

Gold-Preserving Synthetic Augmentation for Verbatim Swedish Automatic Speech Recognition

Giuseppe Chiriatti¹, Dimitrios Kokkinakis², Flavio Bertini¹

Ricardo Muñoz Sánchez² and Herbert Lange²

¹University of Parma, Italy

{giuseppe.chiriatti@studenti.unipr.it}, {flavio.bertini@unipr.it}

²Språkbanken, University of Gothenburg, Sweden

{dimitrios.kokkinakis, herbert.lange, ricardo.munoz.sanchez}@svenska.gu.se

Abstract

Verbatim transcripts, which keep the hesitations and interactional particles that ordinary automatic speech recognition (ASR) removes, are a prerequisite for speech-based cognitive assessment, and they are hardest to obtain for languages with little annotated speech. We study this problem for Swedish with a 1.02-hour corpus of manually transcribed spontaneous speech, too small for full fine-tuning of KB-Whisper-large. We adapt the model with a low-rank adapter and compare three ways of extending the training data: pseudo-labeling of untranscribed in-domain audio, unguided synthesis with a language model and a text-to-speech system, and a *gold-preserving* pipeline that reproduces the disfluency distribution of the gold corpus by construction and can be audited before training. Fine-tuning reduces disfluency-segment word error rate (WER) by 32% relative, with blockwise-bootstrap confidence intervals disjoint from the zero-shot baseline, while the three augmentation strategies are not separable by WER across five seeds. The gold-preserving pipeline needs no untranscribed audio, and its disfluency distribution can be verified in advance.

1 Introduction

Speech carries information that its words alone do not. The density and distribution of hesitations, repetitions, and reformulations are documented markers of cognitive decline and are among the earliest available signals for neurodegenerative conditions (Jia et al., 2025; Liu et al., 2024b). Automatic analysis of this kind, therefore, requires *verbatim* transcripts that retain disfluencies.

State-of-the-art automatic speech recognition (ASR) is built to discard this material. Whisper (Radford et al., 2023) inherits a fluent implicit language model from web-scale normalized transcripts and suppresses filler tokens at decoding time, so that what it returns is a tidied-up version

of what it heard. The behaviour is learned rather than architectural: fine-tuning on verbatim data reverses it (Akinrintoye et al., 2025). That correction is hardest to apply where it is most needed. Languages with little annotated speech are also the ones for which verbatim material is scarcest, and Swedish is a case in point. Its disfluency inventory (Eklund, 2004; Nilsson Björkenstam et al., 2013) comprises filled pauses (*öh, ee, mm*), discourse markers (*alltså, liksom, typ*), modal particles (*ju, va, nä*) and backchannels (*ja, jo*), several of them positionally constrained and without direct English counterparts. KB-Whisper (Vesterbacka et al., 2025), the Swedish adaptation trained on 32,500 hours, deepens the suppression. About 60% of its first-stage data is parliamentary speech, and its second-stage quality filter discards verbatim segments, because the reference transcripts it scores them against contain no disfluencies. The gold corpus available to us, 30 recordings and 1.02 hours, lies far below the three hours under which full fine-tuning of Whisper models above 244M parameters degrades (Liu et al., 2024a), so the training set has to grow without new manual annotation.

In this paper we study to what extent that missing material can be manufactured rather than collected, a direction we are also exploring for neuropsychological research data (Kokkinakis et al., 2026). We fine-tune KB-Whisper-large with a low-rank adapter on the gold corpus and compare three ways of extending it: pseudo-labeling of in-domain untranscribed audio (Li et al., 2024; Wang et al., 2025), unguided synthesis with a language model and a text-to-speech (TTS) system (Mizumoto et al., 2025), and a *gold-preserving* pipeline, which is our proposal. The latter profiles the gold transcripts token by token and in syntactic context, injects disfluencies into generated text under morphosyntactic constraints, and synthesizes and re-aligns the audio, so that the distribution of the resulting corpus can be audited against the gold pro-

file before training. Controlling that distribution matters here because a teacher that still suppresses *öh* and *alltså* produces equally sanitized pseudo-labels, propagating the bias it is meant to correct.

The research question is whether such a distribution-controlled pipeline can match pseudo-labeling without requiring any untranscribed audio. Our contributions are:

- a gold-preserving augmentation pipeline for disfluent speech, combining a per-token gold profile, a part-of-speech (POS)-aware Planner–Generator injector with a morphosyntactic constraint for Swedish post-verbal particles, voice-design TTS and forced alignment, whose output distribution is verifiable before training;
- evidence that in this regime the three strategies are not separable by word error rate (WER), so that the choice among them is decided by the data each one requires rather than by accuracy;
- a negative transfer result: the 2:1 synthetic-to-real ratio recommended by Mizumoto et al. (2025) does not carry over to verbatim low-resource fine-tuning.

2 Corpora

Gold corpus. The dataset comprises 30 mono 16 kHz recordings, one per speaker, of “Trip to Stockholm”, an adaptation of the “Trip to New York” planning task (Fleming and Harris, 2008) administered to older Swedish speakers. Each recording was manually transcribed verbatim by trained transcribers, which is what makes it our *gold* reference. Disfluencies are written in their surface form with no special tags, thereby preserving their lexical identity and interactional function, consistent with evidence that acoustically precise verbatim annotation outperforms generic tags such as <hesitation> by 23.6% relative (Lin et al., 2025). After forced alignment and 0.5–25 s segmentation, the corpus counts 1,248 segments (mean 2.94 s). Ten of the 16 target types occur in this corpus. Their distribution is heavily skewed: *ju* (114 occurrences) and *ja* (100) dominate, followed by *ee* and *alltså* (10 each), *jo* (9), *va* (6), *liksom* (5), *eh* (4), *typ* (3) and *nä* (1); the remaining six forms (*mm*, *hmm*, *hm*, *öh*, *öhm*, *ehm*) do not occur in this narrative register.

Silver corpus. A further 30 recordings of spontaneous Swedish speech from a comparable acous-

tic and demographic domain, without transcripts: never ground truth, only material to be transcribed automatically, hence *silver*.

3 The Proposed Method

All variants start from the same 1.02-hour gold corpus and the same base model, and differ only in what is added to the training data. We describe the generator that is our contribution, then the two alternative extensions used for comparison, and finally the shared adaptation and evaluation setup.

3.1 Gold-preserving synthetic pipeline

The pipeline generates text whose disfluency profile matches the gold corpus, then synthesizes and aligns audio. (1) *Profiling*: per-token densities (occurrences per 100 words) and POS contexts are extracted from the gold transcripts, together with high-filler-density fragments used as few-shot examples. (2) *Base text*: a Swedish-instruction-tuned Llama-3-8B (4-bit, local inference for privacy and exact reproducibility) generates spontaneous-register narratives, with the gold profile and authentic fragments injected into the system prompt and explicit bans on formal connectives absent from the gold. (3) *Planner–Generator injection*: a Planner selects insertion sites subject to positional constraints, notably a V2 *lookback* rule requiring a finite verb or auxiliary within the three preceding tokens for the post-verbal particle *ju*; a Generator selects the token form compatible with the local POS context. A *global-proportional floor* guarantees that rare gold tokens receive nonzero targets, and a chi-square verifier checks the resulting distribution against the gold profile. (4) *TTS*: OmniVoice (Zhu et al., 2026) in voice-design mode with four rotating speaker profiles; disfluency tokens are mapped to nonverbal symbols (e.g., [sigh], [thinking]) where literal pronunciation would be unnatural, a mapping selected by UTMOS (Saeki et al., 2022) A/B comparison. (5) *Alignment*: WhisperX (Bain et al., 2023) with a Swedish wav2vec2 connectionist temporal classification (CTC) model (Kürzinger et al., 2020). This step matters. Dynamic time warping (DTW) alignment via cross-attention lost 72% of *ja* occurrences (ratio 0.29), because Whisper attention does not localize short low-energy tokens; CTC alignment recovers it to 0.96. (6) *Filtering*: a generator–verifier check (Perrin and Boullianne, 2025) transcribes each synthetic file with the zero-shot baseline and quarantines files above

25% WER. The resulting set, **syn-gp**, counts 647 segments, 0.59 h (0.58:1 ratio to gold).

3.2 Comparison corpora

In the pseudo-labeling variant, the model fine-tuned on gold data alone, which we call the gold model, transcribes the silver set; its output is filtered by average token log-probability (threshold -0.8), by regex-based hallucination detection and by duration. At most 50% of each training batch is pseudo-labeled, following Li et al. (2024), and self-training is capped at the rounds where it remains beneficial (Wang et al., 2025); after filtering, the silver set contributes 3,225 segments (2.93 h). In the unguided variant, the same language model and TTS system are used without profiling, injection or distributional check.

3.3 Adaptation and evaluation

We fine-tune KB-Whisper-large with low-rank adaptation (LoRA) (Hu et al., 2022) at rank $r=32$, $\alpha=8$, with rank-stabilized scaling (Kala-jdziewski, 2023), targeting the four attention projections (~ 31.4 M trainable parameters, 2% of the model; feed-forward network (FFN) modules exceeded the 12 GB VRAM budget). Learning rate is 7×10^{-5} following Lin et al. (2025), effective batch 16, up to 6 epochs with early stopping on validation WER. Evaluation uses beam size 5, forced Swedish transcription, an emptied suppress_tokens list, and no prompt or logit biasing. All WER computation uses Whisper’s multilingual BasicTextNormalizer: the English normalizer removes filler words from both hypothesis and reference and inflated baseline WER by 7–8 points in preliminary experiments, a metric artifact that must be fixed before any modeling conclusion.

4 Experimental setup

Seven configurations share the same base model, LoRA recipe, and gold data, differing only in the added training data: **Base** (zero-shot), **FT-gold** (gold only), **FT-pseudo** (+pseudo-labeled silver), **FT-syn-llm** (+ the unguided language-model set syn-llm), **FT-syn-pg** (+ the Planner–Generator set syn-pg, DTW-aligned), **FT-syn-gp** (+ the gold-preserving set syn-gp, WhisperX-aligned), and **FT-syn-2:1** (syn-gp scaled to 2.40 h / 2.35:1 ratio, syn-2:1). Every FT-syn-* variant is trained on the gold set plus the synthetic set of the same name. Each is trained and evaluated on five independent seeds,

Table 1: WER over 5 seeds (mean $\pm$ std, beam 5, BasicTextNormalizer). Boldface indicates the best result in each column; underlining indicates the second-best result.

Model	WER tot. (%)	WER disfl. (%)
Base	27.58 $\pm$ 2.90	33.28 $\pm$ 4.87
FT-gold	19.83 $\pm$ 4.59	22.52 $\pm$ 4.49
FT-pseudo	19.64 $\pm$ 4.31	22.10 $\pm$ 4.73
FT-syn-llm	19.84 $\pm$ 4.76	22.60 $\pm$ 4.83
FT-syn-pg	19.88 $\pm$ 4.56	21.75 $\pm$ 4.47
FT-syn-gp	19.79 $\pm$ 4.63	22.08 $\pm$ 4.10
FT-syn-2:1	<u>20.32 $\pm$ 4.28</u>	<u>23.00 $\pm$ 4.66</u>

yielding 35 WER data points. The 18/6/6 split partitions the 30 gold recordings only, at file level and speaker-disjoint, so that no speaker appears in more than one partition. Validation and test therefore consist exclusively of held-out real speech: pseudo-labeled and synthetic material enters the training partition alone and is never evaluated, and the same six real test files are used for every configuration within a seed. We report *total WER* and *disfluency WER* (segments containing at least one target token as a whole word in the reference). Significance uses blockwise bootstrap at file level (1,000 resamples) (Liu and Peng, 2020), which preserves intra-speaker correlation.

5 Results

Fine-tuning reverses the bias. All fine-tuned models reduce total WER by ~ 7.7 points (-28% relative) and disfluency WER by ~ 10.8 points (-32% relative) over the zero-shot baseline (Table 1), consistently on every seed. The 95% blockwise-bootstrap confidence intervals (CIs) of the baseline do not overlap those of any fine-tuned model on either metric, on both the hardest and the easiest seed analyzed. The larger relative gain on disfluency WER confirms that fine-tuning corrects the anti-disfluency bias rather than merely improving general accuracy.

Among the six fine-tuned configurations, mean differences span 0.24 points (total) and 0.85 points (disfluency, excluding FT-syn-2:1), an order of magnitude below the 4–5 point inter-seed standard deviation; bootstrap CIs overlap by more than 70% for every pair. With six test files per seed, differences under two WER points are indistinguishable from sampling noise. This includes FT-gold: adding pseudo-labeled or synthetic material on top of the gold corpus buys at most 0.19 points of total WER and 0.77 points of disfluency WER over

Table 2: Paired per-seed difference against FT-gold across the five seeds, with 95% CI from a paired t test. Negative values favour the augmented system. Pairing removes the inter-seed variance that dominates Table 1.

Model	Δ WER tot.	Δ WER disfl.
FT-pseudo	-0.20 (-0.74, 0.35)	-0.43 (-2.01, 1.16)
FT-syn-llm	0.01 (-0.56, 0.57)	0.07 (-0.94, 1.08)
FT-syn-pg	0.04 (-0.51, 0.59)	-0.78 (-1.27, -0.28)
FT-syn-gp	-0.04 (-0.53, 0.44)	-0.45 (-1.89, 1.00)
FT-syn-2:1	0.49 (-0.24, 1.22)	0.47 (-0.97, 1.91)

fine-tuning on gold alone, so what this comparison unambiguously supports is the effect of verbatim fine-tuning itself; the paired analysis below qualifies that for one configuration. The strategies differ instead in what they require and what they guarantee: FT-syn-gp needs no untranscribed audio, is exactly reproducible at fixed seed, verifies its output distribution explicitly, and is the most stable of the six, with a disfluency-WER standard deviation of 4.10 against 4.47–4.83 and the project-wide minimum (16.59% on seed 789). Section 6 draws the practical consequence.

A paired comparison sharpens this. Table 1 compares configurations through their means across seeds, where the 4–5 point inter-seed spread dominates everything else. Since every configuration is evaluated on the same five splits, a paired comparison against FT-gold removes that variance (Table 2). On total WER the four augmentation strategies have intervals contained within ± 0.75 points of gold-only, so any true difference larger than that margin is excluded at this sample size; only FT-syn-2:1 reaches +1.22. On disfluency WER the intervals are wider, with one exception. FT-syn-pg improves on all five seeds, by 0.78 points on average, with an interval that excludes zero ($p = 0.012$). Ten comparisons were made and this one does not survive a Bonferroni correction, so we report it as suggestive rather than established. It is nonetheless the only evidence in our data that augmentation can improve on gold-only fine-tuning, and its direction is consistent on every seed.

Distributional control works. Table 3 traces the audit. Unguided generation (syn-llm) overproduces English-mediated fillers (*eh* 8.98 $\times$, *lik-som* 8.18 $\times$) and nearly drops the most frequent gold particle (*ju* 0.15 $\times$). The Planner–Generator fixes *ju* but DTW alignment destroys *ja* (0.29). In

Table 3: Synthetic-to-gold density ratio (target 1.0) for the three synthetic sets: unguided language model (syn-llm), Planner–Generator with DTW alignment (syn-pg), and gold-preserving with WhisperX (syn-gp). Each set is the training data of the correspondingly named FT-syn-* model. The table lists nine of the ten active tokens; *ee* reaches 1.39 in syn-gp (within band).

Token	syn-llm	syn-pg	syn-gp	Gold n
<i>ja</i>	0.49	0.29	1.04	100
<i>ju</i>	0.15	1.08	1.16	114
<i>alltså</i>	3.07	3.07	0.96	10
<i>liksom</i>	8.18	0.20	0.86	5
<i>eh</i>	8.98	0.00	0.27	4
<i>nä</i>	0.00	0.00	1.07	1
<i>jo</i>	n.d.	0.14	1.31	9
<i>typ</i>	3.10	0.30	2.14	3
<i>va</i>	0.11	0.18	0.71	6

syn-gp, 8 of the 10 active tokens fall within the 0.7–1.4 $\times$ fidelity band (*ee* among them, at 1.39); the two outliers, *typ* (2.14) and *eh* (0.27), have gold samples of only 3–4 occurrences, where one occurrence shifts the ratio by 25–33%. Positional fidelity is also high where it is constrained. For *ju*, cosine similarity with the gold preceding-POS distribution is 0.893, with 99% of occurrences in the gold top-3 classes. Per-token prediction ratios after fine-tuning mirror this: *ju* and *ja* reach 0.99–1.12 across all models, while *alltså* stays near 0.80 and *va*, a phonetically reduced utterance-final particle, at 0.34–0.47.

The 2:1 ratio does not transfer. FT-syn-2:1 scales the syn-gp pipeline to a 2.35:1 synthetic-to-real ratio, near the threshold recommended by Mizumoto et al. (2025), changing nothing else. It is worse than FT-syn-gp on four of five seeds on total WER (+0.53 mean) and on three of five on disfluency WER (+0.92), within one inter-seed σ but directionally consistent. Per-token ratios distort at the margins (*alltså* 0.75, *jo* 1.33). We attribute this to the acoustic gap of synthetic speech, which a 31M-parameter adapter must absorb in growing proportion. The knee of the ratio–performance curve in this verbatim low-resource regime lies below 1:1, not at 2:1.

6 Discussion and Conclusion

LoRA fine-tuning of KB-Whisper-large removes a large share of its learned anti-disfluency bias, and distribution-controlled synthetic augmentation reaches that level without any untranscribed audio.

The three augmentation strategies are statisti-

cally indistinguishable on WER, and none of them measurably improves on fine-tuning with gold data alone. This says more about the regime than about the methods, since with one hour of verbatim speech extra data of any provenance moves the mean by less than the seed-to-seed spread. In practice this yields a decision rule. Where in-domain untranscribed audio exists, pseudo-labeling is the cheapest route. Where it does not exist, or where the teacher that would produce the pseudo-labels still suppresses the very phenomena we want to recover, the gold-preserving pipeline reaches the same accuracy without any additional audio, and it is the only one of the three whose output distribution can be audited before training. The comparison matters because it makes untranscribed audio optional.

Stability points the same way: because mean differences are dominated by which speakers fall into the test split, the lower inter-seed variance of FT-syn-gp matters more than its mean in a cognitive-assessment setting, where the speaker population is not controlled.

Portability. The pipeline’s structure is language-agnostic: a per-token gold profile, POS-constrained injection, forced alignment, and voice-design TTS. The V2 *lookback* rule is specific to Swedish post-verbal particles, but conditioning insertion on local morphosyntax transfers to other Germanic languages with comparable particle systems. The binding constraint is the TTS model and the forced aligner, which decide feasibility for genuinely lower-resource languages.

Limitations

The gold corpus is small: with six test files per seed, inter-seed standard deviation is 4–5 WER points and blockwise-bootstrap CIs span 20–30 points, so the evaluation can separate fine-tuned models from the baseline but cannot rank augmentation strategies; we therefore report that their differences are not statistically detectable rather than claiming true equivalence. Five seeds are below the $n \geq 10$ recommended for stable variance estimates. Results come from a single backbone (KB-Whisper-large) and a single narrative elicitation task whose register lacks vocalic filled pauses (*öh*), so recall on those forms is untested. The 12 GB VRAM budget excluded FFN modules from LoRA, likely leaving headroom (Lin et al., 2025). UTMOS is calibrated on English and is used only compara-

tively for Swedish. Two further caveats concern the argument itself. The claim that pseudo-labels inherit the teacher’s anti-disfluency bias motivates the whole approach but is not verified here, since we did not measure disfluency density in the pseudo-labels themselves, which would be the direct test of that premise. The three synthetic sets also differ in more than one component at a time, and they are not matched in size, so the comparison across syn-llm, syn-pg and syn-gp is not a controlled ablation and part of any difference may reflect the amount of added data rather than its distribution. The alignment step is the one component we do isolate, since the same generated text was aligned with DTW and with WhisperX, but the contributions of the Planner–Generator and of the global-proportional floor are not separated from each other, and a component-level ablation is the most useful thing future work could add. On alignment we compared only these two options: HMM-based forced aligners such as the Montreal Forced Aligner remain the reference for the task, and we chose a CTC aligner because it integrates directly with WhisperX and a Swedish wav2vec2 model was readily available, not because it had been shown superior here. Finally, downstream clinical utility (agreement of fluency features extracted from ASR vs. manual transcripts) has not yet been measured, and the corpus, collected under a clinical protocol, cannot be made openly available. However, access to de-identified data may be considered upon reasonable request, subject to ethical approvals, institutional requirements, and the execution of an appropriate data-use agreement.

Ethical Considerations

The gold corpus consists of older adults’ recordings produced within a cognitive-assessment research protocol, and contains identifiable voice data. The study involves human participants, and the Regional Ethical Review Board approved the recordings in Gothenburg (approvals 2016-04-14, Reference No. 2016-16 and 2018-01-05, Reference No. T021-18). The corpus is not publicly released; the conditions under which derived data may be requested are stated above. All processing that touches the recordings, including the language model used to generate synthetic text, was run locally on our own hardware, so no participant audio or transcript was sent to an external service.

The synthetic speech in this work is produced

in voice-design mode, that is, from textual descriptions of speaker attributes rather than by cloning the voice of a real speaker, and none of it reproduces the identity of a corpus participant. We note nonetheless that TTS systems with controllable prosody carry a general risk of misuse for impersonation, and injecting disfluencies makes synthetic speech harder to distinguish from real recordings.

Finally, the system described here is a transcription model, not a diagnostic instrument. Lower disfluency WER makes downstream clinical analysis possible in principle, but we have not shown that the resulting transcripts support valid clinical inference, and they should not be used to inform decisions about individuals.

Acknowledgments

The research presented here is supported by the Swedish Research Council (grant nr 2025-00765), the Swedish national research infrastructure Språkbanken – jointly funded by its 10 partner institutions and the Swedish Research Council (grant nrs 2017-00626 and 2023-00161).

References

- Emmanuel Akinrintoyo, Nadine Abdelhalim, and Nicole Salomons. 2025. [WhisperD: Dementia speech recognition and filler word detection with Whisper](#). In *Proc. Interspeech 2025*.
- Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. [WhisperX: Time-accurate speech transcription of long-form audio](#). In *Proc. Interspeech 2023*.
- Robert Eklund. 2004. *Disfluency in Swedish Human–Human and Human–Machine Travel Booking Dialogues*. PhD dissertation, Linköping University. Linköping Studies in Science and Technology, Dissertation No. 882.
- Valarie B. Fleming and Joyce L. Harris. 2008. [Complex discourse production in mild cognitive impairment: Detecting subtle changes](#). *Aphasiology*, 22(7–8):729–740.
- 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 *Proc. ICLR 2022*.
- Kaichen Jia, Jinpeng Li, Ke Li, and Wei-Qiang Zhang. 2025. [Whisper-based multilingual Alzheimer’s disease detection and improvements for low-resource language](#). In *Proc. Interspeech 2025*.
- Damjan Kalajdzievski. 2023. [A rank stabilization scaling factor for fine-tuning with LoRA](#). *Preprint*, arXiv:2312.03732.
- Dimitrios Kokkinakis, Giuseppe Chiriatti, Herbert Lange, and Ricardo Muñoz Sánchez. 2026. [Longitudinally informed synthetic speech and language data for neuropsychological research – a pilot study](#). In *Proceedings of the 36th Alzheimer Europe Conference (AEC 2026)*, Dublin, Ireland.
- Ludwig Kürzinger, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll. 2020. [CTC-segmentation of large corpora for German end-to-end speech recognition](#). In *Speech and Computer (SPECOM 2020)*, pages 267–277.
- Jinpeng Li, Yu Pu, Qi Sun, and Wei-Qiang Zhang. 2024. [Improving Whisper’s recognition performance for under-represented language Kazakh leveraging unpaired speech and text](#). In *Proc. Interspeech 2024*, pages 2514–2518.
- 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 *Proc. ISCA Workshop on Speech and Language Technology in Education (SLaTE 2025)*.
- Wei Liu, Ying Qin, Zhiyuan Peng, and Tan Lee. 2024a. [Sparsely shared LoRA on Whisper for child speech recognition](#). In *Proc. ICASSP 2024*.
- Yin-Long Liu, Rui Feng, Jia-Hong Yuan, and Zhen-Hua Ling. 2024b. [Leveraging prompt learning and pause encoding for Alzheimer’s disease detection](#). In *Proc. ISCSLP 2024*, pages 486–490.
- Zhe Liu and Fuchun Peng. 2020. [Statistical testing on ASR performance via blockwise bootstrap](#). In *Proc. Interspeech 2020*, pages 596–600.
- Tomoya Mizumoto, Atsushi Kojima, Yusuke Fujita, Lianbo Liu, and Yui Sudo. 2025. [Is synthetic data truly effective for training speech language models?](#) In *Proc. Interspeech 2025*.
- Kristina Nilsson Björkenstam, Mats Wirén, and Robert Eklund. 2013. [Disfluency in child-directed speech](#). In *Proc. Fonetik 2013*, Linköping, Sweden.
- Yanis Perrin and Gilles Boulianne. 2025. [Towards improved speech recognition through optimized synthetic data generation](#). *Preprint*, arXiv:2508.21631.
- 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.

- Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. [UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022](#). In *Proc. Interspeech 2022*.
- Leonora Vesterbacka, Faton Rekathati, Robin Kurtz, Justyna Sikora, and Agnes Toftgård. 2025. [Swedish whispers: Leveraging a massive speech corpus for Swedish speech recognition](#). In *Proc. Interspeech 2025*, pages 758–762.
- Shiyao Wang, Jiaming Zhou, Shiwan Zhao, and Yong Qin. 2025. [A self-training approach for Whisper to enhance long dysarthric speech recognition](#). In *Proc. Interspeech 2025*.
- Han Zhu, Lingxuan Ye, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhifeng Han, Weiji Zhuang, Long Lin, and Daniel Povey. 2026. [OmniVoice: Towards omnilingual zero-shot text-to-speech with diffusion language models](#). *Preprint*, arXiv:2604.00688.