ABSTRACT

Title of Dissertation: ADVANCING SPEECH RECOGNITION FOR LOW-RESOURCE DOMAINS: A CASE STURY I EDUCATIONAL AND CLASSROOM SPEECH

Ahmed Adel Attia
Doctor of Philosophy, 2026

Dissertation Directed by: Prof. Carol Espy-Wilson
Department of Electrical And Computer Engineering

Automatic Speech Recognition (ASR) technology has significantly advanced in recent years, largely driven by innovations in Artificial Intelligence (AI) and transformer-based models. ASR models' performance, like all AI models, is heavily data-dependent, causing them to underperform in many low-resource settings. Many critical applications remain data scarce and low-resource for various reasons, including annotation costs, privacy preservation concerns, and challenging acoustic conditions. Classrooms and other educational settings exemplify such domains with a large potential for useful speech AI applications that is largely unrealized due to the adverse acoustic environment and the data scarcity in that field. ASR has the potential to support inclusive, adaptive learning by providing real-time transcription for students who are hard of hearing or English Language Learners (ELLs) and by offering educators valuable insights into classroom dynamics. Classrooms represent a particularly challenging low-resource English domain, presenting unique complexities, including high variability in speech patterns, disfluencies,

overlapping conversations, and background noise, which often cause ASR models to perform poorly, underscoring the need for specialized approaches in educational ASR.

This dissertation examines data-centric and training-centric approaches for enhancing ASR under adverse low-resource conditions, utilizing classroom speech as a motivating and representative application domain. We study the adaptation of transformer-based ASR models, including Whisper and Wav2vec2.0, to children's speech and classroom environments. We start by curating and analyzing public and otherwise available children and classroom speech datasets, introducing filtering, preprocessing, and error analysis techniques that improve their utility for model training and evaluation.

The majority of research in low-resource speech models is concerned with low-resource languages, where the primary challenge lies in adapting to novel linguistic structures that differ substantially from English and other high-resource languages used in large-scale pretraining. Classroom speech, by contrast, presents a different type of low-resource problem. The linguistic structure remains within English, a high-resource language, but the domain is characterized by highly adversarial acoustic conditions, including multi-speaker overlap and persistent background noise. Although English itself is high-resource, the classroom domain is low-resource due to limited publicly available data.

To this end, we start with a proposed adaptation technique through adapting the acoustic models through Continued Pre-training (CPT) on unlabeled in-domain audio. By adapting existing speech embedding models using unlabeled in-domain data. This approach demonstrates significantly improved robustness in multi-speaker, noisy conditions using real classroom datasets, even under limited labeled data.

We follow this work by introducing techniques to leverage weak, inaccurate, and imprecise

transcription. This approach stems from a practical need to increase the utility of available data and reduce the demand for high-cost, accurate verbatim transcriptions. We propose Weakly Supervised Pre-training (WSP), a training paradigm that utilizes large amounts of weak transcriptions as an intermediate supervision step prior to supervised fine-tuning. We validate this approach both with synthetically corrupted transcripts and real weak transcripts from classroom datasets, and under both tests, WSP shows significant improvement, which increases the utility of limited gold-standard transcripts.

A parallel approach to WSP that is also concerned with improving the acoustic modeling capacity of the model and increasing the utility of limited datasets is data simulation. Recent speech models have utilized AI to produce generative speech from text corpora. This approach, while effective, is expensive and is bottlenecked by the naturalness and variability of Text-To-Speech (TTS) models. We propose a simulation-based approach that uses available natural speech and simulates the acoustic environment itself. This approach's novelty lies in its usage of game engines to recreate the acoustic environments, allowing for fast, cheap, and scalable data simulation. Using these virtual environments, we simulate babble noise and capture Room Impulse Responses (RIRs) in simulated classroom environments. Paired with semantically-paired adult and child educational speech from different datasets, we create RealClass, the first and largest shareable classroom speech corpus. Benchmarks on this dataset show that it outperforms off-the-shelf speech datasets on classroom test data and works best when paired with real classroom speech, increasing the utility of available limited data.

Our final contribution to improve the acoustic modeling of ASR models is concerned with the reintegration of speech articulatory parameters in ASR. Mainly, we add an auxiliary acoustic-to-articulatory Speech Inversion (SI) task, which predicts the speech articulatory parameters from

the raw speech signal, into the ASR model. This approach significantly improves the performance of ASR models under low-resource conditions.

Finally, we compare our approaches to emerging technologies, namely Large Language Model (LLM) driven ASR models. Our benchmarks highlight the great potential of prompt-based contextual biasing as a cheap and effective adaptation technique. We also show that the contributions of this dissertation, applied to a smaller, acoustic ASR model without LLM components, outperform LLM-based models, even with contextual biasing and supervised fine-tuning on in-domain classroom data. This finding underscores that even with the superior linguistic modeling capacity of LLM-based approaches, improved acoustic modeling is more effective in acoustically adversarial environments. Collectively, these contributions advance the understanding and practical performance of ASR systems under data scarcity, offering generalizable methodologies for low-resource speech recognition beyond educational domains.

ADVANCING SPEECH RECOGNITION FOR LOW-RESOURCE DOMAINS: A CASE STUDY IN EDUCATIONAL AND CLASSROOM SPEECH

by

Ahmed Adel Attia

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2026

Advisory Committee:
Professor Carol Espy-Wilson, Chair/Advisor
Professor Dinesh Manocha
Professor Sanghamitra Dutta
Professor Shihab Shamma
Professor Ramani Duraiswami (Dean's Representative)

Dedication

To Hind Rajab, a five-year-old girl whose life was senselessly taken. Her voice and her legacy stand as a reminder of the countless children denied the chance to learn, play, and grow up in a world free from bombs, hunger, and displacement, or even to simply grow up at all.

To all those whose lives were taken before they could truly begin.

May this work be a small step toward a world where every child can learn and thrive in safety and hope.

Acknowledgments

"He who is not grateful to people, is not grateful to God." —Prophet Muhammad

The wisdom of this saying resonates deeply with me at the culmination of this Ph.D. journey. I would be remiss to consider this achievement solely the result of my own efforts and skills, rather than the generosity of the many people who lent me their expertise, advice, wisdom, or simply a caring and listening ear.

Words cannot describe my deepest and most sincere gratitude to Professor Carol Espy-Wilson. Carol has shown both great leadership and friendship over the past five years. She is an exceptional scientist whom I am incredibly lucky to have the privilege to study under, and a true, caring, and compassionate mentor. It is thanks to her that I am the researcher that I am today.

I sincerely thank Professor Shihab Shamma, Professor Sanghamitra Dutta, Professor Dinesh Manocha, and Professor Ramani Duraiswami for contributing their time and expertise by serving on my committee and for their feedback. I want to thank all of my professors at Alexandria University for teaching me the academic basis this work needed.

I am indebted to Professor Jing Liu, Professor Dora Demszky, Professor Wei Ai, Professor Mark Tiede and Tolúlópé Ògúnrèmí and all the people who collaborated with me, for lending me their expertise and guidance.

My past and present Speech Communication Lab researchers, Nina, Gowtham, Chigozie, Thanushi, Avishka, Shuubham, Saba, Cherif, Yashish, Nadee, and Rahil, have enriched this research and deserve a special mention.

Being an international student, I felt that home is much closer thanks to my friends at Maryland, Ahmed Ashry, Ahmed Atwa, Faisal Hamman, Mohamed Elnoor, Hazem Al-Bulqini, Marawan Gomaa, and Cherif Aidara. I am also thankful to my closest friend from Egypt whom distance did not make far away, Ahmed Abdelkader, Ahmed Fakhry, Ahmed Draz, Akram Hesham, Ahmed Hamza, Moemen Amin, and Moaz Nassar.

Finally, I am incredibly thankful to my infinitely patient and most loving wife, Nada. This is her achievement and her moment as well. She has given up so much so that we can reach this point together. Being married to a Ph.D. student, especially an international one away from family and friends, is hard, but her grace, patience, and mere presence have made this journey much easier. I also want to thank my son, Yasseen. Being your father is my greatest achievement, and it has driven me greatly to be someone you can be as proud of as I am proud of you. I want to thank my parents, Adel Attia and Hanan Badr, who have instilled in me the importance of science and education and have guided me and started this journey with me since day one, before anyone else. I am grateful to my sister and brother, Mennatallah and Mohammed, who set the standard that I aim to follow every day.

In thanking all of these people, I am first and foremost thanking God, who placed them in my path, granted me the strength and patience to see it through, and grounded me in faith when it became difficult.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The past few years have witnessed a rapid development of large AI models and architectures that can leverage enormous amounts of data that allow such models to achieve competitive results. Chief among these novel architectures is the Transformer [Vaswani, 2017], which shows competitive performance in natural language tasks by leveraging large datasets.

In the field of Automatic Speech Recognition (ASR), two main approaches have been shown to achieve competitive results. The first is supervised learning, where models are trained on large corpora of transcribed speech [Radford et al., 2023]. The second approach is self-supervised learning (SSL), which pre-trains the model first on untranscribed speech to learn contextual speech representation, which can be later fine-tuned on a smaller transcribed dataset for ASR [Baevski et al., 2020]. Even though off-the-shelf models from either family approach human-level performance in clean English speech, they underperform on low-resource tasks, like children's speech, especially in classrooms.

1.1 Motivation

The development of ASR systems that are robust to children's speech and classroom conditions is a critical part of the development of educational AI tools for classrooms. These tools can utilize sufficiently accurate classroom transcriptions to aid teachers and students in the teaching

and learning process. For example, these transcripts can aid hard-of-hearing students or English Language Learners (ELL) get involved in the classroom and improving their learning outcomes. Additionally, classroom transcripts generated by ASR can be analyzed to understand the dynamics of the classrooms and provide automated feedback for teachers to help them improve their techniques[Demszky et al., 2023, Kraft et al., 2018, Link and Guskey, 2022].

1.2 Barriers to Effective ASR in Children's Speech and Classroom Settings

1.2.1 Challenges Facing Children ASR

State-of-the-art ASR models struggle with children's speech, even in clean conditions. These models are mainly trained on publicly available datasets which mainly comprise adult speech. However, children's speech is distinct from adult speech in many ways.

Children generally speak less clearly than adults [Lee et al., 1999] and the basic acoustic and linguistic characteristics of children's speech differ from that of adults [Gerosa et al., 2009]. We showcase some examples of the unique linguistic characteristics of children's speech in [Attia et al., 2023] and later in Chapter 3. We mainly show how children tend switch topics in the middle of sentences. Additionally, we show how disfluencies are more common in children speech. Children's speech exhibits more intra-speaker variability due to underdeveloped pronunciation skills [Koenig and Lucero, 2008, Lee et al., 1999, Smith, 1992, Vorperian and Kent, 2007], or inter-speaker variability due to varying developmental rates and diverse cultural backgrounds. In the United States, a significant percentage of students are classified as ELLs. ELLs make up 18% in California, 20% in Texas, 10% nationally of the total student population.

1.2.2 Challenges of Classroom Environments

Classrooms are by nature noisy environments. In the US, the average number of students per classroom is around 20 [NCES, 2018]. This results in the most prevalent type of noise in classrooms being babble noise, which is the noise from overlapping background speech. Babble noise is considered one of the most challenging noises, even in adult English speech with adult babble [Simic and Bocklet, 2024]. However, in classrooms, babble noise is a result of overlapping children's speech, which makes it a distinct problem from adult babble, as it is unlikely to exist in public datasets used to train these models. The severity of babble noise is proportionate with the number of students in the classroom which disproportionately affects overcrowded classrooms.

Additionally, classrooms are a multi-speaker environment by nature, where the target speech can come from an unpredetermined number of speakers, possibly overlapping with each other. Multi-speaker ASR is an open area of research in and of itself, and a wide gap still exists between single and multi-speaker ASR [Chang et al., 2019, 2020].

Another factor that affects the quality of classroom ASR is far-field speech, for example, if a student speaks while being far from the microphone. Far-field speech is another open area of research, and previous research has shown recognition to be a much more difficult task than in the case of near-field speech [Zhu et al., 2024].

1.2.3 Data Scarcity

All the previous challenges are compounded by the fact that transcribed children's speech, clean or within classrooms, are scarce, and those that exist are not public, partially due to the sensitivity of data from minors.

Chapter 2: Automatic Speech Recognition

An ASR system is a system that transcribes the acoustic speech signal into word sequences. Classically, the best-performing ASR systems employed Hidden Markov Models (HMMs), but with the advent of deep learning, the best-performing systems comprised neural network architectures. More recently, Transformers have allowed ASR systems to achieve impressive results due to their ability to contextualize information over longer sequences.

2.1 Transformers

Transformers Vaswani [2017] were originally designed for natural language tasks. The key behind the transformer's strong performance is the attention mechanism which allows efficient parallelization and scalability for large datasets.

2.1.1 Attention

Transformers employ two kinds of attention layers, self-attention and cross-attention. In self-attention, the layer captures the relevance of each part of the input sequence in relation to other parts of the sequence. Cross-attention captures the relevance of each part of one sequence in relation to another sequence. In the task of language translation which the model was first used, simplified, self-attention would capture the relevance between each word and other words within

an English sentence, while cross-attention would capture the relevance between each word in an English sentence and the words in its translation.

In self-attention, the model extracts different representations of the input sequence, namely Queries Q, Keys K, and Values V. Formally,

$$Q = XW_Q \tag{2.1}$$

$$K = XW_K \tag{2.2}$$

$$V = XW_V \tag{2.3}$$

Where Q, K, and V are the query, key, and value matrices, X is the input activation, and W_Q, W_K, and W_V are the learned weight matrices for queries, keys, and values, respectively.

Attention is then calculated using the following equation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.4}$$

Where d_k as the dimensionality of the key vectors used for scaling.

From equation 2.4, we can see that attention is a weighted sum of the values, where the weights are computed by the inner product of the keys and queries which acts as a compatibility function.

Figure 2.1 shows the attention from select heads in a self-attention layer in the transformer model introduced in Vaswani [2017]. In some multi-head attention layers, different heads capture different lingustic structures. In this particular examples, it can be seen that certain heads capture the long term dependencies between the verb "making" and the adverb "more" and the adjective

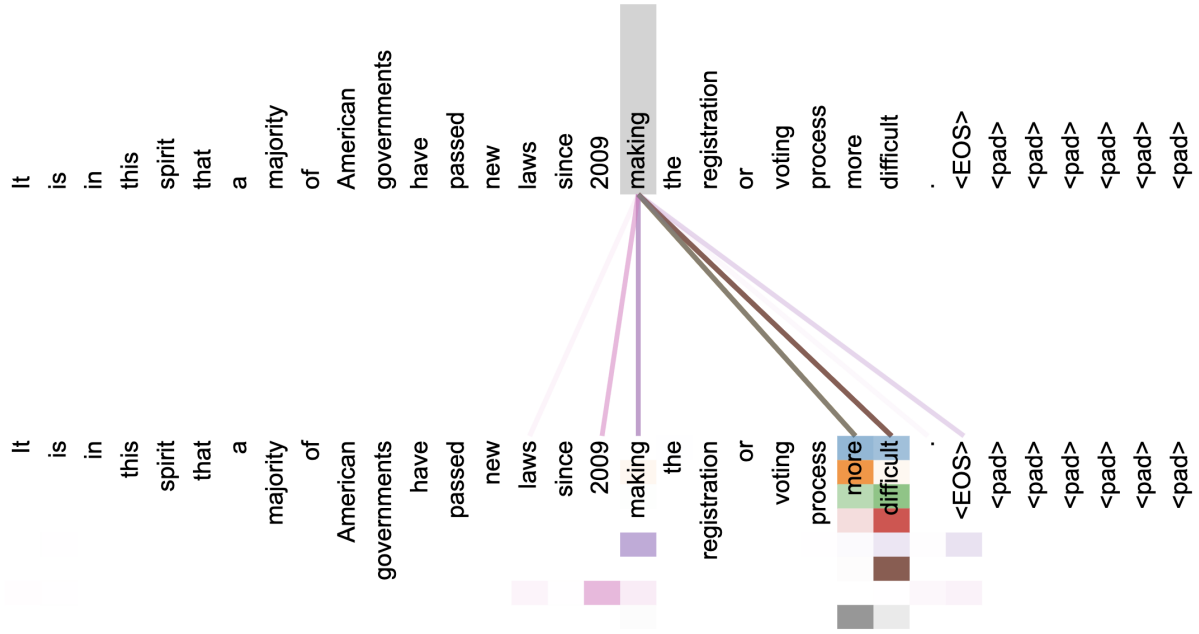

Figure 2.1: An example of the attention mechanism a self-attention in the transformer. Different colors represent different heads. Notice how the attention between the word *making* at the top captures the dependencies in the phrase "*making* ... *more difficult*" at the bottom, by exhibiting high attention with the adjective of the phrase. From Vaswani [2017].

"difficult".

Cross-attention is similar to self-attention in structure, but since it computes the attention between two different sequences instead of within the same sequence, it takes two inputs. The first input is used to compute the queries while the second is used to compute the keys and values. In other words, cross-attention computes the weighted sum of some projection of a sequence, where the weights are based on the compatibility of both sequences together.

2.1.2 Multi-head Attention

Transformers use Multi-Head Attention (MHA) layers. Simply put, instead of calculating attention once using one set of queries, keys, and values, the input is linearly projected h times to produce h sets of queries, keys, and values, and h different attention matrices are computed.

These attentions are then concatenated and projected again.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W_O \tag{2.5}$$

$$\tag{2.6}$$

Where $head_i$ is a single attention head.

2.1.3 Encoder-Decoder Structure

The transformer follows a sequence-to-sequence architecture, with two main components: the encoder and the decoder.

The encoder is made up of several blocks, each containing a self-attention mechanism and a feedforward neural network. The decoder mirrors the encoder but adds an extra component: cross-attention. This layer helps the decoder focus on relevant parts of the input sequence when predicting the next token in the output. It also includes a self-attention layer to ensure that the model pays attention to tokens already generated in the output sequence, which helps maintain coherence.

In addition to the encoder output, the decoder also takes the previously generated tokens as input when predicting the next token in the sequence. This autoregressive nature of the decoder helps it generate sequences step-by-step, allowing it to adapt based on the context of what it has already generated. The decoder's self-attention mechanism ensures that it considers all the tokens generated so far, while the cross-attention layer aligns these tokens with the encoded input sequence, ensuring the output remains coherent and relevant to the input.

2.1.4 Teacher-Forcing

During training transformers for sequence-to-sequence tasks, giving the decoder its previous outputs as an input can increase training time due to the autoregressive nature of inference. Additionally, this may inhibit training, especially in early steps when the decoder output is very inaccurate which can lead to compounding errors. Instead, teacher-forcing feeds the correct previous token from the target sequence from the ground truth.

2.2 Supervised Speech Recognition: Whisper

Transformers have unlocked the potential to utilize large datasets in sequence models. In particular, Whisper [Radford et al., 2023] was trained on 680,000 hours of weakly supervised data by scrapping the internet. Whisper is considered by many to be the state-of-the-art in ASR, as it achieves competitive results in many benchmark datasets. Whisper is also versatile, as it was trained on 96 languages other than English, and on different tasks, including, English-to-English transcription, X-to-English, and X-to-X transcriptions, where X is a language other than English. Additionally, Whisper can be prompted to produce time-stamps along with the transcriptions.

Given the size of the dataset and the method of curation, quality control over the training data was limited. However, the authors report finding subpar transcripts upon manual inspection. Among these are transcripts generated by other, older ASR systems. Training on a mixture of human-transcribed and machine-transcribed examples can impair the model's performance [Ghorbani et al., 2021] and the model learns what's known as "transcript-ese". To remedy that, simple text-based filtration was done to remove machine-transcribed examples, like removing examples without punctuation, or mixed case letters as ASR systems at that time could not handle

these cases. However, modern ASR systems have since learned these grammatical and linguistic structures.

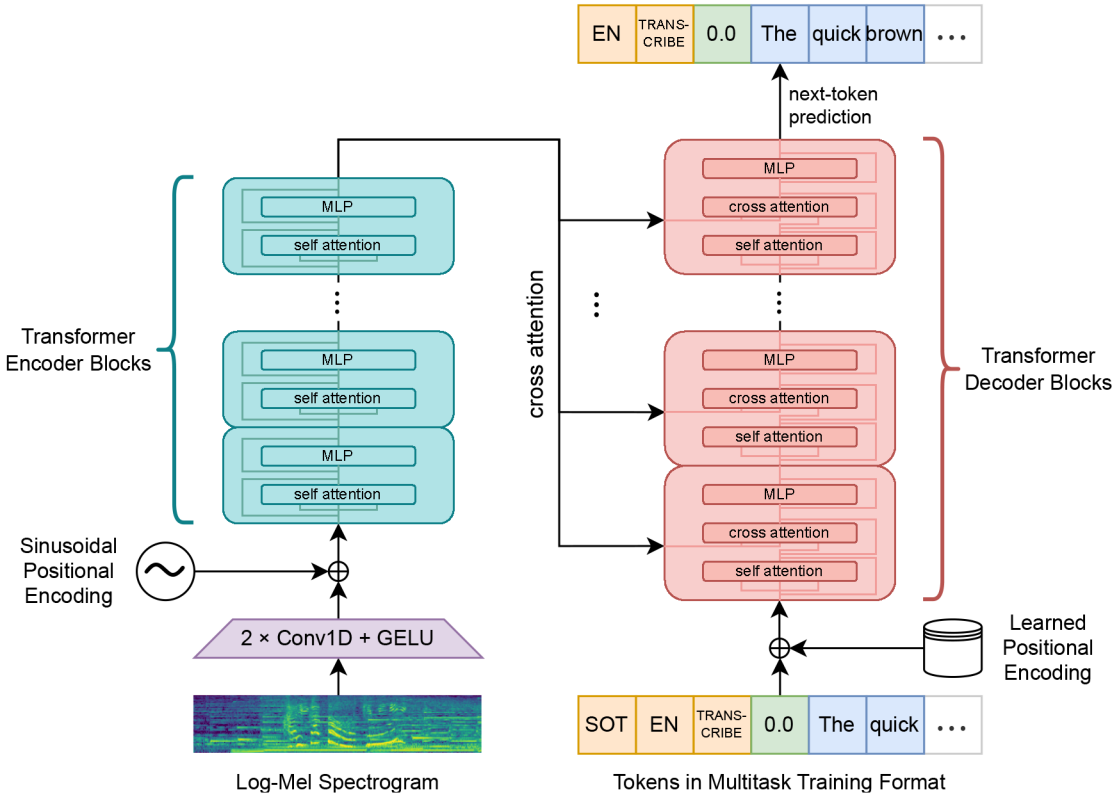

Figure 2.2: Whisper's architecture. The log-mel spectrogram of the audio signal is fed into the encoder, which extracts audio embeddings through a series of self-attention transformer blocks. In the decoder, a shifted version of the target output is fed into a series of cross-attention blocks, as well as the extracted audio embeddings. From Radford et al. [2023].

The main contribution of Whisper is its large scale training. However, the training data is unfortunately not available to the research community, and little information about it is known. To highlight the efficacy of large web-scale weakly supervised datasets, the authors opted for off-the-shelf encoder-decoder transformers. The decoder employs teacher-forcing during training, meaning that the correct transcription, shifted one step into the future, is used as an input, as well as the audio embeddings from the encoder. For this reason, the decoder is considered an

audio-conditional language model, as it takes audio embeddings and natural language as inputs, and outputs natural language. This has been hypothesized to be the main contributor to Whisper's robustness. However, this audio-conditional language model suffers from hallucinations, which is Whisper's main weakness and failure mode.

In deep learning, hallucinations refer to instances where a model generates outputs that are "nonsensical, or unfaithful to the provided source input" [Ji et al., 2023]. In Whisper, hallucinations occur when the ASR model generates speech transcriptions that are entirely fabricated, especially in challenging language environments or when handling low-resource languages. This may be explained as the model decoder receiving noisy or weak audio embeddings, thus relying too much on its own previous text outputs, causing it to act as a language model, producing text unrelated to the audio, or in some cases, repeating a word or a phrase over and over.

Whisper comes in several variations. In terms of size, models were trained, namely tiny, base, small, medium, and large. Additionally, for each of the tiny to medium models, two versions were trained, English-only and multi-lingual, while the large models were only multi-linugual. The English-only models can only transcribe English speech and produce English text, while the multi-lingual models were trained in 96 other languages.

2.3 Self-supervised Speech Representations: Wav2vec2.0

2.3.1 Self-Supervised Learning in Language Modeling

As with most recent advancements in AI, SSL first demonstrated impressive results in language modeling tasks. GPT[Radford and Narasimhan, 2018], for instance, used next-word prediction as a self-supervised pre-training step for its language model. Subsequently, BERT's

[Devlin et al., 2018] SSL pre-training tasks included predicting randomly masked words within a sentence (Masked Language Modeling) and determining whether two sentences follow each other (Next Sentence Prediction). The combination of large training corpora and the transformer's ability to model complex contextual relationships allowed these models to learn rich language representations without requiring extensive human labeling for the pre-training phase. Following pre-training, a fine-tuning step is applied using significantly smaller labeled datasets, offering much better efficiency with labeled data.

2.3.2 Wav2vec

Following the success of SSL in language modeling, Wav2vec[Schneider et al., 2019] and its next iteration Wav2vec2.0 [Baevski et al., 2020] introduced a similar approach to speech tasks. However, predictive training objects like those used in language modeling tasks are less effective in speech due to the continuous and varying nature of the speech signal, unlike text with can be discretized into tokens with inherit structured boundaries. For this reason, Wav2vec uses a contrastive pre-training task.

For the sake of simplicity, we'll stick to discussing the later iteration, Wav2vec2.0. Wav2vec2.0's architecture consists of a convolutional feature extractor that extracts latent representations from raw audio and a transformer network that produces contextual representations of speech. During self-supervised pretraining, the latent representations generated by the convolutional feature encoder are quantized and the unquantized latent features are fed into the transformer, with some frames masked. The constrastive learning objective is to distinguish the quantized representations corresponding to the masked frames from a set of negative distractors. Additionally, a **diversity**

loss is applied to encourage the quantized representations to be as diverse as possible, ensuring the model captures rich, varied speech features.

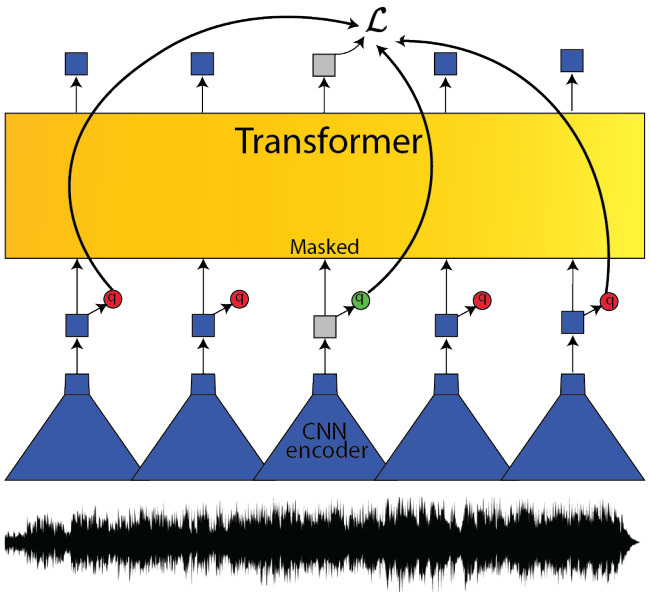

Figure 2.3: Wav2vec2.0 pretraining architecture. Lowercase q in circles represents the quantization networks, with the green circle representing the positive sample and the red circles representing the negative samples. Adapted from Baevski et al. [2020]

2.3.3 ASR Fine-tuning and Connectionist Temporal Classification (CTC) in Wav2vec2.0

Once Wav2vec2.0 is sufficiently pre-trained, it can be **fine-tuned** for automatic speech recognition (ASR) using a smaller labeled dataset. A single randomly initialized classification layer is added on top of the transformer to predict graphemes, and the entire model is trained on this smaller dataset with Connectionist Temporal Classification (CTC) loss [Graves et al., 2006]. This process teaches the model to transform the contextualized speech representations into transcriptions. Crucially, the pretrained Wav2vec model helps reduce the amount of labeled data needed, significantly benefiting ASR tasks in low-resource languages or noisy conditions, where

collecting large labeled datasets is impractical. Fine-tuning on as little as 10 hours of labeled data has shown to achieve strong ASR results, making Wav2vec 2.0 an extremely efficient model for real-world applications.

The strength of Wav2vec 2.0 lies in its **contextual** representations, as the transformer is able to capture long-range dependencies across the input. This allows even a simple classification layer to produce accurate transcriptions by leveraging the broader context around each frame of audio.

2.3.3.1 Connectionist Temporal Classification (CTC)

Connectionist Temporal Classification (CTC) plays a vital role in aligning audio frames with text transcriptions, even when the audio and text sequences differ in length and lack precise alignment. In CTC, each frame can predict either a character or a 'blank' symbol, allowing for flexibility with repeated characters and timing variations. The blank symbol enables the model to adapt to variable-length phonemes, accommodating differences in speaker, speech rate, and phoneme duration. Additionally, the blank symbol serves to represent repeated characters: when a blank appears between identical characters, it prevents them from collapsing into a single character during decoding.

For example, if the target word is *"CAT"* and the model outputs *"—C—AATT—"*, CTC interprets this by collapsing repeated characters and ignoring blanks, yielding the final transcription *"CAT."* Formally, CTC maximizes the probability of the target transcription by summing over all valid alignments of the frame outputs to the target text. Given an input x and target transcription y, the CTC loss is defined as:

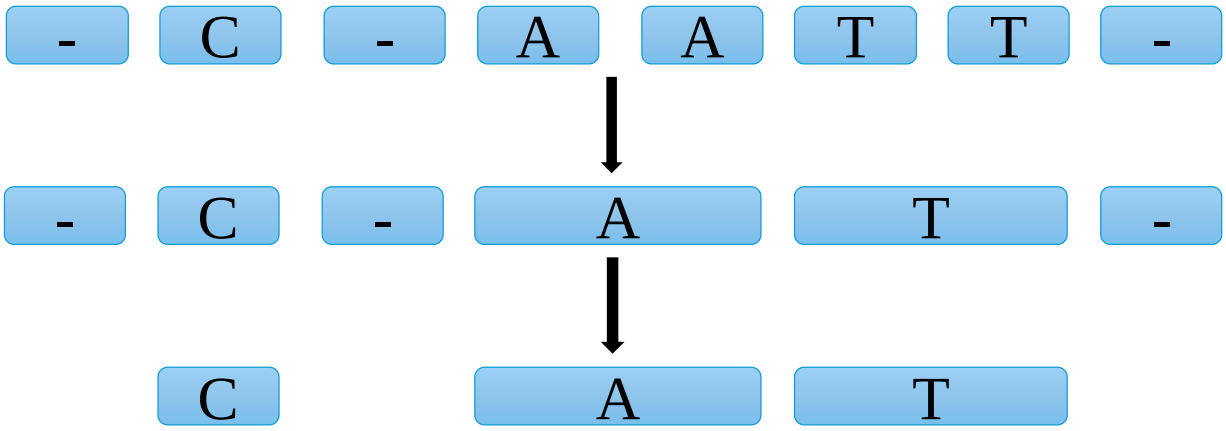

Figure 2.4: CTC decoding of the model output *"—C—AATT—"*.

$$\mathcal{L}_{\text{CTC}} = -\log \sum_{\pi \in \text{alignments}(y)} P(\pi|x) \tag{2.7}$$

Here, π represents all possible alignments of y across the frame sequence of x. The potential alignments for the target *"CAT"* can be large, and the number of combinations will increase exponentially as the length of the input increases. CTC focuses on feasible paths that can collapse to *"CAT"*, using dynamic programming to compute probabilities without explicitly generating every possible alignment.

2.3.3.2 Decoding Methods in Wav2vec2.0

During inference, Wav2vec2.0 employs decoding methods to convert framewise outputs into readable text. **Greedy decoding** selects the most probable character at each frame and collapses the sequence directly, which can lead to errors when contextual information is limited.

To improve accuracy, **Beam-search decoding** with an external language model (LM) is often used during beam search. This LM-assisted decoding balances the probability of acoustic

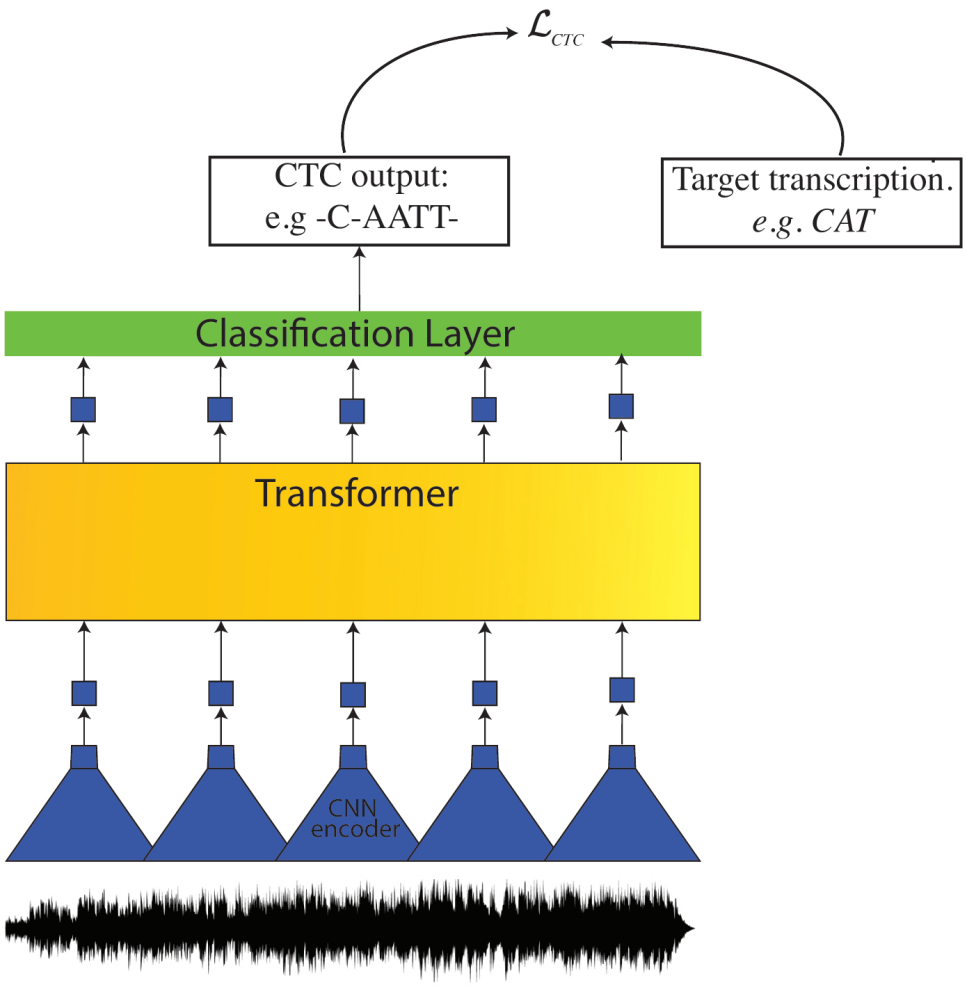

Figure 2.5: Wav2vec2.0 ASR fine-tuning process. A single projection layer is added on top of the Transformer architecture, which maps the learned contextual representations of the audio input into CTC graphemes. This layer enables the model to predict character sequences from the unsegmented audio data during fine-tuning.

matches with the likelihood of coherent word sequences. By weighing sequences based on both framewise probabilities and language coherence, LM-assisted decoding provides more accurate transcriptions, especially for complex or noisy inputs.

2.3.4 Language Models in Wav2vec2.0 Decoding

Language models (LMs) play a crucial role in enhancing the decoding process of Wav2vec2.0 by providing contextual information that helps improve transcription accuracy. Two prominent types of language models used in this context are n-gram models and transformer-based models.

2.3.4.1 N-gram Language Models

N-gram language models are statistical models that predict the likelihood of a word sequence based on the occurrence of n preceding words. For instance, in a bigram model, the probability of a word is based solely on the preceding word, while a trigram model considers the two previous words. This probabilistic approach helps to enforce grammatical structures and common word sequences, aiding the decoding process by reducing ambiguity in transcriptions generated by Wav2vec2.0.

Incorporating an n-gram model during decoding allows the Wav2vec2.0 system to select sequences that are more coherent and linguistically plausible. For example, when faced with multiple possible transcriptions, the n-gram model can evaluate the likelihood of each sequence and guide the decoder towards more probable combinations, thereby enhancing the overall accuracy of the output.

2.3.4.2 Transformer Language Models

Transformer language models, such as BERT or GPT, utilize self-attention mechanisms to capture long-range dependencies and contextual information across the entire input sequence. Unlike traditional n-gram models, transformer models are capable of understanding context at a deeper level, enabling them to provide richer predictions for word sequences.

In the context of Wav2vec2.0 decoding, transformer language models can be integrated into the decoding process to refine predictions based on the broader context of the input speech. By leveraging attention mechanisms, these models can weigh the significance of different parts of the audio input, resulting in more informed transcription decisions. For example, if the model is uncertain between the words "bat" and "cat," a transformer language model can utilize context from the surrounding words to predict the more appropriate choice.

Both n-gram and transformer language models serve to complement the capabilities of Wav2vec2.0, especially in scenarios where the acoustic model faces ambiguity or noisy conditions. By incorporating these language models during the decoding phase, Wav2vec2.0 can produce more accurate and contextually appropriate transcriptions, significantly improving performance in automatic speech recognition tasks. Transformer language models enhance the accuracy of Wav2vec2.0's ASR performance more effectively than n-gram models; however, this improvement comes at the cost of longer decoding times. Both transformer and n-gram models provide greater accuracy than greedy decoding, while n-gram models offer significantly faster decoding with a trade-off in accuracy.

2.3.5 Continued Pre-training of Self-Supervised Speech Models

Continued pretraining (CPT) refers to performing additional self-supervised pre-training on a model that was already pretrained before fine-tuning. Hsu et al. experiments showcase the effect of CPT on domain adaptation. They show that performing CPT with more unlabeled in-domain data and then finetuning on out-of-domain data improves the performance on an in-domain test set.

Several research papers have attempted to investigate the effectiveness of CPT in adapting good multi-lingual self-supervised ASR systems like XLSR53 [Conneau et al., 2020] and XLS-R [Babu et al., 2021]. The research by Nowakowski et al. [2023] aimed to develop an ASR system for Ainu, a critically endangered and low-resource language local to northern Japan, Sakhalin, and Kuril Islands. Starting from XLSR53 which was already pretrained on 56K hours from 53 languages, they performed CPT on 234 hours of Ainu recordings. Then they performed multiple fine-tuning experiments using different subsets of their labeled training data. They describe CPT as "clearly the most effective way to adapt a speech representation model for a new language".

Another interesting study in the field of low-resource languages [San et al., 2024] performed CPT on different languages, including the target language (Punjabi) as well as languages similar and different to it starting from XLS-R 128. Their results supplement previous research and found that CPT on related languages like Hindi outperforms CPT on unrelated languages like Malayalam, Bengali, Odia, or Tamil.

2.3.6 Whisper vs. Wav2vec2.0

Whisper is generally considered to be state-of-the-art in ASR. The massive training data, its multi-lingual capabilities, and the integrated language model allows Whisper to top most performance benchmarks out of the box [Radford et al., 2023]. Wav2vec2.0 however, is a strong contender, especially in low-resource settings. The ability of Wav2vec2.0 to learn using limited labeled data given sufficient self-supervised pre-training makes it a reasonable choice for low-resource tasks.

One important point of distinction between Whisper and Wav2vec2.0 is that Whisper has an integrated audio-conditional language model decoder, while Wav2vec2.0 needs an external language model. This makes Whisper more of an End-to-End (E2E) model than Wav2vec2.0. However, the straightforward integration of language models with Wav2vec2.0 inference especially with simply n-gram language models means that Wav2vec2.0 does not suffer from the same hallucination problems that Wav2vec2.0 suffers from.

Additionally, the stage-based nature of Wav2vec2.0 training allows it to leverage data that Whisper cannot. For example, in domains where limited labeled data exists, but plenty of unlabeled data is available, only Wav2vec2.0 can leverage the unlabeled data as well as the labeled data for domain adaptation.

Chapter 3: Children and Classroom Speech Corpora

Since children are considered a protected class, releasing public datasets including children's speech is complicated as speech is considered Personally Identifiable Information (PII) [Federal Trade Commission, 1998]. Consequently, children's datasets available for training and fine-tuning ASR models are a rare commodity. In this chapter, we overview the datasets available to us during this research.

3.1 Children's Speech Corpora

The datasets we showcase in this section represent children's speech in ideal environments, which can be used to benchmark and train ASR models and to understand children in the absence of noise or other factors that negatively affect the performance of ASR systems. Although some of the corpora are conversational, the adult prompts are not recorded, and no overlapping speech are present in the corpora.

3.1.1 CSLU-Kids Dataset

The CSLU Kids speech corpus contains spontaneous and prompted speech from 1100 children between Kindergarten and Grade 10, with approximately 100 children per grade. In the scripted subset of the dataset, each child was prompted to read from a list of 319 scripts, which can

either be simple words, sentences, or digit strings. Each utterance of spontaneous speech begins with a recitation of the alphabet followed by one minute of weakly prompted speech. Examples of weak prompts include "Tell me about your favorite movie"[Shobaki et al., 2000]. Below is a transcription sample.

> "*. . . usually just lay down on my bed. For now, I don't like to, I don't know, uh, football, okay? First, they are like standing on the ground, and then they run, and then they mm, and if the girl passes the whole field, you get six points. Uh, I think it's twenty-four. I don't know, think yeah, they catch block, and uh, one, uh, the quarterback throws, and the runners run, uh, it's blue, uh, and it has a big, big, big electric train set, uh, I have a workshop. . .* "

Notice how this single utterance discusses multiple topics, which is typical in children's speech and provides a unique challenge to ASR systems.

The scripted portion of the dataset consists of 205 isolated words, 100 prompted sentences, and 10 numeric strings. The children were instructed to read these prompts. This subset of the dataset represents an idealized version of children's speech which does not exhibit the linguistic deviations characteristic of children's speech. The prompts either lacked any linguistic structure like isolated words or digit strings, or followed very simple structures. Examples scripts are shown below.

> *"We gathered sticks for the fire."*
>
> *"I collect stamps from Vietnam."*
>
> *"Two six nine zero two two zero."*
>
> *"Cowboys"*

3.1.2 My Science Tutor (MyST) Dataset

The MyST corpus is the largest publicly available children's speech corpus. It consists of 393 hours of conversational children's speech, recorded from virtual tutoring sessions in physics, geography, biology, and other topics. The corpus spans 1,371 third, fourth, and fifth-grade students. Although age and gender information for each student are not available, each student is identified with a unique anonymized ID.

Around 197 hours of the dataset were transcribed, although the quality of transcriptions varies. A portion of the dataset was transcribed using " rich (slow, expensive) transcription guidelines", but another portion was transcribed using cheaper and faster guidelines. This led to variability in transcription quality which has been known to harm ASR performance Ghorbani et al. [2021].

Upon manual inspection, some transcriptions were assigned to the wrong files completely.

Provided Transcription: *"Um, the wires are like a pathway energy goes through it into the motor and makes it work."*

Actual Transcription: *"Um, because it's metal, and metal I think has energy."*

Other files appear to have been automatically transcribed with a lower-quality automatic transcriber.

Provided Transcription: *"No, I don't hearing even a candle burns."*

Actual Transcription: *"No, I don't hear anything when the candle burns."*

Although the authors have performed data cleanup which removed audio files with subpar audio quality, some remaining files still have poor audio quality. Primarily, some files have the children

speaking too close to the microphone, which results in a high level of distortion in the audio. We discuss this in further detail in Chapter 4. Table 4.2 shows the effect of low-quality transcriptions on the performance of Whisper on MyST.

While the speech in the MyST corpus is spontaneous, meaning that speakers do not read from a script, it is not completely unstructured. The dataset collection followed a strict turn-taking system. These recordings were from virtual tutoring sessions, where the tutor presents information, then poses a question and waits for the answer. The student then answers the question while pressing the spacebar which records their answer. Since the speech is strictly an answer to a question about specific subject matters, this resulted in shorter and more structured sentences than other children datasets like CSLU Kids. Below is an example transcription from the MyST dataset.

> *"They absorb nutrients and make all the other things you don't need waste and they absorb all the water through the food that's getting digested."*

While this sentence demonstrates a clearer structure compared to the examples from CSLU Kids, it still highlights the unique linguistic characteristics often found in children's speech. Despite its overall clarity, the sentence features a unique linguistic structure that is common in children's speech. For instance, the expression *make all the other things you don't need waste* departs from the more conventional phrasing *turn ... into waste*. Such deviations highlight the richness and complexity of children's linguistic development. As they learn to navigate the rules of grammar, they often experiment with syntax and semantics, resulting in novel constructions that may not adhere to standard adult usage.

3.2 Classroom Speech Corpora

The dataset we showcase in this section represents children's speech in noisy classroom environments. Within these classrooms, dozens of children plus one or more adult teachers are present, which creates a noisy environment characterized mainly by overlapping speech, children babble noise, which is the noise from multiple people talking at the same time. The difference between overlapping speech and babble is that overlapping speech comes from multiple target speakers, for example when two people are talking and they interrupt each other. Babble however happens when multiple people talk in the background of the target speech which results.

3.2.1 National Center for Teacher Effectiveness (NCTE) Dataset

The NCTE dataset consists of video and audio recordings of 2128 4th and 5th-grade elementary math classrooms collected as part of the National Center for Teacher Effectiveness (NCTE) Main Study [Kane et al., 2022]. The observations took place between 2010 and 2013 across four districts serving historically marginalized students.

Table 3.1 shows the key demographics of the populations of students and teachers in the recordings. The statistics show that the students are balanced by gender, but the vast majority of them come from minority racial backgrounds (72.3%). The majority of the students received free or reduced-price lunches (64.9%) indicating that they likely come from low-income households. A sizable percentage of the students exhibited limited English proficiency (19.7%) or special educational status (12.3%). On the other hand, the teachers were mostly female (82.4%) and White (64.2%). This disparity between students' and teachers' demographics is within the national statistics [Demszky and Hill, 2022].

Table 3.1: Key demographics from the NCTE dataset. LEP indicates that the students have Limited English Proficiency, FRPL indicates that they receive Free or Reduced-Price Lunches in the given school year and SPED indicates Special Education status.

Statistic	Teachers	Students
Number	313	12661
% Male	16.3%	49.4%
% Female	82.4%	49.6%
% No data	1.3%	1%
% **African-American**	22.4%	42.3%
% Asian	2.6%	7.4%
% Hispanic	2.9%	22.6%
% White	64.2%	22.7%
% Other	3.6%	3.9%
% No data	0%	1%
% LEP	-	19.7%
% No data	-	1.2%
% SPED	-	12.3%
% No data	-	1.2%
% FRPL	-	64.9%
% No data	-	1.2%

For each classroom, 2 to 3 video and audio recordings exist, from different angles and microphones, each lasting 45 minutes to an hour. The total duration of the recordings from all microphones and classrooms amounts to 5,235 hours.

While weak transcriptions exist for this corpus, as discussed in 3.3, they are generally not suitable for ASR training. Therefore, we sample and create high-quality transcriptions from the dataset. This process is slow and expensive, resulting in a relatively small transcribed dataset. Throughout our research, we utilized all available transcribed data, which meant that the specific dataset used varied depending on the availability of transcriptions at each stage of the research as the dataset gradually expanded.

3.2.2 M-Powering Teachers (MPT) Dataset

In addition to the NCTE dataset, our team coordinated, recorded, and transcribed our own classroom datasets as part of the M-Powering Teachers project. Similar to the NCTE dataset, the amount of transcribed data expanded throughout the course of our research.

We recorded 6 classrooms from 3 schools. We recorded two 8th-grade classes in a charter school in Washington, DC that serve predominantly African-American and Hispanic low-income students. Classes included at least one special education student and at least one English language learner. We refer to these recordings as **DC-1** and **DC-2**. Two 5th-grade classrooms were recorded in a school in Eastlake, Ohio, with predominantly White students. Classes included special education students. We refer to these recordings as **OH-1** and **OH-2**. The remaining two recordings came from 6th-grade classrooms at a private school in San Jose, California with predominantly White and Asian high-income students with no special education or English language learner students. We refer to these recordings as **CA-1** and **CA-2**.

In each classroom, five microphones were placed at different places in the classroom. This resulted in a good capture of all the audio in the classroom.

3.3 Text Corpus: NCTE-Text

Demszky and Hill [2022] published a dataset of anonymized transcriptions of 1660 classrooms from the NCTE dataset for the purposes of developing computational measures of classroom discourse.

These transcripts were not intended for ASR training and are not readily suitable for it, because all the teachers and students were anonymized to protect the subjects' privacy, replacing

every name with its initial. Additionally, the transcripts are not verbatim and do not capture the exact speech in the classroom word for word. Add to that that they were not properly time-stamped, making preprocessing for ASR training complicated.

However, this corpus holds tremendous value as it captures the language and speech patterns within the classroom, even though it is mostly uncoupled from the audio recordings of speech.

Chapter 4: Optimizing ASR for Children's Speech: Challenges and Adaptation in Controlled Environments

Due to the unique characteristics of children's speech as discussed in Chapter 3, it is considered a distinct domain from adult speech even within the same language. Children's speech can thus be considered a low-resource domain due to the limited size of children speech corpora relative to adult speech..

In this chapter, we explore how Whisper can be adapted for children's speech. We primarily approach this problem by maximizing the utility of the available dataset and proposing best practices for fine-tuning large transformer ASR models for children's speech. Our efforts in this chapter act as a precursor for the more complicated problem of classroom ASR, discussed in Chapter 5. We consider clean children's ASR as an idealized case of classroom ASR, without noise or overlapping speech typically present in classrooms, and the problem becomes purely to adapt the model to the acoustic and linguistic patterns present in children's speech in clean conditions.

4.1 Previous Works

Previous studies on children's ASR have been bottle-necked by the availability of children's speech corpora. For example, prior research has investigated data augmentation, by modulating

the acoustic signal from adult speech to mimic the acoustic properties of children's speech. Thienpondt and Demuynck [2022] assume a source-filter model of speech, where the speech signal is represented as the modulation of an input source signal—a quasi-periodic sound produced by the vocal cords—by a filter, which is the vocal tract. Their method applies a warping function separately to the source and filter components to adapt adult speech to resemble children's speech. This approach significantly reduces word error rates (WER) for children's speech, particularly on the PF-STAR dataset [Batliner et al., 2005], compared to prior methods. Other data augmentation methods include manipulating the speech rate and prosody of existing children corpora to increase the effective size of the training data[Kadyan et al., 2021]. While these methods have shown promising results; they do not capture the linguistic variabilities in children's speech.

However, the release of the MyST speech corpus, which includes around 200 hours of transcribed conversational children's speech has unlocked a large potential for evaluating and improving ASR models' performance for children's speech. However, due to the low quality of some of the transcriptions, the utility of the dataset has been thus far underutilized.

Among the first works to tackle data cleanup for the MyST corpus were Jain et al. [2023a] and Jain et al. [2023b], where the authors manually inspected and extracted 65 hours of high-quality transcriptions from the MyST corpus. They used the MyST corpus and other children datasets for fine-tuning Whisper and Wav2vec2.0. They split the 65-hour MyST subset into a 55-hour training set and a 10-hour test set with no validation set. While their approach showed favorable results, the main drawback of this method is that they did not follow the training/test/validation split proposed by the dataset curators, ensuring no speaker overlap between data splits. In fact, by manually inspecting their splits, we found such overlap present, which clouds the generalizability of their results. Another drawback is the limited scope of manual inspection, which filters out

Table 4.1: Summary of the Datasets Used. Duration in hours.

Dataset	Training Duration	Dev Duration	Testing Duration	Ages
MyST	125	20.9	25.8	8-11 Yrs.
CSLU Kids Scripted	35	4.8	4.8	6-11 Yrs.
LS Test-clean	0	0	5.4	Adult

more than two-thirds of the transcribed corpus. However, their work and positive results with the MyST corpus provide a strong motivation for our research.

4.2 Motivation and Methodology

Based on our review of previous works, our motivation for performing a detailed analysis of the performance of ASR models on children's speech is well-motivated. To avoid the limitations of previous works, we propose a method to filter the MyST corpus, which ensures the maximization of its utility and quality. We chose Whisper as an initial starting point for our experiments. As we mentioned above, Whisper is considered the state-of-the-art in ASR for its versatility and robustness. Additionally, Whisper comes in different sizes and in English-only and multilingual versions, which allows us to examine the effect of these factors on children's ASR adaptation.

4.3 Datasets and Preprocessing

4.3.1 MyST

Our approach relies on maximizing the utility of the MyST corpus. To do so, we avoid manually filtering out the mistranscribed files and opt for a mixture of automatic filtering and manual inspection. For this purpose, we used the Word Error Rate (WER) metric to filter

the MyST corpus, identifying low-quality transcriptions and audio files. WER is a standard metric for evaluating ASR performance. It provides a direct measure of how closely the ASR system's transcription matches the actual spoken words, making it a clear indicator of transcription accuracy. It is calculated as follows:

$$WER = \frac{S + D + I}{N} \tag{4.1}$$

where:

- S is the number of substitutions (when the system transcribes a word incorrectly)
- D is the number of deletions (when the system omits a word that was spoken),
- I is the number of insertions (when the system adds extra words not spoken),
- N is the total number of words in the reference transcription.

To identify low-quality transcriptions or audio files, we follow a similar technique to the one used by Radford et al. [2023], where the entire dataset is transcribed using Whisper-Large and the files with WER larger than 50% are flagged, as a higher WER indicates a worse performance. Additionally, one and two-word files were removed altogether, because they lacked the context to distinguish between homophones, like "to", "too" and "two". All files with no speech activity, i.e. files labeled as <DISCARD>, <NO_SIGNAL>, or <SILENCE> were also removed from the dataset. Table 4.2 shows the effect of different filtering steps on the total dataset duration and WER. According to our results, around 5 hours of the training data is either mistranscribed or has low audio quality and is responsible for increasing the WER on the training data by about

3%. Similar results can be inferred about the test and development sets. Additionally, short files which accounted for only 4 hours of the training data increased the WER by more than 7%.

Table 4.2: WER of Whisper-Large-v1 transcriptions of all three data splits of the MyST corpus before and after different levels of filtration (Duration of splits in hours).

Filtered Files	**Train**	**Test**	**Dev**
None	29.5 (145)	26.2 (28.1)	26.2 (25.5)
WER > 50%	26.8 (140)	22.3 (26.7)	22.3 (25.5)
WER > 50% or w. Less Than 3 Words	19.2 (132.5)	14.2 (25.6)	12.8 (21)

Files longer than 30 seconds in the training and development sets were also removed. That is because Whisper processes files in 30-second chunks, and any files longer than 30 seconds are truncated by default, and any files shorter than 30 seconds are zero-padded up to 30 seconds. However, it is not possible to accurately truncate the transcriptions without any timestamps present, so these files are unsuitable for loss calculation.

Additionally, the majority of the files in the MyST corpus were too short, with the average file length in the training data being 8 seconds. That would mean that training batches are mostly. Such short files would require a lot of zero padding, leading to inefficient training. To remedy this problem, files within a single recording session were concatenated to be close to but not longer than 30 seconds while maintaining the context of the conversation within the recording session.

Our filtering technique removes 17.8 hours from the entire dataset, which leaves us with 179 hours of well-transcribed speech in total. We maintain the train/development/test split provided in the MyST database to avoid any overlap in speakers between the splits. We ended up with 132.5 hours in the training data, 25.6 hours in the development data, and 21 hours in the test data. In contrast to Jain et al. [2023a, 2024], our method eliminates a significant amount of bad

transcriptions while maintaining speaker exclusivity between data splits while tripling the size of usable data. Our filtered file lists were shared with the dataset curators and other researchers working on the MyST corpus [Fan et al., 2024, Graave et al., 2024].

The text of the transcriptions in the MyST corpus was all upper case which destabilized the training. Consequently, all the text was mapped to be lowercase and further normalized using the WhisperNormalizer[1] Python package, which mapped contractions like "you're" to a standard "you are", as well as mapping all digit numbers to be spelled out. In addition, the normalizer maps tokens representing alternate spellings of the same word to standard American English spelling, for example, "colour" is normalized as "color". This ensured that only actual mistranscriptions would be penalized and not contractions or alternate spellings. This reduces the diversity in transcription, which was noted to harm the performance, unlike diversity in audio quality [Radford et al., 2023].

4.3.2 CSLU-Kids

In addition to the MyST corpus, we also use the CSLU Kids dataset. As mentioned in Chapter 3, the CLSU corpus is partitioned into scripted speech and spontaneous speech subsets, which we henceforth refer to as CSLU-Scripted and CSLU-Spontaneous.

The majority of the recordings in the CSLU-Spontaneous corpus were longer than 30 seconds, and are thus unsuitable for training. Instead, we use the CSLU-Scripted corpus for training and reserve CSLU-Spontaneous as an out-of-sample test set.

The transcriptions were of a high enough quality and filtering was not necessary, but they were all normalized to ensure a standard transcription style. Files in the scripted portion of the

[1] https://pypi.org/project/whisper-normalizer/

dataset were shuffled and split into train, development, and test sets with an 80/10/10 split. The training set was 35 hours long, and the development and test sets were both around 4.8 hours long. Short files were combined to be close to 30 seconds as we did with the MyST corpus.

4.3.3 Librispeech: Test-clean

The test-clean subset of the Librispeech corpus was used to test the ASR model's performance on adult speech and ensure that the fine-tuned model does not suffer from catastrophic forgetting, which is the phenomenon where the model "forgets" about earlier tasks when finetuned for new tasks [Ramasesh et al., 2020]. The test-clean subset contains about 5.4 hours of speech read from Audiobooks from the LibriVox project. Since Librispeech was not used for training, we didn't combine the files, and we also didn't filter out any transcriptions to allow for reproducible and contrastable results. All transcriptions were normalized.

4.4 Training Details and Hyperparameters

We followed the Huggingface Whipser finetuning tutorial [2]. Our evaluation script, which calculates the WER, was adapted from a code snippet by OpenAI[3] to ensure reproducible and contrastable results. All models were trained on Nvidia A6000 50GB GPU. Our training scripts are available on Github [4], and the checkpoints are available on Huggingface[5].

For all models, regardless of size, we used a learning rate of 1×10^{-5}, 1 gradient accumulation step. For the Small models, we used a batch size of 64 for the English-only model, and 32 for

[2]https://huggingface.co/blog/fine-tune-whisper
[3]https://github.com/openai/whisper/discussions/654
[4]https://github.com/ahmedadelattia/Kid-Whisper
[5]https://huggingface.co/aadel4

the multilingual model. For the Medium models, a batch size of 32 for both the English-only and multilingual models. All models were finetuned until convergence and the best checkpoints were used for evaluation.

4.5 Results and Discussion

4.5.1 Whisper Zero-shot Models

Table 4.3 shows the WER for different Whisper models without any finetuning. Looking at these results, the gap between children and adult speech is immediately clear. The WER for the scripted part of CSLU Kids is between 6 and 10 times that of Librispeech, and the WER for MyST is between 3 and 5 times larger. In general, English models perform better than multi-lingual models, with the exception of the Medium model where the results are mixed. That could be because the Medium model is big enough to benefit from seeing more data in different languages. The bigger the model, the better the performance, with the exception of Large-V1 being worse than Medium. In fact, the performance seems to saturate beyond Medium and the difference in performance between Medium and Large-V2 is negligible.

Comparing the performance among the three children corpora tested, we note that the worst-performing subset is CSLU-Scripted. This could be due to the short duration of these utterances, and the lack of context in many of them, which hurts the linguistic contextualization of Whisper's audio-conditional language model decoder. However, MyST significantly outperforms CSLU-Spontaneous. As discussed in Chapter 3, while both of these datasets contain conversational spontaneous children's speech, there are two main differences between them. First, MyST is more structured, as it is in a strictly turn-based question-answering environment. CSLU is weakly

Table 4.3: Zero-shot WER on different test sets for different Whisper Models Without Finetuning.

Model	MyST	CSLU Kids Scripted	CSLU Kids Spontaneous	LS Test-clean
Tiny	21.16	74.98	57.01	7.49
Tiny.en	18.34	61.04	45.29	5.59
Base	18.54	40.20	43.71	4.98
Base.en	15.57	33.18	38.57	4.15
Small	14.06	25.15	36.36	3.39
Small.en	13.93	21.31	32.00	3.05
Medium	12.90	18.62	37.04	2.76
Medium.en	13.23	18.57	31.85	3.02
Large-V1	14.15	21.50	45.18	2.98
Large-V2	12.80	17.22	29.39	2.82

prompted, where the prompts are open-ended questions like "Tell me about your favorite movie" Shobaki et al. [2000]. The second difference is the age range. While MyST's participants were in 4th or 5th grade, CSLU's participants ranged from Kindergarten to 10th grade. Research from Yeung et al. [2021] showed that ASR performance varies significantly with age, and they noted that kindergarten speech performed significantly worse than even 1st grade, with gradual improvement as the students aged.

We note that the zero-shot WER reported here for MyST is smaller than that reported in Jain et al. [2023a]. We attribute this to the fact that they used a different normalizer than the one Whisper was trained with, which we validated by inspecting their datasets which are publicly accessible on Huggingface. Since the performance saturates beyond Medium, we finetune the Small and Medium models, both the English and multilingual variants.

Table 4.4: WER on different test sets for different Whisper Models. EN stands for English-only model and ML stands for multilingual model. MyST55H refers to the 55-hour training set from Jain et al. [2023a].

Model	Training Data	MyST	CSLU Kids Scripted	CSLU Kids Spontaneous	Librispeech testclean
Small					
ML - zero-shot	-	14.06	25.15	36.36	3.39
EN - zero-shot	-	13.93	21.31	32.00	**3.05**
EN - Jain et al. [2023a]	MyST55H	13.23	31.26	28.63	5.40
ML	MyST	11.80	55.51	28.53	6.23
ML	MyST + CSLU	12.11	2.74	32.72	7.97
EN	MyST	**9.11**	33.85	28.47	**4.18**
EN	MyST + CSLU	9.21	**2.59**	**27.16**	4.74
Medium					
ML - zero-shot	-	12.90	18.62	37.04	**2.76**
EN - zero-shot	-	13.23	18.57	31.85	3.02
EN - Jain et al. [2023a]	MyST55H	14.40	28.31	26.76	8.66
ML	MyST	**8.61**	30.10	24.26	5.32
ML	MyST + CSLU	8.99	**1.97**	20.28	4.28
EN	MyST	8.91	47.94	25.56	3.95
EN	MyST + CSLU	8.85	2.38	**16.53**	**3.52**

4.5.2 Finetuned Whipser Models

In this section, we showcase the performance of our finetuned models and contrast them with the models from Jain et al. [2023a], whose models are publicly available on Huggingface. We report the best-performing variants here. We tested all models on test sets from the four corpora, listed in Table 4.1.

Looking at the results in Table 4.4, it is clear that fine-tuning Whisper on children's speech improves its performance on the MyST and CSLU, proving that transformer ASR models have the capacity to recognize children's speech. We establish a strong SOTA performance of 8.61% for the MyST test set. Our best performance on the CSLU scripted dataset of 1.97% beats the current state-of-the-art of 5.52% [Yeung et al., 2021]. We also show improvement on unseen datasets,

since our models trained on just MyST, or a combination of MyST and CSLU scripted data show improvement on CSLU Spontaneous speech without any training on speech from this dataset. Our best-performing model on the CSLU Spontaneous dataset scores 16.53% WER, which is about half the WER of zero-shot Whisper. Additionally, our models "forget" less about the adult speech than the models presented in Jain et al. [2023a], with our models seeing a degradation of only about 1%.

Medium models outperformed Small models and generalized better to unseen datasets. The English-only variant of the Small model showed significant improvement over the multilingual variant in seen and unseen datasets. The Medium multilingual variant performed slightly better on the MyST dataset when finetuned exclusively on it, but the English-only variant generalized better to unseen data. Multilingual models in both sizes had higher WER for Librispeech.

Looking at the results for the scripted portion of the CSLU corpus, it is clear that the lack of context in these scripts harms the performance of the models that weren't trained on speech from this dataset. However, the performance improved significantly when speech from this dataset was included in the training data, mainly because of the lack of variability of the scripts, unlike the more diverse MyST or CSLU Spontaneous datasets.

Performance on CSLU-Spontaneous improves significantly with fine-tuning even though it is explicitly excluded in the training data. Performance improves by 10% or more in both Medium models when training with MyST exclusively, with an additional 4% and 9% improvement when training on MyST and CSLU-Scripted. This shows that the improvement exhibited by fine-tuning is generalizable.

Table 4.5: Example transcriptions from the Medium.en model trained on MyST + CSLU Scripted and the zero-shot Medium.en model along with ground truths from the MyST and the CSLU Spontaneous. All the transcriptions were normalized using the WhisperNormalizer package.

ID	Model	Transcription	WER
		MyST	
1	**Ground Truth**	makes like the thing turn on and gets like a magnet so the little washers will stick	-
	Medium.en Fine-tuned	makes like the thing turn on and gives like a magnet it the little washers will stick	**11.8**
	Medium.en zero-shot	makes the thing turn on it gives a magnet the little washers will stick	29.4
2	**Ground Truth**	the water is plain and the grape it the grape tastic thing is plain the salt	-
	Medium.en Fine-tuned	the water is plain and the grape is the grape tastic thing is plain the salt	**6.3**
	Medium.en zero-shot	the water is plain and the grape the grape tastic thing is plain dissolved	18.8
3	**Ground Truth**	and it is pretty much into m turned into it is pretty much turned into a magnet	-
	Medium.en Fine-tuned	and it is pretty much into ma turned into it is pretty much turned into a magnet	**5.9**
	Medium.en zero-shot	and it is pretty much turned into a magnet	47.1
		CSLU Spontaneous	
4	**Ground Truth**	a b c d e f g h j i k l m n o p q r s t u v w x y and z uhm he is like a grumpy man	-
	Medium.en Fine-tuned	he is like a grumpy man	82.4
	Medium.en zero-shot	he is like a grumpy man	82.4
5	**Ground Truth**	uhm i have the lion king toys uhm no uhm i forgot the babies have a pink dress on	-
	Medium.en Fine-tuned	i have the lion king toys no i forgot my babies have a pink dress on	**21.1**
	Medium.en zero-shot	it has a lion king toys a pink dress on	63.2
6	**Ground Truth**	...q r s t u v w x y z tar i do not know just rock i guess pretty strong probably making music	-
	Medium.en Fine-tuned	...q r s t u v w x y z i do not know just rock i guess pretty strong probably making music	**2.5**
	Medium.en zero-shot	yeah probably making music	92.5

4.6 Error Analysis

In this section, we analyze some example transcriptions from the Zero-shot Medium.en model as well as the Medium.en model trained on MyST and CSLU Scripted to gain insight into how fine-tuning Whisper can improve the performance on different domains. Table 4.5 shows such examples.

4.6.1 MyST

The examples from the top half of the table exemplify the main modes of error made by both our fine-tuned model and the zero-shot Whisper model.

Example 1 shows the main improvement of our model over the zero-shot model: accurately

transcribing filler words (like 'like' and 'uh') and disfluencies (such as repeated or corrected words). Whisper is known to perform implicit disfluency removal as a result of its data normalization [Radford et al., 2023]. Such filler words are common in children's speech [Gósy, 2019] and can provide useful insight into the development of children's speech. In this example, our model learned not to skip the filler word "*like*" but still made a mistake with the word "*gets*" which it transcribed as "*gives*". This gives insight that these errors are caused by the audio-conditional language model in Whisper's decoder, which still struggles with unusual or incorrect grammatical structures.

Another example of this is Example 2, where the disfluency in "*...the grape is, the grape tastic thing is...* is mistranscribed by both models. The zero-shot model completely skips the word "*is*" while our model mistranscribes is as "*it*". Our model correctly transcribes the word "*the salt*" which is added on to the end of the sentence and does not belong to any sound grammatical structure, while the zero-shot model mistranscribes it as "dissolved" which makes more grammatical sense but is not what was said, as confirmed by listening to the audio file. Both of these examples show that the audio-conditional language model might provide some further performance improvement possibly by further fine-tuning of the decoder.

Example 3 shows another example of disfluent speech that our model was better able to handle than the zero-shot model. The disfluency in "*...and it is pretty much into a m.. turned into a magnet*" where the child forgets the phrase "*turned into*" and starts to say "*magnet*" but corrects themself as they start to say it - is correctly captured by the fine-tuned model. The zero-shot model completely discards the disfluency.

4.6.2 CSLU Spontaneous

When examining example transcriptions from the CLSU Spontaneous dataset, it is important to note that while the model was trained on the CSLU Scripted dataset, it was not trained on any examples from the CSLU Spontaneous dataset. That means that while the model might have been conditioned to the acoustic properties in this dataset, it was not exposed to the linguistic properties in CSLU Spontaneous, which is unique from both CSLU Scripted (since it is not read speech) and MyST (since it is not structured spontaneous speech).

Example 4 shows the main source of WER in both our fine-tuned model and the zero-shot model's transcription, which is that sometimes both models completely skips the recitation of the alphabet at the beginning of every recording. This is not a consistent problem, since sometimes our model transcribes the recitation like example 6 and it generally skips it less often than the zero-shot model. However, we did not see a pattern in utterances where it skips versus utterances where it does not.

Example 5 shows that like in MyST, fine-tuning improves Whisper's capacity to handle disfluencies in children's speech. While our model consistently skips "*uhms*" which were mistakenly removed during normalization, the zero-shot transcription tries to completely remove the disfluency while also mistranscribing "*i have*" to "*it has*", and as a result, completely alters the sentence.

Example 6 shows an utterance where our model transcribes the alphabet recitation while the zero-shot model removes it and a large chunk of the sentence, starting the transcription more than 20 seconds into the recording.

4.7 Conclusions

In this chapter, we outlined how Whisper can be finetuned on children's speech using MyST, the largest publicly available conversational children's speech corpus. We showcased a method to filter out mistranscribed files from the corpus that requires minimal manual inspection, maximizing its utility and quality. We also established a strong baseline for children's speech recognition. Our finetuning reduced the WER by 4 to 5% in MyST, and by upwards of 15% in CSLU Spontaneous. In doing so, our efforts mark a step towards bridging the gap between adult and children's speech.

We also outlined some of the challenges that face adapting Whisper for children's ASR, namely the fact that audio-conditional language models are not well adapted to the variability in children's speech. Our error analysis shows that fine-tuning improves Whisper's ability to handle acoustic variabilities in children's speech. Fine-tuning also improves to an extent the model's ability to handle the unique linguistic characteristics in children's speech, like disfluencies, and unusual grammatical structures. However, there is still room for improvement in the latter.

While having an embedded audio-conditional language model in Whisper makes it more versatile and simplifies the training process for large corpora, in low-resource settings like children's speech, it limits our ability to directly inspect and improve the language model. In these scenarios, the stage-based inference process of models without embedded language models like Wav2vec2.0 might be better suited for the task.

Chapter 5: Challenges and Limitations of Classroom ASR in Noisy, Low-Resource Settings

After establishing the performance of state-of-the-art models with children's speech and the challenges of adapting these models to children's speech, this chapter tackles the more practical and complex task of classroom ASR. Classroom speech presents additional difficulties—not only due to the challenges posed by children's speech but also because of the unique noises and diverse acoustic and linguistic conditions found in this environment. As previously discussed in Chapter 1.2.2, classroom environments pose several challenges for ASR systems. These include babble noise from overlapping background speech, the complexity of multi-speaker interactions, and far-field speech, where the distance between speakers and microphones degrades the audio quality, and intelligibility of speech, even to human listeners. Combined with the variability of children's speech, these factors create an acoustically unpredictable and linguistically diverse environment that significantly complicates ASR performance.

5.1 The Status of Classroom ASR

Classroom ASR systems consistently face challenges with high WER, particularly when handling children's speech in noisy, multiparty environments. Studies report WERs as high as 78%, primarily due to deletion errors where the system fails to detect entire words [Southwell

et al., 2022]. These errors can severely degrade performance in downstream tasks like natural language understanding (NLU), complicating the use of ASR for accurate transcription in real-time classroom settings [Cao et al., 2023].

A critical technical limitation is the microphone placement and recording setup in classrooms. While closer microphones improve ASR accuracy, they are often impractical in real-world settings, especially when capturing multiple simultaneous speakers. Studies emphasize that distant microphones, though more practical, introduce significant signal degradation due to reverberation and background noise, which directly impacts ASR performance [Southwell et al., 2022].

Despite these limitations, the feasibility of deploying ASR in classrooms remains promising. Even with high WER, ASR can still contribute to collaborative learning models and provide useful educational feedback. For instance, models trained on noisy ASR transcripts have shown reasonable success in recognizing classroom discourse, suggesting that perfect transcription may not be necessary for many educational applications [Southwell et al., 2022]. This opens up possibilities for adaptive AI systems that can function effectively under less-than-ideal conditions.

5.2 Datasets

Our analysis in the previous chapter revealed that the performance of children's ASR with Whisper may be limited by its integrated language model. Furthermore, as established in Chapter 3.2, there is a scarcity of labeled classroom recordings, in stark contrast to the abundance of untranscribed classroom data.

This scenario motivates our decision to use Wav2vec2.0 as the foundational ASR model for classroom applications. The relative abundance of unlabeled data, compared to labeled data,

aligns well with a self-supervised learning (SSL) paradigm. Moreover, Wav2vec2.0's reliance on an external language model—unlike Whisper's embedded language model—enables us to utilize text from the NCTE-Text corpora, as described in Chapter 3.3, for separate language model training. This approach leverages one of Wav2vec2.0's key strengths: its ability to utilize uncoupled data (i.e., data from one domain, such as audio, that cannot be directly mapped to the target domain, such as text). This capability is particularly beneficial in low-resource scenarios like the one we are addressing.

Particularly, we propose CPT as an efficient method of domain adaptation of self-supervised speech models like Wav2vec2.0 to low-resource tasks like classroom ASR. As discussed in Chapter 2.3.5, CPT has shown promising results in efficiently adapting multi-lingual Wav2vec models, particularly XSLR-53 and XLSR-128, to low-resource languages. However, little research has been done to examine the application of CPT to noise robustness tasks.

We consider three different 300M parameter variations of Wav2vec2.0: Wav2vec2.0 pre-trained on LibriVox-60, XLS-R [Babu et al., 2021] pretrained on speech from 128 languages, and Robust Wav2vec2.0 pretrained on noisy English speech. We perform CPT on unlabeled classroom data for each model and then finetune the resultant models on labeled classroom data for ASR. In addition, we perform finetuning of the off-the-shelf Wav2vec2.0 models without CPT. We also train Wav2vec2.0 from scratch on the same data used in CPT as it is the baseline used in similar publications [Zhu et al., 2022]. To validate our choice of Wav2vec2.0 as a foundation model, we also fine-tune Whisper on labeled classroom data and contrast the results.

5.2.1 NCTE

We use the NCTE audio recordings for self-supervised pre-training of our Wav2vec2.0 models, adding up to 5235 hours. This amount of data is about an order of magnitude lower than the pre-training data foundational Wav2vec2.0 model, which was pre-trained on 60K hours of LibriVox. However, this corpus is the largest available classroom corpus. We resampled the audio from the original 44.1KHz sampling rate to 16KHz and cut it into 20-second chunks. About 10% of the data was reserved for validation, and the rest was used as unlabeled pre-training data.

Out of these classroom recordings, six were initially randomly chosen to be transcribed to create a low-resource unbalanced problem. The duration of this subset is about 5.15 hours, with the duration of each file between 45-60 minutes. Key statistics about the transcribed recordings are in Table 5.1. All teachers are in the training-validation dataset are White women. The student's gender is mostly balanced throughout the entire dataset, with about 56% of the students being male. All classes but one have a between 45% to 65% male population. Class 144 has 85% male students. The majority of the students in the dataset come from minority racial backgrounds (66%), with African-American students making up 42% of the total student population, and both Asian and Hispanic students making up 22%. White students make up the remaining 34%. This racial imbalance in the dataset is a result of the imbalance present in the larger NCTE dataset and classrooms within the US in general. Such imbalance presents a unique challenge and an interesting question to be answered: will pre-training on a racially diverse dataset improve the fairness and decrease the racial bias of the ASR system?

In all but one recording in the dataset, a near-field microphone was used. The recording of class 2619 was obtained with a far-field microphone. This distinction is by design, to test how

Table 5.1: Key statistics from the transcribed portion of the NCTE dataset. % Teacher refers to the percentage of speech duration attributed to the teacher. AFAM refers to African-American. FF indicates a far-field microphone and NN indicates a near-field microphone.

	Teacher		Students								
Used for training and validation											
Class ID	**Gender**	**Race**	**% Male**	**% AFAM**	**% Asian**	**% Hispanic**	**% White**	**% Other**	**# Students**	**% Teacher**	**Microphone**
144	Female	White	85%	62%	0%	8%	31%	0%	13	84%	NF
622	Female	White	53%	16%	11%	47%	26%	0%	19	79%	NF
2619	Female	White	48%	76%	5%	0%	19%	0%	21	84%	FF
2709	Female	White	63%	13%	29%	13%	42%	4%	24	68%	NF
2944	Female	White	54%	46%	4%	8%	38%	4%	26	84%	NF
4724	Female	White	46%	46%	0%	11%	43%	0%	28	92%	NF
Total	-	-	56%	42%	8%	14%	34%	2%	131	-	
Used for testing and analysis											
13	Male	Asian	61%	9%	65%	13%	9%	4%	23	78%	NF
4106	Female	AFAM	50%	14%	50%	21%	0%	14%	12	81%	NF
4352	Male	White	41%	34%	10%	7%	48%	0%	29	88%	NF
4651	Female	AFAM	55%	27%	14%	9%	50%	0%	22	80%	NF
Total	-	-	51%	23%	32%	11%	31%	3%	88	-	

well the ASR system generalizes to microphone configurations unseen in the training data and whether pre-training on different microphone set-ups improves this generalization, as this has been highlighted as a problem by Southwell et al. [2022].

To fix the imbalance issue, we sampled a balanced subset from the NCTE dataset, as described later in Chapter **??**. However, accurate human transcription is a slow process. During late stages of the project, 4 classroom recordings from the balanced subset were transcribed, and reserved for testing. In total, this subset is about 3.7 hours. The racial make-up of this subset is different than the one trained on, with white students representing roughly 31% of the population, African-American students representing 23%, and Asian students being the most represented group with 32% while being the least represented group in the training-validation subset, making up 65% of one class and 50% of another. Additionally, this subset differs from the train-validation subset in the presence of two male teachers, one of them being Asian. We use this dataset to answer the question: how well does the ASR system generalize to different make-ups than the ones seen in training? Does pre-training on a diverse dataset improve this generalization?

5.2.2 MPT

We use this dataset to test the effect of CPT on improving performance in slightly different but related domains than the one seen during CPT. In each classroom, five microphones were placed at different places and the audio streams were then added together. This resulted in a good capture of all the audio in the classroom, as well as extremely noisy audio in one particularly noisy classroom, CA-1. We keep this configuration to test the model's ability to handle extremely noisy conditions.

5.2.3 NCTE-Text

Using this corpus for training presents a challenge, as it is de-identified, with teacher and student names replaced with initials such as "Student K" and "Mrs. D". However, this corpus can be useful for language model training and fine-tuning as it captures the interactions with diverse classroom environments.

To process this data for language model finetuning, we replaced the initials with randomly chosen names. To ensure that the names are unbiased towards race or gender, we referenced a list of the most popular baby names by race between 2011 and 2019 in New York City [1]. For each classroom transcription and every initial, we uniformly sample gender, and then sampled a race uniformly from White, African-American, Hispanic, or Asian. After sampling the race and gender, we referred to the list of popular names of that gender and race based on its popularity according to the list. We then replaced all instances of that anonymized initial in that transcription. Race-aware deanonymization ensures that all races and both genders are equally represented in

[1] https://data.cityofnewyork.us/Health/Popular-Baby-Names/25th-nujf/data

the text corpus.

5.3 Experiments

5.3.1 CPT Experiments

We performed three CPT experiments on three 300M parameter Wav2vec2.0 models. The first was initially pretrained on LibriVox [Kearns, 2014] which totals up to 60K hours. We denote that model as **W2V-LV60**. The second model, **W2V-Robust** [Hsu et al., 2021] was pretrained on LibriVox, as well as noisy English recordings. The third model is **XLS-R** 300M which was pretrained on 436K hours of speech from 128 languages, including English. The contrast between the performance of W2V-LV60 and W2V-Robust will showcase the effect of additional noisy data from a different domain to the target domain during initial pre-training on the efficacy of CPT domain adaptation. The performance of XLS-R will showcase the impact of additional pre-training data from completely different domains and languages. We also pretrain a Wav2vec2.0 model from scratch with the NCTE dataset, which we denote as **W2V-SCR**, representing the baseline configuration from previous works. The contrast in performance between this model and other CPT models will showcase the efficacy of CPT versus randomly initializing the model before pre-training.

We implemented our model using the official fairseq library [2]. Each of the models was pretrained for 1M steps on 3 NVIDIA-A100 GPUs using the configuration file obtained from the Fairseq repository.

[2] `https://github.com/facebookresearch/fairseq`

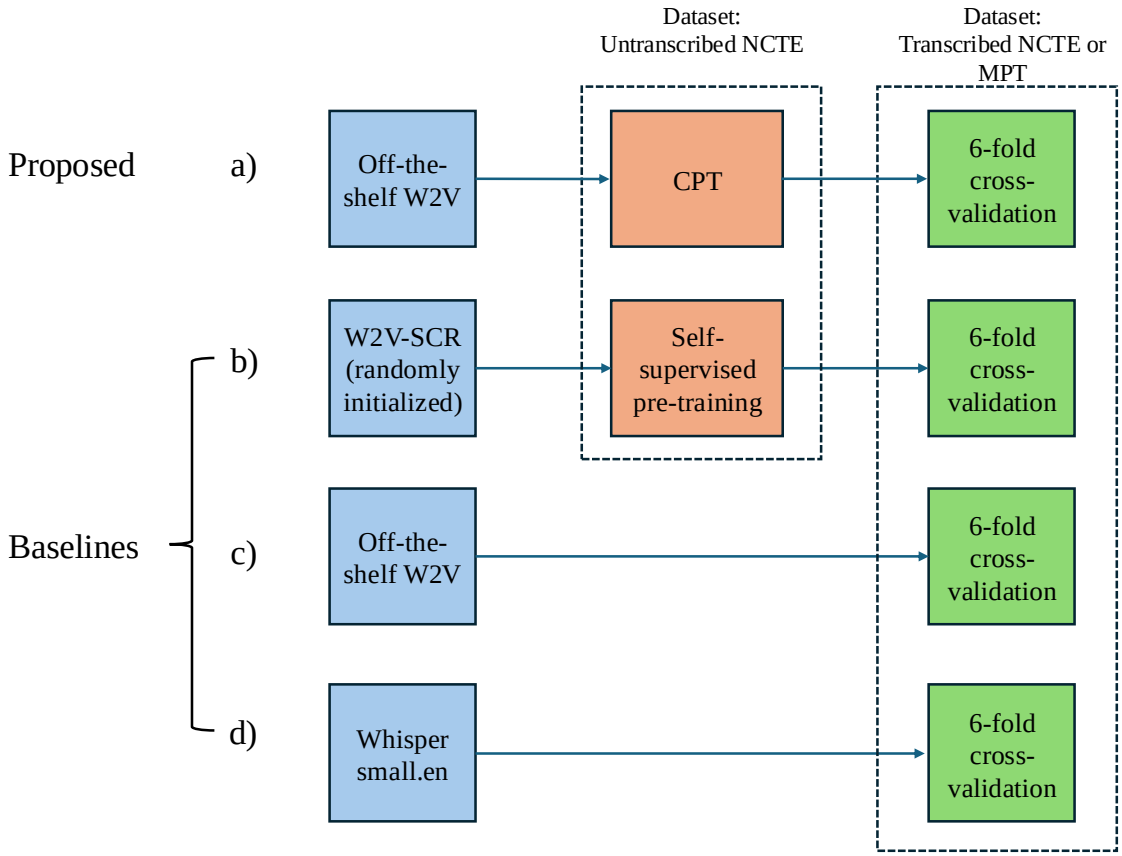

Figure 5.1: Overview of the CPT methodology for domain adaptation in Wav2vec2.0, with comparisons to multiple baselines. (a) **CPT**: Off-the-shelf pretrained Wav2vec2.0 models (W2V-LV60, XLS-R, or W2V-Robust) initially trained on large, general datasets, further adapted through self-supervised pre-training on untranscribed NCTE classroom audio to handle classroom-specific noise and overlapping speech. (b) **W2V-SCR (Randomly Initialized)**: a Wav2vec2.0 model with randomly initialized weights, serving as a control to evaluate the impact of pretraining. (c) **Off-the-shelf W2V**: the same pretrained models used in CPT, but evaluated without additional domain-specific adaptation. (d) **Whisper small.en**: an open-source transcription model included as a general ASR baseline. Each model undergoes six-fold cross-validation on both the NCTE and MPT datasets, with CPT achieving superior performance in noisy, low-resource classroom environments.

5.3.2 Finetuning

We perform two separate 6-fold cross-validation finetuning experiments on each of the two datasets used for training, leaving one classroom recording for validation in each fold. For each fold, we train using a configuration file adapted from the 10-hour configuration file found on the Fairseq repository. Each fold was trained on one NVIDIA-A6000 GPU using the Fairseq library. We finetune each off-the-shelf Wav2vec2.0 model (W2V-LV60, W2V-Robust, and XLS-R), and their CPT counterparts as well as W2V-SCR.

We preprocess our data by cutting the audio into chunks of 30 seconds or less depending on the timestamps of the transcription. We also normalize the text using the WhipserNormalizer Library [3].

5.3.3 N-Gram LM Training

We train our 5-gram LM using the KenLM[4] library. Our training data is the deanonymized NCTE-Text dataset as described above. We normalize the training data using the Whisper Normalizer library to match the transcription text. We use this LM for LM decoding of Wav2vec2.0 in all our experiments.

5.4 Results and Discussion

In this section, we showcase our finetuning results of different Wav2vec2.0 checkpoints as well as the results from the small English-only Whisper checkpoint both off-the-shelf and

[3] https://pypi.org/project/whisper-normalizer/
[4] https://github.com/kpu/kenlm

finetuned in the leave-one-out cross-validation fashion. Whisper-small is comprised of 244M parameters which is the closest Whisper model in size to Wav2vec's 300M parameter models. Additionally, our results from the previous chapter in Tables 4.3 and 4.4 show that English-only Whipser-small generally performs better with the children's speech, and shows more improvement with fine-tuning.

Table 5.2: Average cross-validation WER finetuning results starting from different Wav2vec2.0 checkpoints with and without continued pretraining, as well as comparison with the small English-only checkpoint of Whisper. Whisper-FT refers to finetuning Whisper on the target dataset. The standard deviation of the WER is between brackets. All models decoded with an n-gram LM use our LM trained on race-aware deanonymized NCTE-Text corpus.

Model	**LM**	**WER(STD)**	
		NCTE	**MPT**
Whisper	-	24.46(12.37)	30.15(12.99)
Whisper-FT	-	19.14(6.77)	28.53(14.07)
pre-training from Scratch			
W2V-SCR	None	47.34(5.73)	51.39 (6.83)
	5-gram LM	30.25(15.44)	38.59(12.93)
No Continued Pretraining			
W2V-LV60K	None	39.11(13.01)	37.82(12.30)
	5-gram LM	30.39(14.48)	33.56(10.86)
XLS-R	None	38.19(10.39)	39.12(13.60)
	5-gram LM	29.02(10.96)	32.49(10.76)
W2V-Robust	None	35.07(11.85)	36.36(11.54)
	5-gram LM	27.99(13.28)	31.49(9.97)
Continued Pretraining			
W2V-LV60K (CPT)	None	22.52(4.89)	32.26(8.92)
	5-gram LM	18.13(5.50)	26.72(7.72)
XLS-R (CPT)	None	26.53(5.13)	32.16(10.78)
	5-gram LM	19.37(4.91)	26.80(8.00)
W2V-Robust (CPT)	None	25.04(5.28)	30.97(9.99)
	5-gram LM	**17.71(5.06)**	**26.50(8.09)**

5.4.1 Whisper

The first section of Table 5.2 shows the average cross-validation WER across all folds for off-the-shelf and fine-tuned Whisper. Looking at the results, we see that the WER with classroom recordings is significantly higher than with clean children's speech. Comparing the results in this table versus the results of the unadapted Whisper models with children's speech corpora in Table 4.3, we see that the English-only Whisper-small has a WER of 24.46% and 30.15% on NCTE and MPT respectively, compared to 13.93% on MyST which represents clean spontaneous children's speech under clean conditions. This shows that even though children's speech is considered a challenge, the largest obstacle towards robust and accurate classroom ASR lies in noisy and variable environments, not children's speech.

We also notice a significant gap between the NCTE and MPT datasets, showing that the MPT dataset is more challenging. This was validated by manual inspection, as the MPT recordings exhibited higher levels of noise.

5.4.2 Wav2vec2.0: Overall performance

5.4.2.1 pre-training from scratch on target domain data

Pre-training Wav2vec2.0 from scratch, we get a high WER of 30.25% and 38.59% in NCTE and MPT, respectively. The same performance gap between the two datasets can be seen in the results form Whisper, which further supports that the MPT dataset presents a more difficult task.

The standard deviation in the results is quite high, indicating that each recording has unique characteristics. As a result, the folds perform differently on each one. Although we

consider classroom recordings to be a single domain, the type of classroom environment—whether collaborative or instructional—affects the noise level and the ratio of teacher-to-student speech. All of these factors influence performance.

5.4.2.2 Off-the-shelf pre-trained models

Previous work [Hsu et al., 2021] shows that pre-training on target domain data improves more than pre-training on out-of-domain data. However, our results indicate that pre-training from scratch on target data underperforms off-the-shelf models, including W2V-LV60K which is pre-trained entirely on clean speech. Unlike in the work of Hsu et al., where in-domain and out-of-domain pre-training datasets were the same size, the out-of-domain pre-training datasets used in the off-the-shelf Wav2vec2.0 models, are at least 12 times larger than the target domain pre-training dataset. More precisely, Librivox has 60K hours of speech used to pre-train both W2V-LV60 and W2V-Robust, while our in-domain pre-training data is only 5235 hours. .This highlights the crucial role of pre-training dataset size in learning high-quality speech embeddings.

Without CPT, fine-tuning W2V-Robust outperforms both XLS-R and W2V-LV60K. W2V-Robust was pre-trained on the same data as W2V-LV60K, plus additional noisy English speech, showing that adding out-of-domain noisy data improves performance. XLS-R, despite being trained on a much larger, cross-lingual dataset, performs worse than W2V-Robust, but better than W2V-LV60K. This supports previous findings [Babu et al., 2021, Hsu et al., 2021] that larger cross-domain corpora can help. However, smaller, domain-relevant data (noisy adult English) performs better. All models still outperform training from scratch on smaller in-domain data.

To summarize, pre-training on larger out-of-domain datasets generally yields better perfor-

mance than small in-domain pre-training. The closer the pre-training data aligns with the target domain, the better the results—though more data still proves beneficial regardless.

5.4.2.3 CPT on target domain data

Looking at the CPT results in the third section of Table 5.2, we can see that for NCTE with LM decoding, CPT improves WER by up to 12.26%. This shows that CPT is a powerful tool for domain adaptation when labeled data is scarce and more unlabeled data is available. For the MPT dataset which is different from the pre-training data, CPT is still effective, yielding an improvement of up to 6.84%, proving that CPT is effective in domain adaptation in a generalizable fashion. It is also noticeable how the standard deviation of the WER decreases in both datasets. This shows that performing CPT on diverse classroom environments and noise conditions improves the ability of the model to generalize to these conditions.

In terms of the choice of starting point for CPT, the order of performance with CPT does not follow the order observed with off-the-shelf models. W2V-LV60K outperforms XLS-R with CPT, meaning that the performance edge XLS-R had initially from large cross-lingual pre-training does not carry forward with CPT, and initial pre-training on English-only datasets is superior. However, W2V-Robust still provides the best performance, showing that initial pre-training on out-of-domain noisy English data yields better speech representations when adapted to the target domain.

The gap between LM-decoded and non-LM-decoded results is much larger when pre-training from scratch, suggesting the model learns acoustics but lacks sufficient linguistic representation, which LM decoding compensates for. This indicates that initial pre-training on clean

adult speech, even in other languages, captures useful linguistic features not learned from smaller, noisy in-domain data. Notably, there's a 25% gap in non-LM decoded WER between W2V-LV60K (CPT) and W2V-SCR, highlighting that CPT benefits from clean out-of-domain speech and further refinement on in-domain data with CPT.

5.4.2.4 Comparison with Whisper

Finally, we can see that with CPT, W2V-Robust outperforms the Whisper checkpoint of comparable size, even with finetuning using both NCTE and MPT datasets. The nature of self-supervised speech models breaks down the problem into three parts: pretraining/CPT, finetuning, and LM decoding. This gives us more flexibility to utilize unmappable representations like unlabeled audio or text that doesn't correspond well to said audio. Without CPT or LM decoding, Wav2vec2.0 performs much worse than Whisper, with WER being higher by 15-19% in the NCTE dataset. The flexibility that Wav2vec2.0 allows beyond supervised finetuning improved the performance by up to 19% through a combination of CPT and LM decoding.

5.4.3 Detailed analysis of cross-validation results

In this section, we do a more detailed analysis of the results. We start by discussing the results in Table 5.3 which shows the WER of each fold of the cross-validation in Whisper-FT and W2V-Robust with and without CPT.

Table 5.3: Detailed WER results on each fold of the cross-validation on both the NCTE and MPT datasets, for the off-the-shelf and the CPT version of W2V-Robust and the finetuned small English-only Whisper checkpoint.

NCTE			
Fold	**Whisper-FT**	**W2V-Robust**	**W2V-Robust (CPT)**
144	**14.78**	19.86	15.20
622	**16.97**	26.31	18.77
2619	28.4	54.03	**27.15**
2709	**12.74**	19.09	13.12
2944	26.98	28.07	**17.61**
4724	14.95	20.55	**14.43**
Average	19.14	27.99	**17.71**
MPT			
Fold	**Whisper-FT**	**W2V-Robust**	**W2V-Robust (CPT)**
OH-1	19.76	18.55	**15.42**
OH-2	**16.42**	19.89	16.88
DC-1	28.79	29.42	**24.90**
DC-2	**26.42**	37.32	30.34
CA-1	55.77	49.03	**41.54**
CA-2	**24.04**	34.47	29.91
Average	28.53	31.45	**26.50**

5.4.3.1 NCTE

Looking at the results, it is apparent that CPT significantly improves the performance in each fold of the cross-validation with one notable example recording 2619. This recording is the only one that comes from a far-field microphone, with the entirety of the training data in this cross-validation fold coming from near-field microphones. It is thus no surprise that without CPT,

the model performs much worse in this fold than the others. However, with CPT, the error is cut in half, proving that CPT helps the model generalize to microphone configurations unseen in the labeled data but present in the unlabeled pre-training data.

In terms of comparison with Whisper, looking at the NCTE results, CPT allows the model to achieve close performance or improve upon Whisper in every fold, with one major improvement noticed with recording 2944. Upon manual inspection, it was noted that this class started as an instructional class for 15 minutes and then the teacher assigned a set of questions to the students and started making rounds in the class. This resulted in a much noisier environment than other classrooms with a higher degree of children's babble noise. Whisper was unable to deal with children's babble noise, often interleaving the target speaker with whatever it could discern from the background noise and sometimes exhibiting characteristic Whisper hallucinations by repeating a single word or phrase, for example, *"how do you know that what what is the denominator what is the denominator the the the the the..."* with the word *"the"* repeating 206 times. Wav2vec does not suffer from the same hallucination problem, and with CPT, it's much more capable of dealing with children's babble noise and focusing on the target speaker correctly.

5.4.3.2 MPT dataset

The differences in WER between Whisper and W2V-Robust with CPT on each fold are higher in the MPT dataset. W2V-Robust with CPT outperforms Whisper in 3 folds, with the main improvement coming from the recording of class CA-1. This recording is perhaps the noisiest of all the datasets used in this study, as discussed in Section 5.2.2. Both W2V-Robust and Whisper have high WER for this class recording, but even the non-CPT W2V-Robust outperformed Whisper.

Upon manual inspection, we again see that Whisper suffers from extreme hallucinations with high children babble noise. On the other hand, W2V-Robust with CPT is more capable of handling extremely noisy conditions, outperforming Whisper in this fold by 14%.

5.4.4 Performance on an unseen test set with different demographics

Table 5.4: Test set WER results on the held-out files from the NCTE dataset not used in cross-validation training. Each entry is the average WER of all the cross-validation versions of a particular model when tested on a particular recording in the test set, with the standard deviation in brackets. Average shows the total average WER of each model.

Model	Whisper	Whisper-FT	W2V-Robust	W2V-Robust(CPT)
13	17.55	**14.58(1.29)**	20.24(0.52)	16.10(0.58)
4106	**14.66**	18.89(4.88)	21.76(0.96)	15.53(0.27)
4352	22.14	14.91(3.50)	17.69(0.36)	**12.98(0.24)**
4651	25.88	29.40(4.81)	25.71(0.44)	**18.59(0.40)**
Average	20.06	19.45	21.35	**15.80**

In this section, we discuss the results from the unseen NCTE test set. As previously discussed, and according to Table 5.1, the test set has a different racial makeup with a higher representation of Asian students and the presence of two male teachers, one of them being Asian, while all the teachers in the train-validation subset are White women. This test shows how well the models generalize to different demographics, that are unseen during training.

From Table 5.4, CPT improves the WER on average, from 21.35% to 15.80%, with the decreased standard deviation showing less variation in performance with different training data. This shows that CPT smoothes out the differences caused by different representations in the training data. When compared to Whisper, both the finetuned and non-finetuned versions of Whisper perform worse than both W2V-Robust models and finetuning seems to even harm the

performance on one file, indicating that limited generalizability.

5.5 Conclusion

This chapter demonstrated that Continued Pretraining (CPT) is an effective and robust strategy for adapting Wav2vec2.0 models to domain-specific and acoustically challenging settings. Across multiple experimental configurations, CPT consistently outperformed existing domain adaptation approaches, substantially improving generalization to unseen noise conditions and far-field recordings. In particular, when fine-tuning exclusively on near-field data, CPT reduced the word error rate of far-field classroom recordings by up to 50%, indicating that the model learns domain-relevant acoustic representations beyond those present in the labeled data.

We further showed that CPT improves ASR performance on noisy classroom speech by up to 12.26% on average and up to 27% under specific conditions, establishing it as a strong baseline for low-resource domain adaptation. Our results suggest that in practice, recreating large scale pre-training from scratch is impractical for most researchers. CPT provides an efficient way to make use of the large scale pre-training on OOD data, by continuing the pre-training process on available in-domain data.

While CPT effectively leverages unlabeled audio to learn domain-adapted acoustic representations, it remains limited by the availability of gold-standard transcriptions during fine-tuning. In many real-world settings, including classroom ASR, obtaining large amounts of gold-standard transcription may not be financially viable, but obtaining large-scale weak-transcriptions is. This observation motivates the next chapter, which investigates whether imperfect textual supervision, when carefully paired with gold-standard transcription can be exploited as an intermediate training

signal. Building on the strengths of CPT, Chapter 6 introduces Weakly Supervised Pretraining (WSP), a paradigm that incorporates noisy transcriptions to further increase data efficiency under extreme supervision constraints.

Chapter 6: A Weakly Supervised Pretraining Paradigm for Data-Constrained Classroom ASR

Chapter 5 demonstrated that Continued Pretraining (CPT) is a significant step toward domain adaptation, enabling models to exploit large amounts of unlabeled data to learn domain-specific acoustic representations. Our experiments relied on supervised fine-tuning on the limited amounts of data available to us. While the CPT significantly improved the performance by exposing the model to more data, even if unlabeled, it is reasonable to assume that the performance is bottlenecked by the amount of labeled data.

Ideally, one would fine-tune the model on large amounts of diverse high-quality human-transcribed in-domain classroom data, on the order of hundreds or thousands of hours. While in this chapter, we have more high-quality labeled data, by transcribing more classes from the NCTE dataset, we are still in a low-resource paradigm, having only a few hours of training data.

In our experience commissioning gold-standard transcription for the NCTE data, accurate transcription is a costly and slow process. First, in terms of cost, accurate transcriptions can cost between \$3-\$5 per minute. This means that while obtaining human transcriptions for dozens of hours might be financially feasible, hundreds or thousands of hours is not. Additionally, to transcribe even a few hours at that rate this process took a few months. We also needed to constantly vet the transcriptions, which added human hours on our end. This raises the question

of how to scale the model to ensure its robustness to diverse acoustic and linguistic conditions if we are limited by the size of the gold-standard labels we can get.

In this chapter, we attempt to provide a practical solution to this problem by utilizing weak transcriptions. Weak transcriptions refer to transcripts that are imprecise or inaccurate. One could obtain weak transcripts from a cheaper transcription service tier without proper vetting allowing for faster transcription at scale. In the case of the NCTE, we rely on weak transcripts that already exist for the NCTE dataset that are primarily intended for analysis, not ASR.

While weak and inaccurate labels are traditionally associated with unreliable performance, this chapter explores a different framing: can weak supervision, when used strictly as an intermediate training stage, facilitate the learning of stronger domain-adapted audio embeddings? This perspective is analogous to the role of self-supervised pretraining in data-constrained ASR, where representation learning from imperfect or indirect objectives enables models such as Wav2vec2.0 to outperform fully supervised alternatives. Building on this intuition, we introduce Weakly Supervised Pretraining (WSP), a novel training paradigm that incorporates weak labels as an intermediate supervision signal to improve robustness under limited access to gold-standard data. We show that, even under severe label corruption, weak supervision substantially increases the utility of limited gold-standard data. Through controlled synthetic experiments and a real-world classroom case study, this chapter demonstrates that WSP provides a principled and data-efficient alternative for training ASR systems in data-constrained settings.

6.1 Background: Weakly Supervised Learning

The central question we explore in this chapter is how to effectively integrate weak transcriptions with limited gold-standard transcriptions to improve ASR performance in low-resource domains. This approach falls under the domain of Weakly Supervised Learning (WSL) [Zhou, 2018]. Unlike fully supervised learning, where the model is trained on a precisely labeled dataset, and self-supervised or unsupervised learning, where the model is trained on unlabeled data, WSL is a machine learning paradigm where models are trained on partially labeled or imprecisely labeled data.

There are three types of WSL [Zhou, 2018]:

- **Incomplete Supervision** where only a subset of the training data is labeled. Examples include self-training, where an intermediate model is first trained on the labeled subset, and then used to generate pseudo-labels for the available unlabeled data for subsequent training [Amini et al., 2025].
- **Inexact Supervision** where only coarse-grained labels are given. For example, in image categorization tasks, where image-level labels instead of object-level labels are used.
- **Inaccurate Supervision** where labels are not always precise.

6.1.1 Weakly Supervised Learning in ASR

In the task of ASR, WSL, particularly inaccurate supervision [Brabham, 2008, Frénay and Verleysen, 2013], has enabled the use of imprecise or inaccurate transcriptions.

Perhaps the most prominent large-scale example of weak transcription usage in ASR is

Whisper. Whisper was primarily trained on weak web-scale transcriptions sourced from the internet. Although the full composition of Whisper's training data is not publicly disclosed, we know that it also contained a substantial amount of off-the-shelf high-quality transcriptions and that the weak web-scale transcripts were filtered in a way to remove the most inaccurate transcriptions.

However, the use of weak transcriptions has been shown to affect Whisper's accuracy and reliability, as Whisper suffers from a well-documented issue with hallucinations, which are erroneous, nonsensical transcriptions not related to the input audio [Ji et al., 2023]. A recent study [Barański et al., 2025] investigating Whisper's hallucinations induced by non-speech audio found that among the most common hallucinations are phrases like ”subtitles by the amara org community" and ”thanks for watching", which reveals the effect of Whisper being trained on video transcriptions from movies or TV shows.

Even with the hallucinations, Whisper has been considered one of the most robust and versatile off-the-shelf ASR models, thanks in part to its utilization of weak transcription in a very large training corpus. However, this paradigm does not translate to other domains. Given the filtering of egregiously weak transcriptions described in the Whisper paper, it is safe to assume that at least a substantial amount of the training data is sufficiently accurate. This balance between accurate and inaccurate transcriptions is difficult to achieve in many domains, including the classroom ASR domain. For instance, in the NCTE dataset, there are about 5000 hours of weakly transcribed data, and about 13 hours of gold-standard transcriptions. When combined, the accurate transcriptions would be about 0.25% of the full dataset, which is not enough to drive the model training in any meaningful way. Figure 6.1 illustrates the difference between the balance of precise to imprecise transcription in Whisper's training data vs. classroom data.

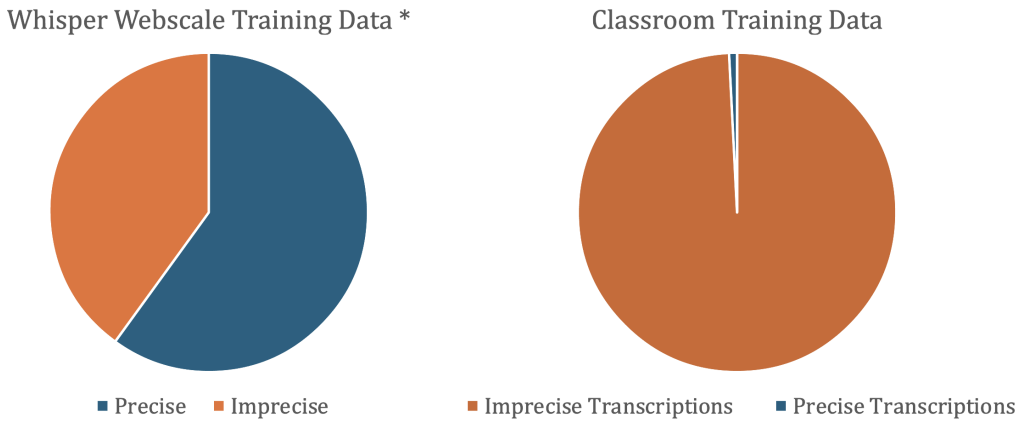

Figure 6.1: An illustration of the ratio of precise to imprecise transcription in Whisper's training data vs available classroom data.

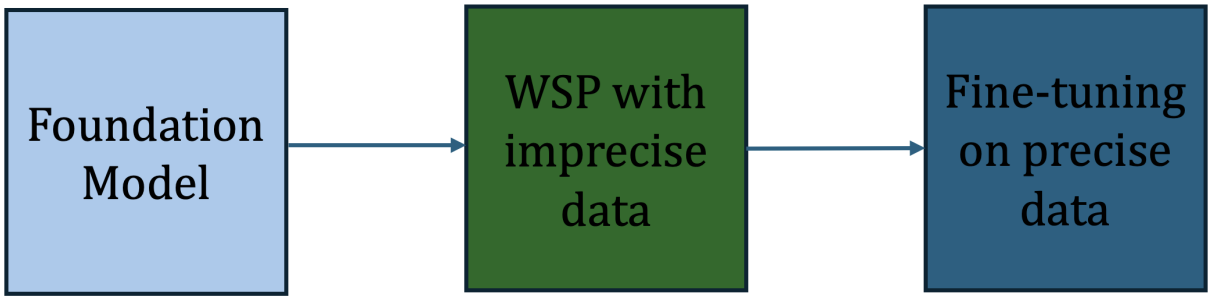

Figure 6.2: Diagram showing WSP as an intermediate step between self-supervised pretraining and accurate supervised finetuning.

6.2 Proposed Approach: Weakly Supervised Pretraining (WSP)

We propose Weakly Supervised Pretraining (WSP), an intermediate pretraining step that utilizes weak labels on large datasets. WSP is effective when there exists a large, noisily labeled dataset and a small, accurately labeled dataset, an assumption that holds for many resource-constrained ASR domains. Figure 6.2 shows WSP as an intermediate supervision step before supervised finetuning.

Given an ASR model, a large weakly labeled dataset and a small gold-standard dataset,

we initially train the model on the large weak dataset, obtaining an intermediate model. This model carries over the error modes present in the dataset, but also learns a general, albiet noisy, representation from the accurate parts of the dataset. For instance, if a dataset has a lot of misspellings in the transcriptions, the model will likewise spell words inconsistently, resulting in a high WER during validation. However, the model also learns a rough representation of the sound of every letter. In noisy settings such as classrooms, the model learns useful acoustic representations, including increased noise robustness and improved sensitivity to primary speakers in multi-speaker environments.

This rough but useful representation can then be refined by training on a small but accurate dataset, resulting in better performance than directly training on the accurate transcriptions. In fact, our experiments with synthetically added mistranscription show that even if we corrupt up to half of the WSP training data and fine-tune with as little as 10 minutes of accurately labeled data, the model achieves a performance that nearly matches training on the full, uncorrupted dataset. We discuss these experiments and results in more detail later in this chapter.

6.3 Experimental Design and Motivation

In this section, we provide a high-level overview of the experiments designed to evaluate the impact of weak transcriptions on ASR performance and to assess the effectiveness of WSP. We begin with an ablation study using an off-the-shelf adult speech dataset with high-quality transcriptions, in which we synthetically corrupt increasing portions of the training data with common transcription errors. This controlled setup allows us to systematically analyze how the type and degree of transcription noise affect WSP, and to compare against a strong baseline trained

on fully gold-standard transcriptions.

We then apply our findings to real-world scenarios using the NCTE dataset with actual weak transcription. This experiment tests the efficacy of WSP in real-world scenarios.

For both experiments, we train Wav2vec2.0-based models, using the fairseq [Wang et al., 2020] implementation.

6.4 Datasets

6.4.1 TEDLIUM

We use the TEDLIUM 3 [Hernandez et al., 2018] dataset for our synthetic experiments. TEDLIUM is a collection of transcribed recordings of TED talks amounting to 540 hours from 2028 speakers. We chose TEDLIUM as it is a high-quality and public dataset that provides a good benchmark for our experiments. Unlike Librispeech [Panayotov et al., 2015], TED talks are a better representation of everyday spontaneous speech, making them more suitable for experiments with transcription errors.

6.4.2 NCTE

As we mentioned in Chapter 3.2, the NCTE dataset contains 5235 hours of recordings. These recordings have been transcribed, without ASR in mind, with weak, inaccurate transcription for the purpose of analysis. Previously, in Chapter 5, we considered these transcripts to be unusable and used the dataset as unlabeled for CPT. In this chapter, we attempt to utilize these transcripts in WSP.

6.4.2.1 NCTE-Weak

The original transcriptions of the NCTE dataset contain a large number of substitutions and deletions, and all the names of students and teachers were omitted to protect their privacy. Most importantly, the transcripts were not properly time-stamped. Loose, inaccurate minute marks are provided that are often off by tens of seconds. In addition, in many instances, utterances have been labeled as “unintelligible" or “side conversation" that are still perfectly intelligible to both ASR models and humans. We call this dataset **NCTE-Weak**. We pass the dataset through a forced aligner [Ashraf, 2025, Wolf, 2020] to obtain more accurate timestamps for preprocessing the dataset before ASR training. While the forced aligner gave better timestamps than the ones provided, deletion errors and the mismatch between the forced alignment model and the classroom speech domains resulted in errors in timestamps.

Ground Truth: *So what would your final answer be?*

Provided Transcription: *line, he is using a number line. Good. So then what would your final answer be?*

6.4.2.2 NCTE-Gold

In Chapter 1.2.2, a small subset of the recordings was re-transcribed using gold-standard human transcriptions. In this chapter, we transcribe 11 more classrooms raising the number of gold-standard transcribed classroom recordings to 17 classes, amounting to 13 hours of recordings. We call this subset **NCTE-Gold** and we split it into a training set (10 hours from 13 classes) and a validation set (3 hours from 4 classes).

6.5 Experiments

6.5.1 Synthetic Corruption - TEDLIUM

For the synthetic corruption experiments, we fine-tune Robust-wav2vec [Hsu et al., 2021], as it is the Wav2vec model pre-trained with the largest amount of English speech audio.

6.5.1.1 Weakly Supervised Pre-training Step

For the intermediate WSP step, we corrupt the transcriptions to model and approximate common human transcription mistakes. We consider 3 types:

- **Deletion:** We randomly drop words from the sentence.
- **Misspellings:** We use the Datamuse API [Datamuse, 2025] to replace a word with another word that is either similar in pronunciation to model hearing mistakes, or one that is a common misspelling, to model common spelling mistakes.
- **Timestamp Inaccuracies:** Long recordings are often partitioned into smaller utterances; however, the timestamps for the beginning or end can be inaccurate. We model this inaccuracy by dropping a few words from the beginning and/or end of a sentence or adding a few random words.

We use two methods to corrupt each sentence:

- **Random Corruption:** Where at least one or more of the above mistakes are randomly applied, but the rest of the sentence is mainly correct. Each word had a 5% probability of being deleted, a 20% probability of being replaced by a soundalike, or a common

misspelling, and a 5% probability of being repeated. Each sentence had a 50% probability of timestamp inaccuracies.

- **Full Corruption:** To model extreme corruption, every word in the sentence is either misspelled or deleted. The remaining types of corruption follow the same probabilities as in random corruption.

Below is an example of random and full corruption as compared to the ground truth transcription.

Ground Truth: *Was offered a position as associate professor of medicine.*

Random Corruption: *Okay was offeree a position as associate of medicine.*

Full Corruption: *Coffered a exposition exposition ass assonate professore off medicines.*

For each corruption method, we interpolate the corrupted transcriptions with gold-standard accurate transcriptions to create different training sets with different degrees of corruption. We create 4 training configurations for each corruption method at 25%, 50%, 75% and 100% corruption, where the percentages refer to the number of corrupted utterances in the training set. These configurations model the case where we mix weakly transcribed data with high-quality transcripts, similar to Whisper's training data. We also train the same model using full, uncorrupted transcriptions as a baseline. We trained the models for 150,000 steps, and all models converged around 70,000 steps.

6.5.1.2 Precise Fine-tuning

All the models in the previous sections are then fine-tuned using 10 minutes of precisely transcribed and uncorrupted data. This approximates an extremely low-resource scenario.

6.5.2 Real World Case Study - Classroom Data

For a real-world case study, we consider the NCTE dataset. We use our CPT-adapted Wav2vec2.0 from Chapter 5, which was specifically adapted to classroom speech using the NCTE dataset.

6.5.2.1 Weakly Supervised Pre-training Step

We first train the intermediate model using the **NCTE-Weak** dataset. We partition the dataset into training and validation splits with no test data using a 90/10 split after removing all the recordings that were re-transcribed in the NCTE-Gold dataset to prevent data leakage. We test that model on the test set from NCTE-Gold. We trained the model using the configuration file used for training Wav2vec on all 960 hours of Librispeech which trains the model for 320,000 steps.

6.5.2.2 Precise Fine-tuning

We set the model trained on NCTE-Weak as initialization for our precise fine-tuning. We fine-tune that model using the NCTE-Gold training data for 500 steps using a few-shot learning scenario. We compare that model against two base-lines:

- **Direct Fine-tuning:** we directly fine-tune the CPT-Wav2vec2.0 model for 20,000 steps on NCTE-Gold, which is the same approach we followed in Chapter 5.
- **Self-training: [Amini et al., 2025, Grill et al., 2020]** We use the **Direct Fine-tuning** model to transcribe a portion of the NCTE-Weak dataset and add it to the NCTE-Gold

dataset. This adds 40 hours to the NCTE-Gold dataset. We fine-tune the CPT-Wav2vec2.0 model using this dataset for our second baseline.

6.6 Results and Discussion

6.6.1 Synthetic Corruption

We test each model regardless of corruption type or percentage on the uncorrupted TEDLIUM test set. We consider the results with and without LM beam-search decoding. Table 6.1 shows the WER results for each model.

The first row shows the performance of the model trained on the original dataset without any corruption. The results from the randomly corrupted training sets in the second column show that while random inconsistent transcription mistakes negatively affect the model's performance, they are not detrimental overall. We can see that even when half of the training data had transcription mistakes, the performance was barely affected and the degradation in WER was less than 1%. Even when all the transcriptions in the training data had random mistakes, the performance was degraded by about 2% with greedy decoding, and less than 1% with LM decoding.

In contrast, in fully corrupted training sets, where every word in the corrupted sentence was misspelled, the degradation in performance is noticeably higher. In the 25% and 50% corrupted training sets, the model could still perform reasonably well **with LM decoding**, however, the performance with non-LM greedy decoding is noticeably worse. This highlights one important distinction between the two corruption methods, where the gap between greedy decoding and LM beam search decoding is significantly larger with full corruption. This points to the fact that most of the misspelling mistakes are usually caught by the LM and corrected accordingly. While

Table 6.1: WER results from training Wav2vec2.0 with corrupted TEDLIUM transcriptions. "% Corr." indicates the proportion of utterances corrupted in the training set. Results before the slash (/) are from greedy decoding (no LM), while those after the slash use LM decoding.

% Corr.	Random Corr.	Full Corr.
0	**7.40/6.77**	
25	7.50/6.82	12.72/8.16
50	7.89/7.06	46.73/10.57
75	8.76/7.03	72.45/ 37.76
100	9.73/7.72	80.36/52.40

this doesn't work as well with higher degrees of corruption (75% and 100%), the improvement in performance with LM decoding is still significant, although the utility of the models at this level of performance is limited. Notably, the model still learns even under severe corruption, likely because our method replaces words with phonetically similar alternatives or common misspellings. While this maintains some phonetic alignment with the gold transcription, the resulting WER remains too high for practical use.

The results of fine-tuning the models trained on corrupted data on 10 minutes of accurate gold-standard data are in Table 6.2. We also fine-tuned the model that was trained on the full original uncorrupted training data as a baseline, which is shown in row 1. We can see that fine-tuning with a small amount of labeled data seems to cancel out the effect of random inconsistent corruption in the original training data with less than 50% random corruption. Further fine-tuning these models matches the performance of the model trained without any corruption in the first row of Table 6.1. With higher degrees of corruption, we still see some improvement with greedy decoding, but we also see, interestingly, that with LM decoding, the model originally trained with some random corruption in the transcription in 100% of the training labels only slightly underperforms the model trained without any corruption.

Table 6.2: WER results from finetuning models from Table 6.1 on 10 minutes of accurate uncorrupted TEDLIUM data. "% Corr." indicates the proportion of utterances corrupted in the training set of the pretrained model.

% Corr.	Random Corr.	Full Corr.
0	7.27/**6.48**	
25	**7.11**/6.50	10.31/7.53
50	7.49/6.65	13.30/8.33
75	8.10/6.55	33.91/19.0
100	8.31/6.86	43.90/ 24.49

With models that were originally trained with fully corrupted labels, fine-tuning with 10 minutes of accurate data helps them become more usable. With LM decoding, even with up to 50% full corruption of the training labels, fine-tuning on a small amount of labeled data reduces the performance gap caused by corruption to below 2%. Even with 100% corruption in the training data, fine-tuning on 10 minutes of accurate data relatively reduces the WER significantly, by around 45.37% with greedy decoding, and 53.26% with LM decoding.

To understand the impact of such results, we note that correcting 10 minutes of weak transcription by a human can take less than an hour, as each minute of transcription can take between 5-8 minutes [Cieri et al., 2004, Fox et al., 2021] and can cost around $25 [Williams et al., 2011]. A model trained with 100% of its training data fully corrupted, a very extreme case, can get a WER as low as 24.49% with fine-tuning on just 10 minutes of accurate data. While 24.49% is a high WER, it can be usable for many low-resource tasks and matches some baselines in classroom ASR as seen in Table 6.3.

We attempted to train the model directly using 10 minutes and 1 hour of labeled data without prior initial training as an additional baseline, but the model did not converge. We did not include that model in the table, but the reader can assume its error to be 100% since it did not learn to output any characters. While Wav2vec2.0 has been successfully trained with just 10 minutes of

Librispeech before, Librispeech is a much cleaner and more straightforward task than TEDLIUM. This highlights an important finding: not only does further fine-tuning on accurate data cancel out the effect of moderate corruption in the transcription, but even severe corruption significantly improves the utility of small precise data. While it remains potentially possible to train a Wav2vec model using limited amounts of data from TEDLIUM, the fact remains that prior training on severely corrupted data makes this training much more straightforward.

6.6.2 Real World Case Study - Classroom Data

Looking at Table 6.3, we see that the model trained on TEDLIUM performs poorly with both test sets which corroborates previous research [Attia et al., 2024, Southwell et al., 2024] that shows that ASR models trained on off-the-shelf data struggle with classroom speech. While training on the NCTE-Weak improves the results, it is outperformed by training on NCTE-Gold, a precisely transcribed dataset that is 500 times smaller. Using the resultant model to generate self-training data further improves the results on both datasets. Unlike our synthetic experiments, we do not mix NCTE-Weak with NCTE-Gold because NCTE-Gold is too small to make a measurable

Table 6.3: WER results on the NCTE and MPT test sets with different training configurations. NCTE-Weak → TED 10-Hr and NCTE-Weak → NCTE-Gold refer to models initially trained on the NCTE-Weak dataset and then fine-tuned on 10 hours of TEDLIUM and the NCTE-Gold dataset respectively.

Training Data	**NCTE**	**MPT**
TEDLIUM	55.82/50.56	55.11/50.50
NCTE-Weak	36.23/32.30	50.84/46.09
NCTE-Gold	21.12/16.47	31.52/27.93
NCTE-Self Training	17.45/15.09	27.42/26.24
NCTE-Weak → TED 10-Hr	25.59/21.14	42.62/37.22
NCTE-Weak → NCTE-Gold	**16.54/13.51**	**25.07/23.70**

impact, being only 0.25% the size of NCTE-Weak.

We ran two WSP experiments, both relying on first training on NCTE-Weak followed by precise fine-tuning. First, we consider the case where no in-domain accurate data exists. We use 10 hours of TEDLIUM, which we have already seen to be a poor match for classroom speech as evident from the first row of Table 6.3. However, few-shot fine-tuning on this dataset improves the results significantly from NCTE-Weak training, by about 10% in both configurations, while still underperforming direct precise fine-tuning on in-domain data. If in-domain accurate data exists, such as NCTE-Gold, we get further improvements, resulting in our best configuration for both NCTE and MPT test sets. In any case, this configuration outperforms both direct fine-tuning and self-training, further showing that initial imprecise training increases the utility of limited precise data.

6.7 Error Analysis

In this section, we discuss sample transcription from models trained with weak supervision, and from models further fine-tuned on small accurate data. With this analysis, we attempt to understand the effect of transcription errors on performance, and why initial large-scale weak pre-training improves performance.

6.7.1 Synthetic Corruption

The example below shows a sample transcription from the extreme case of 100% of the training data being fully corrupted as the WSP model, and model that was subsequently fine-tuned on 10 minutes of precise data. WER is shown in parentheses.

Ground Truth: *Everybody talks about happiness these days.*

100% Full Corr. (WSP): *e bod tal abou hapne thel da.* (116.67%)

WSP → 10 min FT: *Ever body talks about hapines thees das.* (83.34%)

At first glance, the transcription from the WSP model might seem complete gibberish. However, we can see that the model correctly detects some phonemic features from the sentence. For example “bod " matches “everybody", “ta" matches “talks", etc. This can be attributed to the fact that even though the majority of words in the training set were replaced by their soundalikes or common misspellings, these alternative words still match a lot of the sounds in the audio, which the model partially succeeds in detecting. Further precise fine-tuning builds on this weak foundation to output a legible transcription with a few spelling and structure mistakes. These mistakes are usually fixed by LM beam search decoding. In fact, LM decoding (transcription not shown) reduces the WER for this example to 16.67% with a single substitution error (day → das).

6.7.2 Real World Case Study - Classroom Data

The example below shows a sample transcription from three models, the model directly trained on NCTE-Gold, the model trained on NCTE-Weak (WSP), and the model initially trained on NCTE-Weak and then fine-tuned on NCTE-Gold. The models do not predict casing or punctuation, however they were added to improve legibility.

Reference: *33. You are right. Yours is correct okay. I won. Clean up. Go back to your seat. Do not put anything away. You are going to need it. 54. Faithful, you have to stay there, you can not move because of the camera.*

NCTE-Weak (WSP): *33. You re right. Yours is correct. 54. You have to stay there. n me csthe me.* (75.00%)

NCTE-Gold: *33. You are right. Yours is correct. kay. on one. Cleen up. Go back to your sseat. Do not put anything away, you are going to need it. 54. Faithfuel, you have to stay there, you can not move because of the camera.* (15.91%)

WSP → NCTE-Gold: *33. You are right. Yours is correctokay. On one. Clean up. Go back to your seat. Do not put anything away, you are going to need it. 54. Faithful, you have to stay there, you can not move because the camera.* (11.36%)

The weakest performance occurs when training directly on NCTE-Weak. Similarly to how the training data suffered from deletion errors, the transcription skips over large portions of the audio and picks up later. For instance, the phrases ”*Clean up ... You are going to need it.*" are completely missing, followed by correctly transcribing the number 54, and so on. By inspecting the audio, there does not seem to be any pattern to the deleted portions in tone or noise level. However, the model picks up the initial phrases of the audio correctly. The model often gets 0% error for shorter instances. However, even with the deletion mistakes, the model often captures portions of the audio correctly, which explains why fine-tuning this model with NCTE-Gold outperforms direct fine-tuning on the same dataset. By comparing both transcriptions, we see that initial WSP improves several aspects of the transcription, such as the misspelling of ”*clean*" as ”*cleen*", and the substitution of the name "Faithful". Fine-tuning the WSP model on NCTE-Gold

not only fixes the deletion mistakes caused by timestamp inaccuracies in NCTE-Weak but also utilizes the exposure to large amounts of weak data, resulting in improved robustness.

6.8 Conclusion

This chapter introduced Weakly Supervised Pretraining (WSP) as a complementary paradigm to CPT for domain adaptation in data-constrained ASR. CPT demonstrated in Chapter 5 that large amounts of unlabeled audio can be exploited to learn domain-adapted acoustic representations. WSP extends this idea by showing that imperfect textual supervision can provide additional structure when used as an intermediate training signal. WSP has the distinct advantage of practically expanding the amount of labeled data the model is exposed to, given the impracticality of obtaining large amounts of gold-standard data.

Through controlled synthetic corruption experiments on TEDLIUM, we showed that models trained on severely corrupted transcriptions can still learn useful acoustic representations, and that subsequent fine-tuning on as little as 10 minutes of accurate data yields usable ASR performance. The results even match training on the full, uncorrupted training data under certain conditions. We further validated this insight in a real-world classroom setting using the NCTE dataset, where WSP consistently outperformed direct fine-tuning on limited in-domain data and surpassed self-training baselines.

This chapter presents one practical response to the impracticality of large-scale gold-standard transcription by reducing the amount of accurate labeled data required to obtain robust performance. In the next chapter, we discuss a complementary approach that addresses the same bottleneck by generating supervision through simulation rather than annotation.

Chapter 7: RealClass: A Scalable and Shareable Classroom Speech Dataset via Simulation with Public Data

So far, we have established data scarcity as a major obstacle to improving performance of ASR models. This scarcity stems from the general rarity of classroom corpora, and critically, the corpora that exist are not publicly shareable. As a result, researchers working on similar problems often rely on different private datasets, making experimental results difficult to reproduce and slowing progress across the field. Access to protected classroom recordings is typically restricted due to privacy concerns, and even when data is collected, it cannot be redistributed [Attia et al., 2024, Southwell et al., 2022]. While Chapters 5 and 6 demonstrated that this limitation can be partially mitigated through unlabeled data and weak supervision, these unlabeled and weakly labeled corpora still have the same shareability limitations. In this chapter, we propose a solution to the data scarcity that also tackles shareability.

Although several children's ASR datasets are publicly available [Pradhan et al., 2023, Shobaki et al., 2000], they do not capture the defining characteristics of classroom speech, as we have previously discussed how classroom speech is distinct 1.2.2. Existing data augmentation techniques, such as pitch perturbation applied to adult speech [Fan et al., 2024], can approximate some acoustic properties of children's voices but fail to model the linguistic structure of child speech [Attia et al., 2023] or the acoustic complexity of classroom environments. In particular,

classroom speech is dominated by children's babble noise, which is defined as overlapping background speech from multiple speakers. Yet no publicly available corpus of children's babble noise exists. Adult babble corpora [Font et al., 2013, Snyder et al., 2015] are not an adequate substitute, as they differ substantially in both acoustic and linguistic characteristics.

In this chapter, we introduce *RealClass*, a scalable and publicly shareable classroom speech dataset. We present our novel dataset curation framework that aims to solve both issues: the scarcity of classroom data and the lack of children's babble noise corpora. We create a clean, noise-free classroom corpus by combining the My Science Tutor (MyST) corpus with instructional adult speech from lectures from MIT OpenCourseWare (OCW) and Khan Academy. These channels' videos are licensed under Creative Commons BY-NC-SA [Creative Commons, 2019], which permits non-commercial use, adaptation, and redistribution with proper attribution. In addition, we synthesize classroom noise, primarily children's babble, through the Unity Game Engine. The Unity Game Engine provides high-fidelity audio simulation in 3D virtual space, which accurately simulates the acoustic characteristics of the classroom environment, as well as the spatial characteristics of different audio sources. We use the same software to calculate Room Impulse Responses (RIRs) from the virtual classrooms to create the first classroom-specific RIR bank. We discuss this in detail in Chapter 7.2.2.

We call our dataset **RealClass**.[1] We plan to make RealClass and the tools developed to create it publicly available at no cost to researchers. RealClass is the largest and only public classroom speech dataset. In addition, the existence of a clean and noisy version of the dataset, which is not possible with real classroom recordings with naturally occurring noise, opens the

[1] Also known as SimClass in a preprint. We opted to rename the dataset to emphasize that the speech in this dataset is not generative, and to avoid confusion with other techniques that share the same name.

door to many tasks that were not possible, such as speech enhancement. The reason for that is that, unlike real noisy data, RealClass provides clean-noisy pairs for the same audio, required to train speech enhancement models. RealClass mainly simulates elementary school STEM classes based on the datasets used to construct it, but our methodology outlined in this paper paves the way for simulating other kinds of classrooms given different kinds of data.

Our work in this chapter is the first to demonstrate that classroom noise and RIRs can be synthesized at scale, validated through ASR experiments, and released publicly, filling a gap that neither augmentation nor private corpora can address.

The remainder of this chapter details the construction of the RealClass dataset, the simulation pipeline used to generate classroom acoustics and noise, and ASR experiments that validate the effectiveness of the proposed framework.

7.1 Datasets

7.1.1 Datasets Used To Create RealClass

7.1.1.1 MyST

We previously discussed in Chapter 3 our efforts to clean up the mistranscriptions in the dataset. We utilize the untranscribed portion of the dataset for noise generation. We also utilize the transcribed portion of the dataset, after clean-up, to act as a source of labeled children's speech in the data.

7.1.1.2 Khan Academy

Khan Academy is a popular YouTube channel hosting thousands of educational videos targeted at different school-age grades. Although Khan Academy videos provide a good match in both topic and target audience, as there is a variety of elementary STEM playlists, they mainly have the same primary speaker, which can limit the diversity of the data. However, to benefit from the agreement on the subject matter and topics, we include videos from this channel in our dataset.

We chose 7 playlists discussing topics in second and third-grade mathematics. The total duration of the videos in these playlists is about 7.5 hours, and they include high-quality audio accompanied by human transcripts.

7.1.1.3 MIT OpenCourseWare (OCW)

MIT OCW is an online repository of recorded lectures from 2,500 MIT courses. Although the classes in OCW are college-level and graduate-level classes, they still provide a useful resource for elementary school STEM classes, as the acoustic properties of instructional speech are evident in the videos. However, to help bridge the gap between the linguistic components of college-level courses and STEM classes, we picked videos from 10 courses in calculus and linear algebra, totaling around 174 hours.

These videos include high-quality, clear audio from a variety of lecturers with different accents, accompanied by high-quality human transcriptions.

7.1.2 Classroom Datasets Used For Testing: NCTE-Gold Dataset

We use the same NCTE-Gold dataset from Chapter 6 in this chapter. We use the training data to provide a baseline for training on real classroom data, and we use the testset to measure the performance on actual classroom data at test time.

7.2 Methodology

7.2.1 Creating The Clean Classroom Speech Base

To create the clean version of RealClass, we combine individual tracks from the adult corpora with tracks from the MyST dataset. To make the resultant tracks more semantically robust, we perform semantic matching between the adult and child corpora. Specifically, we encode the transcripts from both datasets with the SentenceTransformer [Reimers and Gurevych, 2019] model[2], and normalize the embeddings for similarity search. A FAISS[Douze et al., 2025] index greedily matches children-adult utterance pairs based on semantic similarity. This process produces semantically consistent clean speech pairs that approximate realistic classroom interactions. Below is an example of a semantically matched transcription:

Child: It flows to the positive end to the negative end.

Adult: What flows in the opposite direction?

To create variety, for each combination, the file can either start with child speech followed by adult speech, or vice versa. In either case, we allow for a random overlap between 0.5 seconds and 1 second for 20% of the files, simulating cut-off during turn-taking in normal conversations.

[2] `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`

The total duration of the RealClass dataset is 391 hours, making it the largest classroom speech corpus and the only publicly available one. As for the train/test/validation partition, we partitioned each constituent dataset before the combination. We follow the splits created by the authors of the MyST dataset, which ensures that no speaker exists in two splits. For the YouTube datasets, we partitioned them by channels to create a roughly 80/10/10 partition across both OCW and Khan Academy. As a result, we ended up with 313 hours of training data, 37 hours of development data, and 41 hours of test data.

7.2.2 Simulating Classroom Noises in Unity Game Engine

To create our classroom babble noise, we use the Unity Game Engine. While game engines are typically associated with graphics, modern frameworks like Unity also support high-fidelity acoustic simulations through plug-ins such as Steam Audio [Valve Corporation, 2025]. By leveraging the environment's geometry, Steam Audio models effects such as diffraction, occlusion, and reverberation, and supports "acoustic materials" that define surface properties, such as absorption, transmission, scattering, for objects in the scene [Nieminen, 2021].

We design a classroom environment by acoustically modeling all objects and surfaces, including chairs, desks, doors, windows, ceilings, and carpets. Within the classroom, we place 25 spatially directive audio sources playing tracks from the non-transcribed portion of MyST, representing 25 students. Each source has its own orientation and emission pattern, creating realistic overlapping speech that captures the character of children's babble. To add further realism, we include randomly placed chair noises and ambient playground sounds, sourced from YouTube, at random intervals. Noise is captured using a moving virtual microphone ("audio

listener") to record from different angles, resulting in 50 hours of physically simulated classroom noise.

7.2.3 Measuring RIR from Virtual Environments

An RIR characterizes how an acoustic environment transforms sound. Convolving an RIR with a clean audio signal produces a realistic rendering of how the signal would be perceived within that space. Building on this principle, we construct, to the best of our knowledge, the first dataset of classroom RIRs using a Unity-based simulation framework. This approach offers key advantages over directly re-recording audio in virtual environments: convolution is highly efficient, enabling hundreds of hours of audio to be simulated within minutes, and the resulting RIRs are easily shareable, allowing flexible reuse across diverse signals.

We follow the Exponential Sine Sweep (ESS) technique [Farina et al., 2000] to calculate RIRs. The system is excited with a sinusoidal signal whose frequency increases exponentially across the audible range (20 Hz–20 kHz). The response is recorded and then deconvolved using the inverse of the excitation signal to obtain the RIR.

To construct a diverse RIR bank, we simulate eight distinct classroom environments. In each room, five source positions are selected (center and corners), and for each source, we record responses at the remaining positions. This source–receiver pairing yields 20 unique RIRs per room (160 in total), capturing variability in both sound emission and listening position.

7.3 Validation

7.3.1 ASR Experiments

In this section, we validate our dataset through experiments on the classroom ASR benchmark. We use the Fairseq implementation of Wav2Vec2.0. Unless otherwise noted, models are initialized from the Robust-Wav2vec2.0[Hsu et al., 2021] checkpoint. We also use our CPT Wav2vec2.0 model from Chapter 5. We reserve the CPT model for training our final models that are compared to training on real classroom data.

7.3.1.1 Comparison with Non-Classroom Baselines

To contextualize the performance of our proposed RealClass dataset, we compare against Librispeech, a non-classroom baseline that is commonly used in ASR research, as well as the individual components of RealClass. In addition, we evaluate simple data combinations of these components, as well as our proposed semantic pairing method that aligns adult and child speech into synthetic dialogues. Results are in Table 7.1

Table 7.1: WER (%) on classroom test data for models trained on non-classroom corpora. *"Online Lectures"* refers to data from OCW and Khan Academy YouTube channels. RealClass outperforms all alternatives, showing the value of synthetic classroom simulation.

Training Data	**WER on NCTE**
LibriSpeech	40.64%
Online Lectures	38.33%
MyST	45.82%
Online Lectures + MyST	37.55%
RealClass	**35.52%**

As expected, training on LibriSpeech yields the worst performance on classroom speech,

with a WER of roughly 40%. This result highlights the severe domain mismatch: although LibriSpeech is a large and clean corpus, it lacks the overlapping speech, noisy conditions, and mixed child–adult speech distributions that characterize real classrooms.

Training only on instructional lecture data from OCW and Khan Academy (referred to as *"Online Lectures"* in Table 7.1) provides some improvement over LibriSpeech. This is likely due to a closer match in speaking style and domain vocabulary, as these lectures often resemble teacher discourse in classrooms. Although almost entirely dominated by teacher speech, incidental multi-speaker events (e.g., students asking questions) and the room acoustics present in the recordings introduce variability that overlaps with classroom conditions.

Training exclusively on MyST, a corpus composed entirely of children's speech, results in worse performance than LibriSpeech. This outcome is intuitive: although children's speech is part of the classroom domain, MyST consists of single-speaker tutoring sessions and lacks the adult-dominated discourse that characterizes most classroom speech. Thus, training on children's speech alone yields poor generalization to classroom test data. However, by simply concatenating instructional speech with MyST with no semantic matching, performance improves over either dataset alone, as well as over LibriSpeech. This result suggests that combining adult instructional data with children's speech provides a more balanced proxy for classroom conditions. Taking it a step further, our proposed RealClass methodology further improves performance by semantically pairing utterances from adult and child datasets into synthetic dialogues, with randomized overlaps to mimic interruptions. This curation process not only balances the linguistic roles of adults and children but also more closely approximates the turn-taking dynamics of real classrooms. RealClass serves as a strong baseline for the subsequent addition of simulated room acoustics and noise, validating our approach to dataset construction.

7.3.1.2 Ablation Studies

This section examines both the individual contributions of synthetic realism (RIR, noise) and the cumulative contributions when combining synthetic and real classroom data. Results are shown in Table 7.2.

Table 7.2: Ablation study of RealClass with synthetic realism and continued pretraining. ΔWER (abs.) is the absolute WER reduction compared to RealClass alone. The top block reports independent ablations of RealClass with RIR, additive noise, or both, while the bottom block reports cumulative improvements when further adding real classroom data (NCTE) to the training data and using the CPT Wav2vec2.0 model from Chapter 2.3.5.

System	**WER (%)**	**Δ WER (abs.)**
RealClass	35.52	–
RealClass + Children Babble Noise	32.51	3.01
RealClass + Room Acoustics (RIR)	33.52	2.00
RealClass + RIR + Children Babble Noise	32.76	2.76
Cumulative additions:		
+ CPT	26.24	9.28
+ Additive Real Classroom Data (NCTE)	**19.98**	**15.53**
Baseline: Only Real Classroom Data (NCTE)	*21.12*	*14.40*

As shown in Table 2, convolving RealClass with simulated room impulse responses (RIRs) reduces WER by about 2% absolute, indicating that the addition of classroom-like acoustics makes the synthetic data more consistent with the test conditions. Similarly, adding simulated classroom babble noise improves WER by about 3%, highlighting the effectiveness of modeling the noisy conditions typical of real classrooms. These results suggest that both acoustics and noise individually push the synthetic data closer to real classroom distributions. However, when RIR and noise are combined, performance slightly degrades. We attribute this to an artifact of the current pipeline: the clean speech is first convolved with an RIR, while the noise track already contains reverberation from the Unity simulator. Mixing the two may produce mismatched

acoustics, which may affect the performance. This is a current limitation of the dataset, and we leave matched-acoustics rendering for future work.

When switching from the robust Wav2Vec2.0 model to our CPT model, which has been adapted to classroom data in Chapter 2.3.5, performance improves substantially (from 32.76% to 26.24% WER). This is in line with earlier findings that CPT is the most effective initialization for classroom ASR. We kept earlier ablations fixed to the robust model, to isolate the effect of CPT. Comparing against the NCTE baseline, RealClass+RIR+Noise+CPT achieves 26.24% WER, while NCTE training achieves 21.12% WER. RealClass provides a strong approximation to training on real classroom data when such data is unavailable.

Finally, combining RealClass with NCTE yields the best performance of 19.98% WER, surpassing the use of NCTE alone. This demonstrates a second use case for our dataset: not only can it serve as a standalone substitute when real classroom data is scarce, but it can also act as a complementary resource to real data, further enhancing performance when the two are combined.

7.3.2 Speech Enhancement Experiments

The separation between noise and clean speech in the RealClass dataset opens the door for new tasks in classroom speech AI not previously possible with real classroom recordings, such as speech enhancement. To showcase that, we fine-tune the StoRM diffusion-based speech enhancement model [Lemercier et al., 2023]. We fine-tune the StoRM model, initially trained on the Voicebank/DEMAND[Valentini-Botinhao et al., 2016] dataset, using the RealClass-noisy train set mixed at different SNRs. We chose this pre-trained model because Voicebank/DEMAND is a well-established speech enhancement dataset that includes babble noise in its training set,

making it a suitable match for our RealClass speech enhancement task.

Given the slow inference of the diffusion model, we evaluate only 100 randomly sampled files from the RealClass test set.We ensure diversity in the selection and confirm stability across multiple runs. Table 7.3 shows key metrics from both the fine-tuned and off-the-shelf models on the sampled RealClass test set mixed with simulated classroom noise at SNRs from -5 to 15 dB at 5 dB intervals. We evaluate speech quality using Perceptual Evaluation of Speech Quality (PESQ) [Rix et al., 2001], intelligibility with Extended Short-Term Objective Intelligibility (ESTOI) [Jensen and Taal, 2016], and overall signal fidelity with Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [Le Roux et al., 2019]. Table 7.3 demonstrates that fine-tuning StoRM with RealClass data

Table 7.3: Key speech enhancement metrics of the StoRM diffusion speech enhancement model both off-the-shelf and fine-tuned on the RealClass training data, when tested on the RealClass test data at different SNRs.

StoRM Off-The-Shelf			
SNR	**PESQ**	**ESTOI**	**SI-SDI**
-5	1.085	0.308	-2.815
0	1.167	0.501	4.288
5	1.376	0.644	9.132
10	1.605	0.743	12.827
15	1.923	0.805	15.530
StoRM Finetuned on RealClass			
SNR	**PESQ**	**ESTOI**	**SI-SDI**
-5	**1.388**	**0.596**	**8.093**
0	**1.678**	**0.710**	**11.131**
5	**2.034**	**0.787**	**14.157**
10	**2.453**	**0.842**	**16.988**
15	**2.861**	**0.882**	**19.795**

leads to consistent and significant improvements across all metrics and SNR levels. While the performance at low SNRs still has low integrability and perceptual quality, our results highlight

Table 7.4: ASR WER performance with the W2V-Classroom model fine-tuned on RealClass Noisy and NCTE training sets, evaluated on a subset of the RealClass test set. The test set is either mixed with noise at different SNR levels, enhanced using our speech enhancement model, or a mixture of enhanced and noisy signals.

SNR	Noisy	Enhanced	Mixed
-5	33.01	29.63	**22.35**
0	17.17	17.06	**13.38**
5	12.65	12.35	**10.63**
10	9.85	9.19	**8.82**
15	9.49	8.79	**8.24**
Clean		7.68	

the value of RealClass as a realistic and challenging dataset for classroom speech enhancement.

Furthermore, we test our enhanced speech signals with the ASR system trained on the RealClass noisy training data and the NCTE dataset. Table 7.4 shows the results. While speech enhancement improves the results, the improvement is marginal at SNRs above -5. It is a documented phenomenon that speech enhancement introduces artifacts that can negatively impact ASR systems [Lemercier et al., 2023]. To that end, we also mix the noisy signal with the enhanced signal, a technique previously shown to improve ASR robustness by mitigating enhancement artifacts [Fujimoto and Kawai, 2019]. This combination shows significant improvement in the performance especially under low SNR.

Our goal with this analysis is not to introduce a groundbreaking classroom speech enhancement model but rather to demonstrate the difficulty of the task and how RealClass makes such research and analysis possible. By making RealClass available, we hope to pave the way for deeper investigations into classroom-specific enhancement techniques.

7.4 Conclusion

In this chapter, we introduced RealClass, a simulated classroom dataset that integrates clean adult and child speech with room impulse responses and classroom noise. Our experiments demonstrate that these components, when combined, provide a strong approximation to real classroom speech.

The results highlight two key use cases. (1) In the absence of real classroom data, which is a common challenge due to privacy constraints and collection costs, RealClass serves as a close analog, achieving performance comparable to models trained on real data. (2) In scenarios where limited classroom data is available, combining RealClass with real classroom recordings yields further improvements, surpassing training on the real data alone. This establishes RealClass as a valuable companion resource to mitigate data scarcity. More broadly, RealClass reframes how classroom ASR research can be conducted. Rather than relying on non-shareable datasets collected under restrictive conditions, researchers can now work from a common, extensible foundation that supports reproducibility, systematic ablations, and fair comparison across methods. This capability is particularly important for low-resource educational domains, where data collection is costly, access is limited, and annotation is often impractical.

With this foundation in place, the remainder of this dissertation shifts focus from training paradigms and data constructions to learning strategies. In the next chapter, we examine how the proposed approaches compare with the latest state of the art Large Language Model (LLM) based ASR models, and how they can be integrated with them.

Chapter 8: Large Language Model–Based ASR in Classroom Environments: A Stress Test of Scale and Context

Recent years have witnessed the disruptive force of LLMs in the majority of AI fields, and ASR is no exception. LLMs, such as GPT [Achiam et al., 2023], Gemini [Team et al., 2023], and Claude [Anthropic, 2024] have evolved from being primarily text-based models, into multi-modal models that can process audio and images as well as text. This evolution led to models such as GPT-4o [Hurst et al., 2024] which can understand speech audio and respond accordingly.

This evolution caused another paradigm shift in ASR, in which LLM-ASR models inherit LLM-based decoders attached to ASR speech encoders. Examples of such models include Omnilingual ASR [Omnilingual et al., 2025] and Qwen-ASR [Alibaba Group, 2024]. These models have shown strong performance on general ASR benchmarks. They benefit from larger-scale training, advances in text-to-speech models that allow the creation of large synthetic datasets, and a wealth of world knowledge thanks to their LLM decoders.

Another key advantage of building the model with an LLM decoder is that it allows for flexible and native contextual biasing [Bhattacharjee et al., 2024]. In ASR, contextual biasing refers to providing the model with extra context distinct from the speech signal that allows for higher performance. This context allows for better disambiguation of similar words under adversarial conditions. For example, if we know that the recording came from the classroom, the

transcript "We'll talk about times tables next" is more plausible than "We'll talk about Thames tables next", even though both utterances might sound similar in certain acoustic conditions and are both grammatically plausible. Older models, such as Wav2vec2.0 and Whisper, will attend to the context from the same sentence and from the training data, but do not natively allow for additional context to be added at test time, unlike LLM-based ASR models.

In this chapter, we ask the question, how do the latest LLM-based ASR models compare to our best model adapted through CPT, WSP, and supervised fine-tuning in classroom benchmarks? In other words, do the vast world knowledge of LLMs and the power of contextual biasing outperform careful domain adaptation of an encoder-only model such as Wav2vec2.0?

We conduct experiments using Qwen-Omni2.5, an open-source multimodal LLM with an audio encoder and an LLM decoder. We evaluate both off-the-shelf performance and parameter-efficient fine-tuning, and we explicitly test contextual prompting using realistic classroom meta-data. We find that, while contextual prompting and fine-tuning improve the LLM-based system, the domain-adapted classroom ASR model remains substantially more effective in these conditions. We conclude by discussing how the domain adaptation strategies developed in this dissertation could be integrated into future LLM-centric ASR systems to improve robustness in low-resource settings.

8.1 Background: LLM-Centric ASR

8.1.1 Limitations of Language Modeling in Speech-Native ASR

Recall from Chapters 2.2 and 2.3, we discussed how the design of transformer-based ASR models such as Whisper and Wav2vec2.0 is inspired by successful architectures in language

modeling. For instance, the SSL pre-training design of Wav2vec2.0 is inspired by the acausal masked language modeling training objective of BERT [Devlin et al., 2018]. Furthermore, Whisper's decoder is explicitly trained as an audio-conditional language model. However, in either model's design, the models are only exposed to textual language during their speech recognition training. This means that the amount of language modeling the model can learn is upper-bounded by the amount of transcribed speech data that exists. The success of LLMs that scrape the entire internet for training textual data has proven that language modeling is a data-hungry task that requires substantially more training data than exists in coupled speech-transcription data.

8.1.2 Multi-Modal LLMs in ASR: Qwen-Omni2.5

Multi-modal LLMs leverage the abundance of textual training data to achieve robust performance across a number of tasks in text, audio, and video modalities. To understand how multi-modal LLMs work, we take a look at Qwen-Omni2.5 [Xu et al., 2025], a powerful and open-source multi-modal LLM. Qwen-omni2.5 can process inputs in audio, video, and text modalities and can respond either in text or in audio, through its speech synthesis capability.

The model consists of three main parts:

- **Audio and Visual Encoders**, that encode the audio and video into representations that can condition the language model.
- **Thinker**, which is the core part of the model and acts as an LLM
- **Talker,** which is responsible for the speech generation based on both high-level representations and embeddings of the text tokens sampled by the Thinker.

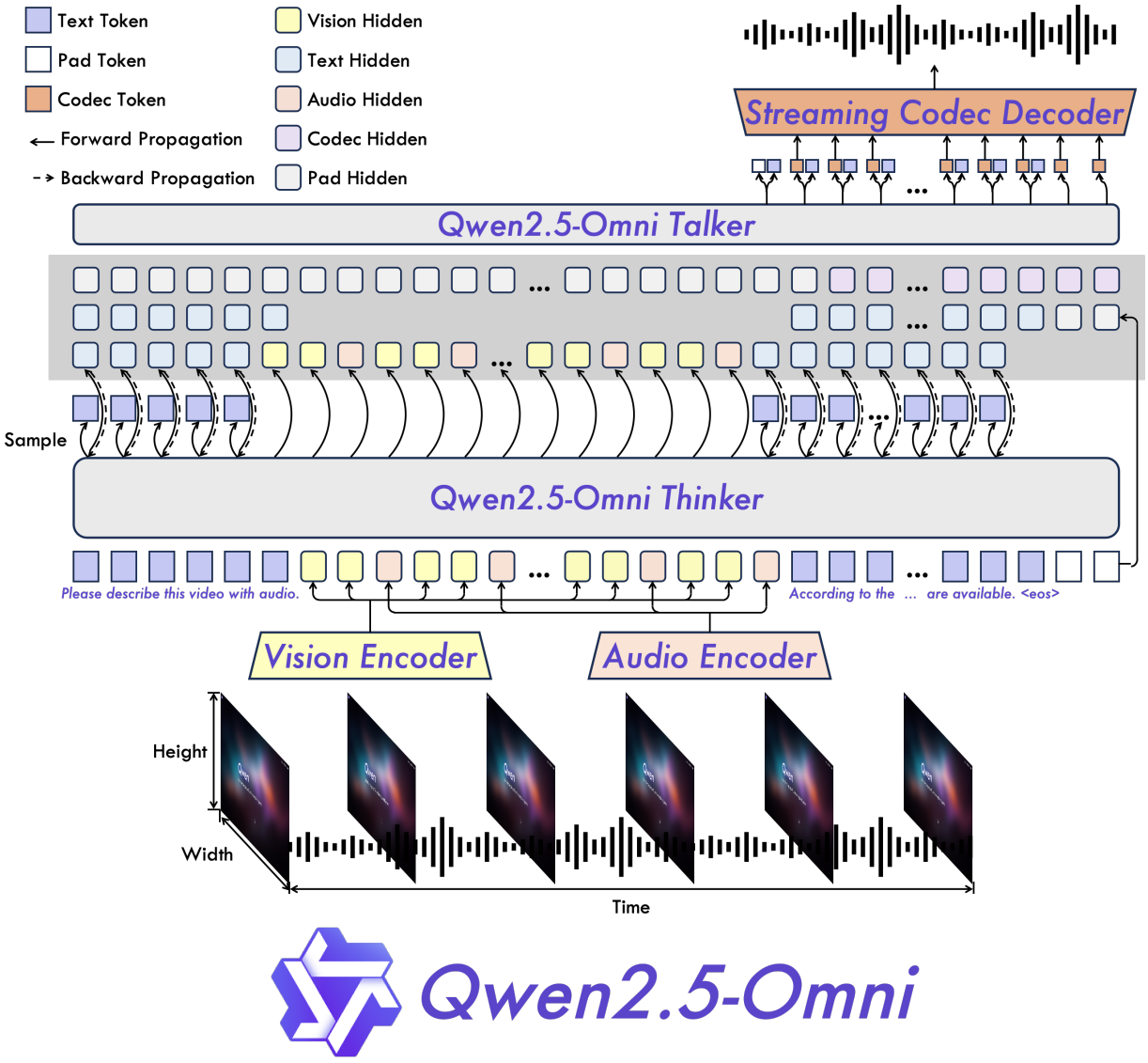

Figure 8.1: Qwen-Omni2.5 Architecture. Figure from Xu et al. [2025]

Both the Talker and the video encoder are out of context for this thesis, so instead we will focus on the Thinker and the audio encoder.

8.1.2.1 The Thinker

The Thinker is initialized from Qwen2.5, a text-only LLM [Qwen et al., 2025]. In other words, we can consider the Thinker to be pre-trained as a text-only language model before introducing other modalities. This helps the model utilize the vast amount of text pre-training data available, which ultimately improves the multi-modal performance and strengthens the model's language modeling priors. As we will later discuss in contextual biasing, this is precisely why we can improve the model ASR performance by simply mentioning "This is a geography class", and this will bias the model towards domain-consistent vocabulary such as "valleys, rivers, borders, etc".

8.1.2.2 Audio Encoder

Audio capabilities are added to the model through an audio encoder initialized from the encoder of *Whisper-large-v3*. This encoder is then trained to map the input audio to the text token space. To ensure this, the first stage of the pre-training process (after pre-training the Thinker as a text-only LLM) freezes the Thinker and only trains the encoders. The model is then trained end-to-end after unfreezing the Thinker in the second stage. In both these stages, the model is trained on audio-text, video-text, and audio-video-text tasks to evolve its multi-model perception abilities, but "pure text data plays an essential role in maintaining and improving language proficiency." [Xu et al., 2025].

While the technical report does not explicitly mention which audio-text task the model was trained on, it does mention "a wide variety of tasks". However, they perform comprehensive evaluations on ASR, Speech-to-Text Translation (S2TT), Speech Entity Recognition (SER), and Vocal Sound classification (VSC).

8.1.2.3 ASR with Qwen-Omni2.5: Contextual Biasing

To perform ASR with Qwen-Omni2.5, we feed it the audio file alongside a textual prompt, such as "Transcribe the English audio into text". Evaluations of Qwen-Omni2.5 show that it matches or exceeds the performance of Whisper-large-v3 and other ASR and LLM models on multiple benchmarks, such as Librispeech, Fleurs, and Common Voice.

We can perform contextual biasing naturally by adding the context to the textual prompt. For example, the textual prompt can be "Transcribe this audio recorded in a Geography class into text. The audio contains multiple speakers. Their names are John, Jane, and Doe."

8.2 Motivation

The goal of this chapter is not to establish state-of-the-art results for LLM-based ASR, but to evaluate whether the advantages of LLM-centric architectures, such as scale, strong language modeling priors, and flexible contextual biasing, translate into improved robustness in classroom environments. Classroom ASR presents a particularly challenging benchmark due to noise, overlapping speech, long-form discourse, and the presence of children's speech.

We compare Qwen-Omni2.5 against our best-performing Wav2vec2.0 model, which was adapted via CPT, WSP, and supervised fine-tuning on the NCTE-Gold training set. This compari-

son tests whether the language modeling capacity of the LLM Thinker, combined with contextual biasing, can match the performance of a carefully domain-adapted speech-native ASR pipeline.

We evaluate the ASR performance of Qwen-Omni2.5 on the NCTE-Gold dataset along two axes:

- Whether the model is adapted to classroom speech.
- Whether additional contextual information is provided alongside the audio input.

This comparison is designed to measure the effectiveness of contextual biasing, and to assess whether inference-time context can partially or fully substitute for domain-specific fine-tuning.

8.3 Experiments

8.3.1 Datasets

8.3.1.1 Training Data

We utilize both the NCTE-Gold and the RealClass Noisy training data for the adaptation of Qwen-Omni2.5 into the classroom ASR domain. We experimented with only using NCTE-Gold, but achieved better results by including RealClass Noisy in the training. The total amount of training data was about 323 hours (313 hours from RealClass, 10 hours from NCTE-Gold).

8.3.1.2 Test Data

We tested all model configurations in this chapter on the NCTE-Gold test data used in Chapters 6 and 7.

8.3.2 Context Construction

In the experiments with context construction, we generated realistic classroom metadata for each classroom in the NCTE-Gold training and test sets. We focused on metadata that would be readily and realistically available for each class at test time. More specifically, we generated the lesson plan and the student name list by feeding GPT-4o the full transcript from every classroom and prompted it to generate a context prompt for LLM-ASR based on the transcription. This generated context is kept static per classroom. To prevent unintentional data leakage, we ensure that the context prompts do not contain any ground-truth transcript text.

Below is our exact prompt to GPT-4o and a sample response.

Prompt: Given this ground truth transcript of classroom recordings, generate a prompt that feeds the transcriber context about the class. Include topic, short lesson plan, and names of students.

...Full Transcript...

GPT:

The recording is from a 5th-grade math lesson focused on understanding and applying positive and negative integers using real-life contexts and interactive gameplay.

Short lesson plan:

- Review prior learning from a class-created anchor chart about negative and positive numbers.
- Reinforce key concepts such as temperature, bank balances, game scores, and number lines.
- Connect real-world scenarios to number values: golf scoring, thermometers, and banking.
- Introduce and model the "Number Line Tug of War" game, including setup and objectives.
- Conduct a guided group game round to model gameplay, including student participation.
- Allow students to partner up and play several rounds, switching team goals (positive vs. negative).
- Conclude with a written reflection about fairness, strategy, and operations used in the game.

Students in this class include: Anthony, Miguel, Orion, Isaiah, Segala, Aaron, Michael, Bobby, Rocco, Alexander, Noisha, Kobe, Alicia, Nathan, Samantha, Angel, and others.

8.3.3 Models

We perform our experiments with Qwen-Omni2.5 using the Huggingface implementation of the 7 Billion parameter variant of the model. We disable the talker and the video encoder to reduce the memory footprint of the model. All experiments are conducted using text output only, with speech synthesis and video encoding disabled.

For Wav2vec2.0, we use our best performing model from Chapter 6, that was adapted using CPT, WSP and fine-tuned on NCTE-Gold. We note here that the size of this model is 300 Million parameters. This highlights a key advantage of this approach as it has a significantly lighter memory footprint than the LLM approach.

8.3.4 Qwen-Omni2.5 Adaptation Startegy

8.3.4.1 Background: LoRA

In the experiments where we adapt the Qwen-Omni2.5 model, we adapt it using Low Rank Adaptation (LoRA) [Hu et al., 2021]. LoRA is a Parameter Efficient Fine-Tuning (PEFT) that allows the adaptation of very large models under GPU and memory constraints. Instead of training all the billions of parameters in the model, LoRA freezes the original model weights and adds small, trainable adapter layers that can adapt the model to new domain data. Specifically, the weight update ΔW is factorized into two small, low rank matrices B, and A. More concretely,

$$\Delta W = A \cdot B \tag{8.1}$$

$$W' = W + \Delta W = W + A \cdot B \tag{8.2}$$

Consider W, a large matrix, with a shape of R_1 by R_2, i.e $R_1 * R_2$ parameters. In LoRA adaptation with rank $r <<< R_1, R_2$ instead of training the full $R_1 * R_2$ parameters, we only train $R_1 * r + R_2 * r$.

For example, if $R_1 = R_2 = 768$, and we set LoRA rank $r = 32$, LoRA reduces the trainable parameters from $589,824$ to $49,152$, which is 8.3% of the original size.

During inference, these adapters are fused with the original weights, causing no additional latency. LoRA provides a practical trade-off between adaptation capacity and computational feasibility in large multimodal models/

8.3.4.2 LoRA Implementation and Limitations

We utilize the LLaMA-Factory [Zheng et al., 2024] library to fine-tune Qwen-Omni2.5 using LoRA. This library provides PEFT recipes for a range of modern open-source LLMs. In this context, a recipe refers to the insertion of LoRA adapter layers into selected components of the model. At the time of our experiments, LoRA adapters were supported only for the Thinker component.

Restricting adaptation to the Thinker isolates the contribution of linguistic adaptation and contextual biasing, which aligns with the primary goals of this chapter. Fine-tuning the audio encoder is a natural extension of this work; however, it is not currently supported in LLaMA-Factory, and full fine-tuning is computationally expensive and infeasible within the available GPU memory. We therefore reserve audio-encoder adaptation for future work.

Table 8.1: Word Error Rate (WER) on the NCTE-Gold test set. FT refers to the model being LoRA fine-tuned on RealClass and NCTE-Gold training sets. Context refers to contextual biasing via text prompting.

Model	WER
Qwen-Omni2.5	33.19%
Qwen-Omni2.5 + Context	27.02%
Qwen-Omni2.5-FT	25.11%
Qwen-Omni2.5-FT + Context	22.47%
Baseline: Wav2vec2.0-CPT-WSP	**13.51%**

8.4 Results and Discussion

8.4.1 LLM-ASR Ablation: Contextual Biasing vs Fine-tuning

Table 8.1 shows the results from our experiments. The first thing we notice is that contextual biasing is significantly effective in improving the model performance, absolutely reducing the WER by 6.17% and 2.64% in the off-the-shelf and fine-tuned model respectively . Fine-tuning also significantly improves the performance, absolutely reducing the WER by 8.08% and 4.55% without and with contextual biasing respectively.

When we compare the model that was fine-tuned but not provided context, and the model that was provided context without any fine-tuning, we see that contextual biasing only slightly under-performs fine-tuning. In realistic test time settings, contextual biasing is free, requiring no training data, and only slightly increasing inference time. Considering that classroom data is rare, and in many cases non existent, contextual biasing can serve as a strong alternative to fine-tuning when labeled data is scarce or unavailable. It is also worth noting here that we did not run any ablation studies on what is the most effective prompt structure for contextual biasing, meaning that there may be room for improvement, further narrowing the gap. In any case, if fine-tuning is feasible, contextual biasing also adds more robustness to the model performance.

8.4.2 LLM-ASR vs Adapted Wav2vec2.0

When we compare the best performing Qwen-Omni2.5 configuration against our best performing Wav2vec2.0 adapted model, we see that the Wav2vec2.0 model still outperforms the much larger Qwen-Omni2.5 model, by about 9% WER (absolute reduction). This highlights the effect of the work laid out in this dissertation so far. We attribute that to a number of key reasons:

- The Wav2vec2.0 model has undergone acoustic adaptation through CPT, which, as we have seen previously, is the biggest contributor to its performance. The linguistic characteristics of classrooms are among the bottlenecks causing classroom speech to remain an open area of research. We saw this in how adapting Qwen-Omni2.5 Thinker (its linguistic brain) improves the performance. However, this result shows acoustics of the classrooms are by far the more challenging problem. Add to that, that the Wav2vec2.0 model has undergone some linguistic adaptation as well, though WSP and supervised fine-tuning. It is clear that the language modeling in Wav2vec2.0, either learned implicitly in the model itself or through the n-gram language model decoding, is far weaker than a modern LLM. However, when combined with careful acoustic domain adaptation, Wav2vec2.0 outperforms a large multimodal model in the classroom ASR setting.

- The main advantage that the large multi-modal model has is its strong linguistic priors. These are achieved through large scale pre-training of its Thinker LLM, mostly on scrapped internet data, and on proprietary data as well. However, the pattern that keeps emerging is that the protected nature of classrooms results in very little leakage of classroom data into public and even proprietary data. This results in many off-the-shelf models, powerful as

they may be under performing domain adapted smaller models.

8.5 Future Directions

The experiments in this chapter show both the promise of LLM inspired ASR solutions in how they allow for native contextual biasing, but also the challenges of such solutions. This leads naturally to an important future experiment: how much would these results change if we fine-tuned the audio encoder as well? This is a feasible question to answer with enough GPU hardware. However, we take this a step further by asking: is it possible and feasible to replace Whisper-large initialized audio encoder with our CPT Wav2vec2.0 audio encoder? The answer to this question relies on the amount of data available to this experiment, which does not necessarily need to be all from the classroom domain, and the available computing resources. Unfortunately, such datasets and computing power are out of reach for most academic research labs.

However, we also want to highlight more feasible solutions. In the recent few months prior to writing this dissertation, an important ASR model has been published, namely **Omnilingual ASR**. Unlike other recent models like **Qwen-ASR**, an exclusively ASR model from the Qwen family of models, which is available only through API access without public weights, Omnilingual ASR is open-weights, making it available to experiment with and fine-tune. Furthermore, Omnilingual ASR is based on the Wav2vec2.0 architecture, making all the solutions presented in this dissertation directly applicable to it. Omnilingual ASR adds LLM decoding allowing for contextual biasing, similar to Qwen-Omni2.5.

8.6 Conclusions

In this chapter, we evaluated whether recent LLM-centric ASR models, which combine large-scale language modeling with native support for contextual biasing, offer improved robustness for classroom speech recognition. Using Qwen-Omni2.5 as a representative open-source multimodal LLM, we conducted a controlled comparison against a carefully adapted speech-native ASR model developed earlier in this dissertation.

Our results show that both contextual biasing and parameter-efficient fine-tuning substantially improve the performance of the LLM-based system. Contextual prompting alone yields large gains without requiring any labeled speech data, while fine-tuning further improves recognition accuracy. These findings highlight the practical value of LLM-based ASR systems, particularly in settings where domain-specific training data is scarce and inference-time metadata is available.

However, despite these improvements, the LLM-based system remains significantly less effective than the speech-native Wav2vec2.0 model adapted through continued pretraining and weakly supervised learning. This performance gap persists even when the LLM-based model is provided with realistic contextual biasing and in-domain supervision. The results suggest that, in classroom environments, acoustic domain mismatch remains the dominant challenge, and that careful acoustic adaptation is more impactful than large-scale language modeling alone.

Taken together, these findings indicate that scale and contextual biasing do not alleviate the need for domain-specific speech modeling in challenging, low-resource settings. While LLM-centric architectures introduce powerful new capabilities, particularly for integrating external context, they do not yet replace the benefits of targeted acoustic and linguistic adaptation. Instead, the results of this chapter point toward hybrid approaches that combine the strengths of speech-

native adaptation with LLM-based decoding and contextual reasoning, a direction that is explored further in the subsequent discussion of future work.

Chapter 9: Articulatory-Informed ASR: Model-Based Improved Robustness of Acoustic Modeling Through an Auxiliary Speech Inversion Task

The results in the previous chapter highlight an important and consistent theme across this dissertation: robust performance in low-resource settings like classrooms remains fundamentally constrained by acoustic modeling, more so than linguistic modeling. Chapter 8 demonstrates that even large, context-aware LLM-based ASR systems, despite strong language priors and inference-time contextual biasing, struggle in classroom environments characterized by noise, overlap, and speaker variability. These findings reinforce the conclusion that improvements in language modeling and scale alone are insufficient to overcome acoustic mismatch.

Thus far, this dissertation has explored primarily data-driven approaches to improving acoustic modeling, including CPT, WSP, and large-scale simulated data generation. While effective, these methods rely on increasing the quantity, diversity, utility or quality of training data, without modifying the acoustic modeling itself. In other words, any improvements in robustness remain constrained by the representational limitations of the acoustic model..

This observation motivates the exploration of a complementary class of methods: model-driven approaches that improve acoustic modeling through auxiliary supervision. Rather than relying solely on data to implicitly learn phonetic structure, these approaches explicitly guide the model toward representations that reflect properties of speech production.

In this chapter, we investigate articulatory-informed acoustic modeling as a principled example of such a model-driven approach. Although this work is not evaluated directly in classroom settings, it targets the same acoustic challenges identified throughout this dissertation and provides a foundation for future extensions to classroom ASR.

The main research goal in this chapter is to integrate articulatory variables in ASR. Articulatory variables track how different articulators (lips, tongue tip, tongue body, velum and glottis) move during speech, and are highly correlated with the speech signal. In fact, articulatory representations have been successfully used in earlier generations of speech recognition models before data-driven, transformer architectures took over. In this chapter, we investigate a novel, scalable approach to reintroduce speech articulation into ASR.

9.1 Background

9.1.1 Speech Articulation

Under the classical the source-filter model of speech production [Stevens, 2000], where the source is the air propelled by the lungs through the vocal cords, the filter is vocal tract. Vocal Tract Variables (TVs) are trajectories that show the motion and position of the articulators along the vocal tract. These signals have been shown to be a very effective representation of speech in a number of fields, including speech pathology [Benway et al., 2025a,b], mental health diagnosis [Premananth and Espy-Wilson, 2025, Premananth et al., 2025, Siriwardena et al., 2021a,b], speech synthesis [Ling et al., 2013, Richmond and King, 2016], and relevantly, Automatic Speech Recognition (ASR) [Mitra et al., 2010, 2017].

9.1.2 Speech Articulation in ASR

In speech recognition, previous works [Mitra et al., 2010, 2017] have shown TVs to be strong and robust input signals, particularly under noise. However, these works have been confined to shallow feed-forward networks or Convolutional Neural Networks (CNNs). Additionally, these works only worked with synthetic rather than natural speech. Nonetheless, they demonstrated that TVs, being less redundant and noisy than acoustics, are effective for ASR. The motion of different parts of the vocal tract (lips, tongue tip, tongue body, etc) is a much more concise representation of speech than the raw acoustic signal.

Despite that promise, TVs have seen little use in modern transformer-based ASR models such as Whisper and Wav2vec2.0. These models require large amounts of training data. On the other hand, the process of collecting articulatory data involves the use of technologies like X-Ray Microbeam (XRMB) [Westbury, 1994], or Electromagnetic articulography (EMA) [Schönle et al., 1987]. These recording sessions are expensive, and long sessions can be uncomfortable and unsafe for the subjects. As a result, articulatory datasets are only a few hours long, which is not enough to train transformer ASR models.

9.1.3 Acoustic-to-Articulatory Speech Inversion

Acoustic-to-Articulatory Speech Inversion (SI) is the process of extracting the TVs from the acoustic signal [Sivaraman et al., 2019]. Using SI models, one can theoretically obtain enough TV data to train large transformer models. Figure 9.1 shows how speech inversion models infer TVs from the acoustic signal of speech. SI systems are trained on limited paired acoustic-articulatory datasets like the XRMB dataset [Westbury, 1994]. They can then be used to extract articulatory

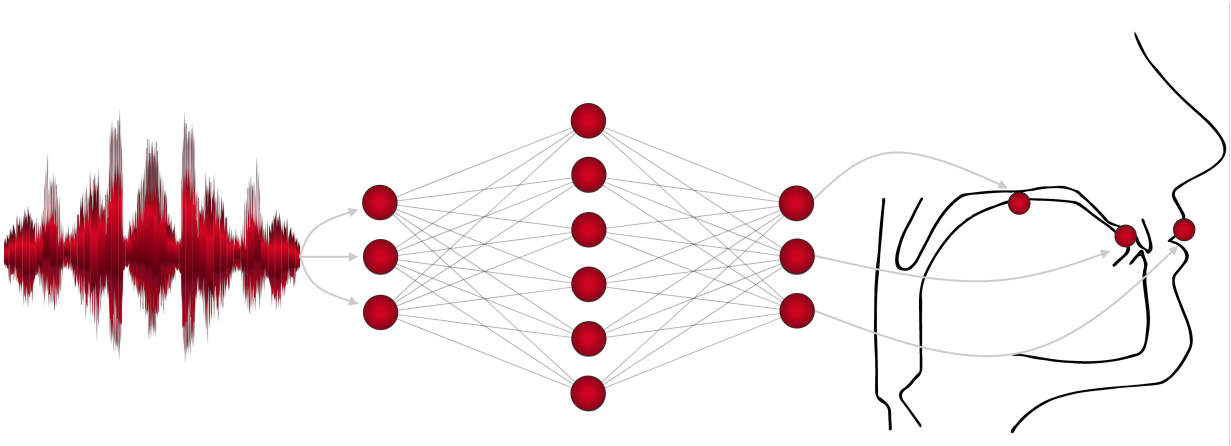

Figure 9.1: Illustration of acoustic-to-articulatory speech inversion. A neural network maps acoustic speech features to continuous articulatory variables representing vocal tract movements during speech production. Points on the left side of the diagram represent tongue body, tongue tip and lip constrictions.

parameters in place of using expensive technologies like XRMB or EMA.

9.2 Methodology

9.2.1 The Impracticality of Pure Articulatory ASR

The goal of this chapter is to investigate the reintroduction of TVs into ASR models. We first consider the practicality of a pure articulatory ASR model. Training a large transformer ASR model, like Whisper or Wav2vec2.0, from scratch is expensive and time-consuming, which has unfortunately become out of reach for many academic institutions. Many researchers rely on fine-tuning pre-trained off-the-shelf models to sidestep the massive resources required to train these models from scratch. Using pre-trained models, however, prevents any changes in the input side of the model, because doing so shifts the distribution of the activations for later layers.

Importantly, this limitation is not conceptual but practical: current state-of-the-art ASR systems rely on large-scale acoustic pretraining, making pure articulatory input incompatible

without retraining models from scratch.

9.2.2 Multi-task Acoustic-Articulatory ASR

[Siriwardena et al., 2022] proposed a Multi-Task Learning (MTL) approach to improve the performance of SI systems by adding an auxiliary ASR task. Inspired by this approach, we propose Articulation-Informed ASR, a framework that integrates articulatory information into speech recognition models through a dual strategy:

- An auxiliary SI task that encourages articulation-aware acoustic speech representations
- A cross-attention fusion block that injects predicted articulatory trajectories as an auxiliary conditioning signal before CTC decoding.

The entire model is trained through a MTL paradigm. While we base our design on Wav2vec2.0, in line with our previous experiments in this dissertation, our methods here can be applied to ASR models, like Whisper. We leave that implementation to future work. An overview of our model is shown in Figure 9.2.

At inference time, articulatory trajectories are predicted by the auxiliary SI branch, and no external articulatory measurements are required.

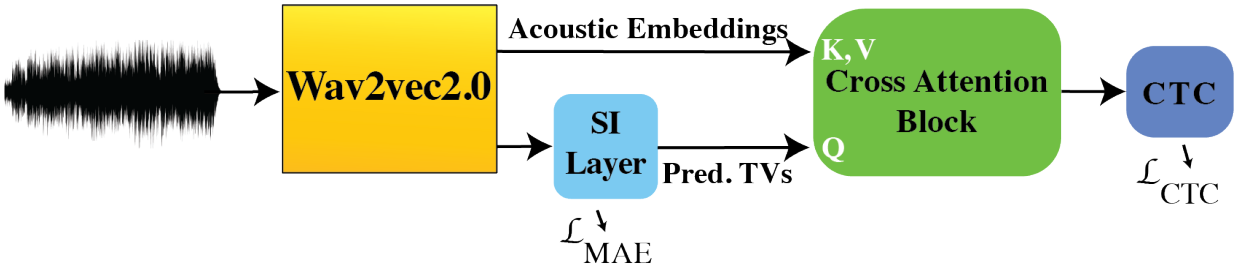

Figure 9.2: The proposed Articulation-Informed ASR framework.

9.2.3 Auxiliary SI Task

Given an acoustic signal, the Wav2vec2.0 transformer extracts contextual speech embeddings. In vanilla Wav2vec2.0, these embeddings are directly classified into graphemes. We follow a similar approach in our model, where we add a single SI feed-forward layer that maps the embeddings into TVs. This task is learned using a simple Mean Absolute Error (MAE) loss between the predicted and target TVs. The MAE loss trains the SI layer, and also contributes to updating the Wav2vec2.0 backbone. This allows for positive transfer between the SI and ASR tasks.

9.2.4 Cross-Attention Fusion for Acoustic-Articulatory CTC

To further tie the articulatory information with the ASR task, we add a cross-attention block, otherwise identical to the self-attention blocks in Wav2vec2.0 transformers. The difference is that the keys and values are derived from the Wav2vec2.0 acoustic embeddings, and the queries are derived from the predicted TVs. This weighs the acoustic embeddings based on the agreement between the acoustic embeddings and the TVs.

Let $\mathbf{H}_a \in \mathbb{R}^{T \times d}$ denote the acoustic embeddings produced by the Wav2vec2.0 encoder, and let $\mathbf{Z}_{tv} \in \mathbb{R}^{T \times D_{tv}}$ denote the predicted articulatory trajectories output by the auxiliary speech inversion branch. Since the dimensionality of the TVs differs from the acoustic hidden size, the TVs are first projected into the acoustic embedding space using a learned linear projection:

$$\mathbf{H}_{tv} = \mathbf{Z}_{tv}\mathbf{W}_P, \tag{9.1}$$

where $\mathbf{W}_P \in \mathbb{R}^{D_{tv} \times d}$.

The cross-attention operation is then defined as:

$$\mathbf{Q} = \mathbf{H}_{tv}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{H}_a\mathbf{W}_K, \quad \mathbf{V} = \mathbf{H}_a\mathbf{W}_V, \tag{9.2}$$

$$\text{Attn}(\mathbf{Z}_{tv}, \mathbf{H}_a) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}, \tag{9.3}$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are learned projection matrices and d is the attention dimension.

This late fusion of the TVs with the deep acoustic embeddings results in a mixture of top-down and bottom-up representations for the ASR task. The resulting representations are directly fed into the CTC layer, similar to vanilla Wav2vec2.0. This task is learned using the CTC loss, and the gradients backpropagate into the CTC and SI layers, cross-attention block and the Wav2vec2.0 backbone.

9.2.5 Loss Functions

Multi-task learning often requires manually tuning individual task losses and their contributions to the total loss:

$$\mathcal{L}_t = \alpha_{\text{CTC}}\,\mathcal{L}_{\text{CTC}} + \alpha_{\text{MAE}}\,\mathcal{L}_{\text{MAE}}. \tag{9.4}$$

The main objective of this tuning process is to ensure training stability and adequate learning for each task. However, in early experiments we found this approach rigid, as weights that work in early stages may not be optimal later.

Uncertainty-based weighting (UBW) provides a more adaptive method of combining regression and classification losses [Kendall et al., 2018]. Instead of hand-crafted weights, UBW

introduces two learnable parameters that adjust dynamically based on the progression of each loss. In our experiments, this method outperformed manual weighting, though it proved somewhat sensitive to initialization.

For classification (CTC) and regression (MAE), the combined objective is:

$$\mathcal{L}_t = \frac{1}{\sigma_{\text{CTC}}^2}\mathcal{L}_{\text{CTC}} + \frac{1}{2\sigma_{\text{MAE}}^2}\mathcal{L}_{\text{MAE}} + \log\sigma_{\text{CTC}} + \log\sigma_{\text{MAE}}, \tag{9.5}$$

where σ_{CTC} and σ_{MAE} are learnable parameters corresponding to task uncertainty.

9.3 Datasets and Addressing Articulatory Data Scarcity

Training our multi-task model requires a speech dataset with both articulatory labels and orthographic transcriptions. However, articulatory datasets are typically limited in size and often lack accurate orthographic transcriptions, while ASR datasets do not contain articulatory labels. To combine both tasks, one must either use a pre-trained ASR system to label articulatory datasets with transcriptions or a pre-trained SI system to generate articulatory features for ASR datasets. We adopt the latter approach, as our primary focus is the ASR task, and it is preferable to retain high-quality human transcriptions. Moreover, articulatory datasets span only a few hours, compared to ASR datasets that span hundreds or thousands of hours. Our approach is also motivated by prior work [Premananth et al., 2025], which has shown that machine-predicted articulatory labels can still provide effective supervision in downstream tasks.

We conduct experiments on LibriSpeech, a high-quality ASR corpus of read speech. A pre-trained SI system [Tabatabaee et al., 2025] is employed to predict TVs for the LibriSpeech corpus. The system predicts a set of TVs, including Lip Aperture (LA), Lip Protrusion (LP),

Tongue Body Constriction Location (TBCL), Tongue Body Constriction Degree (TBCD), Tongue Tip Constriction Location (TTCL), Tongue Tip Constriction Degree (TTCD), and Velopharyngeal Port (VP). In addition, it extracts two source features: Periodicity and Aperiodicity. In total, we use nine TVs.

The predicted TVs are sampled at 100 Hz. Wav2Vec2.0 representations are sampled at 50 Hz, matching the 20 ms stride of Wav2Vec2.0. To simplify alignment in the cross-attention module, we downsample the TVs to 50 Hz, ensuring temporal consistency.

9.4 Experiments

9.4.1 Experimental Setup

We evaluate our proposed method on the LibriSpeech corpus using both Wav2Vec2.0 *base* and *large* models to assess scalability across encoder sizes. We also train otherwise identical single-task vanilla Wav2vec2.0 models with the same data and hyperparameters.

To investigate data efficiency, we consider subsets of {10 minutes, 1 hour, 10 hours, 100 hours} of labeled training data from the `train-clean-100` partition of LibriSpeech, following common practice in low-resource ASR evaluations. This setup allows us to examine the contribution of articulatory features under varying amounts of supervision.

The 10-minute and 1-hour models were trained for 13k steps with a learning rate of 5×10^{-5}, the 10-hour models were trained for 20k steps with the same rate, and the 100-hour models were trained for 80k steps with a learning rate of 3×10^{-5}. All models were trained on A100 or H100 GPUs.

Table 9.1: WER (%) on LibriSpeech with Wav2Vec2.0 *Base* and *Large*. Baseline = CTC only; Proposed = SI+ASR with cross-attention fusion. Values shown as No LM / LM.

Hours	Test-Clean		Test-Other	
	Baseline	Proposed	Baseline	Proposed
Base				
10m	57.86 / 50.85	**45.83 / 35.83**	69.52 / 63.74	**56.68 / 46.28**
1h	23.66 / 19.46	**20.42 / 17.70**	35.98 / 30.53	**31.37 / 27.92**
10h	9.75 / 7.59	**9.05 / 5.57**	20.20 / 16.53	**18.75 / 15.10**
100h	5.53 / 4.46	**5.43 / 4.32**	14.28 / 11.85	**12.80 / 10.38**
Large				
10m	39.21 / 35.01	**37.94 / 31.94**	51.03 / 46.73	**45.66 / 39.40**
1h	20.32 / 17.56	**16.07 / 13.62**	30.08 / 26.51	**23.69 / 20.43**
10h	8.62 / 7.46	**6.59 / 6.90**	16.26 / 12.79	**13.58 / 11.96**
100h	4.00 / **3.28**	**3.89** / 3.41	9.56 / 8.22	**9.17 / 8.08**

9.4.2 Evaluation

We evaluate our models on `test-clean` and `test-other` sets using WER. While the SI task is typically evaluated using the Pearson Product-Moment Correlation (PPMC), we focus here on ASR performance. We note that all models achieve high correlation with the machine-generated labels, although none outperform the dedicated SI model [Tabatabaee et al., 2025] on real articulatory data.

9.4.3 Language Model Beam-Search Decoding

We report results with and without n-gram Language Model (LM) beam-search decoding, following [Baevski et al., 2020]. Results without LM integration reflect the raw model performance, while results with LM reflect performance in more practical settings.

9.5 Results

9.5.1 Analysis

Table 9.1 shows the results of our experiments. Overall, the results demonstrate that incorporating articulatory supervision through our multi-task model improves ASR performance across both the *base* and *large* Wav2Vec2.0 configurations. In every training regime, the proposed method either matches or outperforms the baseline, underscoring the robustness of the approach.

For the *base* model, the general trend is that the largest relative gains appear in the data-constrained settings. When training with only 10 minutes of labeled speech, the proposed model produces substantial improvements, with relative WER reductions **20.79%** and **29.52%** for the `test-clean` testset with and without LM, respectively, and **18.47%** and **27.39%** for the `test-other` testset with and without LM respectively. This suggests that the articulatory information is particularly valuable when the acoustic encoder alone cannot generalize well from such limited data and that the added supervision increases the utility of small datasets.

At 1 hour and 10 hours of training data, we continue to observe consistent improvements across both `test-clean` and `test-other`, showing that articulatory features remain beneficial in the low-resource regime. Even at 100 hours, where the baseline is already strong, the multi-task model still provides measurable gains. Specifically, on the `test-clean` test-set we see a relative improvement of **1.81%** and **3.09%** with and without an LM. For `test-other`, we observe a **10.33%** relative improvement without an LM and a **12.41%** relative improvement with LM decoding. These results indicate that the proposed approach is not only effective in extremely low-resource settings, but also continues to provide complementary information at larger scales.

The story is somewhat different for the *large* model. Here, improvements are observed across most training conditions, but the trend is less uniform. At the lowest data regime of 10 minutes, the relative improvements are modest compared to the *base* configuration. This outcome is expected, since the larger encoder capacity allows the model to extract stronger acoustic representations even with limited data, reducing its dependence on auxiliary supervision. However, as we increase the amount of labeled training data, the articulatory supervision begins to play a more important role. At 1 hour and 10 hours, the *large* configuration with the multi-task model shows clear and consistent gains across both test sets, highlighting that articulatory features can reinforce and stabilize training once sufficient supervision is available. This pattern suggests that the utility of articulatory supervision interacts with model capacity: smaller encoders benefit most when data is scarce, while larger encoders leverage the supervision more effectively in mid-resource regimes.

One outlier evident in the results is that the *large* model with LM decoding shows a slight degradation on `test-clean`, from 3.28% WER with the baseline to 3.41% with the proposed model. However, since the proposed model still improves without LM decoding and also on `test-other`, we attribute this anomaly to suboptimal LM decoding rather than a fundamental limitation of our method. It is also worth noting that our LM decoding results for the baseline model do not fully match those reported in the seminal work [Baevski et al., 2020], which may further explain this discrepancy.

Taken together, these findings suggest that the proposed method provides robust improvements in both small and large model configurations, with the nature of the improvements depending on the amount of available labeled data. In the *base* model, articulatory supervision provides strong gains in extremely low-resource conditions, demonstrating its ability to supplement weak

acoustic representations. In the *large* model, improvements are less dramatic at the lowest data scales, but grow more consistent as more data becomes available. The overall takeaway is that articulatory supervision is particularly effective in low- and mid-resource regimes, particularly with smaller models, while still offering measurable benefits even when training with larger amounts of data.

An interesting pattern emerges when comparing the proposed MTL *base* model with the *large* baseline, particularly in the medium-resource settings (1 and 10 hours) and, to a lesser extent, in the low-resource regime. The proposed *base* model closely matches the performance of the much larger *large* baseline model, despite having roughly one third of the parameters. In other words, incorporating articulatory information as both an auxiliary task and an additional input stream effectively narrows the capacity gap between *base* and *large*, yielding performance comparable to a model three times its size with only minimal added complexity.

9.6 Ablation Study

To better understand the contribution of each component in the proposed Articulation-Informed ASR framework, we conduct a comprehensive ablation study. The goal of this analysis is to isolate the effects of (1) auxiliary speech inversion supervision, (2) uncertainty-based multi-task loss weighting, and (3) the effect of fusing articulatory and acoustic features.

Table 9.2 summarizes the results of this study. All experiments are conducted using the Wav2vec2.0-base model trained on 10 minutes of `train-clean-100` and are evaluated on the `test-clean` and `test-other` test sets.

Table 9.2: WER(%) on LibriSpeech with Wav2vec2.0 Base trained with different configurations on 10 minutes of Librispeech for the ablation study. *UBW* refers to Uncertainty Based Weighing. *X-Attn Fusion* refers to the proposed cross-attention fusion.

SI Input Feature	SI Loss	Loss Weighting	X-Attn Fusion	Test-clean	Test-Other
None (Baseline)	–	–	–	57.86	69.52
Transformer emb.	✓	Fixed	None	57.15	68.68
Transformer emb.	✓	UBW	None	55.87	68.67
Transformer emb.	✓	UBW	✓	**48.07**	**60.05**

9.6.1 Baseline vs. Naive Multi-Task Learning vs. Uncertainty-Based Weighting

The first row in Table 9.2 corresponds to the baseline Wav2vec2.0 model trained with the CTC objective only. The second row introduces the auxiliary SI task in a naive MTL setup, without feature fusion and using fixed, manually tuned loss weights. This configuration does not improve performance significantly and results in only a slight improvement on both test sets.

The fixed loss weights are determined by briefly training the model on a small validation set and scaling the CTC and SI losses to have approximately equal magnitude early in training. However, inspection of the training logs reveals that the SI task, being substantially simpler than ASR, converges and saturates within the first few training steps. As a result, the SI loss contributes little to learning for the remainder of training, limiting positive transfer.

This observation motivates the use of UBW, shown in the third row. UBW improves performance relative to fixed loss weighting. Analysis of the training dynamics shows that UBW stabilizes optimization by slowing the effective convergence of the SI loss, allowing both tasks to be learned at comparable rates throughout training.

9.6.2 Articulatory–Acoustic Cross-Attention Fusion

We examine whether explicitly fusing articulatory and acoustic representations prior to CTC decoding provides additional benefit beyond auxiliary supervision alone.

The strongest gains are achieved using the proposed cross-attention fusion mechanism. Cross-attention allows the model to selectively reweight acoustic representations based on articulatory trajectories while preserving acoustic features as the primary signal for decoding. This structured interaction enables more effective integration of articulatory information than static feature concatenation.

9.7 Qualitative Example

To illustrate the effect of articulatory supervision beyond quantitative WER scores, we present a read speech example, completely outside of the Librispeech dataset..

Ground Truth*:"i can say this is thomas gibbs gee my one and only child and when he finished high school we had always planned to send him to princeton but his father had been called back into the service as a reserve officer and was stationed in washington"*

At 10 hours of training data, the *base* model without LM decoding produced the following output:

Baseline: *"i can say this is toms gibsgi mynonly choildand when he finishd hyscol we had always pland to send tim to prinsto but his father had been called back into the servecs as a reserve offiseor and was statient in wasingt"*

Proposed: *"i can say this is tomes gibs ge my one and only child and when he finished hiyschoul we had always pland to send him to prinstomn but his father had been called back into the servis as a reserve offiser and was stationd in washington"*

The proposed model reduces errors substantially, improving WER from 40.43% to 25.53% (36.84% relative improvement). Beyond the numerical score, the transcription from the proposed model is much more legible and semantically faithful to the reference, whereas the baseline contains multiple word merges and unintelligible segments. At 10 minutes of training data, the baseline model produced mostly gibberish, making detailed analysis impractical, whereas the proposed model yielded partially intelligible output.

9.7.1 Limitations and Applicability to Classroom ASR

While the proposed articulation-informed ASR framework demonstrates improved acoustic modeling under controlled evaluation conditions, it does not yet translate directly to classroom ASR settings. This limitation stems primarily from the current state of speech inversion (SI) models, which are largely developed and evaluated under non-overlapping-speakers, clean, or near-clean acoustic conditions.

Most existing SI systems are trained on datasets collected using specialized hardware, such as XRMB or EMA, where speech is produced in quiet environments with a single speaker and minimal background interference. Although there has been early work exploring speech inversion under noise conditions [Siriwardena et al., 2023], these studies remain limited in scope and do not address the challenges posed by overlapping speech, reverberation, and spontaneous multi-speaker interaction typical of classroom recordings.

In the proposed framework, articulatory trajectories are used as auxiliary supervision signals through predicted pseudo-labels. Recent work has improved the performance of SI systems in environments [Tabatabaee and Espy-Wilson, 2026]. However, in multi-talker environments, current SI models are unlikely to produce sufficiently accurate articulatory estimates, which limits the effectiveness of positive transfer to the ASR task. Consequently, directly applying this approach to classroom data without improved SI models would likely yield diminished gains.

An important direction for future work is the development of robust, multi-talker speech inversion models capable of disentangling articulatory dynamics from complex acoustic mixtures. Progress along this axis would enable the generation of higher-quality articulatory pseudo-labels in realistic classroom conditions, thereby strengthening the auxiliary supervision signal and improving the effectiveness of articulation-informed ASR.

Despite these limitations, the results presented in this chapter demonstrate that incorporating articulatory structure as an inductive bias can improve acoustic modeling in modern ASR systems. This suggests that articulation-informed objectives remain a promising complementary direction, and that advances in speech inversion, particularly in multi-speaker and noisy settings, could unlock their applicability to classroom ASR in future work.

9.8 Conclusion

This chapter investigated articulatory-informed acoustic modeling as a model-driven approach to improving robustness in automatic speech recognition. Motivated by the limitations of purely data-driven strategies identified throughout this dissertation, we explored whether incorporating articulatory speech inversion can serve as a complementary inductive bias for modern

ASR systems.

Our experiments demonstrate that incorporating articulatory supervision through a multi-task framework yields consistent improvements in ASR performance across both *base* and *large* Wav2Vec2.0 configurations. The gains are particularly pronounced in the low-resource regimes, where relative WER reductions exceed 25%, but remain measurable even with larger training sets and larger encoder models. Qualitative analysis further shows that the proposed approach improves transcription legibility, reducing merged words and unintelligible segments compared to the baseline.

An additional observation is that the proposed *base* model, when augmented with articulatory supervision, closely matches the performance of the much larger *large* baseline model in the medium-resource settings (1 and 10 hours). In effect, adding articulatory information as both an auxiliary task and an input stream narrows the capacity gap, providing performance comparable to a model three times its size with only minimal added complexity. This highlights not only the accuracy benefits of our approach but also its efficiency.

While the proposed approach does not yet extend directly to classroom ASR due to limitations in current speech inversion models, the results highlight the continued importance of acoustic modeling in low-resource and high-mismatch settings. More broadly, this chapter shows that introducing principled inductive biases into acoustic encoders remains a viable and promising direction, even in the era of large-scale, data-driven speech recognition systems.

Together with the preceding chapters, this work reinforces the central thesis of this dissertation: robust ASR in challenging environments requires a combination of data-centric adaptation, careful modeling choices, and structured inductive bias. Advances in speech inversion, particularly in multi-speaker and noisy conditions, may enable articulation-informed objectives to play

a practical role in future classroom ASR systems.

Chapter 10: Conclusions and Future Directions

10.1 Conclusions

This thesis investigated adapting pre-trained ASR models under data constraints. After examining the performance of current models on the challenging low-resource domain of classroom speech, we introduced a training-centric three-step domain adaptation pipeline, suitable for domains with an abundance of unlabeled and weakly labeled data, but a small amount of gold-standard data, as little as a dozen hours of transcribed speech. This approach is summarized in Figure 10.1, and is detailed below.

1. Continued Pre-training: This first step in the adaptation pipeline is CPT on unlabeled in-domain data. In chapter 2.3.5, we discussed how CPT can significantly improve the performance of self-supervised ASR models, with limited gold-standard supervised fine-tuning. This step alone causes a 55% relative reduction in WER on classroom test sets. Notably, it also improves the performance on acoustic conditions unseen in the supervised

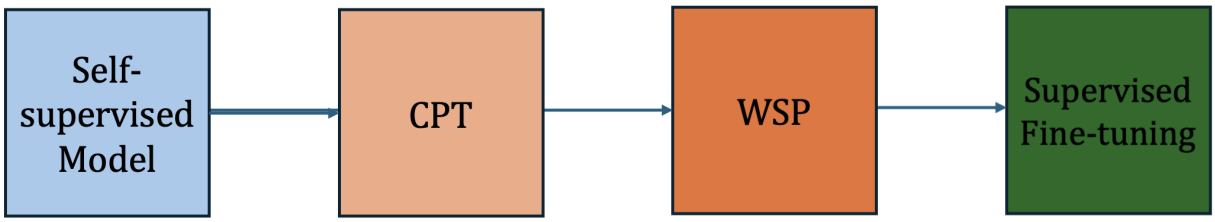

Figure 10.1: A flowchart of the proposed approach applied to self-supervised ASR models.

fine-tuning training set, like far-field microphones or exceedingly high noise.

2. Weakly Supervised Fine-tuning: The second step in the pipeline utilizes weak, imprecise transcription to learn a strong general, albeit noisy prior. This step, discussed in Chapter 6, while weakly supervised, functions as an intermediate pre-training stage. This intermediate model learns strong acoustic and linguistic priors, but also carries the error modes of the weak transcripts, resulting in weak benchmarks. However, when combined with supervised fine-tuning, it becomes evident that this step causes a further relative reduction in WER by 18%.

3. Supervised Fine-tuning: The final step in that pipeline is the classical supervised fine-tuning. This step remains unchanged from classical approaches, but the prior two steps increase the utility of limited labeled data.

In parallel, we explored data-centric approaches. In Chapter 7, we investigated increasing the actual size of the data through realistic, scalable acoustic data simulation in game engines. This novel approach generated the first realistic classroom noise and RIR corpora. Paired with semantically matched children's recordings from the MyST corpus and instructional adult speech from Khan Academy, the resultant RealClass training set is the largest and first shareable classroom corpus to date, to the best of our knowledge. Our benchmarks show that RealClass outperforms public alternatives on classroom test sets in the absence of classroom training data. When paired with the limited available classroom training data, we see a relative reduction in WER by 5.4%.

We compared our proposed adaptation pipeline against LLM-based ASR models in Chapter 8. First, we investigated the effect of prompt-based contextual biasing and found it to be a cheap and effective way to utilize the strong linguistic priors of the LLMs in low-resource settings. However,

even with LoRA adaptation and contextual biasing, our proposed approach still outperformed the LLM-based approach. This shows the significance of the acoustic adaptation carried out in the CPT and WSP steps. This suggests that future work may benefit from placing greater emphasis on acoustic adaptation than linguistic adaptation.

Lastly, in Chapter 9 we proposed Articulatory-Informed ASR, a novel technique to reintroduce speech articulation into ASR. Our approach uses an auxiliary output layer to infer the TVs and re-feed them into the CTC input stream. We also utilize pre-trained SI systems to pseudo-label the training data, requiring no large labeled articulatory datasets. This approach proved to be especially effective in low-resource settings with small models, being most effective in experiments with an hour or less of training data.

10.2 Limitations

Despite the substantial improvements demonstrated in this thesis, several limitations remain.

1. While the proposed methods lead to large relative gains in ASR performance, the resulting error rates are still not sufficient for production-level deployment. On the NCTE benchmark, the best-performing configuration achieves a WER of 13.51%, and 23.70% on the MPT benchmark. These results represent a significant improvement over models trained under comparable low-resource constraints and establish the effectiveness of the proposed training strategies. However, in absolute terms, these error rates remain high for many real-world classroom applications, particularly those that require fine-grained discourse analysis or reliable real-time transcription. This highlights the intrinsic difficulty of classroom ASR and suggests that further advances are needed before such systems can be robustly deployed

in practice.

2. All ASR experiments in this work are evaluated using uncased and unpunctuated transcripts. This abstraction is standard in ASR research and allows for controlled comparison across models and training paradigms. However, real classroom transcription systems require punctuation and casing for readability and downstream processing. Reintroducing these elements would likely inflate effective error rates, especially in long-form classroom speech that contains frequent disfluencies, interruptions, and incomplete utterances. As a result, the reported WER values should be interpreted as optimistic estimates of end-user transcription quality.

3. This thesis focuses exclusively on automatic speech recognition, addressing what was said but not who said it. In practical classroom systems, speaker diarization is a critical component, as it enables attribution of speech to teachers versus students and supports the analysis of classroom interaction patterns. Diarization in classrooms is substantially more challenging than ASR due to overlapping speech, far-field microphones, and a large and variable number of speakers. While we conducted preliminary exploratory experiments on diarization [Khan et al., 2025], achieving diarization performance suitable for classroom deployment remains an open problem and is beyond the scope of this work.

4. The evaluation in this thesis is constrained by the available classroom datasets and benchmarks. Experiments are conducted primarily on the NCTE and MPT test sets, which are necessarily small due to the overal data scarcity. While we have attempted to diversify the test sets as much as possible, these test sets cannot fully capture the diversity of real classrooms, including variation in grade level, subject matter, classroom layout, instructional

style, student demographics, and acoustic conditions. These benchmarks provide a valuable and realistic evaluation setting, however, performance on these datasets may not generalize uniformly across all classroom environments.

10.3 Future Work

10.3.1 LLM-Based ASR Models

Chapter 8 showed the great promise of prompt-based contextual biasing, but also showed that acoustic adaptation is more effective. Future work should combine both approaches. The recent Omnilingual ASR model, built on self-supervised speech encoders, is a promising LLM-ASR model that can be adapted with the same pipeline introduced in this dissertation. This would combine the strong linguistic priors of LLMs with the improved acoustic modeling through self-supervised and weakly-supervised adaptation.

10.3.2 Articulatory-Informed ASR in Adversarial Settings

Our research in Chapter 9 shows that Articulatory-Informed ASR is promising in controlled settings, but has limited applicability in noisy, multispeaker environments like classrooms. This is bottlenecked by the performance of SI models in these environments. Recent work has shown promising results under noisy environments with babble noise [Tabatabaee and Espy-Wilson, 2026]. Future work should investigate how to improve the robustness of SI models in adversarial environments, which will, in turn, unlock the potential of Articulatory-Informed ASR in such environments.

10.3.3 Multi-Source Consistency for ASR

We previously mentioned that the NCTE recordings have multiple recordings from the same classroom. Thus far, we have either considered one recording from each classroom, like in the case of our precise transcription of NCTE, or as separate recordings, as in our self-supervised pre-training experiments without exploiting their relationship to each other.

We propose training a multi-source ASR system. During training, given a multi-source recording, we train the audio encoder to generate consistent representations for each audio source. Another approach would be to process all the different audio sources and merge their representations to process the audio from different points within the room.

For controlled experiments, RealClass would serve as an excellent data source as we can control the placement and number of microphones within the virtual space. We can then apply the findings for the multi-source recordings from the NCTE dataset.

10.3.4 Biomimetic Self-Supervised Pre-training

Previous work by Famularo et al. [2024] has established that biological front ends significantly improve the robustness of speech models. Famularo et al. developed a differentiable audio front-end inspired by human auditory processing, which integrates cochlear and cortical features into a deep learning framework. Their methodology focuses on modeling auditory perception using biologically plausible signal processing techniques, making the model more robust to noise and computationally efficient. Their results demonstrated that this biomimetic approach outperformed traditional black-box models in both classification and enhancement tasks, especially in low-resource settings, highlighting its potential for speech recognition applications.

To the best of our knowledge, relatively little research has explored the design of the convolutional feature extractor in Wav2vec2.0. Future work can investigate replacing this component with biomimetic models or speech inversion models. Biomimetic models, inspired by the human auditory or speech production system, offer robustness and interpretability advantages due to their biologically plausible design.

Bibliography

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Open ai. *arXiv preprint arXiv:2303.08774*, 2023.

Alibaba Group. Qwen-asr: Hear clearly, transcribe smartly. https://qwen.ai/blog?id=41e4c0f6175f9b004a03a07e42343eaaf48329e7&from=research.latest-advancements-list, 2024. Accessed: 2025-01-08.

Massih-Reza Amini, Vasilii Feofanov, Loïc Pauletto, Liès Hadjadj, Émilie Devijver, and Yury Maximov. Self-training: A survey. *Neurocomputing*, 616:128904, 2025. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2024.128904. URL https://www.sciencedirect.com/science/article/pii/S0925231224016758.

AI Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1(1):4, 2024.

Mahmoud Ashraf. Mms-300m-1130 forced aligner. https://huggingface.co/MahmoudAshraf/mms-300m-1130-forced-aligner, 2025. Accessed: 2025-02-19.

Ahmed Adel Attia, Jing Liu, Wei Ai, Dorottya Demszky, and Carol Espy-Wilson. Kid-whisper: Towards bridging the performance gap in automatic speech recognition for children vs. adults. *arXiv preprint arXiv:2309.07927*, 2023.

Ahmed Adel Attia, Dorottya Demszky, Tolulope Ogunremi, Jing Liu, and Carol Espy-Wilson. Cpt-boosted wav2vec2. 0: Towards noise robust speech recognition for classroom environments. *arXiv preprint arXiv:2409.14494*, 2024.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*, 2021.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Mateusz Barański, Jan Jasiński, Julitta Bartolewska, Stanisław Kacprzak, Marcin Witkowski, and Konrad Kowalczyk. Investigation of whisper asr hallucinations induced by non-speech audio. *arXiv preprint arXiv:2501.11378*, 2025.

Anton Batliner, Mats Blomberg, Shona D'Arcy, Daniel Elenius, Diego Giuliani, Matteo Gerosa, Christian Hacker, Martin Russell, Stefan Steidl, and Michael Wong. The pf_star children's speech corpus. In *Proc. Interspeech 2005*, pages 2761–2764, 09 2005. doi: 10.21437/Interspeech.2005-705.

Nina R Benway, Saba Tabatabaee, Benjamin Munson, Jonathan Preston, and Carol Espy-Wilson. Subtyping speech errors in childhood speech sound disorders with acoustic-to-articulatory speech inversion. In *Proc. Interspeech 2025*, pages 2800–2804, 2025a.

Nina R Benway, Saba Tabatabaee, Dongliang Wang, Benjamin Munson, Jonathan L Preston, and Carol Espy-Wilson. Perceptual ratings predict speech inversion articulatory kinematics in childhood speech sound disorders. *arXiv preprint arXiv:2507.01888*, 2025b.

Mrinmoy Bhattacharjee, Iuliia Nigmatulina, Amrutha Prasad, Pradeep Rangappa, Srikanth Madikeri, Petr Motlicek, Hartmut Helmke, and Matthias Kleinert. Contextual biasing methods for improving rare word detection in automatic speech recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12652–12656. IEEE, 2024.

Daren C Brabham. Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence*, 14(1):75–90, 2008.

Jie Cao, Ananya Ganesh, Jon Cai, Rosy Southwell, E Margaret Perkoff, Michael Regan, Katharina Kann, James H Martin, Martha Palmer, and Sidney D'Mello. A comparative analysis of automatic speech recognition errors in small group classroom discourse. pages 250–262, 2023.

Xuankai Chang, Yanmin Qian, Kai Yu, and Shinji Watanabe. End-to-end monaural multi-speaker asr system without pretraining. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6256–6260. IEEE, 2019.

Xuankai Chang, Wangyou Zhang, Yanmin Qian, Jonathan Le Roux, and Shinji Watanabe. End-to-end multi-speaker speech recognition with transformer. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6134–6138. IEEE, 2020.

Christopher Cieri, David Miller, and Kevin Walker. The fisher corpus: A resource for the next generations of speech-to-text. In *LREC*, volume 4, pages 69–71, 2004.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020.

Creative Commons. Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0), 2019. URL `https://creativecommons.org/licenses/by-nc-sa/4.0/`.

Datamuse. Datamuse api. `https://www.datamuse.com/api/`, 2025. Accessed: 2025-02-19.

Dorottya Demszky and Heather Hill. The ncte transcripts: A dataset of elementary math classroom transcripts. *arXiv preprint arXiv:2211.11772*, 2022.

Dorottya Demszky, Jing Liu, Heather C. Hill, Shyamoli Sanghi, and Ariel Chung. Improving teachers' questioning quality through automated feedback: A mixed-methods randomized controlled trial in brick-and-mortar classrooms. *EdWorkingPapers*, 2023. doi: 10.26300/8pnw-5q67.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library, 2025. URL https://arxiv.org/abs/2401.08281.

Ruolan Leslie Famularo, Dmitry N Zotkin, Shihab A Shamma, and Ramani Duraiswami. Biomimetic frontend for differentiable audio processing. *arXiv preprint arXiv:2409.08997*, 2024.

Ruchao Fan, Natarajan Balaji Shankar, and Abeer Alwan. Benchmarking children's asr with supervised and self-supervised speech foundation models. *arXiv preprint arXiv:2406.10507*, 2024.

Angelo Farina et al. Simultaneous measurement of impulse response and distortion with a swept-sine technique. *Preprints-Audio Engineering Society*, 2000.

Federal Trade Commission. Children's online privacy protection rule (coppa), 1998. URL https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa. Accessed: 2024-10-12.

Frederic Font, Gerard Roma, and Xavier Serra. Freesound technical demo. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 411–412, 2013.

Carly B. Fox, Megan Israelsen-Augenstein, Sharad Jones, and Sandra Laing Gillam. An evaluation of expedited transcription methods for school-age children's narrative language: Automatic speech recognition and real-time transcription. *Journal of Speech, Language, and Hearing Research*, 64(9):3533–3548, 2021. doi: 10.1044/2021_JSLHR-21-00096. URL https://pubs.asha.org/doi/abs/10.1044/2021_JSLHR-21-00096.

Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2013.

Masakiyo Fujimoto and Hisashi Kawai. One-pass single-channel noisy speech recognition using a combination of noisy and enhanced features. In *Interspeech*, pages 486–490, 2019.

Matteo Gerosa, Diego Giuliani, Shrikanth Narayanan, and Alexandros Potamianos. A review of asr technologies for children's speech. In *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, pages 1–8, 2009.

Behrooz Ghorbani, Orhan Firat, Markus Freitag, Ankur Bapna, Maxim Krikun, Xavier Garcia, Ciprian Chelba, and Colin Cherry. Scaling laws for neural machine translation. *arXiv preprint arXiv:2109.07740*, 2021.

Mária Gósy. Filler words in children's and adults' spontaneous speech. In *The 9th Workshop on Disfluency in Spontaneous Speech*, page 89, 2019.

Thomas Graave, Zhengyang Li, Timo Lohrenz, and Tim Fingscheidt. Mixed children/adult/childr-enized fine-tuning for children's asr: How to reduce age mismatch and speaking style mismatch. In *Interspeech*, volume 2024, pages 5188–5192, 2024.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings 20*, pages 198–208. Springer, 2018.

Wei-Ning Hsu, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, et al. Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. *arXiv preprint arXiv:2104.01027*, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL `https://arxiv.org/abs/2106.09685`.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Rishabh Jain, Andrei Barcovschi, Mariam Yiwere, Peter Corcoran, and Horia Cucu. Adaptation of whisper models to child speech recognition. *arXiv preprint arXiv:2307.13008*, 2023a.

Rishabh Jain, Andrei Barcovschi, Mariam Yahayah Yiwere, Dan Bigioi, Peter Corcoran, and Horia Cucu. A wav2vec2-based experimental study on self-supervised learning methods to improve child speech recognition. *IEEE Access*, 11:46938–46948, 2023b.

Rishabh Jain, Andrei Barcovschi, Mariam Yahayah Yiwere, Peter Corcoran, and Horia Cucu. Exploring native and non-native english child speech recognition with whisper. *IEEE Access*, 12:41601–41610, 2024.

Jesper Jensen and Cees H Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2009–2022, 2016.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Virender Kadyan, Hemant Kathania, Prajjval Govil, and Mikko Kurimo. Synthesis speech based data augmentation for low resource children asr. In *Speech and Computer: 23rd International Conference, SPECOM 2021, St. Petersburg, Russia, September 27–30, 2021, Proceedings 23*, pages 317–326. Springer, 2021.

Thomas Kane, Heather Hill, and Douglas Staiger. National center for teacher effectiveness main study. Inter-university Consortium for Political and Social Research [distributor], 2022. URL `https://doi.org/10.3886/ICPSR36095.v4`.

Jodi Kearns. Librivox: Free public domain audiobooks. *Reference Reviews*, 28(1):7–8, 2014.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

Ali Sartaz Khan, Tolulope Ogunremi, Ahmed Adel Attia, and Dorottya Demszky. Multi-stage speaker diarization for noisy classrooms, 2025. URL `https://arxiv.org/abs/2505.10879`.

Laura L Koenig and Jorge C Lucero. Stop consonant voicing and intraoral pressure contours in women and children. *The Journal of the Acoustical Society of America*, 123(2):1077–1088, 2008.

Matthew A Kraft, David Blazar, and Dylan Hogan. The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of educational research*, 88 (4):547–588, 2018.

Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. Sdr–half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630. IEEE, 2019.

Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, 105(3):1455–1468, 1999.

Jean-Marie Lemercier, Julius Richter, Simon Welker, and Timo Gerkmann. Storm: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

Zhen-Hua Ling, Korin Richmond, and Junichi Yamagishi. Articulatory control of hmm-based parametric speech synthesis using feature-space-switched multiple regression. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(1):207–219, 2013. doi: 10.1109/TASL.2012.2215600.

LJ Link and TR Guskey. Grading and assessment survey, 2022.

Vikramjit Mitra, Hosung Nam, Carol Y Espy-Wilson, Elliot Saltzman, and Louis Goldstein. Articulatory information for noise robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):1913–1924, 2010.

Vikramjit Mitra, Ganesh Sivaraman, Hosung Nam, Carol Espy-Wilson, Elliot Saltzman, and Mark Tiede. Hybrid convolutional neural networks for articulatory and acoustic information based speech recognition. *Speech Communication*, 89:103–112, 2017.

NCES. National teacher and principal survey. *National Center for Education Statistics (NCES) Home Page, a part of the U.S. Department of Education*, 2018.

Topi Nieminen. Unity game engine in visualization, simulation and modelling. B.S. thesis, Tampere University, 2021.

Karol Nowakowski, Michal Ptaszynski, Kyoko Murasaki, and Jagna Nieuważny. Adapting multilingual speech representation model for a new, underresourced language through multilingual fine-tuning and continued pretraining. *Information Processing & Management*, 60(2):103148, 2023.

ASR Omnilingual, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, et al. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages. *arXiv preprint arXiv:2511.09690*, 2025.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

Sameer S Pradhan, Ronald A Cole, and Wayne H Ward. My science tutor (myst)–a large corpus of children's conversational speech. *arXiv preprint arXiv:2309.13347*, 2023.

Gowtham Premananth and Carol Espy-Wilson. Speech-based estimation of schizophrenia severity using feature fusion. In *2025 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5, 2025. doi: 10.1109/ICASSPW65056.2025.11011243.

Gowtham Premananth, Philip Resnik, Sonia Bansal, Deanna L Kelly, and Carol Espy-Wilson. Multimodal biomarkers for schizophrenia: Towards individual symptom severity estimation. *arXiv preprint arXiv:2505.16044*, 2025.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL `https://arxiv.org/abs/2412.15115`.

Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018. URL `https://api.semanticscholar.org/CorpusID:49313245`.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

Vinay V Ramasesh, Ethan Dyer, and Maithra Raghu. Anatomy of catastrophic forgetting: Hidden representations and task semantics. *arXiv preprint arXiv:2007.07400*, 2020.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

Korin Richmond and Simon King. Smooth talking: Articulatory join costs for unit selection. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5150–5154, 2016. doi: 10.1109/ICASSP.2016.7472659.

Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pages 749–752. IEEE, 2001.

Nay San, Georgios Paraskevopoulos, Aryaman Arora, Xiluo He, Prabhjot Kaur, Oliver Adams, and Dan Jurafsky. Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens. *arXiv preprint arXiv:2402.02302*, 2024.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

Paul W. Schönle, Klaus Gräbe, Peter Wenig, Jörg Höhne, Jörg Schrader, and Bastian Conrad. Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain and Language*, 31(1):26–35, may 1987. ISSN 10902155. doi: 10.1016/0093-934X(87)90058-7.

Khaldoun Shobaki, John-Paul Hosom, and Ronald Cole. The ogi kids' speech corpus and recognizers. In *Proc. of ICSLP*, pages 564–567. Citeseer, 2000.

Christopher Simic and Tobias Bocklet. Self-supervised adaptive av fusion module for pre-trained asr models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12787–12791. IEEE, 2024.

Yashish M. Siriwardena, Carol Espy-Wilson, Chris Kitchen, and Deanna L. Kelly. Multimodal approach for assessing neuromotor coordination in schizophrenia using convolutional neural networks. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, ICMI '21, page 768–772, New York, NY, USA, 2021a. Association for Computing Machinery. ISBN 9781450384810. doi: 10.1145/3462244.3479967. URL https://doi.org/10.1145/3462244.3479967.

Yashish M Siriwardena, Chris Kitchen, Deanna L Kelly, and Carol Espy-Wilson. Inverted vocal tract variables and facial action units to quantify neuromotor coordination in schizophrenia. In *Proc. 12th International Seminar on Speech Production (ISSP 2020)*, pages 174–177, 2021b.

Yashish M Siriwardena, Ganesh Sivaraman, and Carol Espy-Wilson. Acoustic-to-articulatory speech inversion with multi-task learning. *arXiv preprint arXiv:2205.13755*, 2022.

Yashish M. Siriwardena, Ahmed Adel Attia, Ganesh Sivaraman, and Carol Espy-Wilson. Audio data augmentation for acoustic-to-articulatory speech inversion. In *2023 31st European Signal Processing Conference (EUSIPCO)*, pages 301–305, 2023. doi: 10.23919/EUSIPCO58844.2023.10290069.

Ganesh Sivaraman, Vikramjit Mitra, Hosung Nam, Mark Tiede, and Carol Espy-Wilson. Unsupervised speaker adaptation for speaker independent acoustic to articulatory speech inversion. *The Journal of the Acoustical Society of America*, 146(1):316–329, 2019.

Bruce L Smith. Relationships between duration and temporal variability in children's speech. *The Journal of the Acoustical Society of America*, 91(4):2165–2174, 1992.

David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*, 2015.

Rosy Southwell, Samuel Pugh, E Margaret Perkoff, Charis Clevenger, Jeffrey B Bush, Rachel Lieber, Wayne Ward, Peter Foltz, and Sidney D'Mello. Challenges and feasibility of automatic speech recognition for modeling student collaborative discourse in classrooms. *International Educational Data Mining Society*, 2022.

Rosy Southwell, Wayne Ward, Viet Anh Trinh, Charis Clevenger, Clay Clevenger, Emily Watts, Jason Reitman, Sidney D'Mello, and Jacob Whitehill. Automatic speech recognition tuned for child speech in the classroom. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12291–12295. IEEE, 2024.

Kenneth N Stevens. *Acoustic phonetics*, volume 30. MIT press, 2000.

Saba Tabatabaee and Carol Espy-Wilson. Towards noise-robust speech inversion through multi-task learning with speech enhancement, 2026. URL https://arxiv.org/abs/2601.14516.

Saba Tabatabaee, Suzanne Boyce, Liran Oren, Mark Tiede, and Carol Espy-Wilson. Enhancing acoustic-to-articulatory speech inversion by incorporating nasality. *arXiv preprint arXiv:2506.09231*, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Jenthe Thienpondt and Kris Demuynck. Transfer learning for robust low-resource children's speech asr with transformers and source-filter warping. *arXiv preprint arXiv:2206.09396*, 2022.

Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. Investigating rnn-based speech enhancement methods for noise-robust text-to-speech. In *SSW*, pages 146–152, 2016.

Valve Corporation. Steam Audio: Real-time Physics-Based Spatial Audio, 2025. URL `https://valvesoftware.github.io/steam-audio/`. Accessed: 2025-02-09.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Houri K. Vorperian and Ray D. Kent. Vowel acoustic space development in children: A synthesis of acoustic and anatomic data. *Journal of Speech, Language, and Hearing Research*, 50(6): 1510–1545, 2007. doi: 10.1044/1092-4388(2007/104). URL `https://pubs.asha.org/doi/abs/10.1044/1092-4388%282007/104%29`.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Sravya Popuri, Dmytro Okhonko, and Juan Pino. Fairseq s2t: Fast speech-to-text modeling with fairseq. *arXiv preprint arXiv:2010.05171*, 2020.

John R Westbury. Speech Production Database User ' S Handbook. *IEEE Personal Communications - IEEE Pers. Commun.*, 0(June), 1994.

Jason D Williams, I Dan Melamed, Tirso Alonso, Barbara Hollister, and Jay Wilpon. Crowdsourcing for difficult transcription of speech. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 535–540. IEEE, 2011.

Thomas et. al. Wolf. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics, October 2020. doi: 10.18653/v1/2020.emnlp-demos.6.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025. URL `https://arxiv.org/abs/2503.20215`.

Gary Yeung, Ruchao Fan, and Abeer Alwan. Fundamental frequency feature normalization and data augmentation for child speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6993–6997. IEEE, 2021.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models, 2024. URL https://arxiv.org/abs/2403.13372.

Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5 (1):44–53, 2018.

Qiu-Shi Zhu, Jie Zhang, Zi-Qiang Zhang, Ming-Hui Wu, Xin Fang, and Li-Rong Dai. A noise-robust self-supervised pre-training model based speech representation learning for automatic speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3174–3178. IEEE, 2022.

Qiushi Zhu, Jie Zhang, Yu Gu, Yuchen Hu, and Lirong Dai. Multichannel av-wav2vec2: A framework for learning multichannel multi-modal speech representation. *arXiv preprint arXiv:2401.03468*, 2024.