized interactional phenomena (fillers, repetitions, and repairs) for searchable text corpora, leaving acoustic timing and overlap modeling to future work. We use it to test whether such supervision transfers to heterogeneous Australian and New Zealand web speech.

2.2 Models

Verbatim full FT. We updated the complete encoder–decoder model of Whisper large-v3 against the verbatim target. Training used bfloat16 on four NVIDIA GH200 GPUs, an effective batch size of 16, a learning rate of 10^{-5} , 10% warmup, at most ten epochs, and early stopping with patience three. The selected checkpoint minimized verbatim validation WER.

Dual-policy LoRA. Adapting one of the techniques described in (Wagner et al., 2026), we used the CA-transcript-derived text as the verbatim target and created a corresponding intended-output version by deterministically removing fillers, repetitions, and other disfluencies (for example, transforming verbatim “*we uh we had to d-decide*” into intended “*we had to decide*”). The model was trained on both versions, with each audio–transcript pair marked by a new decoder-prefix token: either <verbatim> or <intended>. Training otherwise followed a standard PEFT/LoRA procedure. Adapters were applied throughout the encoder and decoder attention layers (rank 32, alpha 64, dropout .05). Because the original Whisper parameters re-

mained frozen, the embeddings for the two new policy tokens were trained separately alongside the LoRA adapters.

This procedure differed from the staged method of Wagner et al. (2026), who first trained only newly introduced token embeddings and then unfroze the decoder while retaining a frozen encoder. Because our adapters operated in both the encoder and decoder, they could adapt both output policy and acoustic/domain representations; the experiment did not isolate these two effects. We evaluated only the verbatim mode in the main target-domain comparison.

Supplementary adaptation experiments. In addition to evaluating out-of-the-box CrisperWhisper2 as a baseline in our primary benchmark, we applied the same LoRA configuration directly to the CrisperWhisper2 large model in a supplementary experiment, using the La Trobe and Santa Barbara verbatim targets and CrisperWhisper2’s verbatim decoder prefix. As in our dual-policy model, rank-32 adapters were applied to the query and value projections throughout both the encoder and decoder; the original CrisperWhisper2 parameters remained frozen. This experiment tested whether the specialist model would benefit from the same conversational supervision. Finally, we repeated verbatim full fine-tuning of Whisper large-v3 after adding GCSAusE. This three-source model used the same full-training hyperparameters and selected a checkpoint by the unweighted validation-WER macro across the three source corpora. It stopped after five epochs and restored the epoch-two checkpoint.

In the primary target-domain evaluation, we

compared both principal adapted systems (verbatim full FT and dual-policy LoRA) against unmodified Whisper large-v3 and out-of-the-box CrisperWhisper2. Each system received the identical mono 16-kHz target waveform and no textual context; decoding used standard greedy search (temperature 0).

2.3 Independent web-speech evaluation

The evaluation audio comes from original source recordings re-collected for a candidate future CoANZSE version; it is not yet part of the released CoANZSE v1.0 index. While published by local-government channels, the source recordings in CoANZSE span diverse communicative genres (Coats, 2024): beyond council meeting debates, they include public consultations, citizen interviews, community event presentations, and informational videos, capturing both formal deliberations and spontaneous conversational interaction. We selected 50 source videos from 50 local-government channels, 25 from Australia and 25 from New Zealand. From each video we took one 20-second clip whose YouTube caption contained a filler marker and one comparison clip without such a marker. YouTube captions were used only to stratify candidates, not as evaluation references.

Because conversational fillers and repair particles are frequently missed in high-volume crowd-sourced annotation, we prioritized expert judgment over large manual transcript volumes. A single linguist (the author, a native English speaker trained in corpus linguistics) transcribed all 100 clips with access to extended acoustic context and video metadata. To resolve local names and low-context fragments, annotation was context assisted: the annotator could listen to a longer window and inspect the video title, council and location, and neighboring YouTube captions. The target transcript nevertheless describes only the exact 20-second clip. Annotation preceded evaluation and did not display system predictions.

Before inference, we froze a primary quality rule: intact boundaries, good or usable audio, and no more than minor overlap. Fifty-seven clips from 39 videos met the rule (3,188 scored words; 19 minutes): 27 Australian and 30 New Zealand, with 27 filler-marked and 30 unmarked candidates. We report all 100 clips as a sensitivity analysis.

2.4 Metrics and uncertainty

Verbatim WER (vWER) retains audible fillers, repetitions, fragments, contractions, and reduced forms. Normalized WER removes fillers and fragments, collapses adjacent repetitions, and normalizes selected reduced/full forms. Both ignore punctuation and case and normalize for inaudible orthographic choices, including Australian/UK versus US spelling, titles, initialism formatting, and macrons (common in some Māori borrowings). Reference [unclear] tokens mask one uncertain unit rather than forcing a possibly incorrect gold word.

We additionally report F_1 for fillers and adjacent repetitions. Confidence intervals use 20,000 paired bootstrap replicates clustered by source video, preserving dependence between the two sampled clips. The clean references contain 129 fillers, 25 adjacent repetitions, and six word fragments. Fragment results are consequently too sparse for a substantive comparison and are not included in the main table. Reporting both WER variants is essential: vWER measures fidelity to conversational form, whereas normalized WER asks whether a system still recognizes lexical content after policy-related differences are removed.

3 Results

3.1 Target-domain results

Table 1 reports the results on the quality-controlled CoANZSE candidate subset. All three verbatim-oriented systems produced lower vWER point estimates than base Whisper. The full verbatim model performed best numerically, reducing vWER from 16.06% to 11.39%: an improvement of 4.67 percentage points, or 29.1% relative. The paired difference from base had a 95% interval of -6.32 to -2.99 points. The full model was also the only system for which the normalized-WER improvement over base had a paired interval wholly below zero (-1.92 points, CI -3.50 to -0.39).

The filler and repetition measures reveal a somewhat different pattern. CrisperWhisper2 achieved the highest filler F_1 (.859), narrowly ahead of the dual-policy LoRA (.843) and full model (.819), while the dual-policy LoRA achieved the highest repetition F_1 (.800). Reporting filler precision and recall alongside F_1 shows that CrisperWhisper2’s slight advantage is driven primarily by precision (.843 vs .782 for full FT), whereas recall remains comparable across all three adapted models (.853–

System	vWER ↓	95% CI	norm. WER ↓	filler P / R	filler F_1 ↑	repetition F_1 ↑
Whisper Large-v3	16.06	13.41–18.80	12.11	1.0 / .062	.117	.375
CrisperWhisper2	13.11	9.82–16.88	12.44	.843 / .876	.859	.724
Dual-policy LoRA	12.17	9.78–14.82	11.07	.833 / .853	.843	.800
Verbatim full FT	11.39	9.23–13.78	10.19	.782 / .860	.819	.710

Table 1: Results on the prespecified quality-controlled CoANZSE candidate subset (57 clips). WER values are percentages. vWER intervals are 95% percentile intervals from 20,000 paired bootstrap replicates clustered by source video.

System	Australia	New Zealand
Whisper Large-v3	13.71	18.24
CrisperWhisper2	11.23	14.86
Dual-policy LoRA	11.88	12.44
Verbatim full FT	11.03	11.71

Table 2: Descriptive vWER (%) by country on the clean subset (Australia: 27 clips; New Zealand: 30).

.876). Base Whisper had filler precision of 1.0 but recall of only .062: its eight predicted fillers were correct, but it recovered only 8 of the 129 fillers in the reference transcripts. The three verbatim-oriented systems recovered substantially more of these conversational events.

The paired contrasts nevertheless do not identify a universal winner. The full model’s vWER was 0.78 points lower than that of the dual-policy LoRA (95% CI -1.77 to 0.07) and 1.73 points lower than that of CrisperWhisper2 (CI -4.11 to 0.25), but neither interval excluded zero. Its normalized WER was lower than that of the dual-policy LoRA by 0.88 points (CI 0.03 – 1.88) and lower than that of CrisperWhisper2 by 2.24 points (CI 0.03 – 4.96). Thus, strong recovery of conversational events did not necessarily coincide with the most accurate lexical transcript.

The pattern was not confined to Australian speech. On the 30 New Zealand clips, vWER fell from 18.24% for base Whisper to 11.71% for the full model and 12.44% for the dual-policy LoRA, although neither adaptation corpus contained New Zealand speech. These subgroup results are descriptive because of the small sample.

A typical sequence of omissions by base Whisper is shown below (irrelevant words elided). The full model recovered the repeated articles and all three filler positions, although it rendered the first *uh* as *um*.

Reference: “the ~~the~~ council’s not ... seek to ~~uh~~ replace ... agencies ~~um~~ in that regard ~~uh~~ ... the ~~the~~ proponent ...”

Base Whisper: “the council is not ... seek to re-

place ... agencies in that regard ... the proponent ...”

Full FT: “the ~~the~~ council’s not ... seek to ~~um~~ replace ... agencies ~~um~~ in that regard ~~uh~~ ... the ~~the~~ proponent ...”

A sensitivity analysis using all 100 annotated clips preserved the vWER ranking: 19.55% for base Whisper, 16.25% for CrisperWhisper2, 14.07% for the dual-policy LoRA, and 13.84% for the full model.

The same ranking appeared in both sampling strata. For base Whisper, CrisperWhisper2, the dual-policy LoRA, and the full model respectively, vWER was 15.05%, 12.17%, 11.19%, and 10.93% on the 27 filler-marked clips, compared with 16.99%, 13.98%, 13.07%, and 11.81% on the 30 nominally unmarked clips. Audible fillers occurred in both strata: YouTube captions were useful for retrieving likely examples, but the absence of a captioned filler was not a reliable negative label.

3.2 Supplementary source-corpus experiments

Table 3 compares the two- and three-source full models on all three held-out source-corpus partitions. Adding 0.64 hours of GCSAusE training audio produced small and mixed changes. It lowered vWER on La Trobe from 14.84% to 14.41% and on GCSAusE from 30.29% to 28.87%, while SBCSAE vWER increased from 9.97% to 10.20%. The unweighted three-corpus macro fell from 18.37% to 17.83%. GCSAusE repetition F_1 increased from .400 to .545, but filler F_1 changed only from .408 to .426 and declined slightly on the other two corpora. Additional data therefore produced modest redistribution rather than a uniform improvement.

Adapting CrisperWhisper2 with LoRA similarly improved lexical accuracy without improving every conversational-event measure. On held-out La Trobe, vWER fell from 20.29% to 17.07%, while filler F_1 fell from .757 to .599. On Santa Barbara, vWER fell from 14.11% to 10.65%, while filler F_1

Held-out corpus	System	vWER ↓	filler F_1 ↑	repetition F_1 ↑
La Trobe	Two-source full FT	14.84	.800	.709
	Three-source full FT	14.41	.793	.719
Santa Barbara	Two-source full FT	9.97	.618	.695
	Three-source full FT	10.20	.594	.703
Griffith	Two-source full FT	30.29	.408	.400
	Three-source full FT	28.87	.426	.545

Table 3: Held-out source-corpus results for the data-addition experiment. The two-source model used La Trobe and Santa Barbara; the three-source model additionally used GCSAusE. Values are descriptive point estimates.

fell from .494 to .430. Thus, applying the same adaptation data and method to a specialist verbatim model did not simply amplify its original strengths.

Finally, as a policy-control check, we decoded the held-out La Trobe and Santa Barbara speech using both modes of the dual-policy model. Relative to verbatim mode, intended-mode decoding reduced normalized WER from 19.60% to 13.62% on La Trobe and from 11.65% to 9.26% on Santa Barbara. This confirmed that the prefix tokens affected the model’s output policy. The intended references were derived from the verbatim targets rather than independently annotated; this comparison therefore served as a manipulation check, not as an evaluation of intended-transcription quality. Furthermore, decoding the primary 57-clip CoANZSE test set in <intended> mode effectively suppressed reference fillers ($F_1 = .045$, predicting only three reference fillers) while achieving a normalized WER of 10.94%, statistically indistinguishable from verbatim mode (11.10%; 95% CI $[-1.13, +0.70]$), confirming policy controllability on the target domain without a lexical accuracy penalty on remaining words.

4 Discussion

The findings support a data-centric approach to web speech. Web captions are useful for discovery and rough lexical access, but they are not neutral representations of interaction. In our nominally “unmarked” comparison stratum, manual listening found genuine fillers that YouTube captions had omitted; the verbatim systems correctly recovered many of them. Caption absence should therefore not be treated as a negative label.

This is not a conventional same-domain adaptation setting. Supervision comes from research recordings of naturally occurring Australian and US interaction, whereas evaluation uses heterogeneous institutional web video; more than half of the

clean test clips are New Zealand speech. The particularly large New Zealand reduction suggests that detailed supervision can transfer a transcription policy without requiring a matched corpus for every target variety. Accent adaptation and policy adaptation remain entangled in our encoder–decoder training, however, so the country split should be interpreted as transfer evidence rather than an accent-specific causal result.

The strongest production choice depends on purpose. CrisperWhisper2 is best if filler recovery alone dominates. The dual LoRA provides explicit verbatim/ intended control and the best repetition F_1 in this sample. The full model is the best candidate for rebuilding a searchable corpus because it combines high event recall with the lowest verbatim and normalized WER. In a corpus such as CoANZSE, that balance matters: hallucinated or substituted content damages lexical searches, while deleted disfluencies make interactional searches impossible.

For corpus construction, these errors have different downstream costs. A content substitution creates false positives and false negatives in ordinary lexical searches; deletion of a genuine filler or repetition removes the event from the corpus altogether and prevents it from entering an interactional sample. A production pipeline should therefore retain both a verbatim layer and a normalized search view rather than forcing one transcript to serve both purposes. The present scoring framework operationalizes that distinction even when the deployed model emits only verbatim text.

No single metric captures these requirements. Normalized WER makes base Whisper appear closer to the specialist systems by removing many systematic omissions before scoring; filler F_1 favors CrisperWhisper2 while concealing its greater lexical error. The paired results favor the full model as the current production candidate, but its modest vWER advantage over the dual LoRA is not

conclusive. CrisperWhisper2 achieved the highest filler F1 in our evaluation and its remit addresses a considerably broader problem than the present study, including controllable transcription policy, paralinguistic events, and word-level timing in multilingual data. Our experiment instead considered a narrower practical scenario: whether existing detailed conversational transcripts can improve the transcription of a particular regional web corpus using a comparatively conventional adaptation pipeline. The lower lexical accuracy of CrisperWhisper2 on CoANZSE may reflect differences between its training data and the Australian and New Zealand target domain (CrisperWhisper2 was trained in part on CORAAL, comprising African-American speech), but the present evaluation cannot isolate the contribution of accent, recording conditions, model configuration, or individual training corpora. The systems should therefore be understood as emphasizing different, potentially complementary forms of adaptation.

The supplementary experiments reinforce this interpretation. LoRA adaptation of CrisperWhisper2 reduced source-corpus WER but also reduced filler F_1 , while adding the small Griffith dataset to full fine-tuning improved some corpus-metric combinations and slightly worsened others. Fine-tuning did not operate as a simple accumulation of improvements: the composition and transcription conventions of the supervision affected which properties were retained. This is particularly relevant when the intended output is a multi-purpose linguistic corpus rather than a single benchmark score.

With regards to an update of the CoANZSE resource, a potential v2 could preserve verbatim ASR as its archival annotation while deriving a normalized token layer for ordinary concordance queries. Users could then search for lexical content or interactional forms and inspect the same aligned audio. Unlike a single “clean” transcript, this design would not irreversibly remove conversational events; the transcription policy could be made explicit in the corpus documentation.

5 Conclusion

Base Whisper’s readable output systematically removes conversational evidence needed for some kinds of linguistic research. A few hours of carefully segmented conversation-analysis data were sufficient to reduce verbatim error on independent

Australian and New Zealand web speech while retaining strong lexical accuracy. For web speech corpora, transcription policy should be an explicit design and evaluation decision, not an undocumented side effect of the ASR model.

Limitations

Our evaluation focuses on a targeted, context-assisted benchmark of 100 annotated clips across 50 municipal channels (analyzed as a frozen 57-clip primary subset and a 100-clip sensitivity set). Reference transcripts were produced by a single annotator without inter-annotator agreement measurement, and results are situated within Australian, New Zealand, and US varieties of English. While event sparsity for rare categories like word fragments ($N = 6$) precludes firm claims for those specific subtypes, paired cluster-bootstrap resampling across 20,000 replicates demonstrates that the main reductions in verbatim WER and filler recovery are statistically robust across diverse source channels. Original web recordings can be stereo, whereas all systems were compared on a common mono 16-kHz conversion. The intended targets used by the dual model are deterministic transformations rather than independent human transcripts. Our encoder-decoder LoRA does not isolate output-policy learning from acoustic adaptation; a decoder-only or token-embedding-only ablation would do so more directly. We did not evaluate diarization or word timing, so our conclusions concern transcription only.

Ethical considerations

The benchmark is derived from publicly available institutional web video, but public availability does not eliminate privacy, copyright, platform, or contextual-integrity concerns. The audio was not originally recorded under a research-consent protocol. We therefore plan to release code, scoring tests, source identifiers, time offsets, and aggregate results, while redistributing audio, reference text, or trained weights only where source-corpus licenses and web-platform conditions permit. The benchmark material comes from a candidate future CoANZSE pool and should not be represented as part of the current public CoANZSE v1.0 index.

Acknowledgments

This work was supported by the European Union NextGenerationEU instrument and was funded by

the Research Council of Finland under grant number 358720. Computational resources were provided by Finland’s Centre for Scientific Computing (CSC).

References

- Jennifer E. Arnold, Maria Fagnano, and Michael K. Tanenhaus. 2003. [Disfluencies signal thee, um, new information](#). *Journal of Psycholinguistic Research*, 32(1):25–36.
- Herbert H. Clark and Jean E. Fox Tree. 2002. [Using uh and um in spontaneous speaking](#). *Cognition*, 84(1):73–111.
- Steven Coats. 2024. [Building a searchable online corpus of Australian and New Zealand aligned speech](#). *Australian Journal of Linguistics*, 44(2–3):261–277.
- Steven Coats, Carmelo Alessandro Basile, Cameron Morin, and Robert Fuchs. forthcoming. Fine-tuning ASR for corpus linguistics: Singapore English. *Digital Scholarship in the Humanities*.
- John W. Du Bois, Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson, and Nii Martey. 2000–2005. [Santa Barbara Corpus of Spoken American English, Parts 1–4](#).
- Josef Fruehwald. 2016. Filled pause choice as a sociolinguistic variable. *University of Pennsylvania Working Papers in Linguistics*, 22(2):41–49.
- Michael Haugh and Wei-Lin Melody Chang. 2013. Collaborative creation of spoken language corpora. In Tim Greer, Donna Tatsuki, and Carsten Roever, editors, *Pragmatics and Language Learning*, volume 13, pages 133–159. National Foreign Language Resource Center, University of Hawaii at Manoa, Honolulu, HI.
- Koharu Horii, Meiko Fukuda, Kengo Ohta, Ryota Nishimura, Atsunori Ogawa, and Norihide Kitaoka. 2022. [End-to-End Spontaneous Speech Recognition Using Disfluency Labeling](#). In *Interspeech 2022*, pages 4108–4112.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Kobin H. Kendrick and Francisco Torreira. 2015. [The timing and construction of preference: A quantitative study](#). *Discourse Processes*, 52(4):255–289.
- Colin Lea, Zifang Huang, Jaya Narain, Lauren Tooley, Dianna Yee, Dung Tien Tran, Panayiotis Georgiou, Jeffrey P Bigham, and Leah Findlater. 2023. [From user perceptions to technical improvement: Enabling people who stutter to better use speech recognition](#). In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI ’23, New York, NY, USA. Association for Computing Machinery.
- Jhen-Ke Lin, Hao-Chien Lu, Chung-Chun Wang, Hong-Yun Lin, and Berlin Chen. 2025. [Acoustically Precise Hesitation Tagging Is Essential for End-to-End Verbatim Transcription Systems](#). In *10th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 163–166.
- Kerry Mullan. 2002. [The La Trobe corpus of spoken Australian English](#).
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. [Robust speech recognition via large-scale weak supervision](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.
- Gunnel Tottie. 2011. [Uh and um as sociolinguistic markers in British English](#). *International Journal of Corpus Linguistics*, 16(2):173–197.
- Laurin Wagner, Mario Zúñiga, and Bernhard Thallinger. 2026. [Transcription policy as a latent variable: Activating controllable verbatim ASR with word-level timing](#). *Preprint*, arXiv:2607.18934. Accepted at Interspeech 2026.
- Martijn Wieling, Jack Grieve, Gosse Bouma, Josef Fruehwald, John Coleman, and Mark Liberman. 2016. [Variation and change in the use of hesitation markers in Germanic languages](#). *Language Dynamics and Change*, 6(2):199–234.
- Mario Zúñiga, Laurin Wagner, and Bernhard Thallinger. 2024. [CrisperWhisper: Accurate timestamps on verbatim speech transcriptions](#). In *Proceedings of Interspeech 2024*.