university of groningen

campus fryslân

Improving OOV Word Recognition in End-to-End ASR via Lightweight Domain Adaptation with a TED-LIUM2 N-gram Language Model

XuefeiBian

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

Improving OOV Word Recognition in End-to-End ASR via Lightweight Domain Adaptation with a TED-LIUM2 N-gram Language Model

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Assoc. Prof. Dr.Phat (Voice Technology, University of Groningen)

Xuefei Bian (S5836670)

June 11, 2025

Acknowledgements

I would like to sincerely thank my supervisor, Tan Phat Do, for his kind help and guidance during my thesis. He always replied to my questions quickly and clearly, and helped me organize my thoughts when I was confused. In the beginning, I planned to collect my own speech data and do experiments with it. Phat gave me many useful suggestions and explained the possible problems. Although this part of the project did not work out, he helped me stop in time and change to another research direction. Later, when I met some difficulties with the model, he also helped me solve them so that I could continue working smoothly. Without his support, I would not have been able to finish this thesis. I also want to thank Xiyuan Gao for her help at the early stage of my project. When I was working on collecting data from YouTube and TikTok, she gave me a lot of advice about how to use the APIs and follow the rules about privacy and data use. She also helped me understand how to get transcripts and gave me tips about using language models.

In addition, I am grateful to Dr.Shekhar for recommend the course on fine-tuning models. He answered my questions patiently during class and helped me better understand how ASR systems work. I also want to thank Dr.Matt for his suggestions in the early stage of my thesis, and Dr.Vass for encouraging me and helping me find the right direction for my research. Besides, I am grateful to Tiantian Zhang and Meilin Li. At the beginning of the project, they helped me record more than 40 speech samples. Even though I didn't use the self-collected data in the end because it was too complicated, I really appreciate their support.

Furthermore, I want to express my deepest thanks to my grandfather, who made it possible for me to study for this master's degree. He passed away in the second week after I arrived, but I will always remember his support, love, and financial help. And I am grateful to Dr. Zhang from Peking University Sixth Hospital for giving me important mental support when I needed it the most.

I acknowledge the Center for Information Technology of the University of Groningen for their technical support and for providing access to the Hábrók high-performance computing cluster.

Lastly, I would like to express my gratitude to all the individuals who have played a part, no matter how small, in shaping my academic journey and the successful completion of this thesis.

Abstract

Automatic Speech Recognition (ASR) systems have achieved strong performance on read and clean speech. However, ASR systems still face major challenges when recognizing spontaneous speech, especially for low frequency or domain specific out of vocabulary (OOV) words. Most current research focuses on improving overall recognition accuracy in non-spontaneous speech, but few studies explore whether it is possible to improve OOV recognition and overall performance in spontaneous speech without retraining the acoustic model. This study aims to fill this gap. I trained an n-gram language model (LM) using a small amount of spontaneous speech data from TED-LIUM2 and applied it to a pre trained Wav2Vec2.0 acoustic model using shallow fusion. This study tested three settings: no LM, a TED LIUM2 trained LM, and a pretrained LM based on LibriSpeech. I evaluated the performance in terms of Word Error Rate (WER), OOV recall, and three types of recognition errors: substitution, insertion, and deletion.

Results show that the TED LIUM2 LM increased the OOV recall rate by 21.75% compared to the no LM baseline, confirming its ability to help recognize domain specific words. However, this LM also caused a increase in overall WER from 25.04% to 44.28%, mainly due to substitution errors caused by incomplete sentence structure, redundant content, and missing context in the training data. In contrast, the LibriSpeech LM achieved a better balance between WER (21.22%) and OOV recall (40.41%).

This work contributes by providing the first systematic evaluation of using small scale spontaneous speech to train a language model for ASR without changing the acoustic model. It also highlights how the quality and structure of training text (such as context and sentence completeness) play a key role in ASR performance.

Contents

1 Introduction

Nowadays, automatic speech recognition (ASR) plays an increasingly important role in real-world applications. Among various methods, end-to-end (E2E) models based on self-supervised learning, such as Wav2Vec2, have attracted wide attention due to their simplified architecture and strong performance. These models do not need to rely on acoustic models, pronunciation dictionaries, or external language models. This is different from traditional hybrid systems such as HMM-DNN (Inaguma, Mimura, Sakai, & Kawahara, 2018).

However, although E2E models have many advantages, they still face challenges in practical use. In extremely noisy environments, when recognizing rare words, or when the speech quality is poor or pronunciation varies a lot, the accuracy of recognition often drops. Researchers have tried to solve noise-related problems by adding artificial noise during training or using masking strategies such as SpecAugment (D. S. Park et al., 2019). To address challenges caused by accents and speaker variability, researchers have explored data augmentation with accented speech (Fukuda et al., 2018) or speaker-adaptive training (Saon, Soltau, Nahamoo, & Picheny, 2013). But in spontaneous spoken scenarios, such as in marketing, medicine, or finance, the low recognition rate of OOV words remains a difficult issue. This is because, with the rapid growth of social media, new vocabulary is constantly emerging. Inaguma et al. (2018) found that in sentences containing OOV words, the word error rate (WER) of ASR is three times higher than in sentences without OOVs. Dufour, Jousse, Estève, Béchet, and Linarès (2009) also showed that the error rate of highly spontaneous segments is 41.2%, almost twice that of well-prepared speech. As pointed out by Koto et al. (2014), traditional training corpora such as broadcast news cannot reflect the richness of vocabulary and diverse styles of spontaneous speech. To handle these challenges, researchers have proposed different solutions for out-of-vocabulary (OOV) words. Many of them try to use vocabulary updating, new deep learning models, cross-modal enhancement, or subword modeling (Fox & Delworth, 2022). For example, Jung, Kim, Ryu, and Lee (2024) applied cross-modal enhancement by incorporating visual or textual information to guide speech recognition, Zhou, Zeineldeen, Zheng, Schlüter, and Ney (2021) trained an acoustic model from scratch using a fully data-driven subword modeling approach, and Inaguma et al. (2018) employed external language models (e.g., n-gram or Transformer) with shallow fusion. These methods aim to overcome the limitations of acoustic models in spelling prediction and context handling. However, some approaches, such as cross-modal enhancement, are relatively complex and computationally expensive, which may hinder their broader adoption in real-world applications. In addition, many existing studies are still trained on broadcast news, scripted speech, or other well-structured data (Koto et al., 2014). This means that spontaneous speech, like impromptu presentations or casual conversation, is still underrepresented in current ASR research.

From a practical point of view, more than one billion people visit YouTube every month, and they watch more than six billion hours of videos (Alberti & Bacchiani, 2009). Ensuring subtitle accuracy has become a key goal to help viewers understand and verify the meaning of audio and video content. Guillot (2010) explained from the view of intercultural linguistics that subtitles are a multimodal text system, which has special ways of expressing and interpreting meaning. More specifically, high-quality subtitles can clearly deliver language content, reflect cultural differences, and improve the audience's cross-cultural understanding. With the rapid development of social media and short video platforms, the demand for easier and more accessible video content is also increasing. While some content creators still manually transcribe subtitles to ensure high quality, this process is expensive and time-consuming. It is especially difficult for long videos that last

more than two hours. Generally, YouTube subtitles are divided into three types. The first type is human-generated subtitles, which are very accurate but often slow to produce. The second type is semi-supervised subtitles, where users upload transcripts and the system performs forced alignment. The third type is automatically generated subtitles using real-time ASR systems, which are faster but usually have low to medium accuracy.

Based on this background, this study aims to explore how ASR technology can help creators save time in making accurate subtitles and solve the low recognition rate of spontaneous video content. The research asks the following questions: How can we improve the recognition accuracy for spontaneous speech? How can we better handle the OOV problem in spontaneous speech? Is there a simple method to improve the model's expression ability?

In this context, the study tries to solve the problem of low-frequency words and professional terms (which usually belong to OOV) being misrecognized or ignored. It also aims to improve recognition accuracy in specific domains. I chose to use a lightweight and easy-to-use shallow fusion method. This method combines an n-gram language model trained on the TED-LIUM2 spontaneous speech dataset with the decoding results of a CTC-based ASR system. The goal is to enhance the recognition of domain-specific OOV words without fine-tuning the original acoustic model.

The motivation of this study has been introduced above. The remaining parts of this thesis are structured as follows. Section 1.1 presents the research questions and hypotheses. Section 2 reviews related work and explains the academic background of this study. Section 3 describes the methods used in this research. Section 4 explains the experimental setup in detail. Section 5 reports the findings and analysis. Section 6 discusses the broader implications of the results. Finally, Section 7 summarizes the main contributions and provides suggestions for future research.

1.1 Research Questions and Hypotheses

ASR systems struggle with fast-evolving vocabulary, such as buzzwords and slang, which are common in social media. These terms are often out-of-vocabulary (OOV) for ASR models due to their reliance on static training data (Qu, Weber, & Wermter, 2023). This leads to my central research question: which kind of method can be used to improve the accuracy in spontaneous speech which include a lot of out of vocabulary words and randomized vocabulary?

In this study, I will utilize a dataset sourced from TED, which encompasses diverse content categories such as news, entertainment, and other spoken materials. Previous research has demonstrated that both contextual and lexical adaptation techniques can effectively enhance out-of-vocabulary (OOV) word recognition. For instance, Xu et al. (2023) introduced the CATT model for context-aware transduction, achieving significant improvements, including a 6.7% reduction in Word Error Rate (WER) on LibriSpeech and up to a 20.7% Character Error Rate (CER) improvement on proprietary in-house names. These results underscore the potential of phrase-list-based bias injection methods. Similarly, Zhou et al. (2021) proposed ADSM, an acoustically-driven subword segmentation approach, which outperformed BPE baselines with a WER reduction of 0.9–1.2%. However, their method required retraining the entire ASR model.

In 2014, Koto et al. (2014) first demonstrated that combining semantic and acoustic features in MMR (Maximal Marginal Relevance) could improve the summarization of TED Talk transcripts. However, their work primarily focused on identifying representative sentences for human reading rather than enhancing speech recognition accuracy. Their method employed TF-IDF and MMR techniques on ASR-generated text. Sarı, Moritz, Hori, and Le Roux (2020) addressed the issue of

recognition accuracy degradation caused by speaker mismatches. They evaluated their approach on the TED-LIUM 2 corpus, using the WSJ (Wall Street Journal) as the training set. And achieved frame-level speaker adaptation by introducing an external M-vector memory into the encoder. This method still need to modify the model architecture and retraining.

Therefore, this study investigates a key challenge in end-to-end (E2E) automatic speech recognition (ASR): improving the recognition of rare and out-of-vocabulary (OOV) words in spontaneous speech. Unlike prior work that requires retraining the entire model, my approach focuses on enhancing performance through domain-specific language modeling. Specifically, I explore whether integrating an n-gram language model and trained on in-domain text into Connectionist Temporal Classification (CTC) decoding can boost recall for OOV words without modifying the acoustic model or tokenizer. The experiments are conducted using modern neural E2E ASR architectures Wav2Vec 2.0 and evaluated on spontaneous speech datasets. Performance is measured using standard ASR metrics, including Word Error Rate (WER) and OOV recall rate. The goal is to demonstrate that targeted language model adaptation, rather than full model retraining, can significantly improve recognition of dynamically emerging terms (e.g., 'consoles', 'controllers', or 'CTOs') that are often missed by conventional ASR systems.

> **My main research question: How can the integration of an external n-gram language model during decoding improve the recognition accuracy of spontaneous speech and the handling of out-of-vocabulary words in end-to-end ASR systems without modifying the acoustic model?**

This main research question can be further divided into two specific inquiries. First, can training a language model solely on spontaneous speech data without fine-tuning the acoustic model help reduce the overall Word Error Rate (WER)? Second, does the use of an external language model contribute to improving the recall of Out-of-Vocabulary (OOV) words in spontaneous speech domains?

After comparing the testing results of the Wav2Vec2.0 model with and without an external language model, the OOV recall increased from 31% to 40%. Based on this, I hypothesize that by leveraging an adapted language model—without retraining the acoustic model or tokenizer—it is possible to achieve a relative WER reduction of at least 10%, while increasing the recall rate of OOV and rare words to over 50%. The baseline model for comparison is Wav2Vec2.0 (facebook/wav2vec2-base-100h) fine-tuned on the TED-LIUM2 dataset. The key innovation of this approach lies in improving recognition accuracy through targeted language model adaptation rather than full model retraining.

This method is expected to significantly enhance the recognition of OOV terms (e.g., 'esolar', 'etruscans', or 'foodie'), which are often underrepresented in conventional ASR training corpora. If validated, the results will demonstrate that domain-specific linguistic modeling can effectively compensate for the limitations of the acoustic model without requiring retraining. In other words, an ASR system can adapt to evolving or unpredictable vocabulary in spontaneous speech through lightweight training of the language model. This has practical implications for low-resource or real-time deployment scenarios, where retraining large neural models is often infeasible.

Conversely, if the hypothesis is falsified, it may suggest an incompatibility between phrase biasing and subword representations during decoding, or that subword granularity is insufficient for representing novel terms. In such cases, more advanced mechanisms—such as NAM (Munkhdalai et al., 2022), OCR-guided semantic reranking (Kuhn, Kersken, Reuter, Egger, & Zimmermann, 2024),

or acoustic-driven unit retraining (Zhou et al., 2021) may be necessary to better balance contextual specificity and model generalization.

2 Literature Review

Automatic Speech Recognition (ASR) systems have rapidly evolved over the past two decades, yet their performance continues to be challenged by fast-changing or rare vocabulary. For example, terms such as TikTokification, growth hacking, and prompt engineering are often out-of-vocabulary (OOV) and poorly represented in ASR training corpora. This literature review surveys existing research in four major areas relevant to this challenge: (1) the evolution of ASR models; (2) spontaneous speech; (3) the definition of and recognition difficulties associated with OOV terms; and (4) methods for handling OOV vocabulary.

2.1 Search Strategy and Selection Criteria

This literature review covers four key thematic areas relevant to the current research. First, concerning the evolution of Automatic Speech Recognition (ASR), the review focuses on the development of major architectures such as end-to-end ASR systems, GMM-HMM, DNN-HMM, Transformer-based models, CTC (Connectionist Temporal Classification), RNN-Transducer (RNN-T), Wav2Vec 2.0, and Whisper. The relevant search terms include "end-to-end ASR," "GMM-HMM," "Transformer ASR," "CTC," "DNN-HMM," "RNN-T," "Whisper," and "Wav2vec." Second, in relation to spontaneous speech, this review examines challenges and approaches specific to recognizing unscripted, conversational speech, using keywords such as "spontaneous speech recognition" and "spontaneous speech." Third, the discussion of Out-of-Vocabulary (OOV) words explores the definition and recognition difficulty of domain-specific terms, particularly in dynamic environments such as social media. Key search terms include "popular vocabulary in social media" and "out-of-vocabulary (OOV) terms." Lastly, the review highlights OOV handling methods in ASR, focusing on approaches to improve recognition of rare and unseen terms. The associated search terms include "OOV speech recognition" and "Special Noun Recognition".

Inclusion criteria will cover end-to-end ASR models (CTC, RNN-T, AED, or Transformer-based ASR), studies on OOV word recognition in social media. All of criteria will be search from IEEE Xplore, ACM Digital Library and Google Scholar. These research time are between 2000 to 2025.

Studies centered on ASR applications in unrelated fields such as healthcare and legal transcription were also excluded due to limited relevance to social media or buzzword contexts. In addition, works written in languages other than English, non-peer-reviewed sources such as blog posts, news stories, or opinion pieces were also excluded.

2.2 The Evolution Of ASR Models

In the field of research on Automatic Speech Recognition (ASR),early systems around 2009 primarily used GMM-HMM models with MFCC features and forced alignment but suffered from high word error rates (WER), often exceeding 50%. The shift began with Google's adoption of fully supervised DNN-HMM systems for YouTube captioning in 2012. Jaitly, Nguyen, Senior, and Vanhoucke (2012) found that the WER below 50% for news-style videos but less than 10% lift relative to the GMM model. Seide, Li, Chen, and Yu (2011) further demonstrated that feature engineering with context-dependent DNNs significantly improved ASR for conversational speech. Meanwhile, Liao (2013) and Sainath, Kingsbury, Sindhwani, Arisoy, and Ramabhadran (2013) highlighted the challenges and gains of deep networks trained on GPUs, noting that while DNNs offer over 20%

relative improvements on tasks like voice search, gains on YouTube were still limited (6%), likely due to content complexity and poor speaker adaptation. Recent studies attempt to bridge this performance gap by tailoring ASR systems to downstream tasks like subtitle generation. D. Liu, Niehues, and Spanakis (2020) explored adapting Transformer-based ASR systems with output length constraints for generating readable German TV subtitles which model fine-tuned on news data, improved ROUGE scores and better recognized rare words. Compared with the Baseline ASR The Word Erros Rate has dropped dramatically 19.2%, but faced challenges with excessive paraphrasing when trained unsupervised. Similarly, Wan et al. (2021) addressed segmentation errors in low-resource ASR by training LSTM taggers on subtitle-derived synthetic data, boosting downstream translation and retrieval performance, though limited by the absence of acoustic signals in subtitle data. Moreover, Thara et al. (2024) integrated OpenAI's Whisper model for subtitle synchronization, showing improvements in aligning mismatched timestamps, albeit still lacking support for semantic-level cues and character identification. While modern ASR systems have become more robust and adaptive, current research continues to reveal limitations in handling content compression, domain shift, segmentation, and synchronization questions. Especially for applications like subtitling and spoken language translation. However, significant challenges remain when applying these systems to more dynamic, real-world environments such as YouTube subtitling or caption generation.

Review of the state of the art ASR systems, I found that the advancement of self-supervised learning has revolutionized speech representation learning. In 2018, Devlin, Chang, Lee, and Toutanova (2019) introduced masked language modeling (MLM) to natural language processing (NLP), enabling deep contextualized representations, which later inspired similar techniques in speech processing. RoBERTa (2019) enhanced BERT by using dynamic masking and larger training data, showing that MLM remains effective with improved design (Y. Liu et al., 2019). After that researcheers base on the BERT structure, Schneider, Baevski, Collobert, and Auli (2019)proposed Wav2Vec, marking significant progress in unsupervised pretraining for speech recognition. This model learned meaningful speech representations without transcribed text. Likewise, VQ-Wav2Vec (2019) based on Wav2Vec which integrated vector quantization to align speech signals with discrete units, facilitating cross-domain adaptation from NLP (e.g., BERT) to ASR (Baevski, Schneider, & Auli, 2019). In October 2020, Baevski, Zhou, Mohamed, and Auli (2020)introduced Wav2Vec 2.0, which borrowed BERT's MLM strategy and employed a multi-layer transformer encoder to model temporal dependencies in masked audio frames. This was another milestone following the success of BERT pretraining in NLP. Wav2Vec 2.0 demonstrated exceptional performance even in low-resource datasets. When fine-tuned with only 10 minutes of labeled data, it achieved a word-error-rate (WER) of 4.8% on test-clean and 8.2% on test-other. With 1 hour of labeled data, performance further improved to 2.9% and 5.8%.

Wav2Vec 2.0 significantly reduced WER in self-supervised ASR and also performed well in tasks beyond speech recognition. For instance, Pepino, Riera, and Ferrer (2021) showed that the change the intermediate transformer layers to make optimal results. Additionally, Yi, Wang, Cheng, Zhou, and Xu (2020) demonstrated that fine-tuning Wav2Vec 2.0 significantly improved ASR accuracy for low-resource languages, outperforming previous MFCC-based or multilingual ASR systems.

Although some of models like HuBERT (Hsu et al., 2021) improved representation learning through changing the cluster based masking, or Whisper Radford et al. (2023) expanded to multilingual and multitask end-to-end speech modeling which have better performance in unsupervised training. Wav2Vec 2.0 still keep a preferred choice for domain adaptation. It can keep a balance between model complexity and resource efficiency. Made it particularly suitable for tasks with limited

labeled data.

In this study, I utlize the Wav2Vec2-Base-100h pretrained model due to its effectiveness in CTC-based decoding. The Connectionist Temporal Classification (CTC) framework allows flexible alignment between variable-length audio and text, while shallow fusion with an external n-gram language model further enhances domain-specific vocabulary recognition. I try to assume that in the presence of different domain oov vocabularies, character-level language models have good performance in recognition and do not require a lot of specialized vocabulary training. Through this way, this study focus on improving out-of-vocabulary (OOV) word detection. Particularly in spontaneous speech, the Wav2Vec 2.0 framework provides a robust, modular, and well-established foundation.

2.3 Spontaneous Speech

Spontaneous speech refers to naturally occurring, unscripted, and unrehearsed spoken language. Unlike read speech, which is intentionally produced and typically follows grammatical conventions, spontaneous speech exhibits the following characteristics: frequent pauses (e.g., “uh” and “um”), incomplete or truncated phrases, interjections, ungrammatical structures, and vague or telegraphic expressions. Ward (1989) notes that these linguistic phenomena are inherent manifestations of spoken language systems, whereas traditional ASR systems often assume well-structured input and require all words to be part of a predefined vocabulary. From an acoustic perspective, Nakamura, Iwano, and Furui (2008) found that spontaneous speech exhibits spectral reduction in the Mel-Frequency Cepstral Coefficients (MFCC) feature space which means the distribution of speech signal characteristics becomes more compact. This effect is particularly pronounced in conversational speech, leading to reduced phoneme discriminability and, consequently, degraded recognition performance. Despite the advancements in end-to-end (E2E) architectures like Wav2Vec2, accurately transcribing spontaneous speech remains challenging. Informal speech often includes incorrect forms, word expansions or reductions, all of which complicate phoneme boundary detection. Additionally, disfluencies can confuse sequence models such as repetitions and false pronunciations. Another major obstacle is out-of-vocabulary (OOV) words, as spontaneous speech frequently includes names, slang, or rare terms absent from the training data.

In research on spontaneous speech recognition, early work by Maekawa et al. (2003) created the manually annotated Corpus of Spontaneous Japanese (CSJ) to construct a high-quality dataset for evaluation and validation. Furui, Nakamura, Ichiba, and Iwano (2005) study shows that when the training text increased eightfold from 0.86 million to 6.84 million words, the corresponding WER dropped by 17%. Under the training condition of 510 hours of acoustic data and 6.84 million words of text, the best WER achieved was 25.3%. Zhou et al. (2021) used a hybrid BLSTM-HMM system enhanced with SpecAugment for acoustic robustness and i-vectors for speaker adaptation. Performance improved further by integrating three language models during decoding and applying sequence discriminative training (sMBR). This approach achieved state-of-the-art results on TED-LIUM2, reducing test set WER from 9.8% to 5.6% when they added a Transformer LM. The study validated SpecAugment’s effectiveness for this domain. In other research, Sarı et al. (2020) proposed an end-to-end CTC and Attention model with a Neural Turing Machine-style memory module (M-vector) for unsupervised, frame-level speaker adaptation. Their method improved robustness to speaker variability without requiring test set speaker information, reducing WER from 16.7% (E2E baseline) to 11.0%. Though the absolute WER remained slightly higher than Zhou et al. (2021), Sarı et al. (2020) use unsupervised approach demonstrated significant gains. However, both methods are

computationally intensive and complex.

2.4 OOV words Definition and Difficulty

Out-of-vocabulary (OOV) words can be defined spoken words that are not present in the predefined vocabulary of a speech recognition system. Most automatic speech recognition (ASR) systems are designed with a fixed vocabulary during training. And as a result, they are only capable of recognizing words included in that vocabulary. But when users speak words outside of this set—such as rare names, new terminology or foreign words, the system cannot recognize them correctly. Instead, the recognizer often substitutes them with similar-sounding in-vocabulary (IV) words, which leads to recognition errors (Bazzi, 2002).

The presence of OOV words remains one of the main challenges in speech recognition, especially in real-world, spontaneous speech settings where vocabulary coverage cannot be guaranteed. As language is constantly evolving, it is practically impossible to construct a vocabulary that includes all possible words. Even in large-vocabulary ASR systems (e.g., with over 60,000 words), OOV words may still occur, particularly when there is a mismatch between the training data and the target domain (Schaaf, 2001).

The difficulty of handling OOV words lies not only in their recognition failure, but also in their tendency to degrade the recognition of nearby words. When an OOV word is misrecognized, it often causes neighboring words in the utterance to also be misrecognized, resulting in a significant increase in the overall word error rate (WER) (Zue et al., 2002).

Several studies have proposed various approaches to detect or recover OOV words. One approach is to build explicit OOV models that operate on subword units such as phones or syllables, enabling the system to construct possible word hypotheses beyond its fixed vocabulary (A. Park & Glass, 2005). Subword-based models allow open-vocabulary recognition, as subwords are from a closed set but can be concatenated flexibly to cover novel words. However, these models must be carefully designed to avoid misclassifying in-vocabulary speech segments as OOVs, which would harm recognition accuracy.

Another direction is to improve confidence scoring mechanisms, which help detect when the recognizer is uncertain about a word (Wessel, Schluter, Macherey, & Ney, 2002). Acoustic confidence measures and semantic pragmatic confidence scores can be used together to identify likely OOV regions (Klakow, Rose, & Aubert, 1999). For example, if the system assigns a low confidence score to a word, it may be flagged as a potential OOV, and either discarded or processed with additional models.

Despite these efforts, detecting and handling OOV words remains an open problem in ASR, particularly for applications involving domain mismatch, spontaneous conversation, or under-resourced languages.

2.5 Methods For Handling OOV Terms

In previous research on OOV (out-of-vocabulary) words, the focus has primarily been on domain-specific terms, neologisms, and proper nouns such as place names and personal names. Methods to improve OOV recognition accuracy mainly fall into four categories: vocabulary updating, deep model innovation, cross-modal enhancement strategies, and subword modeling for OOV.

Early ASR systems relied on outdated language models (e.g., GigaWord corpus). Jouvet et al. (2018) collected contemporaneous news and textual data from the internet to update vocabulary and language models, thereby improving ASR transcription accuracy, particularly for personal and place names. This approach reduced OOV rates from 16.4% to 2.0% in Arabic, 5.5% to 1.5% in English, and 1.8% to 0.2% in French. Additionally, on French videos, the updated vocabulary significantly improved overall ASR accuracy, reducing WER from 17.4% to 14.9%, and lowering WER for personal names from 40% to 27%. In 2020, Y. Liu, Bao, Wang, Weng, and Zhao (2023) introduced an interaction mechanism incorporating a hot-word model, but it did not yield significant accuracy improvements.

In terms of architectural innovation, Xu et al. (2023) proposed the CATT (Context-Aware Transducer with Triggered biasing) model, built upon the RNN-Transducer (RNN-T) framework. The model includes an Entity Detector that determines whether context biasing should be triggered. When activated, it dynamically inserts a bias phrase list to influence the prediction path. Results showed a 6.7% reduction in WER on the LibriSpeech dataset and a 20.7% reduction in CER, with a maximum reduction of 96.7% on proprietary corporate data-all while maintaining real-time performance (RTF closed 1). However, due to the complexity of context structure construction, the NAM module still requires explicit context graphs to be loaded during inference. Its generalization capability also remains limited in emergent scenarios. Munkhdalai et al. (2022) enhanced the traditional LAS (Listen, Attend and Spell) architecture by introducing a NAM (Neural Associative Memory) module, which consists of an attention- based memory encoder and an associative retrieval mechanism to rapidly activate potential hot-word information from a predefined context phrase list. During training and inference, NAM associatively activates semantically similar words, improving recognition of rare terms and proper nouns. On the Wiki-Names dataset, this approach reduced WER by 12% and improved F1-score by 15.7%, with fine-tuning further increasing F1 to 83.4%. However, due to the complexity of context structure construction, the NAM module still requires explicit loading of context graphs during inference, and its generalization capability remains limited in emergent scenarios. Other advancements, such as BERT-based deep clustering (BDC) have demonstrated promising results in enhancing the recognition of OOV words and proper nouns, particularly in Chinese text (Ma, He, & Niu, 2023). These approaches leverage contextual embeddings and synthetic data to improve recall rates and reduce word error rates (WER), making them suitable for real-time subtitling applications, especially on platforms like YouTube.. Specifically, BERT's reliance on pre-trained static datasets often struggles to adapt to the rapid emergence of new terms and slang, which are prevalent in dynamic environments like social media and digital marketing. Additionally, BERT's computational complexity and slower processing times for large vocabularies can hinder its effectiveness in real-time applications, such as YouTube subtitling.

In cross-modal enhancement strategies, Kuhn et al. (2024) extracted OCR text from video slides, compared it with the Google corpus to obtain Relative Frequency (RF) values, and incorporated these as bias terms into ASR beam search rescoring. Results showed that in videos containing academic keywords, ASR accuracy for technical terms improved by up to 3.22%. However, this method depends on OCR accuracy, requires external lexical resources (e.g., Google Books), and demands empirical parameter tuning.

In subword modeling, Sennrich, Haddow, and Birch (2015)proposed Byte Pair Encoding (BPE), which iteratively merges the most frequent character pairs to generate fixed-size subword units. However, BPE relies solely on character frequency and does not account for acoustic structure. Zhou et al. (2021) introduced ADSM (Acoustic Data-driven Subword Modeling), which combines

G2P initialization and a GramCTC loss function to train subword units entirely based on acoustic features, constructing subword lexicons that better align with phonetic logic. On LibriSpeech, ADSM outperformed BPE and PASM, achieving WERs of 5.0/12.6 (clean/other) under an RNN-T architecture. However, training is time-consuming (1 week per epoch), and limitations remain for extremely low-frequency words. Dockes (2022) compared BPE performance in English and German ASR, highlighting the impact of linguistic structure on subword modeling effectiveness. German performed better than English due to its compound word structure, but BPE for English required additional symbols (e.g., underscores) for optimal results.

Research has evolved from early corpus-based word distribution synthesis (2018) and rapid vocabulary updates (2019–2022) toward dynamic context-aware modeling (2023) and multimodal fusion (2024), alongside the development of subword-level modeling. Current challenges include: Rapid emergence of new vocabulary outpacing model adaptation; Dependency on pre-defined context phrases; Difficulty in modeling semantic ambiguity and user variability in real-world.

In conclusion, Automatic Speech Recognition (ASR) systems have made a lot of progress. Modern models such as Wav2Vec 2.0 and Whisper are examples of this development. However, problems still exist, especially when it comes to understanding spontaneous speech and recognizing out-of-vocabulary (OOV) words. Spontaneous speech often includes disfluencies, ungrammatical sentences, and new or fast-changing vocabulary, which makes it harder for ASR systems to work well. To deal with these problems, some researchers have tried different methods. These include updating the vocabulary, using deep context models like CATT, NAM, and BDC, or splitting words into subword units using techniques like BPE and ADSM. However, most of these methods need to retrain the acoustic model, use extra data like images or semantics, or require large amounts of computing power. Also, many of these methods do not work well across different speech domains and are not suitable for real-time tasks such as live subtitles. Therefore, there is still a need for simple and flexible solutions that can help ASR systems work better in real-life situations, especially when dealing with spontaneous speech and unknown words. One possible way to solve this is by using domain-specific n-gram language models during the decoding stage of ASR. However, this idea has not been fully explored in current end-to-end ASR systems.

Some earlier studies have tried combining language models with ASR. For example, Inaguma et al. (2018) applied shallow fusion with an RNN-based language model on an end-to-end A2W ASR system. Their approach reduced WER from 22.1% to 20.5% on the Switchboard dataset, demonstrating the benefit of integrating external LMs for OOV resolution and general decoding accuracy. Zhou et al. (2020) achieved a 5.6% WER using a 4-gram language model with TED-LIUM2 data, but their system needed a complex setup, including data augmentation, speaker adaptation, and transformer-based rescoring. Koto et al. (2014) also showed that TED Talks include spontaneous speech features that are often not covered by regular training data. This shows that vocabulary in such domains is hard to model without special adaptation.

This study will suggest a different approach. It uses n-gram language models trained on spontaneous speech data and applies them only at the decoding stage. This does not change the acoustic model and does not require a lot of resources. It focuses on improving OOV word recognition in dynamic and real-world situations. This method is light, easy to apply, and can help ASR systems perform better without extra training.

3 Methodology

This study evaluates the effectiveness of training n-gram language models on a spontaneous speech dataset to improve Out-of-Vocabulary (OOV) word recognition and overall performance, as measured by Word Error Rate (WER), without modifying the end-to-end acoustic model (e.g., wav2vec2-base-100h). The experiments aim to improve the recognition of rare and OOV words.

The baseline models include facebook/wav2vec2-base-100h (without a language model) and patrickvonplaten/wav2vec2-base-100h-with-lm (with a built-in language model), both available on Hugging Face and pre-trained on LibriSpeech. These models are tested on the TED-LIUM2 dataset without any parameter fine-tuning. The decoding is performed using Connectionist Temporal Classification (CTC) with letter-level outputs.

For language modeling, the TED-LIUM2 training and development transcripts are used to train a 4-gram language model using the KenLM Toolkit (Heafield, 2011). This trained 4-gram model is then integrated into the wav2vec2 decoding process via beam search rescoring using the pyctcdecode library. The final evaluation compares WER and OOV recall to assess the effectiveness of the domain-specific language model.

The experimental design consists of one baseline and two additional test conditions. The baseline involves decoding using the wav2vec2-base-100h pretrained model without incorporating any language modeling information. This used to measure the Word Error Rate (WER) and Out-of-Vocabulary (OOV) recall on the TED-LIUM2 test set. Experiment 1 employs the official patrickvonplaten/wav2ve base-100h-with-lm model, which integrates a generic English language model for probabilistic rescoring during the beam search decoding phase. In Experiment 2, the same pretrained acoustic model is combined with a custom 4-gram language model trained on TED-LIUM2 transcripts. Like the previous setups, this configuration is used to evaluate both WER and OOV recall on the same test set, enabling direct performance comparisons.

3.1 Dataset Description

For the training set, validator and test set, I chose to use TED-LIUM2. The full name of this dataset is TED-LIUM Release 2: TED Talks corpus which is English speech recognition training corpus from TED talks, created by Laboratoire d'Informatique de l'Université du Maine (LIUM). TED is a non-profit organization dedicated to spreading worth ideas. Since 1984, TED has been organizing conferences on the themes of Technology, Entertainment, and Design, with a core format that invites speakers to give a "life-changing" talk in 18 minutes (Koto et al., 2014). So the transcriptions in this dataset cover a wide range of topics and show rich vocabulary, spanning fields such as society, science and technology, and the humanities. This diversity is beneficial for training more robust speech recognition models. Moreover, the speakers typically do not read from a script. Instead, their speech contains hesitations, pauses, repetitions, and colloquial expressions, which are more challenging but also more relevant to real-world applications. The TED-LIUM2 corpus is collected from TED Talks, which consist of English-language presentations. Although these talks are generally prepared in advance, speakers often do not strictly follow their scripts and usually speak naturally. As a result, the language displays distinct spoken characteristics such as repetitions, self-corrections, and parenthetical expressions (Koto et al., 2014). This makes TED-LIUM2 distinct from purely read speech corpora and positions it closer to natural semi-spontaneous speech. Features such as filled pauses, restarts, ungrammatical constructions, and elliptical phrases are prevalent throughout

TED-LIUM2 (Ward, 1989). Consequently, it has been classified as exhibiting an "intermediate style between written and conversational" speech.

The dataset includes 1,495 audio talks in NIST Sphere format (SPH), 1,495 transcripts in STM format, a pronunciation dictionary containing 159,848 entries, and selected monolingual data for language modeling from the publicly available WMT12 corpora. Each SPH file has a 16,000 Hz sampling rate, mono channel, and 16-bit precision (Rousseau, Deléglise, & Estève, 2014).

This study uses the same transcription dataset as the researchers in Unsupervised Speaker Adaptation Using Attention-Based Speaker Memory for End-to-End ASR (Sarı et al., 2020). The training set contains 211.1 hours of audio, the validation set 1.6 hours, and the test set 2.6 hours.

3.2 Core Models

This study uses a Wav2Vec 2.0-CTC architecture, in which a self-supervised speech encoder is fine-tuned with a linear projection layer and optimized using the Connectionist Temporal Classification (CTC) criterion. CTC removes the need for frame-level alignments by marginalizing over all possible alignments between input frames and output tokens. During inference, this research performs shallow fusion beam search with an external 4-gram KenLM trained on domain-matched text, which effectively reduces spelling mistakes and recovers many acoustically confusable words. The language model is kept frozen; it does not back-propagate gradients but only re-ranks CTC paths at decode time. Hence, the overall system can be viewed as a Wav2Vec 2.0 encoder fine-tuned with a CTC layer, plus a TEDLIUM2 n-gram LM used at inference.

3.2.1 Wav2vec 2.0

The overall architecture of Wav2vec 2.0 comprises four key components: a convolutional feature encoder, a context network, a quantization module, and a contrastive loss. First, the feature encoder transforms raw speech signals into low-frequency latent representations. Subsequently, a subset of these latent features is randomly masked and fed into a Transformer-based context network to capture long-range contextual dependencies. Simultaneously, the unmasked features are discretized via a Gumbel-Softmax mechanism using a predefined codebook, serving as targets for contrastive learning (Baevski et al., 2020).

The pretraining objective employs a contrastive loss function, where the model learns discriminative representations by distinguishing the correct target representation from multiple negative samples. Additionally, to prevent the model from relying on only a limited set of codewords, an entropy regularization term is incorporated during training to promote diversified usage of codewords. After pretraining, the model can be fine-tuned using a CTC loss, making it adaptable to tasks such as speech recognition, emotion recognition, and speaker recognition.

On the LibriSpeech benchmark, Wav2vec 2.0 significantly outperforms previous systems. For instance, the LARGE model, fine-tuned on just 10 minutes of labeled speech data, achieves WERs of 5.2%/8.6% on the clean/other test sets, respectively (Baevski et al., 2020). When fine-tuned on the full 960 hours of labeled data, the LARGE model further reduces the WER to 1.8%/3.3%. Even without an external language model, it achieves competitive results of 2.0%/4.0%, demonstrating robust speech modeling capabilities.

This "pretraining + fine-tuning" paradigm has established a BERT-like framework in the speech

domain, significantly advancing speech recognition systems toward semi-supervised or even unsupervised learning approaches.

3.2.2 N-gram

An n-gram language model is a probabilistic model that estimates the likelihood of a word based on its $n-1$ preceding words (Jurafsky & Martin, n.d.). Formally, it approximates the joint probability of a word sequence using the Markov assumption:

$$P(w_1, w_2, \ldots, w_N) \approx \prod_{i=1}^{N} P(w_i \mid w_{i-n+1}^{i-1}) \tag{1}$$

This model assumes that the probability of each word depends only on the previous $n-1$ words (Jurafsky & Martin, n.d.). The conditional probability is typically estimated via Maximum Likelihood Estimation (MLE):

$$P(w_i \mid w_{i-n+1}^{i-1}) = \frac{C(w_{i-n+1}^{i})}{C(w_{i-n+1}^{i-1})} \tag{2}$$

where $C(\cdot)$ denotes the count of the given n-gram or history in the training corpus.

In this work, I train a 4-gram language model using the TED-LIUM2 corpus, which primarily consists of spontaneous speech. The model was constructed with KenLM[1], a widely used toolkit for efficient n-gram language modeling. A key advantage of KenLM is its speed and memory efficiency. Benchmark results show that KenLM runs significantly faster and has a lower memory footprint than earlier toolkits such as SRILM and IRSTLM. It supports multi-threaded training, binary query formats, and flexible integration with decoders via C++, Python, or Shell interfaces. In addition, it is compatible with various decoding frameworks such as Moses, cdec and pyctcdecode, which makes it ideal for shallow fusion in end-to-end ASR systems.

The resulting model was used to score ASR hypotheses during decoding, providing linguistic context to improve recognition quality, particularly in the spontaneous speech domain.

3.3 Evaluation Methodology

The Recognition performance is evaluated using the following two master metrics: Word Error Rate (WER) and OOV Recall.

3.3.1 WER

Word Error Rate (WER) is a standard metric used to evaluate the performance of automatic speech recognition (ASR) systems. It measures the minimum number of word-level edits needed to transform the ASR hypothesis into the reference transcription. The edit operations include substitutions (S), insertions (I), and deletions (D), normalized by the total number of words in the reference (N). Formally, WER is defined as:

[1] https://github.com/kpu/kenlm

$$\text{WER} = \frac{S+D+I}{N} \times 100\%$$

A lower WER indicates better ASR performance. WER captures both acoustic errors and language model mispredictions, making it a comprehensive metric for overall system accuracy.

3.3.2 OOV recall

Out-of-Vocabulary (OOV) Recall measures the system's ability to correctly recognize words that are not present in the training set vocabulary. It is particularly important in evaluating ASR performance in dynamic or domain-specific contexts where new terms (e.g., names, emerging phrases) frequently appear.

Let R_{OOV} be the number of correctly recognized OOV words and T_{OOV} be the total number of OOV words in the reference. The OOV recall is defined as:

$$\text{OOV Recall} = \frac{R_{\text{OOV}}}{T_{\text{OOV}}} \times 100\%$$

Higher OOV recall indicates stronger generalization and lexical flexibility of the ASR system.

3.3.3 Error Analysis

To better understand how the integration of an external n-gram language model influences recognition performance, we conduct a fine-grained error analysis based on the classical decomposition of word error rate (WER) into substitution (S), insertion (I), and deletion (D) components. This breakdown is based on the idea from Bahl and Jelinek (2003), which views ASR as a noisy communication process where errors such as substitutions, insertions, and deletions change the correct transcript into the recognized output.

Substitution errors occur when a spoken word is recognized as an incorrect but phonetically similar alternative, often due to acoustic ambiguity or insufficient contextual modeling. For instance, "peace" may be misrecognized as "piece," indicating that although the acoustic model correctly captured the spectral-temporal patterns, the language model either lacked sufficient context or failed to disambiguate homophonic pairs. Such errors are particularly common with minimal-pair words, where subtle phonetic distinctions (e.g., voicing or vowel length) are acoustically ambiguous or when the language model's contextual window is too narrow to leverage syntactic or semantic cues.

Insertion errors arise when the recognizer outputs words that are not present in the reference, typically resulting from overconfident language model predictions. For example, the system might insert "me" in "book me a flight" for the reference "book a flight." This suggests that the language model was weighted too heavily, causing the decoder to favor grammatically plausible sequences over the actual acoustic evidence.

Deletion errors represent missing words that were present in the reference but not recovered during decoding, often reflecting under-confidence or failure to detect weak acoustic cues. A common example is recognizing "She has car" instead of "She has a red car." This can happen because the acoustic model missed weak pronunciations (e.g., the reduced vowel in "a" or nasal blending in "red") or because the decoder's confidence threshold was too strict, leading it to drop low-probability tokens. Such errors become more frequent in fast speech, where short words like "a" or "the" blend into surrounding words and are therefore harder to detect.

This level of analysis, this study will show that whether improvements in OOV recall or word error rate(WER) from better word selection (fewer substitutions), better use of context (suitable insertions), or broader lexical coverage (fewer deletions). In Section 4.3, I will show a detailed breakdown of error types comparing the baseline Wav2Vec2 model and my n-gram-enhanced system. Understanding how the error types shift helps evaluate the trade-offs and limitations of lightweight language model adaptation without retraining the acoustic encoder.

3.4 Ethics Consideration

This research project exclusively uses publicly available data and does not involve the collection of new data from human participants. For alignment with the FAIR (Findable, Accessible, Interoperable, Reusable) data principles, the study employs the TED-LIUM Release 2 corpus Rousseau et al. (2014), which comprises English-language audio recordings of TED Talks and their corresponding transcripts. All audio and textual content originates from the TED website and is owned by TED Conferences LLC. The corpus is easily discoverable and accessible for download via OpenSLR (identifier: SLR19), fulfilling the findability and accessibility criteria. The dataset is provided in standardized formats (e.g., STM and SPH) to ensure interoperability across automatic speech recognition (ASR) toolkits. Furthermore, the corpus's licensing terms (CC BY-NC-ND 3.0) permit transparent reuse for academic research, supporting the reusability principle. All preprocessing scripts, model checkpoints, and evaluation results are publicly released via GitHub and Hugging Face to enhance transparency and reproducibility.

The dataset used does not contain any personally identifiable information (PII) or sensitive data. Consequently, this study does not involve direct participation of human subjects, nor does it require renewed informed consent. All audio recordings in the TED-LIUM2 corpus were originally published as part of TED Talks, and the dataset was curated and processed by the Laboratoire d'Informatique de l'Université du Maine (LIUM) for research purposes.

This study uses this dataset in combination with open-source tools such as KenLM to train a custom 4-gram language model. The model aims to demonstrate whether a domain-specific language model for spontaneous speech can improve the recognition accuracy of ASR systems in this domain, as evaluated through objective performance metrics, including word error rate (WER) and out-of-vocabulary (OOV) word recall. The TED-LIUM2 dataset presents potential limitations related to speaker diversity and topic coverage. As it primarily consists of curated public speeches, it may not fully capture the variability of spontaneous or conversational speech. This could introduce certain biases into the trained model, which will be discussed in the results section.

All code and trained models used in this study are made openly available through public repositories to ensure full reproducibility. While minor variations in results may arise due to differences in hardware or software configurations, the methodology is designed to ensure the results are clear and repeatable. The experiments were conducted using the high-performance computing resources of the Hábrók cluster at the University of Groningen.

In summary, this research complies with the ethical standards of the institution and does not involve human subject research, data sensitivity, or privacy.

4 Experimental Setup

To ensure reproducibility and clarity, this section explains the experimental setup used in this study. The goal is to test whether a domain-specific language model can improve recognition for spontaneous speech and OOV words, without retraining the acoustic model. This is done by adding an n-gram language model through shallow fusion.

I describe the full pipeline, including data preparation, data splitting, and the three main experiments. All experiments use the same base model, Wav2Vec2.0, and compare decoding results with and without added language models. I also explain preprocessing steps, parameter settings, hardware used, and the software tools like pyctcdecode and KenLM. All scripts, trained models, and results are shared on GitHub to support future work. The specific language model can be downloaded and tested from GitHub[2].

4.1 Data Preparation

In this work, I use a pretrained acoustic model (wav2vec2-base-100h) and focus on improving decoding through shallow fusion with an n-gram language model trained on spontaneous speech. The language model is trained exclusively on TED-LIUM2 transcriptions, without any fine-tuning of the acoustic model. This setup enables a lightweight domain adaptation pipeline that relies only on text data, avoiding retraining or augmentation of the audio data. The original audio files are in sphere format (SPH), which I converted to WAV format. Each test audio file was then segmented into approximately 30-second clips for recognition.

For reference transcript preparation, I then removed all lines containing `<o,,unknown>` and `ignore_time_segment_in_scoring`. At the same time, after obtaining the results from Baseline 1 (i.e., wav2vec2-base-100h without language modeling), which were saved in a TXT file with the same number of rows, I confirmed through manual inspection that the original dataset was also segmented to align with the 30-second clips. Therefore, each sentence in the transcript matched the corresponding segment.

Next, I cleaned the Baseline 1 results by removing rows without transcribed content. I also stripped both files (reference and predicted) of metadata such as filenames, speaker IDs, durations, gender, and other comments. The ground-truth transcript was saved as reftext.txt, and the Baseline 1 result as testtext.txt. Both files contain only the transcribed sentences, making them suitable for accurate Word Error Rate (WER) calculation.

4.2 Data Splitting

The test dataset used in this study is the TED-LIUM2 corpus, which contains a total of 1,495 audio TED Talks along with manually transcribed STM texts. To ensure a balanced evaluation, the dataset was divided into three subsets. The training dataset includes 93,480 sentences after removing blank lines from both the training and development sets. The test dataset comprises 2.6 hours of audio and 1,155 transcript sentences, also cleaned of blank lines. In addition, an out-of-vocabulary (OOV) dataset was constructed to identify words that appear in the test set but not in the training set. A Python script was developed to automatically extract OOV words from the STM-format transcripts

[2] https://github.com/bxftan90/ted/tree/main/TEDLIUM_release2

of both sets. The script processes the transcribed text, starting from the seventh field of each STM line, while ignoring metadata such as time stamps and speaker names. All transcriptions were lowercased to avoid mismatches due to case differences. OOV words were then identified by calculating the set difference between the test and training vocabularies.

4.2.1 Baseline 1: Pretrained model without language model

The study by Baevski et al. (2020) evaluated the Wav2Vec 2.0 model on the LibriSpeech benchmark, achieving a Word Error Rate (WER) of approximately 3.4% on the "test-clean" subset and 8.0% on the "test-other" subset. The Wav2Vec BASE model consists of a 12-layer Transformer encoder with a hidden size of 768 and approximately 95 million parameters. The model was first pre-trained in a self-supervised manner on 960 hours of unlabeled speech data from the LibriSpeech corpus (LS-960) to learn generic audio representations. It was then fine-tuned using 100 hours of labeled data from the clean-100 subset to perform automatic speech recognition. In this study, this baseline is used as a reference for establishing the WER on the TED-LIUM2 dataset.

4.2.2 Experiment 1: Pretraining model with language model

To evaluate the effect of integrating an external language model into the decoding process, this experiment uses the wav2vec2-base-100h-with-lm model provided by Hugging Face. This model incorporates a pretrained 4-gram language model built using KenLM. Shallow fusion is applied during decoding, where the acoustic scores predicted by the Wav2Vec2.0 model are combined with language model probabilities. The beam search algorithm then selects the most likely transcription candidates based on this joint scoring. This experiment aims to determine whether integrating such a general-purpose language model can enhance recognition accuracy in the spontaneous speech domain.

4.2.3 Experiment 2: Pretraining model with language model which trained by TEDLIUM2

In this experiment, the objective is to enhance the automatic speech recognition (ASR) system's ability to recognize rare words or out-of-vocabulary (OOV) words that frequently occur in spontaneous speech. These words are typically missing from the training data and may include specific names, locations, or technical terms. Although the ASR model Wav2Vec2.0 (facebook/wav2vec2-base-100h from Hugging Face) already demonstrates strong recognition capabilities, it still struggles with such specialized vocabulary.

To address this issue, I trained a 4-gram language model (LM) using KenLM. The LM was trained on the transcriptions from the training split of the TED-LIUM2 dataset. TED Talks were chosen because they feature natural speaking styles and cover a wide range of topics such as technology, education, arts, and psychology. These characteristics make TED-LIUM2 highly suitable for training an LM adapted to spontaneous speech. Additionally, TED-LIUM2 is a widely-used public dataset with carefully curated transcripts, which is important for effective n-gram-based language modeling. By leveraging these real-world transcriptions, the LM can better capture common word sequences and syntactic patterns, improving ASR decoding accuracy.

During training, Kneser-Ney smoothing was applied, and memory-related settings were tuned to avoid resource issues. The trained model was then converted from .arpa format to .binary to

optimize inference speed. I used PyCTCDecode to integrate the language model into the ASR decoding process. It combines the acoustic output from Wav2Vec2.0 with the language model to improve transcription accuracy. All speech segments were preprocessed into 30-second chunks, and recognition results were later aggregated for evaluation.

All experiments were conducted on the Hábrók high-performance computing (HPC) cluster at my university. I used an NVIDIA A100 GPU node with 20 CPU cores and 48GB of memory. This setup ensured efficient processing for memory-intensive tasks such as loading large ASR models and performing batch inference. The GPU significantly reduced processing time and prevented out-of-memory errors. Additionally, Hábrók's job scheduling system allowed smooth execution of both LM training and large-scale decoding.

To evaluate the effectiveness of the language model, I extracted a list of out-of-vocabulary (OOV) words that appeared in the test set but not in the training set. I then measured how many of these words were correctly recognized in the final ASR output. This enabled the calculation of OOV recall, which reflects the language model's ability to improve the recognition of previously unseen vocabulary. The list of OOV terms was generated by comparing the vocabularies of the test and training sets. Representative examples include esolar, etruscans, and foodie. The full list of OOV words along with their recognition results is provided in Appendix A.

5 Results

To evaluate the effectiveness of external n-gram language models (LMs) in improving spontaneous speech recognition, I compared three results: (1) Wav2Vec2.0 Base with no LM (baseline), (2) Wav2Vec2.0 integrated with a pretrained LM (trained on LibriSpeech), and (3) Wav2Vec2.0 integrated with a 4-gram LM trained on the TED-LIUM2 corpus, which consists of spontaneous speech data.

As shown in Table 1, integrating the pretrained LibriSpeech LM significantly reduced the overall word error rate (WER) from 25.04% to 21.22%, and improved the out-of-vocabulary (OOV) recall from 31.51% to 40.41%. This confirms that a well-trained external LM can enhance both general decoding accuracy and the recognition of rare words, even without modifying the acoustic model.

In contrast, the TED-LIUM2-based 4-gram LM achieved a comparable OOV recall of 38.36%, representing a 21.75% relative improvement over the baseline. This supports the hypothesis that an external LM trained solely on spontaneous speech can still improve the recognition of OOV terms. However, the overall WER increased substantially to 44.28%, a relative increase of 76.77%.

The pretrained language model used in the comparison is part of the model patrickvonplaten/wav2vec2-base-100h-with-lm, which was provided by Hugging Face. While specific details about the text corpus used to train its 4-gram KenLM component are not fully disclosed, it is likely to be based on a relatively large and clean English corpus consistent with LibriSpeech standards. In contrast, my custom 4-gram model was trained using only the TED-LIUM2 training transcriptions, which include approximately 93,480 cleaned sentences. The resulting KenLM binary is about 79.4MB in size. These differences in dataset scale, lexical diversity, and domain characteristics likely contribute to the observed contrast in WER performance.

In conclusion, while domain-matched spontaneous speech LMs can improve OOV recall, they may lack the robustness needed for overall transcription accuracy unless trained on larger and more coherent corpora.

Table 1: WER and OOV Recall with Relative Change Compared to Baseline

System	**WER (%)**	**ΔWER (%)**	**OOV Recall (%)**	**ΔRecall (%)**
Wav2Vec2 Base (no LM)	25.04	–	31.51	–
Wav2Vec2 + Pretrained LM	21.22	−15.26	40.41	+28.23
Wav2Vec2 + TED-LIUM2 n-gram LM	44.28	+76.77	38.36	+21.75

The relative change in word error rate (WER) and OOV recall which allows us to measure the proportional improvement or degradation compared to the baseline is calculate as:

$$\Delta\% = \frac{\text{New Value} - \text{Original Value}}{\text{Original Value}} \times 100\%$$

Table 2: Error Type Breakdown Across Systems (Substitution / Insertion / Deletion / Hits)

System	Substitutions	Insertions	Deletions	Hits
Baseline	6029	283	144	22666
Pretrained LM	4824	278	229	23822
TED-LIUM2 LM	11747	100	247	16482

To further investigate how language modeling impacts decoding behavior, I analyzed recognition errors in terms of substitutions, insertions, deletions, and correctly recognized tokens (hits) across the three systems.

Table 2 presents a breakdown of error types for each system. Notably, the system using the TED-LIUM2-trained language model exhibits a significant increase in substitution errors (11,747) compared to both the baseline system (6,029) and the pretrained LM system (4,824). This suggests that although the TED-LIUM2 LM may better capture domain-specific expressions, it also introduces more word-level mismatches, potentially due to limited vocabulary generalization or overfitting to the training domain.

In contrast, insertion errors are lowest in the TED-LIUM2 LM system (100), indicating that it is more conservative in introducing spurious words, likely due to stricter decoding constraints. However, deletion errors slightly increase (247) compared to the baseline (144), possibly reflecting a trade-off between avoiding insertions and missing low-confidence words.

Interestingly, the pretrained LM system achieves the highest number of correct word matches (Hits: 23,822), suggesting that its broader linguistic coverage, derived from large-scale LibriSpeech training data, results in more accurate recognition—especially for general vocabulary. Although it produces slightly more insertions than the TED-LIUM2 LM, it manages substitutions and deletions more effectively.

Overall, the results indicate that although the domain-specific language model (TED-LIUM2) improves insertion handling, it performs worse in managing deletion errors and shows a substantial increase in substitution errors. These combined effects lead to a significant drop in the total number of correctly recognized words, suggesting that the model may struggle with general vocabulary coverage despite its domain alignment.

6 Discussion

Upon analyzing the results presented in Table 1 and Table 2, it is evident that integrating an external language model trained on TED-LIUM2 led to a significantly higher substitution rate (11,747) compared to both the baseline system (6,029) and the LibriSpeech-pretrained language model (4,824). However, the TED-LIUM2 LM demonstrated the lowest number of insertions (100) and a moderate number of deletions (247). In terms of Word Error Rate (WER), the TED-LIUM2 LM produced the highest WER at 44.28%, compared to 25.04% for the baseline and 21.22% for the pretrained LM. Nonetheless, it also achieved a relative OOV recall improvement of +21.75%, indicating enhanced recognition of rare terms. These results provide valuable insights into the main research question: *How can the integration of an external n-gram language model during decoding improve the recognition accuracy of spontaneous speech and the handling of out-of-vocabulary words in end-to-end ASR systems without modifying the acoustic model?*

6.1 Validation of the First Hypothesis

The first hypothesis anticipated that external language modeling would improve OOV recall in the spontaneous speech domain. This hypothesis is supported. Despite the higher WER, The numbers show that the TED-LIUM2 LM was more successful in recognizing spontaneous vocabulary and domain-specific expressions. Specifically, the TED-LIUM2 LM achieved an OOV recall of 38.36%, reflecting a +21.75% improvement over the baseline (31.51%). While slightly below the pretrained LM (40.41%, +28.23%), it still shows some improvement in recognizing rare or out-of-vocabulary items.

This finding aligns with prior work such as Zhou et al. (2021), which showed that domain-adapted language models, even if small in size, can improve recall for rare and contextually salient terms. The TED-LIUM2 LM, though smaller, better captured the distribution of terms common in unscripted, natural speech. This supports the idea that domain specificity can enhance lexical coverage, even if overall WER increases.

6.2 Validation of the Second Hypothesis

The second hypothesis posited that training a domain-specific language model for spontaneous speech (TED-LIUM2) without fine-tuning the acoustic model could reduce the overall WER. Based on the results, this hypothesis is not supported. The TED-LIUM2 LM produced a WER of 44.28%, representing a 76.77% relative increase compared to the baseline. While it achieved the lowest insertion count, it yielded a substitution rate nearly twice that of the baseline and more than double that of the LibriSpeech-pretrained LM. Furthermore, its hit count (16,482) was markedly lower than that of the baseline (22,666) and the LibriSpeech LM (23,822).

A reason for this finding relates to the characteristics of the training corpora. By comparison, the official LibriSpeech corpus contains approximately 281,000 sentences spanning 960 hours, whereas the TED-LIUM2 training set includes only 93,480 sentences. Consequently, the TED-LIUM2 model's limited corpus size and vocabulary reduce its ability to assign reliable n-gram probabilities during beam-search decoding, impairing its capacity to properly re-rank competing hypotheses at each word position. As discussed in the previous literature review, the study by Furui et al. (2005) also confirmed this finding. Using the Corpus of Spontaneous Japanese (CSJ), a large-scale

spontaneous speech database containing approximately 7 million words and 650 hours of speech, they conducted recognition experiments under various training data conditions. With increasingly larger training corpora, the results demonstrated a clear improvement in recognition performance. Specifically, when the language model training text increased eightfold from 0.86 million to 6.84 million words, the corresponding word error rate (WER) decreased by 17%. Furthermore, under the condition of using 510 hours of acoustic training data together with the full 6.84 million-word text corpus, the best WER achieved was 25.3%. This supports the importance of both data quantity and spontaneous-domain relevance for improving recognition accuracy in spontaneous speech scenarios. However, the TED-LIUM 2 training set is relatively small, with only around 210 hours of speech data. This limitation highlights the potential for future work to explore larger and more diverse spontaneous speech corpora for further improving ASR performance.

The high substitution rate in TED-LIUM2 LM outputs is primarily due to errors involving contractions (e.g., “don’t” and “do not”) and word concatenation (e.g., “andhow” instead of “and how”). These mistakes reflect the n-gram LM’s inability to disambiguate short-range contextual dependencies and space boundary errors during beam search decoding.

To further explore whether connected word errors have too much influence on the overall WER, I made an extra assumption. I believe that although these errors are counted as substitution errors in form, they may not really affect the understanding of listeners. For example, “thankyou” was recognized as one word, but the reference is “thank you.” The meaning is almost the same, so counting it as an error may make the system performance look worse than it really is. To test this idea, I used Python to find all connected word errors and removed 492 of them from the WER calculation (the full list is in Appendix B). This operation simulates a more relaxed evaluation, where small spelling connection mistakes are not treated as recognition errors if they do not change the meaning.

Table 4 show that after removing these errors, the WER of the TED-LIUM2 language model dropped from 44.28% to 38.17%. This shows that connected word errors do affect WER to some extent. But even after fixing them, the WER is still much higher than the baseline system without any language model (25.04%) and also much higher than the system using the LibriSpeech pretrained language model (21.22%). So, the results show that although connected word errors do increase WER a little, they are not the main reason why the TED-LIUM2 language model performs poorly. A more likely reason is that the language model has weak generalization ability. This may be because the training data is small, the sentence structure is not very standard, and the vocabulary is limited. Just fixing the connected words is not enough to solve the performance problem. Future work can try methods like data augmentation, vocabulary expansion, or smoothing strategies to improve the model performance.

The language model was trained and decoded using a single A100 GPU with 16 CPU cores and 48GB of memory. The training and inference pipeline was highly efficient: the n-gram LM training took under 10 minutes, and decoding with shallow fusion remained within real-time constraints. This suggests that although the performance of the LM is not optimal, it provides a lightweight and computationally affordable framework for rapid domain adaptation.

Table 3: Substitution Errors and Correct Forms

No.	Recognized Result	Correct Form
1	andhow	and how
2	andatferst	and at first
3	don't	do not
4	everam	ever am
5	gaveme	gave me
6	howaryou'regoing	how are you're going
7	it's	it is
8	itdoesn't	it doesn't
9	itwas	it was
10	nearto	near to
11	ofcourse	of course
12	thankyou	thank you
13	thatyou	that you
14	you're	you are
15	youknow	you know

Note: Some contractions (e.g., `don't`) may be counted as substitution errors when compared to their full forms in the reference transcript (e.g., `do not`). Certain concatenated words (e.g., `howaryou'regoing`) may involve not only missing spaces but also grammatical or spelling errors.

Table 4: WER with OOV Recall and Impact of Removing Connected-Word Errors

System	WER (%)	ΔWER (%)	OOV Recall (%)
Wav2Vec2 Base (no LM)	25.04	—	31.51
Wav2Vec2 + Pretrained LM	21.22	–15.26	40.41
Wav2Vec2 + TED-LIUM2 4-gram LM	44.28	+76.77	38.36
Wav2Vec2 + TED-LIUM2 4-gram LM			
(Corrected Connected Word Errors)	38.17	+52.44	38.36

6.3 Limitations

This study has several limitations. First, the TED-LIUM2 corpus is significantly smaller, containing only 93,480 sentences. In comparison, the LibriSpeech corpus includes around 281,000 sentences. Besides the difference in data scale, there are also clear differences in text quality and standardization. Specifically, LibriSpeech's transcripts are derived from read books, with highly standardized language. Each sentence in LibriSpeech is semantically complete, syntactically well-formed, and contextually coherent. This makes it ideal for training language models to learn grammar rules and word order patterns. Such clean and formal data helps the language model learn stable word combinations and contextual probabilities, which improves prediction during decoding.

In contrast, TED-LIUM2's transcripts are from spontaneous TED talks. Although the content includes many out-of-vocabulary (OOV) terms, the transcription is segmented by time stamps, so many sentences are not complete. They may lack full semantic content or even basic syntactic structure. For example, some utterances are split into short fragments like "if you've" or "you've been over", which do not provide enough context and make it difficult for the language model to learn sentence-level dependencies. In addition, spontaneous speech contains disfluencies such as repetitions, corrections, and filler words. These result in many redundant items in the transcripts, including repeated words like "so so" or fillers like "you know" and "I mean", which increase the ambiguity of language structure. Cleaning them manually would take a lot of time.

Second, this study did not include experiments with LM weights. In (Tian, Yu, Weng, Zou, & Yu, 2022) study on improving Mandarin end-to-end speech recognition with a word N-gram language model, found that different LM weight values affected the recognition stability. In the Chinese research setting, the optimal weight was 0.4. The influence of the language model on decoding results is usually controlled by a hyperparameter α. If this weight is not properly tuned, the language model may be either too dominant or too weak, which may lead to decoding errors.

7 Conclusion

This thesis explored how the integration of an external n-gram language model during decoding can improve the recognition of spontaneous speech and out-of-vocabulary (OOV) words in end-to-end automatic speech recognition (ASR) systems, without modifying the acoustic model. The conclusion summarizes the key contributions, proposes future research directions, and highlights the relevance of this work.

7.1 Summary of the Main Contributions

This study contributes by demonstrating that a lightweight, word-level n-gram language model trained on a small-scale spontaneous speech corpus (TED-LIUM2) can improve the recognition of out-of-vocabulary (OOV) words in spontaneous ASR tasks. Despite the limited size and informal nature of the corpus, the domain-specific LM achieved a +21.75% increase in OOV recall compared to the baseline. This result highlights the potential of low-resource, domain-adapted LMs to enhance coverage of rare and in-domain terms in end-to-end ASR systems. It also provides evidence that such models can be trained efficiently within 10 minutes on a single A100 GPU, which making them scalable and accessible. However, the same LM also increased overall WER due to structural limitations in the training corpus.

7.2 Future Work

Future work can explore using more spontaneous speech data to further enhance OOV recognition. However, such data should be manually or algorithmically cleaned to remove repetitions and properly segmented into complete sentences, so that both language and acoustic models can better capture contextual and syntactic structures. In addition to expanding the LM training corpus, future research may also consider fine-tuning the acoustic model on spontaneous speech, or incorporating hybrid strategies such as combining n-gram LMs with subword modeling techniques (e.g., BPE, ADSM) to reduce boundary and compound word errors. Furthermore, neural-based context biasing and advanced rescoring methods (e.g., NAM, CATT) could be integrated to better resolve ambiguous or rare phrase predictions. These approaches may help strike a better balance between OOV recall and overall transcription accuracy, especially in low-resource or domain-specific ASR scenarios.

7.3 Impact & Relevance

This research has practical significance in domains where spontaneous, dynamic language dominates, such as education, meetings, video content, and assistive technologies. Its lightweight nature also suggests promise for real-time or on-device speech applications in low-resource environments.

Beyond technical contributions, this work highlights the importance of matching language models to speech domains and balancing between vocabulary flexibility and structural accuracy. These insights are timely given the rapid emergence of new terms in digital communication and the growing demand for accurate, adaptable ASR systems.

In summary, this paper has made progress in enhancing domain-aware automatic speech recognition. By training a language model on spontaneous speech data and analyzing word error rate (WER) and out-of-vocabulary (OOV) word recall, the study demonstrates the model's improved ability to

handle OOV vocabulary. This work lays the foundation for developing a more robust and efficient decoding mechanism in future end-to-end speech recognition systems.

References

Alberti, C., & Bacchiani, M. (2009, December). *Automatic captioning in youtube.* Retrieved from https://research.google/blog/automatic-captioning-in-youtube/ (Google Research Blog)

Baevski, A., Schneider, S., & Auli, M. (2019). vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449–12460.

Bahl, L., & Jelinek, F. (2003). Decoding for channels with insertions, deletions, and substitutions with applications to speech recognition. *IEEE Transactions on Information Theory*, *21*(4), 404–411.

Bazzi, I. (2002). *Modelling out-of-vocabulary words for robust speech recognition* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171–4186).

Dockes, P. (2022). *Recognizing rare words: Experiments with subword units.* Retrieved from https://www.speechmatics.com/company/articles-and-news/recognizing-rare-words-experiments-with-subword-units (Accessed: 2025-04-25)

Dufour, R., Jousse, V., Estève, Y., Béchet, F., & Linarès, G. (2009). Spontaneous speech characterization and detection in large audio database. *SPECOM, St. Petersburg*.

Fox, J. D., & Delworth, N. (2022). Improving contextual recognition of rare words with an alternate spelling prediction model. *arXiv preprint arXiv:2209.01250*.

Fukuda, T., Fernandez, R., Rosenberg, A., Thomas, S., Ramabhadran, B., Sorin, A., & Kurata, G. (2018). Data augmentation improves recognition of foreign accented speech. In *Interspeech* (pp. 2409–2413).

Furui, S., Nakamura, M., Ichiba, T., & Iwano, K. (2005). Analysis and recognition of spontaneous speech using corpus of spontaneous japanese. *Speech Communication*, *47*(1-2), 208–219.

Guillot, M.-N. (2010). Film subtitles from a cross-cultural pragmatics perspective: Issues of linguistic and cultural representation. *The Translator*, *16*(1), 67–92.

Heafield, K. (2011). Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation* (pp. 187–197).

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, *29*, 3451–3460.

Inaguma, H., Mimura, M., Sakai, S., & Kawahara, T. (2018). Improving oov detection and resolution with external language models in acoustic-to-word asr. In *2018 ieee spoken language technology workshop (slt)* (pp. 212–218).

Jaitly, N., Nguyen, P., Senior, A. W., & Vanhoucke, V. (2012). Application of pretrained deep neural networks to large vocabulary speech recognition. In *Interspeech* (Vol. 2012, pp. 2578–2581).

Jouvet, D., Langlois, D., Menacer, M. A., Fohr, D., Mella, O., & Smaïli, K. (2018). Adaptation

of speech recognition vocabularies for improved transcription of youtube videos. *Journal of International Science and General Applications*, *1*(1), 1–9.

Jung, K., Kim, N.-J., Ryu, H. G., & Lee, H.-J. (2024). Enhancing asr performance through ocr word frequency analysis: Theoretical foundations. *arXiv preprint arXiv:2405.02995*.

Jurafsky, D., & Martin, J. H. (n.d.). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.*

Klakow, D., Rose, G., & Aubert, X. (1999). Oov-detection in large vocabulary system using automatically defined word-fragments as fillers. In *Proc. eurospeech 1999* (pp. 49–52).

Koto, F., Sakti, S., Neubig, G., Toda, T., Adriani, M., & Nakamura, S. (2014). The use of semantic and acoustic features for open-domain ted talk summarization. In *Signal and information processing association annual summit and conference (apsipa), 2014 asia-pacific* (pp. 1–4).

Kuhn, K., Kersken, V., Reuter, B., Egger, N., & Zimmermann, G. (2024). Measuring the accuracy of automatic speech recognition solutions. *ACM Transactions on Accessible Computing*, *16*(4), 1–23.

Liao, H. (2013). Speaker adaptation of context dependent deep neural networks. In *2013 ieee international conference on acoustics, speech and signal processing* (pp. 7947–7951).

Liu, D., Niehues, J., & Spanakis, G. (2020). Adapting end-to-end speech recognition for readable subtitles. *arXiv preprint arXiv:2005.12143*.

Liu, Y., Bao, L., Wang, J., Weng, Y., & Zhao, Y. (2023). A study on the vitality of internet buzzwords based on search index image analysis. *Modern Linguistics ()*, *11*, 1998. (In Chinese)

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ma, Y., He, H., & Niu, Z. (2023). Bdc: Using bert and deep clustering to improve chinese proper noun recognition. In *Seke* (pp. 57–62).

Maekawa, K., et al. (2003). Corpus of spontaneous japanese: Its design and evaluation. In *Proc. isca & ieee workshop on spontaneous speech processing and recognition* (Vol. 2003, pp. 7–12).

Munkhdalai, T., Sim, K. C., Chandorkar, A., Gao, F., Chua, M., Strohman, T., & Beaufays, F. (2022). Fast contextual adaptation with neural associative memory for on-device personalized speech recognition. In *Icassp 2022-2022 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 6632–6636).

Nakamura, M., Iwano, K., & Furui, S. (2008). Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech & Language*, *22*(2), 171–184.

Park, A., & Glass, J. R. (2005). Towards unsupervised pattern discovery in speech. In *Ieee workshop on automatic speech recognition and understanding, 2005.* (pp. 53–58).

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Pepino, L., Riera, P., & Ferrer, L. (2021). Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502*.

Qu, L., Weber, C., & Wermter, S. (2023). Emphasizing unseen words: New vocabulary acquisition for end-to-end speech recognition. *Neural Networks*, *161*, 494–504.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning* (pp. 28492–28518).

Rousseau, A., Deléglise, P., & Estève, Y. (2014). Enhancing the ted-lium corpus with selected data for language modeling and more ted talks. In *Proceedings of the ninth international conference on language resources and evaluation (lrec)*. Retrieved from https://www.openslr.org/19/

Sainath, T. N., Kingsbury, B., Sindhwani, V., Arisoy, E., & Ramabhadran, B. (2013). Low-rank matrix factorization for deep neural network training with high-dimensional output targets. In *2013 ieee international conference on acoustics, speech and signal processing* (pp. 6655–6659).

Saon, G., Soltau, H., Nahamoo, D., & Picheny, M. (2013). Speaker adaptation of neural network acoustic models using i-vectors. In *2013 ieee workshop on automatic speech recognition and understanding* (pp. 55–59).

Sarı, L., Moritz, N., Hori, T., & Le Roux, J. (2020). Unsupervised speaker adaptation using attention-based speaker memory for end-to-end asr. In *Icassp 2020-2020 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 7384–7388).

Schaaf, T. (2001). Detection of oov words using generalized word models and a semantic class language model. In *Interspeech* (pp. 2581–2584).

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Seide, F., Li, G., Chen, X., & Yu, D. (2011). Feature engineering in context-dependent deep neural networks for conversational speech transcription. In *2011 ieee workshop on automatic speech recognition & understanding* (pp. 24–29).

Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Thara, P., Azneed, M., Sanas, M., PM, J. P., Naik, H. P., & Divya, B. (2024). Subtitle synchronization using whisper asr model. In *2024 international conference on power, energy, control and transmission systems (icpects)* (pp. 1–6).

Tian, J., Yu, J., Weng, C., Zou, Y., & Yu, D. (2022). Improving mandarin end-to-end speech recognition with word n-gram language model. *IEEE Signal Processing Letters*, *29*, 812–816.

Wan, D., Kedzie, C., Ladhak, F., Turcan, E., Galuščáková, P., Zotkina, E., ... McKeown, K. (2021). Segmenting subtitles for correcting asr segmentation errors. *arXiv preprint arXiv:2104.07868*.

Ward, W. (1989). Understanding spontaneous speech. In *Speech and natural language: Proceedings of a workshop held at philadelphia, pennsylvania, february 21-23, 1989*.

Wessel, F., Schluter, R., Macherey, K., & Ney, H. (2002). Confidence measures for large vocabulary continuous speech recognition. *IEEE Transactions on speech and audio processing*, *9*(3), 288–298.

Xu, T., Yang, Z., Huang, K., Guo, P., Zhang, A., Li, B., ... Xie, L. (2023). Adaptive contextual biasing for transducer based streaming speech recognition. *arXiv preprint arXiv:2306.00804*.

Yi, C., Wang, J., Cheng, N., Zhou, S., & Xu, B. (2020). Applying wav2vec2. 0 to speech recognition in various low-resource languages. *arXiv preprint arXiv:2012.12121*.

Zhou, W., Michel, W., Irie, K., Kitza, M., Schlüter, R., & Ney, H. (2020). The rwth asr system for ted-lium release 2: Improving hybrid hmm with specaugment. In *Icassp 2020-2020 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 7839–7843).

Zhou, W., Zeineldeen, M., Zheng, Z., Schlüter, R., & Ney, H. (2021). Acoustic data-driven subword

modeling for end-to-end speech recognition. *arXiv preprint arXiv:2104.09106*.

Zue, V., Seneff, S., Glass, J. R., Polifroni, J., Pao, C., Hazen, T. J., & Hetherington, L. (2002). Jupleter: a telephone-based conversational interface for weather information. *IEEE Transactions on speech and audio processing*, *8*(1), 85–96.

Appendices

A List of OOV

acai	acrid	adjudicate	administer
affliction	agitation	ambivalently	antioxidant
antonyms	argentinians	arpeggios	astrobiologists
attentuation	azeroth	battled	bismark
bolster	brood	bruckheimer	bruckner
buffing	bushel	calmly	cantaloupe
casking	castronova	cellists	cg
chicanery	cinematic	circumnavigate	consoles
controlers	ctos	damian	daunt
deadening	deficiencies	deja	dembois
dimensionalize	diphtheria	disgraceful	disposition
dosages	downey	drawback	dundee
echinacea	educe	enemas	erudite
esa	esolar	etruscans	exhaustive
exodus	feasting	flamingos	flavorful
flue	foodie	foxx	frankenfoods
gimmicky	ginkgo	goody	guadalquivir
hafiz	hahnemann	halibuts	herodotus
hundreth	ignore_time_segment_in_scoring	impurities	info
innumerable	instructive	intermittent	jockey
jockeying	jovial	kean	khosla
kindergarteners	knuckles	laypeople	lemond
lomborg	lydians	malcom	marshmallowchallenge
marshmallows	mcgonigal	mckay	mendelssohn
methodologically	mikumi	mullet	mumps
nepotistic	offit	orienting	overcooked
palma	pekka	pestered	pizzutillo
ploy	pressurize	procured	prospering
purification	raggedy	realists	redemptive
repetitions	rotavirus	rubella	salmons
salonen	shielding	sibelius	skillman
slogging	spooled	sprig	superstruct
swooning	synopsis	tangerines	telepresense
terrapower	thesaurus	timescale	toledano
unforgiving	unhappier	unsuccessfully	vanquished
veta	vinod	vu	wilmington
wrecked	zoomable		

B List of Connected Word Errors

andmachines	and machines	greatgrandparents	great grandparents
encouragepeople	encourage people	behindthe	behind the
commonstance	common stance	fivebillion	five billion
deepdives	deep dives	theunited	the united
themeabout	theme about	worldlast	world last
thehealth	the health	bottomline	bottom line
decentpeople	decent people	magictrick	magic trick
andcooperation	and cooperation	maybethat	maybe that
thateffect	that effect	justgoing	just going
nextdecade	next decade	notbecause	not because
twoselves	two selves	thevalues	the values
broadbacking	broad backing	backpocket	back pocket
theninety	the ninety	andvaluable	and valuable
theenergy	the energy	worldsaving	world saving
themessage	the message	thepeople	the people
theopposite	the opposite	secondgrade	second grade
justviewing	just viewing	reportcard	report card
productionpossible	production possible	disclosurerules	disclosure rules
andadventures	and adventures	justenough	just enough
burningcoal	burning coal	andplaying	and playing
weretaking	were taking	gamesthan	games than
theextraordinary	the extraordinary	whitecolor	white color
thewireless	the wireless	manythings	many things
haveheard	have heard	southafrica	south africa
giantfields	giant fields	filmmaking	film making
understoodthat	understood that	experiencingself	experiencing self
simpletrick	simple trick	thesedays	these days
summertime	summer time	andlikewise	and likewise
goodeither	good either	everygrain	every grain
happinessself	happiness self	globalcooling	global cooling
mistakenow	mistake now	sleeveback	sleeve back
dicegames	dice games	withserious	with serious
andclimate	and climate	hadtalked	had talked
abouthappiness	about happiness	poliodoes	polio does
secondtrap	second trap	certaindate	certain date
newspaperclipping	newspaper clipping	farmraised	farm raised
adversityand	adversity and	justproducers	just producers
existhardly	exist hardly	aboutgenetically	about genetically
withrealistic	with realistic	dreamteam	dream team
overspecific	over specific	arecompletely	are completely

billionpeople	billion people	closelook	close look
thescientist	the scientist	twopercent	two percent
thedesire	the desire	gregariousway	gregarious way
standback	stand back	thatseems	that seems
offnothing	off nothing	fivehundred	five hundred
diseaseand	disease and	medstudents	med students
thisinvolves	this involves	thisabsurd	this absurd
happenfaster	happen faster	theblueprints	the blueprints
notenough	not enough	theradiation	the radiation
thetechnology	the technology	differentcharacters	different characters
fourthings	four things	effectfor	effect for
fakebecomes	fake becomes	barelybroke	barely broke
otherside	other side	arerecognizing	are recognizing
energymiracles	energy miracles	marchthird	march third
areliving	are living	thecorridor	the corridor
spacemission	space mission	spacemissions	space missions
beefcattle	beef cattle	highlyregulated	highly regulated
sixhundred	six hundred	twentyone	twenty one
foodsystem	food system	twelvepeople	twelve people
thepathology	the pathology	andinterestingly	and interestingly
theplanet	the planet	globalscale	global scale
evenleave	even leave	theregulator	the regulator
fishfarming	fish farming	righthere	right here
demandproof	demand proof	comesfrom	comes from
everychild	every child	dumbenough	dumb enough
takingthem	taking them	talleststructure	tallest structure
creatureand	creature and	fromrapidly	from rapidly
anyonehad	anyone had	liquidwater	liquid water
veryimportant	very important	forgranted	for granted
darnunlikely	darn unlikely	schoolwork	school work
thetitanic	the titanic	askaround	ask around
thinkabout	think about	andrelative	and relative
therehave	there have	persianpoet	persian poet
walkingthe	walking the	withoutarmed	without armed
betterthings	better things	greatbooks	great books
kingdomwide	kingdom wide	theexercise	the exercise
twentyeight	twenty eight	howshould	how should
rememberingself	remembering self	soundlike	sound like
squaremiles	square miles	globalaudience	global audience
fishfarms	fish farms	theresome	there some

publicpolicy	public policy	manyfields	many fields
straightabout	straight about	processthey	process they
whatmight	what might	fullpercentage	full percentage
thethings	the things	pondwater	pond water
couldthrive	could thrive	everybodytalks	everybody talks
thedeadline	the deadline	plantsgoes	plants goes
thankthank	thank thank	theirmost	their most
somebodyand	somebody and	townwhere	town where
himexploding	him exploding	maybeeven	maybe even
therenewable	the renewable	withstudents	with students
theservices	the services	althoughthat	although that
tightsocial	tight social	samplebelow	sample below
highschool	high school	healthcare	health care
haveexplained	have explained	everysingle	every single
grownflesh	grown flesh	theywould	they would
thestudio	the studio	areinvesting	are investing
argumentthat	argument that	thatreleasing	that releasing
theeffects	the effects	figureout	figure out
somethinglike	something like	pullwater	pull water
thestories	the stories	shouldyou	should you
fewpeople	few people	throughour	through our
whitepill	white pill	thereason	the reason
problemsand	problems and	havesignificant	have significant
streetlamps	street lamps	hotelroom	hotel room
shocktherapy	shock therapy	thewinners	the winners
cansimulate	can simulate	youactually	you actually
firstplace	first place	thethrough	the through
thetallest	the tallest	everymonth	every month
reflectiveself	reflective self	thecanals	the canals
helpsthem	helps them	buildingnests	building nests
filmdirector	film director	talkedthem	talked them
hundredyears	hundred years	rightaway	right away
thedoctor	the doctor	thosesame	those same
wouldpick	would pick	everybodywould	everybody would
getdistracted	get distracted	airconditioning	air conditioning
forconsuming	for consuming	believethat	believe that
quietcuriosity	quiet curiosity	globalconflict	global conflict
tastesgood	tastes good	localfood	local food
spacecommunity	space community	butgetting	but getting
theground	the ground	spacestation	space station

certainlyuncertainty	certainly uncertainty	areessentially	are essentially
youprobably	you probably	birdsanctuaries	bird sanctuaries
hundredand	hundred and	thenineteen	the nineteen
couldpick	could pick	startgetting	start getting
thatexists	that exists	aquaticplants	aquatic plants
goodgamer	good gamer	sciencefiction	science fiction
stepinstructions	step instructions	howhappily	how happily
theability	the ability	aboutright	about right
picturehere	picture here	everysprig	every sprig
theseconversations	these conversations	justdeveloped	just developed
onlinegames	online games	withtheir	with their
discussedbroadly	discussed broadly	justkeeping	just keeping
romantickind	romantic kind	environmentand	environment and
conductedabout	conducted about	reasonablepeople	reasonable people
seeingcreatures	seeing creatures	knownabout	known about
weredifferent	were different	forbetter	for better
thelocation	the location	lowenergy	low energy
anddoctor	and doctor	diseasecomes	disease comes
thefuture	the future	betterlives	better lives
breadbasket	bread basket	withamazing	with amazing
hikingand	hiking and	imagedthem	imaged them
everymorning	every morning	superimportant	super important
canexpect	can expect	careabout	care about
worldtoday	world today	rightthere	right there
beenplaying	been playing	whathappens	what happens
exerciseand	exercise and	noteconomically	not economically
withincome	with income	startedplaying	started playing
somethingthat	something that	twentyyears	twenty years
fourwords	four words	deepocean	deep ocean
gettingmore	getting more	andtreating	and treating
thepopulation	the population	mathclass	math class
titanicand	titanic and	moreabout	more about
sixbillion	six billion	everyyear	every year
playinggames	playing games	basedsystem	based system
greataccomplishments	great accomplishments	canignite	can ignite
designstuff	design stuff	overeating	over eating
wherepain	where pain	rightthey	right they
thepenguins	the penguins	achievemore	achieve more
peoplewill	people will	whatshould	what should
birdsanctuary	bird sanctuary	simpleploy	simple ploy

primatehuman	primate human	ninebillion	nine billion
liquidmetal	liquid metal	thevirtual	the virtual
energyand	energy and	themedical	the medical
becausethey	because they	handmoves	hand moves
thechildren	the children	withscience	with science
downthere	down there	thebusiness	the business
fivetimes	five times	workingwith	working with
evergrowing	ever growing	undernear	under near
knowheart	know heart	couldthat	could that
thirtyyears	thirty years	garlicand	garlic and
teddinners	ted dinners	theyexisted	they existed
happypeople	happy people	justabout	just about
freestanding	free standing	planetthe	planet the
effectbut	effect but	worldchanging	world changing
hundredmillion	hundred million	smallteam	small team
theinstitute	the institute	admitcomplexity	admit complexity
fourthfactor	fourth factor	violinlesson	violin lesson
castshadows	cast shadows	manyyears	many years
somethingand	something and	humanability	human ability
termmiracle	term miracle	searchhistory	search history
traineddouble	trained double	knifetrick	knife trick
broughtback	brought back	ownchoosing	own choosing
cometogether	come together	backstageand	backstage and
temperaturewill	temperature will	gamechanges	game changes
thesource	the source	canachieve	can achieve
fiftyyears	fifty years	theinternational	the international
wellbeing	well being	otherpeople	other people
wishfulthinking	wishful thinking	deathnumber	death number
anyoneeven	anyone even	learningabout	learning about
problemsolvers	problem solvers	activeand	active and
farmsthat	farms that	massdestruction	mass destruction
aboutpoint	about point	gameworlds	game worlds
andperhaps	and perhaps	passagework	passage work
amongstothers	amongst others	darkenyour	darken your
areactual	are actual	longbeach	long beach
thepursuit	the pursuit	reasonsare	reasons are
groupedtogether	grouped together	firstthing	first thing
coolingpools	cooling pools	foodproblem	food problem
thekingdom	the kingdom	notnecessarily	not necessarily
differenttypes	different types	anddiving	and diving

problemwith	problem with	andabsolutely	and absolutely
lowerthat	lower that	birdpopulation	bird population
theglobal	the global	eightthousand	eight thousand
everyevening	every evening	themajority	the majority
theirabilities	their abilities	ninetypercent	ninety percent
themoving	the moving	epicmission	epic mission
somepeople	some people	productionand	production and
moregamers	more gamers	themineral	the mineral
justthrough	just through	themagnitude	the magnitude
potablewater	potable water	gettingconfused	getting confused
scubadiver	scuba diver	doingthis	doing this
everydaylife	everyday life	differentapproach	different approach
safetynet	safety net	governmentofficial	government official
beenchanged	been changed	wouldpower	would power
higheststandards	highest standards	storythis	story this
littlebit	little bit	fivepoint	five point
foreveryone	for everyone	theaudience	the audience
drycasking	dry casking	whetherthey	whether they
whatabout	what about	trustthat	trust that
afraidthat	afraid that	thedestruction	the destruction
fishmeals	fish meals	bismarkand	bismark and
everyweek	every week	thecivilization	the civilization
flippedthe	flipped the	everysixty	every sixty
theproblem	the problem		

C Declaration on Use of AI Tools

I hereby affirm that this Master thesis was composed by myself, and the work herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification, nor has it been published. Where other people's work has been used (from any source: printed, internet or otherwise), this has been carefully acknowledged and referenced.

During the preparation of this thesis, I used ChatGPT-4o (OpenAI, June 2025) for the following purposes:

Code generation and fixing: I used ChatGPT-4o to generate Python scripts for WER evaluation, error list extraction, and file merging or connection in the experiments.

Extraction tool professional term: To improve the clarity and completeness of the methodology section, I used ChatGPT-4o to help explain certain technical terms and tools involved in my code logic. After understanding these explanations, I wrote the implementation myself. For example, this applies to the use of PyCTCDecode in Method part.

Equation formatting: In Sections 3.2.2, 3.1.1, and 3.1.2, I used ChatGPT-4o to help extract mathematical expressions from academic papers and insert them into Overleaf in LaTeX format. The equations were selected and interpreted by myself based on the original papers.

Appendix formatting: The word lists in the appendix were formatted and inserted into the LaTeX file with the assistance of ChatGPT-4o. The identification of error types was initially assisted by ChatGPT-4o, which helped organize the output into a structured list. Based on that, I generated Python scripts with its assistance to extract and connect the error information.

Results presentation: I used ChatGPT-4o to assist in presenting and verifying numerical results (e.g., WER values and changes) to ensure consistency with cited studies.

Translation and ethics: For sections such as Ethics Considerations, I used ChatGPT-4o to translate the TEDLIUM2 web. Then summarized by myself. I wrote it by Chinese and translate in ChatGPT-4o to English.

Language polishing: The entire thesis was reviewed using ChatGPT-4o to correct grammatical issues and improve clarity (e.g., replacing informal phrases like "I got the result"). However, the substance, argumentation, and structure of the work remain my own.

Literature review: I used ChatGPT-4o to translate some academic papers into Chinese for easier understanding. Based on these translations, I wrote my own summaries and critical reviews. For numerical information such as WER improvement rates, I verified correctness with ChatGPT before final editing.

All content was subsequently reviewed, revised, and substantially modified by me. The final thesis represents my own understanding, analysis, and academic contribution.

Name: XuefeiBian
Date: 11-06-2025