

Learning to Hear Hesitation: Continual Learning for Disfluency-Aware ASR

Henri-Leon Kordt^{*}, Theresa Pekarek Rosin^{*}, Jae Hee Lee^{*}, Stefan Wermter^{*}

Knowledge Technology, Department of Informatics, University of Hamburg, Germany

{henri-leon.kordt, theresa.pekarek-rosin, jae.hee.lee, stefan.wermter}@uni-hamburg.de

Abstract

Despite advances in large-scale Automatic Speech Recognition (ASR), disfluent speech remains challenging, as state-of-the-art systems are often optimized to omit disfluencies, leading to information loss and hallucinations. Prior work has focused on verbatim transcription and the integration of disfluency markers, but adapting models on limited datasets can lead to catastrophic forgetting of general-domain knowledge. We address this gap by leveraging continual learning (CL) with explicit disfluency tokens. We first introduce these tokens into a pretrained ASR model to establish stable token mechanisms, and then continue training on additional datasets with varying disfluency distributions. Through a detailed analysis of model dynamics during training, we identify a trade-off between marker learning and ASR performance, and a consistent cross-attention head mechanism shared across CL methods.

Index Terms: speech recognition, continual learning, verbatim transcription, disfluent speech recognition

1. Introduction

Although Automatic Speech Recognition has improved substantially in recent years due to the introduction of large-scale models trained on vast amounts of audio data [1], disfluent speech recognition has not seen the same progress [2]. Disfluencies are perturbations in speech that can range in severity from simple filler words (“uh”, “uhm”), hesitations, and word or phrase repairs to slurred speech or stutters caused by underlying conditions. However, state-of-the-art models are often trained to ignore disfluencies in order to produce clean transcriptions [3, 4], which can lead to incorrectly generated utterances, i.e., hallucinations or fabrications, and to a general loss of information and transparency [5, 6].

To address this problem and improve performance on disfluent speech, MacDonald et al. [7] have suggested personalizing models for each speaker, while others have focused on creating more verbatim transcriptions [5, 8]. Such verbatim transcriptions and the subsequent inclusion of disfluencies can capture clinically relevant speech information [9], improve audio-text alignment for timestamping [8], and even improve downstream ASR performance [5, 10]. Previous work has also introduced specific markers or tokens to represent speech disfluencies by training pretrained models on additional annotated speech corpora, both to detect dementia from utterances [9] and to improve performance on disfluent [10] or spontaneous speech [8]. However, when applied carelessly to small-scale datasets, these approaches often lead to a loss of domain-agnostic performance [11], an effect known as catastrophic forgetting [12]. A naive approach to

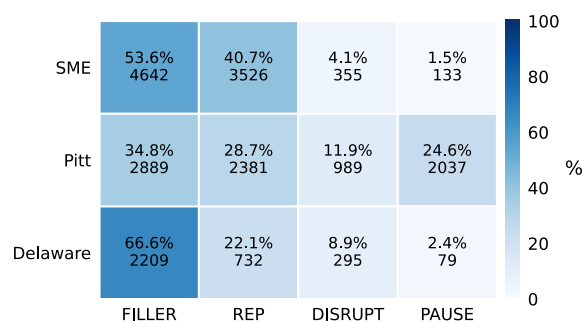


Figure 1: The distribution of disfluencies across the selected datasets. Percentages are calculated with respect to each dataset’s distribution.

address this effect is to jointly train the model on the new data and the entire pretraining data. Since speech patterns change over time and new data becomes available constantly, this process would have to be repeated for every new data instance, which is highly inefficient. Additionally, because training data for disfluent speech is often limited and subject to privacy regulations, this approach is frequently infeasible.

Continual Learning (CL) is a paradigm for mitigating catastrophic forgetting by balancing *learning plasticity* and *memory stability* through regularization, replay, or architectural approaches [11]. Such methods provide an efficient way to integrate heterogeneous datasets and speech patterns without sacrificing the pretrained model’s performance. CL has been successfully applied to domain adaptation for non-standard speech [13, 14] and low-resource languages [15, 16], but it has not yet been utilized for disfluent or impaired speech. Moreover, disfluency events and disfluency transcriptions pose unique challenges that differ from typical ASR domain adaptation tasks, for which evidence of CL effectiveness remains unexplored.

In this work, we address disfluent speech recognition using disfluency markers, which are commonly employed in disfluency-focused ASR research [5, 8, 10] and are also present in the manual transcriptions of datasets collected under the Talk-Bank [17, 18] umbrella. We aggregate these markers into four token types and introduce them to a pretrained model using different CL methods to preserve the model’s performance. We then continue training on two annotated datasets with different disfluency distributions to examine the robustness of the established marker representations for continual adaptation. We examine adaptation differences across CL methods and find that, despite performance differences, successful marker learning largely follows a shared attention-head specialization pattern.

^{*}These authors contributed equally.

2. Methodology

To enable more verbatim transcriptions and increase the amount of information produced by the model, we introduce disfluency markers into a pretrained backbone model. We combine continual learning (CL) with explainability methods across three datasets to investigate internal specialization mechanisms while preserving performance under continual model adaptation.

2.1. Datasets

To cover both healthy and impaired disfluent speech, we use three datasets from the TalkBank repository [17, 18], which were selected based on data availability, number of samples, disfluency distribution, and domain: the Standard Malaysian English (SME) Corpus [19], the Pitt Corpus [20], and the Delaware Corpus [18]. All three use the unified CHAT Transcription Format [21], which ensures consistency in the labeling syntax. Their transcriptions also contain manually annotated disfluency markers, allowing a direct mapping to disfluency tokens during training.

The SME Corpus [19] contains 11.79 h of English speech from Malaysian university students who are second-language learners. This dataset represents a common subgroup of speakers who are generally healthy but may exhibit disfluencies in their speech production due to the language gap. The Pitt Corpus [20] is a collection of 21.30 h of speech from elderly speakers with dementia or Alzheimer’s disease, while the Delaware Corpus [18] contains 9.72 h of speech from people with mild cognitive impairment. Both Pitt and Delaware include speech from healthy individuals in their control groups, which we divide evenly across the training and validation splits. We include only 12.11 h of the Pitt corpus in the CL setup to reduce dataset-size effects. We use these datasets to demonstrate the robustness of our approach across different severities and distributions of disfluencies (see Figure 1). In the CL setup, we also include a small subset of LibriSpeech data [22] for the rehearsal buffer in replay-based methods and as a representation of unperturbed speech.

2.2. Continual learning setup

We use `whisper-small.en`¹, a pretrained monolingual model, as the backbone for the CL setup. To preserve the pretrained performance of the model and handle the different disfluency distributions in the datasets, we use four CL methods that are commonly employed for domain-incremental ASR training: Elastic Weight Consolidation (EWC) [23], Experience Replay (ER) [24], A-GEM [25], and Weight Averaging (WA) [13].

EWC [23] is a regularization-based method that identifies model parameters important to previously learned tasks and penalizes updates to those parameters (using the diagonal Fisher information matrix). For ER [24], a small percentage of the old data is kept in memory (rehearsal buffer) and used to supplement the training data for the new task, preventing drastic gradient updates. A-GEM [25] is a gradient-based method that calculates the dot product between the gradients of the rehearsal buffer and the training data of the current dataset to penalize gradients in opposing directions. For WA [13], the weights of the old model are kept in memory, while another version is trained on the new task. Once training is complete, the model is updated with an average of the old and the new weights.

To evaluate general ASR performance, we use the preprocessed WER (pWER), in which punctuation, special characters, and disfluency markers are removed from the transcripts before

the metric is computed. To assess the model’s memory stability and learning plasticity during CL, we use the metrics defined by Wang et al. [11], with minor modifications to accommodate pWER and marker F1 in place of accuracy. Average Word Error Rate (A-WER) and Average Incremental WER (AI-WER) measure overall pWER performance at the current step and across the full CL trajectory, respectively. For marker prediction, we similarly report Average F1 (A-F1) and Average Incremental F1 (AI-F1), which replace per-task accuracy with per-task marker F1. The F1 scores are calculated using a bag-of-words approach. Backward Transfer (BWT) and Forgetting Measure (FM) assess memory stability, whereas Forward Transfer (FWT) and Intransigence Measure (IM) assess the model’s learning plasticity.

2.3. Explainability methods

Given the tendency of transformer models to develop specialization and redundancy [26, 27], we investigate whether the transcription of disfluency events induces specialized circuits within the model. This is particularly relevant in the CL setting, since CL methods may yield similar aggregate error rates while inducing different adaptation dynamics and internal strategies. By probing attention heads, we aim to assess whether disfluency handling concentrates in a small set of heads (specialized circuitry) and how stable these mechanisms are across CL approaches, thereby providing insight into robustness and potential method-specific trade-offs. We follow the head-masking approach of Michel et al. [26], which introduces a learnable scalar gate ξ_h per attention head h . To estimate token-level head importance I_h , we compute the sensitivity of a token objective to ξ_h and average over a set of token instances X_t :

$$I_h = \mathbb{E}_{x \sim X_t} \left[\frac{\partial \mathcal{L}(x)}{\partial \xi_h} \right] \quad (1)$$

Based on this per-head importance I_h , we rank head attribution for disfluency tokens and compare importance rankings averaged over non-disfluency tokens. Specifically, we calculate the top-10 lift, defined as the difference in the frequency with which a head h appears in the top-10 ranking for disfluency tokens versus its frequency in the baseline distribution (all other tokens):

$$\text{Lift}_h = P(h \in \text{Top-}k \mid X_t) - P(h \in \text{Top-}k \mid X_{\text{base}}) \quad (2)$$

To confirm head attribution, we ablate heads that show high top-10 lift scores using zero-masking. We inspect changes in token emission and control for overall performance using pWER.

2.4. Experiments

To assess the introduction of disfluency tokens into a pretrained model and the resulting task performance in a Continual Learning (CL) setup, we conduct two experiments: (i) disfluency token introduction and (ii) sequential continual adaptation. Following the literature [8, 10], we aggregate similar disfluency events into four distinct token types: filler words (FILLER; e.g., “um”, “uh”), repetitions/revisions (REP; word- and phoneme-level), disruptions (DISRUPT; e.g., coughs, laughs), and pauses (PAUSE).

In the **disfluency token introduction** experiment (cf. Section 3.1), we focus on introducing these tokens into a pretrained model without overfitting to the new dataset, i.e., while keeping performance on non-disfluent speech high. We fine-tune the backbone on the SME dataset with disfluency tokens using the CL techniques discussed above. SME covers the domain closest to the backbone and predominantly contains only two marker

¹<https://huggingface.co/openai/whisper-small.en>

Table 1: *Disfluency token introduction: The preprocessed Word Error Rate (pWER) averaged over 3 seeds after 10 epochs of training for SME and LibriSpeech (LS). We additionally report the micro-F1 score for marker predictions on SME.*

	SME (pWER% ↓)	LS (pWER% ↓)	SME (F1 ↑)
BACKBONE	15.97	3.47	0.00
FT	12.21 ± 0.18	5.06 ± 0.02	0.73 ± 0.01
A-GEM	12.17 ± 0.17	4.39 ± 0.07	0.75 ± 0.01
ER	12.47 ± 0.03	4.55 ± 0.13	0.73 ± 0.01
EWC	10.34 ± 0.47	4.43 ± 0.07	0.21 ± 0.07
WA	9.64 ± 0.08	3.41 ± 0.01	0.00 ± 0.00

types: FILLER and REP. Additionally, we analyze whether the introduction of disfluency tokens is associated with the emergence of internal model mechanisms and whether these mechanisms differ across CL methods.

For the **sequential continual adaptation** experiment (cf. Section 3.2), we choose the best-performing model from the disfluency token introduction stage: a model that shows successful marker integration (high marker-F1) while maintaining the smallest pWER trade-off. We then continue expanding the model with the Pitt and Delaware corpora, two disfluent-speech datasets that vary in domain and disfluency distribution (see Figure 1). In doing so, we approximate lifelong model adaptation to changing domains and speaker variations through sequential training on more difficult datasets. We begin with the Pitt Corpus, which introduces a substantial number of pause markers and also increases the number of disruptions. Afterward, we train the model on the Delaware Corpus, which contains the fewest markers overall and therefore specifically challenges the model to retain them.

We ensure that the training and evaluation splits for both experiments are speaker-disjoint and partition each dataset into an 80/20 split. For the rehearsal buffer, we randomly sample 10% of the training data from each dataset, greedily prioritizing rare disfluency markers so that the token distribution mirrors that of the full dataset. We train on each dataset for 10 epochs with a learning rate of $2e-5$ and a batch size of 16, ensuring convergence in both marker scores and ASR performance. Results are averaged across three runs. Following the literature [23, 24] and our empirical evaluation, we include 25% of old data per batch, sampled from the rehearsal buffer for ER, and set the importance parameter to 1000 for EWC. In addition to the general performance and CL metrics discussed in Section 2.2, we also measure disfluency-marker recognition using micro- and macro-averaged F1 scores. For the disfluency token introduction experiment, we use micro-F1 to summarize overall marker integration for the introduced markers, while macro-F1 is used later to assess per-type balance once all markers are in scope.

3. Results

3.1. Disfluency token introduction

We train on SME using different CL methods to examine how well performance on non-disfluent speech (LibriSpeech test-clean, LS) can be retained while improving performance on the new domain and introducing disfluency markers into the model. Table 1 shows that all CL methods retain LS performance sub-

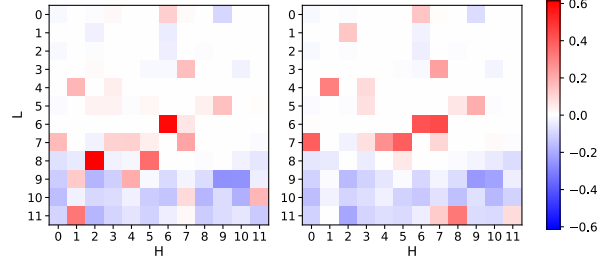


Figure 2: *Top-10 lift scores for FILLER (left) and REP (right) across decoder cross-attention heads computed on the train split averaged across all token emitting methods.*

Table 2: *Zero-mask ablation, averaged across methods. We report the relative change in marker emission (%) and the change in pWER after masking the top-5 cross-attention heads identified for FILLER, REP, their union, and a random-head control.*

	Δ FILLER	Δ REP	Δ pWER%
FILLER: Top-5	-57.0 ± 5.2	-7.1 ± 3.8	+0.42 ± 0.64
REP: Top-5	-2.8 ± 1.1	-15.7 ± 4.2	+0.13 ± 0.46
$\cup$ Top-5	-62.2 ± 5.2	-21.2 ± 9.8	+0.44 ± 0.68
Control ($5 \times r$)	+0.2 ± 0.5	+0.4 ± 1.2	+0.15 ± 0.21

stantially better than the FT baseline while also improving pWER on SME. Weight Averaging (WA) outperforms the other methods by retaining, and even slightly improving, the ASR model performance on LS while achieving the lowest pWER on SME at 9.64%. EWC also improves pWER on SME relative to FT, but shows weaker LibriSpeech pWER retention than WA. Importantly, both methods fail to integrate markers reliably. The FT baseline, ER, and A-GEM consistently produce marker micro-F1 scores between 0.73 and 0.75, whereas EWC remains at a micro-F1 of 0.21 and WA fails to emit markers at all. This means that although A-GEM, ER, and the FT baseline perform worse on the trained domain, improving the pWER only to around 12%, they successfully produce disfluency markers.

These results indicate a trade-off between pWER and marker performance, suggesting that marker training may require parameter updates that conflict with low-cost general ASR domain adaptation. To test whether marker learning corresponds to a consistent mechanism that differs from normal token production, we apply our head-attribution approach (Section 2.3) across all methods. We find that when markers are successfully trained, a few specialized decoder cross-attention heads are consistently assigned to marker emission, as shown in Figure 2.

Notably, the same small set of heads emerges across all marker-emitting methods, suggesting a shared cross-attention mechanism rather than method-specific marker strategies induced by different regularization. In Table 2, we verify their causal role through an ablation study on the top-5 cross-attention heads associated with disfluency markers and show that we can consistently remove around 50% of the emitted FILLER markers while retaining similar pWER values.

3.2. Sequential continual adaptation

For the sequential continual adaptation experiment, we investigate the ability of CL methods to generalize and retain knowledge when the model is continuously trained on more difficult

Table 3: Continual learning metrics after sequential adaptation on Pitt and Delaware, initialized from the selected disfluency token introduction checkpoint. For each metric, we report pWER-based and marker F1-based variants.

	A-WER% / A-F1		AI-WER%/AI-F1		BWT		FM		FWT		IM	
	pWER ↓	F1 ↑	pWER ↓	F1 ↑	pWER ↑	F1 ↑	pWER ↓	F1 ↓	pWER ↑	F1 ↑	pWER ↓	F1 ↓
JOINT	17.95 $\pm$ 0.28	0.47 $\pm$ 0.01	—	—	—	—	—	—	—	—	—	—
FT	20.24 $\pm$ 0.14	0.39 $\pm$ 0.02	19.00 $\pm$ 0.06	0.49 $\pm$ 0.01	-3.18 $\pm$ 0.26	-0.16 $\pm$ 0.02	3.48 $\pm$ 0.35	0.19 $\pm$ 0.02	-3.81 $\pm$ 0.10	-0.05 $\pm$ 0.01	0.44 $\pm$ 0.07	-0.03 $\pm$ 0.01
A-GEM	20.15 $\pm$ 0.68	0.36 $\pm$ 0.02	18.88 $\pm$ 0.24	0.48 $\pm$ 0.01	-3.46 $\pm$ 0.78	-0.19 $\pm$ 0.03	3.50 $\pm$ 0.72	0.21 $\pm$ 0.03	-3.39 $\pm$ 0.24	-0.07 $\pm$ 0.01	0.16 $\pm$ 0.16	-0.02 $\pm$ 0.00
ER	19.71 $\pm$ 0.23	0.49 $\pm$ 0.01	18.17 $\pm$ 0.10	0.53 $\pm$ 0.01	-2.57 $\pm$ 0.48	0.01 $\pm$ 0.01	2.59 $\pm$ 0.46	0.02 $\pm$ 0.00	-3.63 $\pm$ 0.52	-0.06 $\pm$ 0.00	0.32 $\pm$ 0.34	-0.03 $\pm$ 0.00
EWC	19.01 $\pm$ 0.08	0.44 $\pm$ 0.01	18.28 $\pm$ 0.05	0.52 $\pm$ 0.00	-1.40 $\pm$ 0.08	-0.09 $\pm$ 0.02	2.18 $\pm$ 0.16	0.13 $\pm$ 0.02	-3.75 $\pm$ 0.20	-0.05 $\pm$ 0.01	0.40 $\pm$ 0.13	-0.03 $\pm$ 0.01
WA	18.90 $\pm$ 0.19	0.46 $\pm$ 0.00	17.57 $\pm$ 0.04	0.51 $\pm$ 0.01	-0.55 $\pm$ 0.34	-0.01 $\pm$ 0.00	1.35 $\pm$ 0.20	0.05 $\pm$ 0.01	-4.43 $\pm$ 0.13	-0.10 $\pm$ 0.00	0.85 $\pm$ 0.09	0.00 $\pm$ 0.00

disfluency datasets after an initial marker set has been established. We initialize all methods from the same disfluency-token introduction checkpoint (A-GEM Seed 2), which reliably emits markers, selected by SME marker-F1 and SME pWER, and then sequentially train on Pitt and Delaware using each CL strategy. In Table 3, we report common CL metrics (cf. Section 2.2) after fine-tuning on Pitt and Delaware using different CL techniques.

Our results show that all CL methods reduce pWER relative to the FT baseline, as reflected in lower A-WERs across all methods and all methods except A-GEM also improve marker F1. Importantly, the best-performing methods differ for ASR performance and marker retention: WA achieves an A-WER of 18.90%, while ER shows the best marker performance with a macro-F1 score of 0.49.

WA shows the strongest pWER reduction while also exhibiting the most stable pWER performance, as indicated by AI-WER. WA achieves these gains mainly through retention, exhibiting minimal forgetting and stable backward transfer, while showing comparatively limited plasticity in learning new tasks, as reflected by the high IM score. ER, on the other hand, achieves the best performance for disfluency-marker F1. The method also shows strong generalization for markers, with nearly no forgetting as indicated by an FM score of 0.02, and even slight improvements on old tasks, reflected in a slightly positive marker-F1 BWT of 0.01.

To understand which marker types drive the overall F1 differences, we report the final marker-F1 scores for the different disfluency markers in Table 4. FILLER markers are robust across methods, with F1 scores between 0.68 and 0.75, whereas REP markers exhibit the strongest method sensitivity, ranging from 0.35 to 0.63, indicating learning differences or forgetting. PAUSE markers are the most challenging overall, but ER yields a clear advantage, which is consistent with ER achieving the best aggregate marker F1 in Table 3.

We additionally evaluate pWER for LS to isolate non-disfluent speech retention after sequential training and find that the ranking closely matches the disfluent-speech retention trend: WA performs best by a clear margin (4.68%), EWC and ER follow with 6.45% and 7.14% respectively, and A-GEM (8.36%) shows nearly no improvement over the FT baseline (8.37%).

4. Conclusion

We compared several continual learning (CL) methods to investigate their ability to integrate disfluency tokens for more verbatim transcriptions and better performance on disfluent speech. We specifically focused on realistic model-expansion scenarios without joint retraining on large disfluency datasets. Overall, CL improved sequential training performance for both ASR perfor-

Table 4: Marker-F1 ($\uparrow$) on the combination of all three datasets after sequential training on Pitt and Delaware, initialized on the selected SME checkpoint.

	FILLER	REP	DISRUPT	PAUSE
JOINT	0.71 $\pm$ 0.01	0.54 $\pm$ 0.02	0.42 $\pm$ 0.02	0.25 $\pm$ 0.02
FT	0.69 $\pm$ 0.02	0.42 $\pm$ 0.02	0.34 $\pm$ 0.00	0.09 $\pm$ 0.06
A-GEM	0.68 $\pm$ 0.01	0.35 $\pm$ 0.03	0.34 $\pm$ 0.03	0.05 $\pm$ 0.02
ER	0.75 $\pm$ 0.01	0.61 $\pm$ 0.00	0.38 $\pm$ 0.01	0.23 $\pm$ 0.02
EWC	0.72 $\pm$ 0.01	0.49 $\pm$ 0.02	0.37 $\pm$ 0.02	0.16 $\pm$ 0.01
WA	0.73 $\pm$ 0.00	0.63 $\pm$ 0.01	0.37 $\pm$ 0.01	0.09 $\pm$ 0.01

mance and disfluency-marker retention, but we found that the best CL method depended on the primary objective. We evaluated performance in terms of initial marker inclusion, marker retention, and ASR performance on both disfluent and non-disfluent speech.

We found that CL strategies are not equally effective at introducing disfluency markers while preserving ASR quality. Across CL settings, marker F1 and pWER showed a consistent trade-off: methods that maintained the strongest pWER performance tended to under-introduce markers, while those that reliably emitted markers did so at a measurable ASR cost. Further investigation of the effect showed that successful marker learning was linked to a consistent cross-attention pattern: a small set of heads was consistently active during marker production, and we found that targeted ablations reduced marker emission while largely preserving pWER. This points to a reliable learned attention mechanism and offers a mechanistic explanation for why some CL methods struggle more with marker introduction.

When extending a model with additional disfluency datasets, ER achieved the strongest marker generalization and retention, with the best overall marker performance and minimal forgetting. For the preservation of disfluent ASR performance, WA was the most stable across incremental updates, consistently delivering the strongest pWER retention and the lowest forgetting. This stability also transferred to clean speech, where WA performed best in both experiments.

While we limited ourselves to a single backbone model and a single realistic task ordering, overall, this work presented one of the first studies of CL for disfluency modeling, offering both practical guidance for verbatim ASR adaptation and mechanistic insights into how disfluency token behaviors are acquired in pretrained models. These insights provide a foundation for future work to combine and design methods that enable robust domain adaptation and reliable marker retention.

5. Acknowledgements

The authors gratefully acknowledge funding from Horizon Europe under the MSCA grant agreements No 101072488 (TRAIL), No 101226624 (GREET), and No 101168792 (SWEET), as well as from the German Research Foundation (DFG), project number 551629603.

Data collection for the Delaware Corpus was supported by the National Institute of Aging of the National Institutes of Health under award number RF1AG083823 (PIs: MacWhinney and Lanzi). The Pitt Corpus is part of DementiaBank, which is funded by NIH grants NIA AG03705 and AG05133.

6. Generative AI Use Disclosure

Generative AI tools have been used to polish the language in the manuscript.

7. References

- [1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” in *Proceedings of the 40th International Conference on Machine Learning (ICML’23)*, Honolulu, Hawaii, USA, 2023, pp. 28 492–28 518.
- [2] S. Feng, B. M. Halpern, O. Kudina *et al.*, “Towards inclusive automatic speech recognition,” *Computer Speech and Language*, vol. 84, no. C, p. 101567, 2024.
- [3] H. Hwang, E. Jordan, D.-H. Kim-Dufor, C. Lemey, and M. Alrahabi, “Evaluating ASR in a clinical context : What whisper misses,” in *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*. Southern Denmark University, Odense, Denmark: Association for Computational Linguistics, Aug. 2025, pp. 374–378.
- [4] D. Mujtaba, N. Mahapatra, M. Arney, J. Yaruss, H. Gerlach-Houck, C. Herring, and J. Bin, “Lost in transcription: Identifying and quantifying the accuracy biases of automatic speech recognition systems against disfluent speech,” in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 4795–4809.
- [5] J.-K. Lin, H.-C. Lu, C.-C. Wang, H.-Y. Lin, and B. Chen, “Acoustically precise hesitation tagging is essential for end-to-end verbatim transcription systems,” in *10th Workshop on Speech and Language Technology in Education (SLaTE)*, 2025, pp. 163–166.
- [6] T. Pekarek Rosin, J. Gachot, H.-L. Kordt, M. Kerzel, and S. Wermter, “Talking to robots: A practical examination of speech foundation models for hri applications,” in *Workshop on Foundation Model for Social Robotics (FoMoSR) at ICSR 2025*, Aug 2025.
- [7] R. L. MacDonald, P.-P. Jiang, J. Cattiau, R. Heywood, R. Cave, K. Seaver, M. A. Ladewig, J. Tobin, M. P. Brenner, P. C. Nelson *et al.*, “Disordered Speech Data Collection: Lessons Learned at 1 Million Utterances from Project Euphonia,” in *Proceedings of Interspeech 2021*, Brno, Czech Republic, Aug. 2021.
- [8] M. Zusag, L. Wagner, and B. Thallinger, “CrisperWhisper: Accurate Timestamps on Verbatim Speech Transcriptions,” in *Proceedings of Interspeech 2024*, Kos, Greece, 2024, pp. 1265–1269.
- [9] R. Soleimani, S. Guo, K. L. Haley, A. Jacks, and E. Lobaton, “The impact of pause and filler word encoding on dementia detection with contrastive learning,” *Applied Sciences*, vol. 14, no. 19, 2024.
- [10] E. Akinrintoyo, N. Abdelhalim, and N. Salomons, “WhisperD: Dementia Speech Recognition and Filler Word Detection with Whisper,” in *Proceedings of Interspeech 2025*, 2025, pp. 1413–1417.
- [11] L. Wang, X. Zhang, H. Su, and J. Zhu, “A comprehensive survey of continual learning: Theory, method and application,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5362–5383, Aug. 2024.
- [12] G. M. van de Ven, N. Soures, and D. Kudithipudi, “Continual learning and catastrophic forgetting,” in *Learning and Memory: A Comprehensive Reference (Third Edition)*, 3rd ed. Oxford: Academic Press, 2025, pp. 153–168.
- [13] S. Vander Eeck and H. Van hamme, “Rehearsal-free online continual learning for automatic speech recognition,” in *Proceedings of Interspeech 2023*, 2023, pp. 944–948.
- [14] T. Pekarek Rosin and S. Wermter, “Replay to remember: Continual layer-specific fine-tuning for german speech recognition,” in *Artificial Neural Networks and Machine Learning – ICANN 2023*. Springer, Sep 2023, p. 489–500.
- [15] L. Della Libera, P. Mousavi, S. Zaiem, C. Subakan, and M. Ravanelli, “Cl-masr: A continual learning benchmark for multilingual asr,” *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 32, pp. 4931–4944, Oct. 2024.
- [16] Z. Song, J. Zhuo, Y. Yang, Z. Ma, S. Zhang, and X. Chen, “LoRA-Whisper: Parameter-Efficient and Extensible Multilingual ASR,” in *Interspeech 2024*, 2024, pp. 3934–3938.
- [17] B. MacWhinney, “Talkbank and sla,” in *The Handbook of SLA and Corpora*, N. Tracy-Ventura and M. Paquot, Eds. New York: Routledge, 2019.
- [18] A. M. Lanzi, A. K. Saylor, D. Fromm, H. Liu, B. MacWhinney, and M. L. Cohen, “Dementiabank: Theoretical rationale, protocol, and illustrative analyses,” *American Journal of Speech-Language Pathology*, vol. 32, no. 2, pp. 426–438, 2023.
- [19] S. Sie, “Investigating the role of an indigenised variety of english in the acquisitional and sociolinguistic contexts of the malaysian ecology,” Ph.D. dissertation, Apollo - University of Cambridge Repository, 2023.
- [20] J. T. Becker, F. Boller, O. L. Lopez, J. Saxton, and K. L. McGonigle, “The natural history of alzheimer’s disease: Description of study cohort and accuracy of diagnosis,” *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 06 1994.
- [21] B. MacWhinney, *The CHILDES Project: Tools for Analyzing Talk*, 3rd ed. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.
- [22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.
- [23] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, “Overcoming catastrophic forgetting in neural networks,” *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.
- [24] D. Rolnick, A. Ahuja, J. Schwarz, T. Lillicrap, and G. Wayne, “Experience replay for continual learning,” in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.
- [25] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny, “Efficient lifelong learning with a-gem,” in *Proceedings of ICLR 2019*, 2019.
- [26] P. Michel, O. Levy, and G. Neubig, “Are sixteen heads really better than one?” in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.
- [27] K. Ahrens, L. Bengtson, J. H. Lee, and S. Wermter, “Visually Grounded Continual Language Learning with Selective Specialization,” in *EMNLP Findings*. Association for Computational Linguistics, 2023, pp. 7037–7054.