No Label? No Problem: Unsupervised Continual Learning for Adaptive Medical ASR

Meizhu Liu
Oracle AI
meizhu.liu@oracle.com

Tao Sheng
Oracle AI
tao.t.sheng@oracle.com

Abstract

Automatic Speech Recognition (ASR) plays an important role in healthcare but faces unique challenges. Medical audio often contains specialized terminology, such as medication names, which existing ASR systems struggle to transcribe accurately. High error rates arise from pronunciation variability, the continual introduction of new terms, and the scarcity of high-quality labeled data—whose collection is costly and requires medical expertise. Although synthetic datasets partially alleviate this problem, they fail to capture the noise and variability of real-world recordings. Moreover, ASR models trained in controlled environments are highly sensitive to noise, leading to degraded performance in clinical settings. To address these limitations, we propose an unsupervised continual learning ASR framework that adapts to new data while preserving prior knowledge. This enables efficient domain adaptation without extensive retraining. Experiments on real-world medical audio demonstrate significant improvements over state-of-the-art baselines.

1 Introduction

Automatic speech recognition (ASR) converts spoken words into text and has been widely applied across various domains (Han et al., 2020; Gulati et al., 2020; Zeineldeen et al., 2021). In the medical domain, ASR is particularly valuable: it can transcribe conversations between physicians and patients or among healthcare professionals (Ahlawat et al., 2025), reducing documentation time and helping prevent medication errors caused by illegible handwriting or misspellings (Schmidt, 2010), thereby improving patient safety and healthcare efficiency.

Despite its promise, medical ASR faces unique challenges. Medical audio contains specialized entities such as medication names, diseases, symptoms, and procedures, which are often complex and constantly evolving (see Table 1). In addition, speakers exhibit diverse accents (Afonja et al., 2024), dialects, and vocal characteristics, and the pronunciation of medical terms can vary. Patients' voice characteristics may further differ depending on symptoms. Real-world recordings are often noisy, unlike the controlled conditions many ASR models are trained on. These factors make accurate transcription—especially of medication names—extremely challenging, and errors can pose serious risks in clinical settings (Fouda, 2024).

Another major challenge is the limited availability of data. High-quality, labeled doctor–patient conversations are costly and difficult to acquire, leading many studies to rely on synthetic datasets (Kazi et al.). Strict privacy regulations further restrict access to real medical recordings, rendering conventional supervised training approaches impractical. Although continual learning has been investigated for general ASR (Houston and Kirchhoff, 2020; Fu et al., 2021; Eeckt and hamme, 2024), its application to medical ASR remains underdeveloped, primarily due to the scarcity of labeled streaming data (Kessler et al., 2021; Chang et al., 2021; Eeckt and hamme, 2023). Collecting and annotating medical audio is labor-intensive, requires specialized medical expertise, and is often infeasible because of privacy constraints.

To address these issues, we propose MeSR, an unsupervised online continual learning framework that enables medical ASR systems to adapt to new data and domains without requiring labeled examples or storing original datasets. MeSR builds on Whisper (Radford et al., 2023) and incorporates several key advantages:

- No labeled data required: Transcriptions with confidence scores are generated automatically and used directly for training, eliminating costly manual labeling.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track), pages 330–337
March 25-27, 2026 ©2026 Association for Computational Linguistics

medical terms	medication names	symptoms
echocardiogram	Acetaminophen; Hydrocodone Bitartrate	fatigue
biopsied	Bacitracin; Neomycin; Polymyxin B	chronic pain
echocardiogram	Cetirizine Hydrochloride	paresthesia
carotid ultrasound	Dicyclomine Hydrochloride	ringing or hissing in my ears
auscultation	Diltiazem Hydrochloride	defecate abnormally

Table 1: Examples of medical terms, medication names, and symptoms

- Adaptive and robust loss function: An adaptive weighted loss enhances model precision and robustness.
- High resiliency through data augmentation: Training data is enriched with diverse voices, accents, and noise, improving performance across speakers and environments.
- Low computational cost: The model updates efficiently in real-time using minimal resources.
- Privacy-conscious design: No audio data is stored; the model learns continuously while respecting privacy constraints.
- Scalability: The combination of unsupervised learning, adaptive training, and data efficiency allows deployment across large-scale medical datasets.

The following of the paper is organized as following. Section 2 detailed the MeSR pipeline. Section 3 shows the experimental results, and Section 4 concludes the paper.

2 Model pipeline

We present MeSR for medical speech recognition. MeSR is an unsupervised continuous learning model designed to adaptively update ASR systems with new data while avoiding catastrophic forgetting (CF). Built on Whisper-large-v2 (Radford et al., 2023) as the base model, the pipeline comprises the following key steps: 1) Generate transcription along with transcription confidence for new unlabeled audio. 2) Apply filtering to only keep highly accurate transcribed data for training. 3) Enhance audio to enrich training data to ensure the robustness of the model. 4) Apply multiple loss functions to preserve model knowledge. 5) Use adaptive training to increase training efficiency and model robustness. These steps are explained in detail in the upcoming sections.

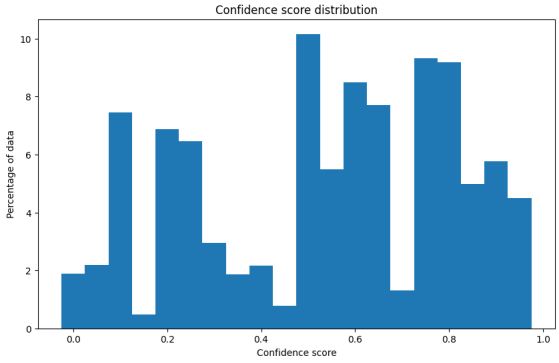

Figure 1: Transcription confidence score distribution.

2.1 Weak supervision signal generation

The model training process does not require labeled data. Instead, it generates transcriptions along with confidence scores, which are used directly during training. This pseudo-labeling approach (Prakash et al., 2025; Zhu et al., 2023) removes the need for costly and time-consuming manual annotation, making it more efficient than conventional Whisper fine-tuning methods (Cheng, 2023).

Iterative pseudo-labeling (Wang et al., 2022; Fan et al., 2023; Likhomanenko et al., 2021) has recently shown promising results in speech recognition, where model-generated transcriptions are refined over multiple rounds of training to progressively improve quality (Xu et al., 2020; Meng et al., 2025). This line of work represents an important direction for future research. However, in medical online continual learning settings, there are additional constraints—models must operate with low latency, and storing user data for repeated training passes is generally prohibited due to privacy concerns. As a result, a one-pass pseudo-labeling strategy is more appropriate in this context.

For each new unlabeled audio input, MeSR generates transcriptions that serve as weak supervision for further training. In addition, the model estimates transcription confidence through the following steps:

Extract logits: The Whisper model outputs logits,

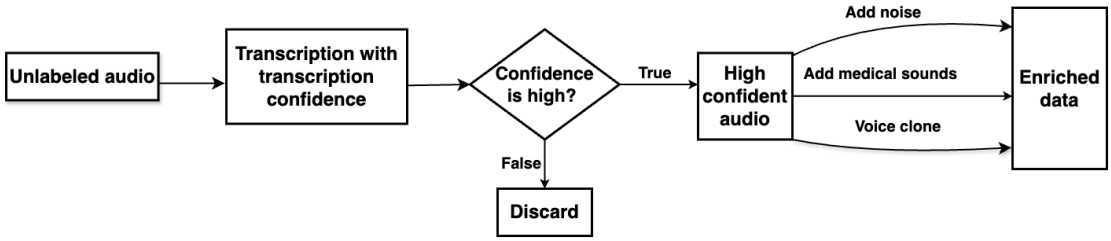

Figure 2: The data augmentation pipeline to enhance the richness of the training dataset. Given an audio sample, we first process it through an ASR model to obtain a transcription along with a confidence score. If the confidence score is high, we enrich the dataset by augmenting the audio with noise, medical sounds, and voice cloning techniques.

which represent unnormalized token probabilities. Higher logits indicate greater confidence in the predicted token.

Convert logits to probabilities: The logits are converted into probabilities using the softmax function. These probabilities reflect the model's confidence in selecting each token.

Calculate overall transcription confidence: The overall confidence score is computed by averaging the confidence values across all tokens in the transcription.

2.2 Data filtering

After getting the transcriptions with confidence, to maintain high-quality training data, we only keep the high-confidence pairs (audio + transcription) for model training. To do this, we established a statistically robust confidence threshold. Only transcriptions surpassing this threshold are included in the fine-tuning process, ensuring the model learns from reliable data. The threshold was chosen in the following way. First, we used the base model to transcribe 22k medical audios (with ground-truth transcriptions) to get the transcriptions and the transcription confidence C. The confidence score distribution is shown in Fig. 1. Then we calculated the transcription word error rate (WER) E for each audio. After that, we looked at the relationship between E and C. We set up this requirement:

$$If \quad C > t_c, \quad then \quad E < t_e, \tag{1}$$

where t_c is the lower threshold for confidence, and t_e is the upper threshold for the WER. Since we only want to use highly accurate transcriptions as training data, we chose t_e to be a very small number (e.g. $t_e = 0.001$ and this can be adjustable for different applications dependent on the application requirement for the transcription accuracy). To choose t_c, we look at every value from 1 to 0, with step size 0.000001, and we chose the smallest value that meets the requirement in Eqn. (1) for at least 95% the 22K medical audios. This means that if the confidence is above t_c, then the transcription error is below t_e for at least 95% the 22K medical audios. This is a very strict filter to ensure the selected data for fine tuning has high-quality and this is feasible given the large volume of real-life medical audio data.

2.3 Data augmentation

State-of-the-art ASR models are highly sensitive to noise, where even minor noise can lead to significantly different transcriptions. To ensure the model produces stable and consistent outputs despite variations in input, we enhanced the training process using three key augmentation strategies: noise injection, voice cloning (Qin et al., 2023), and the addition of medical sound effects. These techniques allow us to augment the training dataset while maintaining consistent transcriptions between the original and augmented data, reinforcing the model's robustness. The data augmentation pipeline is shown in Fig. 2 and detailed explanations are below.

Noise injection: We incorporated different types of noise into the audio inputs, including Gaussian noises and environmental noises commonly found in medical settings. Gaussian noises are random noises generated from the Gaussian distribution. We used 0 as the distribution mean and used different deviations as in (0.001, 0.002, 0.003, 0.004, 0.005, 0.1, 0.2, 0.3, 0.4, 0.5) to generate noises at different levels. The environmental noises include the beep of a heart monitor, the sound of a ventilator, and the noises of blood pressure monitors etc, simulating the acoustic environment of a doctor's office.

Voice cloning: To enrich the dataset with diverse accents and speech characteristics, we recorded the voices of coworkers of various genders and accents. Using voice cloning techniques (Qin et al.,

2023; Shen et al., 2018), we replaced the original voices in the training audio with these diverse voice profiles, creating a more inclusive dataset.

Medical sound effects: We specifically added over 400 types of medical sound effects [1] (e.g. medical ambulance siren, cough voices) to the training data, covering a wide range of scenarios encountered in real-world medical environments. These additions mimic conditions that ASR systems might face in production, such as overlapping sounds and background chatter.

These data manipulations not only improve the robustness of the model but also ensure that it can adapt to real-world medical audio conditions, reducing its sensitivity to noise and improving transcription accuracy.

2.4 Training with adaptive loss for knowledge preservation

To prevent the model from losing knowledge of previously learned tasks while leveraging the richness and reliability of new data, we implemented a reliable loss function based on Elastic Weight Consolidation (EWC) (Kirkpatrick et al.). EWC is a well-established technique to mitigate catastrophic forgetting by preserving critical weights associated with earlier tasks, ensuring the model retains its prior knowledge while adapting to new tasks or data. The EWC loss works by adding a penalty term to the loss function, which constrains the changes to weights deemed important for previously learned tasks. This penalty is determined using the Fisher Information matrix (Ly et al., 2017), which identifies crucial parameters that should remain stable during fine-tuning.

To further improve the reliability of the training process, we integrated a weighting mechanism into the EWC loss. Each training sample is assigned a weight proportional to its transcription confidence score. Samples with higher confidence scores exert a stronger influence on the loss, encouraging the model to prioritize learning from highly reliable data. The loss function is the following.

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda C \sum_i F_i(\theta_i - \theta_i^*)^2 \qquad (2)$$

where $\mathcal{L}_{CE}$ is the cross entropy loss (Mao et al., 2023). i is index over model parameters. θ_i is current model parameters. θ_i^* is previous optimal model parameters. F_i is Fisher Information matrix diagonal values. They represent the amount of information a random variable carries about each individual parameter in the model, essentially indicating how much a small change in that parameter affects the observed data distribution, thus capturing the importance of each parameter. C is the transcription confidence and λ is the regularization strength, and together they two control the balance between new learning and knowledge retention. Higher C means the previous model transcribe more correctly, therefore it gives higher penalty if the current model diverges more from the previous model.

This adaptive weighting approach enhances the overall robustness and precision of the model, especially in domains like medical ASR, where accuracy is critical. This combination of EWC and confidence-based weighting ensures the model achieves a balance between retaining prior knowledge and effectively incorporating new, reliable data, thus maintaining its performance across diverse and evolving datasets.

2.5 Adaptive training

Since efficiency is one major key in medical continual learning, we propose an adaptive training methodology designed to balance the need for model optimization with computational efficiency, ensuring that high-confidence data is utilized most effectively throughout the training process. In the Whisper model (Radford et al., 2023), the early layers are responsible for extracting a general understanding of the audio, while the later layers specialize in adapting the model for specific tasks. To improve training efficiency, we implement a two-tier approach that divides the training dataset into two distinct groups, each with a tailored training strategy. These strategies include fine-tuning the entire model as well as selectively fine-tuning only the later layers. The adaptive training pipeline is illustrated in Fig. 3 and explained in the following.

2.5.1 Fine-tuning the entire model

For this stage, we leverage audio-transcription pairs with exceptionally high confidence scores (typically those above a threshold of 0.996, which corresponds to the top 1% of transcription confidence, based on the 22k audios). These pairs are used to fine-tune the entire model, ensuring that the model learns from the most accurate and reliable data. The high-confidence threshold can be adjusted according to the specific requirements of the task at hand,

[1] https://www.soundsnap.com/tags/doctors_office

WER Whisper	EWER Whisper	WER MeSR$_{200k}$	EWER MeSR$_{200k}$	WER MeSR$_{400k}$	EWER MeSR$_{400k}$	WER MeSR	EWER MeSR
3.13	41.10	2.91	38.89	2.80	37.72	2.11	27.32

Table 2: Comparison between Whisper and MeSR. The metrics are WER, and EWER for medication names. For MeSR, 200k and 400k indicate the number of training examples used. The last two columns present our final results obtained by continually training the model on 2.1 million examples.

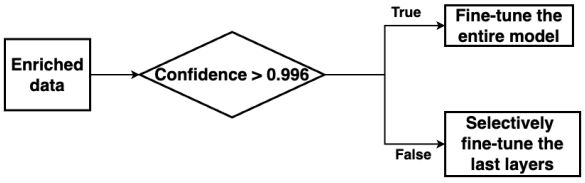

Figure 3: The adaptive training pipeline. Samples with exceptionally high confidence scores will be used to fine-tune the entire model. The rest samples will be used to fine-tune only the last k layers, where k is a randomly chosen integer between 1 and 5.

allowing flexibility in the selection of training data.

2.5.2 Selective fine-tuning of later layers

For the remaining high-confidence audio-transcription pairs, we adopt a more focused approach, fine-tuning only the last k layers of the model. Here, k is a randomly chosen integer where $1 \leq k \leq 5$, ensuring that only the layers responsible for task-specific adaptations are adjusted. This targeted fine-tuning allows the model to refine its capabilities for specific tasks while maintaining the more general audio understanding embedded in the earlier layers. This strategy helps to preserve the broader functionality of the model while optimizing its efficiency and performance for specialized tasks.

3 Experimental Results and Ablation Studies

For model training and evaluation, we used a real-world production dataset of doctor–patient conversations collected over the course of 6 months from multiple clinics. The training set contains 2.1 million medical audio recordings representing a wide range of accents. The testing set consists of 6,000 medical audio clips, each paired with ground-truth transcriptions. All experiments were conducted on eight NVIDIA A100 GPUs with 80 GB of memory each.

The audio samples range from 5 seconds to 10 minutes and span a wide variety of clinical content, including symptoms, procedures, and medication names. We evaluated model performance using two metrics: the overall Word Error Rate (WER), and the Entity Word Error Rate (EWER), which focuses specifically on medical terms. EWER quantifies transcription accuracy at the entity level. For example, if a medical term contains three words (e.g., Ethinyl Estradiol Norethindrone) and only two are transcribed correctly (e.g., Ethinyl Estradiol), the EWER for that entity would be $\frac{1}{3}$. We used the implementation from [2] to compute WER and our internal tool to calculate EWER.

Data	Whisper	Proposed
test-clean	2.71	2.03
test-other	4.96	3.52

Table 3: WER of Whisper and the proposed model on the LibriSpeech test-clean and test-other datasets.

W1	E1	W2	E2	W3	E3
2.93	37.06	2.73	36.75	2.76	36.84

Table 4: W1 and E1: the WER and EWER of removing the data augmentation. W2 and E2: the WER and EWER of removing the confidence-weighted loss. W3 and E3: the WER and EWER of disabling selective fine-tuning.

To choose the hyperparameter λ in Eqn. (2), we tried all values in $[0.1, 0.3, 0.5, 0.8, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30]$ to see which one gives the lowest WER and EWER. We randomly selected 6000 audios from the training data and trained the model for each λ. We evaluated the trained models on the testing set, and found $\lambda = 1$ gave the best performance. Furthermore, using the filtering criteria defined earlier in **subsection 2.2**, we found that 4. 5% of the audios were selected for training.

The results are presented in Table 2. Compared to the Whisper, the MeSR model yields substantial gains, reducing the overall WER from 3.13% to 2.11% and lowering the EWER from 41.10% to

[2] https://github.com/jitsi/jiwer

WERLoRA	EWERLoRA	TLoRA(h)	WERIter	EWERIter	TIter(s)	WER	EWER	T(h)
2.93	37.06	3.2	2.75	29.84	9.7	2.11	27.32	1.6

Table 5: WERLora and EWERLora: the results of fine tuning the whole model with LORA. WERIter and EWERIter: results of iterative pseudo-labeling. TLora(h), TIter(h) and T(h): time (in hours) consumed of the whole pipeline using LORA fine tuning the entire model, iterative pseudo labeling, and proposed method.

27.32%. Intermediate results after training on 200k and 400k audio clips are also included in Table 2, showing a consistent trend: as MeSR is trained on more data, both WER and EWER continue to decrease. This suggests that, MeSR can further improve over time as additional training data becomes available (Indeed the model has been deployed in production and we have seen decreasing WER and EWER).

To assess generalization beyond the medical domain, we evaluated the model — after continual training on 2.1 million medical audio samples — on the open sourced LibriSpeech (Panayotov et al., 2015) test-clean and test-other subsets. As shown in Table 3, MeSR achieves lower WER than Whisper on both datasets. The improvement is especially notable on the more challenging test-other set, which contains noisier and more accent-diverse speech. This is likely because the medical training data includes a substantial number of non-medical words, enabling the model to enhance its performance on general speech as well.

To evaluate the contribution of each core component in our proposed model, we conducted comprehensive ablation studies focusing on three aspects: data augmentation, adaptive loss, and adaptive training. Specifically, we applied the following interventions individually: (1) removing all data augmentation strategies, (2) replacing the confidence-weighted loss in Eqn. 2 with a standard alternative (EWC loss), and (3) disabling selective fine-tuning by updating all layers uniformly across the datasets. The results, summarized in Table 4, indicate that each component plays a meaningful role in improving overall performance.

We further compared both accuracy and efficiency under two additional settings: (1) applying iterative pseudo-labeling to train on unlabeled audio (Xu et al., 2020; Meng et al., 2025), and (2) using selective fine-tuning versus fine-tuning the entire model with LoRA (Hu et al., 2022). As shown in Table 5, full-model LoRA fine-tuning resulted in lower overall accuracy (WER) and incurred higher computational cost. Iterative pseudo-labeling achieved comparable performance but required more training time, making one-pass pseudo-labeling more practical in our setting.

We evaluated our proposed method on several open-source datasets, including LibriSpeech (Panayotov et al., 2015), Common Voice (Ardila et al., 2020), Artie (Meyer et al., 2020), and CORAAL (Kendall and Farrington, 2022). For LibriSpeech, we used both the train-clean and the more challenging train-other subsets (500 hours of difficult audio) for model training, while for the other datasets, we utilized their respective training splits. Evaluation on the test subsets shows that our method consistently outperforms the baseline Whisper model, substantially reducing error rates across all datasets (Table 6).

Data	Whisper	MeSR
LibriSpeech-test-clean	2.71	1.50
LibriSpeech-test-other	4.96	2.13
Common Voice	8.85	3.16
Artie	6.18	3.02
CORAAL	16.23	6.35

Table 6: Resuslts (WER) comparison of Whisper and our proposed method on various datasets (test splits).

4 Conclusions

We introduce MeSR, an unsupervised online continual learning framework for medical audio transcription. It improves ASR models continuously without labeled data. No sensitive patient information is stored. MeSR uses diverse data augmentation to increase training variability. This ensures robust and accurate performance. Its adaptive training is efficient and supports real-time updates. The framework overcomes key ASR limitations. It is privacy-conscious, resource-efficient, and continually adaptable. MeSR outperforms existing models like Whisper in accuracy and cost-efficiency. Deployed in production, MeSR has reduced doctors' documentation time. Transcription performance continues to improve with ongoing monitoring. It is a transformative tool for healthcare AI and workflow optimization.

5 Limitations

This work has several limitations. First, the proposed model focuses on augmenting high-confidence examples and fine-tuning on the augmented data, leaving open whether it can improve performance on more challenging, low-confidence cases—which are currently excluded.

Second, only a small portion of the available medical audio data was used, leading to a high discard rate and substantial data inefficiency. Developing methods to better leverage the full dataset—including the discarded audio—remains an important direction for future research.

Third, ASR models may sometimes assign high confidence to incorrect transcriptions, potentially reinforcing errors during fine-tuning.

Fourth, although the model has been deployed in production for eight months and continuously monitored on selected labeled datasets to guard against performance drift, its strong performance may not generalize as more diverse data is encountered over time.

Fifth, the approach relies on a hyperparameter λ that must be tuned with some labeled data; however, in many domains, even obtaining a small labeled set is challenging.

Sixth, the model remains imperfect on rare syndromes, uncommon drug names, and investigational therapies, often generating hallucinated or similar-sounding substitutions.

Lastly, the production data cannot be made publicly available in a short time due to user privacy and organizational policies.

References

Tejumade Afonja, Tobi Olatunji, Sewade Ogun, Naome A. Etori, Abraham Owodunni, and Moshood Yekini. 2024. Performant asr models for medical entities in accented speech.

Harsh Ahlawat, Naveen Aggarwal, and Deepti Gupta. 2025. Automatic speech recognition: A survey of deep learning techniques and approaches.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Heng-Jui Chang, Hung yi Lee, and Lin shan Lee. 2021. Towards lifelong learning of end-to-end asr.

Xin Cheng. 2023. Openai whisper fine-tuning.

Steven Vander Eeckt and Hugo Van hamme. 2023. Rehearsal-free online continual learning for automatic speech recognition. In *Interspeech*, pages 944–948.

Steven Vander Eeckt and Hugo Van hamme. 2024. Unsupervised online continual learning for automatic speech recognition.

Shiyu Fan, Nurmemet Yolwas, Wen Li, and Jinting Zhang. 2023. Iterative pseudo-labeling methods for improving speech recognition.

Mohammed Fouda. 2024. How ai is mitigating look-alike, sound-alike medication errors.

Li Fu, Xiaoxiao Li, Libo Zi, Zhengchen Zhang, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2021. Incremental learning for end-to-end automatic speech recognition.

A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech*, page 5036–5040.

W. Han, Z. Zhang, Y. Zhang, J. Yu, C.-C. Chiu, J. Qin, A. Gulati, R. Pang, and Y. Wu. 2020. Contextnet: Improving convolutional neural networks for automatic speech recognition with global context. In *Interspeech*, page 3610–3614.

B. Houston and K. Kirchhoff. 2020. Continual learning for multidialect acoustic models. In *Interspeech*.

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 (ICLR)*.

Nazmul Kazi, Matt Kuntz, Upulee Kanewala, and Indika Kahanda. Dataset for automated medical transcription.

Tyler Kendall and Charlie Farrington. 2022. Coraal - online resources for african american language. Accessed: 2025-09-30.

Samuel Kessler, Bethan Thomas, and Salah Karout. 2021. Continual-wav2vec2: an application of continual learning for self-supervised automatic speech recognition.

James Kirkpatrick, Razvan Pascanu, Joel Veness Neil Rabinowitz, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks.

Tatiana Likhomanenko, Qiantong Xu, Jacob Kahn, Gabriel Synnaeve, and Ronan Collobert. 2021. slimipl: Language-model-free iterative pseudo-labeling.

Alexander Ly, Maarten Marsman, Josine Verhagen, Raoul Grasman, and Eric-Jan Wagenmakers. 2017. How ai is mitigating look-alike, sound-alike medication errors.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. 2023. Cross-entropy loss functions: Theoretical analysis and applications.

Qingliang Meng, Hao Wu, Wei Liang, Wei Xu, and Qing Zhao. 2025. Ilt-iterative lora training through focus-feedback-fix for multilingual speech recognition.

Josh Meyer, Lindy Rauchenstein, Joshua D. Eisenberg, and Nicholas Howell. 2020. Artie bias corpus: An open dataset for detecting demographic bias in speech applications. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*, pages 6462–6468.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *ICASSP*.

Jeena Prakash, Blessingh Kumar, Kadri Hacioglu, Bidisha Sharma, Sindhuja Gopalan, Malolan Chetlur, Shankar Venkatesan, and Andreas Stolcke. 2025. Better pseudo-labeling with multi-asr fusion and error correction by speechllm.

Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. 2023. Openvoice: Versatile instant voice cloning.

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 40 th International Conference on Machine Learning*.

Brooke Schmidt. 2010. Look-alike drug name errors.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Yannis Agiomvrgiannakis, and Yonghui Wu. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *ICASSP*, pages 4779–4783.

Mengqian Wang, Ilya Valmianski, Xavier Amatriain, and Anitha Kannan. 2022. Learning functional sections in medical conversations: iterative pseudo-labeling and human-in-the-loop approach.

Qiantong Xu, Tatiana Likhomanenko, Jacob Kahn, Awni Hannun, Gabriel Synnaeve, and Ronan Collobert. 2020. Iterative pseudo-labeling for speech recognition.

M. Zeineldeen, A. Glushko, W. Michel, A. Zeyer, R. Schluter, and H. Neyy. 2021. Investigating methods to improve language model integration for attention-based encoder-decoder asr models. In *Interspeech*, page 2856–2860.

Han Zhu, Dongji Gao, Gaofeng Cheng, Daniel Povey, Pengyuan Zhang, and Yonghong Yan. 2023. Alternative pseudo-labeling for semi-supervised automatic speech recognition.