university of groningen / campus fryslân

University of Groningen

Improving the Performance of Code-Switching Recognition Using Whisper

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
Prof. dr. Tan Phat Do (Voice Technology, University of Groningen)
and
Prof. dr. M. Coler (Voice Technology, University of Groningen)

Yaling Deng (s5666546)

June 4, 2024

Abstract

The intersection of technology and linguistics has given rise to a plethora of challenges and opportunities. Among these, the task of Automatic Speech Recognition (ASR) stands out as a critical area of research and development. ASR systems have become an integral part of our daily lives, facilitating communication and accessibility across various platforms. However, the prevalence of multilingual interactions, particularly code-switching (CS), poses a significant challenge to the accuracy and reliability of these systems. Code-switching, the practice of alternating between two or more languages in the context of a single conversation, is a common phenomenon in many bilingual communities. This research focuses on Mandarin-English intra-sentential CS, a particularly complex scenario due to the stark differences between the two languages in terms of phonetics, syntax, and semantics.

The ubiquity of ASR systems has been propelled by advancements in deep learning and the abundance of data. Deep learning models, with their ability to capture intricate patterns and representations, have revolutionized the field of ASR. However, existing ASR systems often falter when faced with the complexities of multilingual interactions, such as CS. This study is motivated by the linguistic phenomenon of CS and its implications in the accuracy of ASR. It aims to enhance the recognition capabilities of ASR systems in handling Mandarin-English mixed-language speech, a task that is not only technologically challenging but also socially significant.

The Whisper model, developed by OpenAI, serves as the foundation for this research. Whisper is an innovative ASR model that leverages weak supervision and has demonstrated robust performance across various languages and dialects. It is designed to handle the nuances of speech, including different accents and languages, making it an ideal candidate for tackling the challenge of Mandarin-English CS. However, the model's effectiveness in decoding Mandarin-English CS is yet to be fully realized. This research hypothesizes that fine-tuning the Whisper model with a dedicated Mandarin-English CS dataset will significantly improve its performance in recognizing code-switched speech.

To test this hypothesis, a meticulous fine-tuning process was undertaken. The process involved the collection and preparation of a Mandarin-English CS dataset, which was then used to train and refine the Whisper model. The dataset was carefully curated to represent a diverse range of CS scenarios, ensuring that the model would be exposed to a wide variety of linguistic contexts. The fine-tuning process was rigorously evaluated, with a focus on reducing the Mixture Error Rate (MER), a key metric in assessing the performance of ASR systems in handling CS speech.

The results of this study are promising. A substantial reduction in the MER for Mandarin-

English CS speech was observed, validating the hypothesis and highlighting the potential of tailored models in enhancing ASR accuracy. The fine-tuned models, ME-Whisper-small and ME-Whisper-large-v3, exhibited a marked improvement in their MER, showcasing the efficacy of the proposed approach. This improvement is not only statistically significant but also practically relevant, as it translates to better recognition rates and user experiences for Mandarin-English bilingual speakers.

Despite these promising outcomes, the study acknowledges several limitations. One of the primary limitations is the modest size of the Mandarin-English CS dataset. The size and diversity of the dataset are critical factors in the performance of ASR systems. A larger and more diverse dataset could potentially lead to further improvements in model performance. Additionally, constraints on experimental time have influenced the extent of model optimization. More time could allow for more rigorous hyperparameter tuning and additional iterations of model training and evaluation.

Future work is suggested to address these limitations and further refine model performance. Expanding the dataset to include more examples of Mandarin-English CS speech is a priority. This could involve collecting more data from bilingual communities, as well as incorporating a wider range of accents and dialects. Pursuing more rigorous hyperparameter tuning is also recommended, as this could lead to further optimizations in model performance. Furthermore, the exploration of cross-lingual datasets and the development of severity-dependent models are proposed as avenues for future research. Cross-lingual datasets could help the model generalize better across different language pairs, while severity-dependent models could adapt to the varying degrees of CS present in speech. These advancements could foster more equitable and effective ASR systems that are better suited to handle the complexities of multilingual communication.

In conclusion, this pioneering research in Mandarin-English CS speech recognition using the Whisper model sets a benchmark for future exploration and innovation. The findings underscore the importance of adapting ASR systems to the complexities of multilingual communication. By doing so, we can pave the way for more inclusive and responsive technologies that cater to the diverse linguistic landscape of our global community. This research not only contributes to the field of ASR but also has broader implications for the development of technologies that are sensitive to the needs of multilingual speakers. The abstract provided here encapsulates the essence of the research, highlighting its significance, methodology, results, and implications for the field of ASR.

1 Introduction

In recent years, Automatic Speech Recognition (ASR) has seen remarkable advancements, transforming from a niche technology into a ubiquitous tool that powers a multitude of applications, from virtual assistants to transcription services. The development of deep learning techniques and the availability of vast datasets have significantly improved the accuracy and reliability of ASR systems. These systems are now capable of understanding and transcribing speech in a variety of languages with high precision, marking a new era in human-machine interaction.

Parallel to the technological progress in ASR, the linguistic landscape has been evolving due to globalization and the rise of multilingualism. One such phenomenon that has gained prominence is code-switching (CS), which occurs when a speaker alternates between different languages in a conversation. It is a common and informal way of communication by multilingual speakers [1].CS can occur at different levels of language use, and depending on where the language switch happens, there are two primary types of CS. Intra-sentential CS involves switches within an utterance or sentence, while inter-sentential CS occurs at the boundaries of utterances or larger linguistic units [2].The following shows example sentences of inter-sentential and intra-sentential Mandarin/English codeswitching:

- Oh, My God, 我电脑死机了.(The literal translation in English is: “Oh, My God, My computer crashed.)
- 我今天想去 Starbucks 买一杯 ice coffee.(The literal translation in English is: “I, today, am thinking of going to Starbucks to buy a cup of ice coffee”)

Intra-sentential CS poses a greater challenge for ASR systems due to the significant acoustical variations that can occur when languages mix within the confines of a single utterance [3]. Recognizing the complexity and frequency of code-switching, this thesis focuses on tackling the more intricate aspect of CS in ASR: intra-sentential code-switching, particularly in contexts where Mandarin is the dominant language, interspersed with English words. This specific focus is chosen due to the increasing prevalence of such linguistic practices, especially in regions with a strong Mandarin-speaking population but also significant English influence. The decision to concentrate on Mandarin-English code-switching in ASR is driven by several compelling reasons. Firstly, the number of individuals engaging in this form of language use is substantial and continues to grow, reflecting the global impact of both languages. Secondly, it is important to note the presence of proprietary terms or slang in English that do not have direct translations in Mandarin, necessitating accurate ASR to avoid misinterpretation. Enhancing the recognition of Mandarin-English CS can have far-reaching applications across various sectors,

including business, healthcare, and education, thereby making a tangible difference in people's lives.Thirdly, existing ASR models often struggle with the nuances of Mandarin-English code-switching, resulting in lower recognition accuracy. This scenario presents a significant opportunity for improvement, as the correct interpretation of such speech is crucial for maintaining the intended meaning.

The Whisper model [4], developed by OpenAI, represents a groundbreaking advancement in the field of ASR. This model is distinguished by its use of large-scale weak supervision, which leverages an extensive array of audio data, often only partially labeled or even unlabeled. Unlike traditional ASR systems that heavily rely on fully annotated training sets, Whisper ingeniously capitalizes on diverse and readily available internet audio datasets. This approach not only circumvents the labor-intensive process of manual transcription but also enriches the model' s exposure to a wide variety of linguistic phenomena, including code-switching. Built upon the robust framework of the Transformer architecture, the Whisper model employs self-attention mechanisms that allow it to effectively process long-range dependencies within speech. This capability is crucial for handling complex acoustic patterns, particularly those involving intra-sentential code-switching where rapid shifts between languages can occur within single utterances. The model's design enables it to adapt to various acoustic environments and decode speech with remarkable accuracy across multiple languages and dialects. The extensive training on diverse, real-world data also enhances the Whisper model's robustness against background noise and other acoustic variabilities, ensuring reliable performance in everyday applications.

Despite its prowess, Whisper, like many ASR models, encounters limitations when decoding the nuanced dynamics of Mandarin-English code-switching. Recognizing this gap, this research embarks on a pivotal endeavor: to refine the Whisper model through targeted fine-tuning with a dedicated Mandarin-English code-switching dataset. The hypothesis, informed by Xie et al.'s (2023) study [5] that reported a substantial Mix Error Rate (MER) reduction in Cantonese-English code-switching after model adaptation, posits that a similar fine-tuning strategy will significantly lower the Word Error Rate (WER) for Mandarin-English code-switched speech. WER is a commonly used metric in ASR evaluation. It measures the error rate of the sentence predicted by an ASR model compared to a reference sentence, calculating the error rate at a word level. A lower WER indicates a better-performing model that predicts words with minimal errors.

Through meticulous fine-tuning and rigorous evaluation, this thesis aims not only to elevate the Whisper model's performance in mixed-language settings but also to contribute meaningfully to the broader ASR discipline, aligning it more closely with the realities of a diversely multilingual world. By addressing the specific challenges of Mandarin-English code-switching, this work

paves the way for enhanced communication technologies that better serve our interconnected global society.

1.1 Research Questions and Hypothesis

To summarize, this thesis focuses on the following problem and hypothesis:

Question. How does fine-tuning the Whisper model with a Mandarin-English code-switched dataset specifically influence its Mix Error Rate (MER) in recognizing mixed-language speech, compared to its performance with monolingual datasets?

Hypothesis. Based on the findings of Xie et al. (2023) [5], who achieved an impressive MER of 14.28% through fine-tuning the Whisper model for Cantonese-English code-switching, representing a remarkable 35.13% reduction compared to the original Whisper model, we formulate the following refined hypothesis:Fine-tuning the Whisper model with a Mandarin-English code-switched dataset will result in a statistically significant increase in the performance for code-switched speech recognition compared to its pre-fine-tuned state.

1.2 Thesis Outline

Following the brief elucidation of the motivation that propels this research, this thesis is structured in a detailed and systematic arrangement as delineated below:

In the first subsection, the inception of the research takes place with the introduction of the research question and the hypothesis. This part of the thesis aims to provide an academic grounding to the research, familiarizing readers with the intent and focus of the study. The research question is integral as it guides the direction of the study, forming the axis around which the entire research revolves. Furthermore, the hypothesis forms the preliminary and anticipated answer to the research question based on educated guesses and information at hand. The research question and hypothesis will be inspected and deliberated in an exhaustive manner in the succeeding section.

In the second section, the thesis pivots into an extensive literature review pertinent to the research question and hypothesis. It endeavors to survey and synthesize the latest and most relevant research from scholarly articles to define the current state-of-the-art. This deep dive into existing literature provides a robust academic bedrock for the thesis, assisting in clarifying, supporting, or contesting the central argument of the study. It contextualizes the research question and hypothesis within the prevailing discourse and research landscape.

Transitioning into the third section, the thesis shifts its focus on explicating the methodology and experiments adopted during the research process. This section gives a precise description of the procedures undertaken to address the research question, and test the hypothesis, thus promoting transparency and reproducibility of the research. Furthermore, the basis and reasoning behind the choice of any models used within the research are clarified, contributing towards the reliability of the research findings.

In the subsequent fourth section, the thesis encapsulates and summarises the results obtained from the implemented experimental procedures. These results are precisely and objectively contrasted against a predetermined baseline, providing a reference point to judge the relative success or failure of the research. This comparative analysis imbues the study' s findings with greater rigor and reliability.

As the analysis advances to the fifth section, a thorough and comprehensive examination of the aforementioned results unfolds. Here, findings are analyzed meticulously and meticulously interpreted in the light of study objectives, hypothesis, and existing literature. This detailed scrutiny plays an imperative role in solidifying the implications and contributory value of the research findings.

Finally, in the closing section, the thesis recapitulates the entire journey of the research - condensing the study' s essence, recapping the research question and hypothesis, discussing the methodology and experiments, reflecting upon results and their implications. This summary is teamed up with the conclusions drawn, providing a holistic understanding of the research' s outcomes. Additionally, this section also prompts recommendations for future work, thus offering pathways for potential explorations and expansions on the research theme.

2 Literature Review

Automatic Speech Recognition (ASR) software is reshaping the interaction between humans and technology, pioneering a revolution in how machines comprehend human language. Autonomously transforming spoken language into written text, ASR finds myriad niche applications, which can be broadly categorized based on their linguistic capabilities.

These Monolingual applications are renowned for their capability to work wonders with one language exclusively. For example, consider a monolingual application fine-tuned specifically to recognize and process the Mandarin language. Utterances in Mandarin are recognized and processed effortlessly, while utterances in any other language lie outside the application's grasp. Graduating to bilingual applications, these are equipped with dual language prowess, masterfully handling two languages simultaneously. They flexibly recognize and process languages in pairs. To illustrate, a Mandarin-English bilingual application would have the finesse to comprehend and process both Mandarin and English language utterances with equal proficiency. On the more sophisticated end of the spectrum lie multilingual applications. Unflinchingly grappling with three or more languages, such as Mandarin, English, and Cantonese, these applications come equipped with the versatility to recognize and process a wide array of language input.

A stimulating area within bilingual speech recognition is 'code-switching', the practice of alternating between two languages within a single conversation. To shed a brighter light on this intriguing facet, this chapter aspires to furnish readers with an in-depth and nuanced overview of related research and literature from the vantage point of bilingual language recognition.

2.1 DEVELOPMENT OF BILINGUAL ASR SYSTEMS

2.1.1 Traditional ASR

Traditional ASR systems typically consist of multiple components, including signal processing, feature extraction, acoustic modeling, pronunciation modeling, language modeling, and decoding. As shown in Fig.1

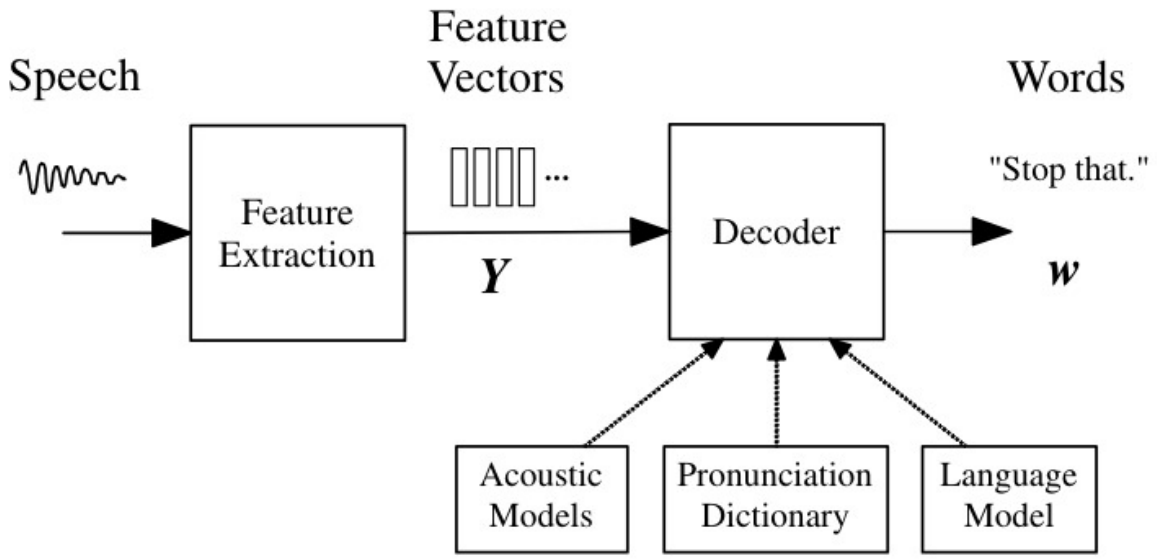

Figure 1: Traditional ASR[6]

In this stage, the raw speech signal is preprocessed to remove noise and irrelevant information. Feature extraction techniques such as Mel-frequency cepstral coefficients (MFCCs) [7], perceptual linear prediction (PLP) [8] or filter bank energies are then applied to convert the speech signal into a feature representation suitable for modeling. Acoustic modeling involves the statistical representation of speech sounds. Gaussian Mixture Models (GMMs) were traditionally used to model the relationship between acoustic features and phonemes or subword units. Hidden Markov Models (HMMs) are often employed to capture temporal dependencies and model the sequential nature of speech. Pronunciation modeling component deals with modeling the pronunciation variations of words. Techniques such as decision trees or finite-state transducers are used to model the mapping between graphemes or phonemes and their corresponding acoustic representations. In the field of language modeling, the traditional statistical language model N-gram [9] is the most commonly used, although N-gram has a simple structure and high training efficiency, but the model does not have the required semantic connection, and the parameters will grow exponentially with the increase of the dataset, which can not get better performance, so it can not have a significant breakthrough. Later, some shallow neural networks were used in language modeling, and the performance was improved, but it can only handle a certain length of historical information, and the prediction of the current word can only be associated with the previous N-1 words, which affects the accuracy of the model results. Decoding involves searching through the space of possible word sequences to find the most likely transcription given the acoustic input and language model. Dynamic programming algorithms

such as Viterbi decoding [10] or beam search [11] are often employed for this purpose.

2.1.2 DNN-BASED ASR

In the late 1980's, the appearance of ANN marked a new direction in speech recognition research, but the effect of shallow neural network is general, until 2006 Hinton [12] proposed Deep Belief Network (DBN) to initialize the nodes in the neural network, so that the problem of falling into the local optimum in the training process has been solved, and proved that the deep neural network structure has better effect in feature extraction and model training, since then deep learning based speech recognition has been widely studied and concerned. In 2009, Hinton and Mohamed [13] introduced DBN into acoustic modeling and achieved success on the TIMIT small dataset. 2011, Deep Neural Network (DNN) achieved success on the large vocabulary dataset [14] , marking the mainstream research direction of deep neural network-based speech recognition. When dealing with more complex signals, it is more suitable to use networks with deep model structure, because deep networks have stronger expressive ability through multi-layer nonlinear transformations [15]. In the field of acoustic modeling, DNN can use features composed of spliced adjacent speech frames and the fusion of multiple features as inputs to give the model longer structural information, and at the same time, using DNN does not need to assume the distribution of speech signals when estimating the posterior probability distribution, but the DNN structure can only learn the mapping relationship of a fixed length, which makes its modeling inflexible, and the deep network that is popular at the beginning is the feed-forward The most popular deep network is Feed-forward Deep Neural Network (FDNN). The emergence of Recurrent Neural Network (RNN) [16] gradually replaced DNN, which added a feedback connection layer on the hidden layer, and used the output of the previous layer and the output of the hidden layer as the current input, so that the RNN can see the previous history information, with a memory function and a longer modeling ability, using this feature of the RNN is more suitable for modeling speech signals, but it is not flexible enough in the modeling of speech signals. This feature of RNN is more suitable for modeling speech signals, but the gradient disappearance it produces during the training process makes the model training very difficult, so Long-short Term Memory (LSTM) RNN [17] has been proposed, and LSTM-RNN uses various types of gates to control the information, including input gates, output gates, and oblivion gates, which can make full use of the history of the future information, and it has a better performance than the unidirectional LSTM. LSTM with better performance than unidirectional LSTM, while bidirectional LSTM-RNN is not suitable for real-time systems due to the use of future information, which generates a large delay when used. Based on the above problems, Feed-forward Sequential Memory Network (FSMN) is developed by KUDA Xunfei [18] , which combines the features of DNN and RNN, and adds a module

with memory on the hidden layer to store the historical future information that is useful for the current speech frame, and the length can be adjusted according to the actual demand. The length can be adjusted according to the actual demand FSMN solves the problem of shaving vanishing while controlling the delay. The convolutional operation of the convolutional layer is the core of Convolutional Neural Network (CNN) [19] , which is also another model that can be trained using contextual information, and the convolutional operation is utilized to extract the features used for training, which can easily deal with high-dimensional data, but with the deepening of the network hierarchy, the gradient descent is used to make the training results gradually converge to the local minimum, resulting in certain errors. local minima, causing some errors, and the pooling layer also loses some useful information.

2.1.3 End-to-End ASR

The end-to-end (E2E) approach to speech recognition, which has emerged in recent years, has greatly simplified the model structure and training process of speech recognition. It has become one of the most mainstream directions in the research of speech recognition. The E2E model can directly obtain the mapping relationship between the input audio features and the output text content through a neural network model, without the need to individually train the individual modules that make up the system. This makes the process of speech recognition simpler and more efficient, and it also enhances the interaction between human and machine. The relationship between the input audio features and the output text content is established through a neural network model, obviating the necessity for the individual modules that comprise the system to be individually trained. This simplifies the process of speech recognition and facilitates a more harmonious interaction between human and machine. Three principal methodologies have been developed for this purpose: the Connected Temporal Classifier (CTC) [20, 21, 22], the Attention-based Model (AM) [23], and the Recurrent Neural Network Transducer (RNN-T). All three methods have been demonstrated to be effective in aligning input sequences with output sequences, obviating the necessity for the traditional manual labeling and forced alignment of input sequences. These three methodologies are described in detail in the following sections. In 2006, Graves et al. [24] proposed the use of the CTC algorithm to address the issue of unequal lengths of input and output sequences. This is a loss function computation method that effectively realizes the automatic alignment between speech signals and text sequences by adding a blank label to indicate the absence of output content and splitting the adjacent characters. Since then, CTC has remained a prominent approach in the field of speech recognition. In 2015, Baidu proposed the Deep Speech 2 [25] system based on CTC, which can be used to achieve end-to-end speech recognition for Chinese and English, respectively. The model trained on the Switchboard dataset demonstrated a reduction in the

word error rate of 16.5%. However, the CTC model suffers from the conditional independence assumption problem and needs to be paired with a deep acoustic model as a coding layer. Attention-based models were initially proposed in the field of machine translation and subsequently introduced into speech recognition tasks with the objective of learning the alignment between speech data and text sequences. These models comprise an encoder and a decoder, which accept the input sequences and output the hidden state to the decoder, which is then used to output the predicted sequences. Chorowski et al. [26] employed the Seq2seq model for speech recognition, marking the first instance of its use in conjunction with an attention-based model. For the first time, an attentiona mechanism was incorporated into a model. The Listen, Attend and Spell (LAS) model structure was proposed by Chan et al. [27] and demonstrated effective performance in large vocabulary conversational speech recognition. This formalized the basic architecture of the LAS model. The LAS model can remove redundant information, and the nonlinear features can be better applied to the deep network. Subsequently, the attention mechanism model was combined with CTC [28]. The attention model is based on the characteristics of the CTC loss function, which can improve the convergence speed and enhance the robustness of the model simultaneously. This reduces the character error rate by 5% to 10% on the CSJ database. The success of the Transformer model structure based on the self-attention mechanism in the field of machine translation has prompted researchers in the field of speech recognition to explore the potential of this structure for improving speech recognition [29]. One such approach is the encoder block processing proposed by Tsunoo et al., which corresponds to the inheritance mechanism of Transformer that can sense the context and utilize global acoustic features and semantic environment. The Transformer-XL [30], proposed by Dai et al., incorporates positional coding based on the Transformer structure to address the context dependency of long speech signals. The Conformer structure, proposed by Gulati et al [31]., combines the advantages of the Transformer's convolutional pooling of local features and the extraction of sufficiently long histories of future information in LibriSpeech with the advantages of the Transformer's convolutional pooling of local features. The proposed method obtains a word error rate of 2.1% to 4.3% on the LibriSpeech dataset. Chui et al [32]. proposed Monotonic Chunkwise Attention MoChA and optimized it using planned sampling, label smoothing, and simultaneous training, which somewhat alleviates the monotonic constraints, effectively narrowing the performance gap between streaming and offline recognition. The Monotonic Multihead Attention mechanism (MMA) [33] was proposed as a solution to the streaming decoding problem of the Transformer. However, the input required for model recognition must be a complete sentence, and there is a delay in the decoding process. This structure of streaming decoding will inevitably deteriorate the model performance. The RNN-T model was proposed by Graves et al [34]. to address the shortcomings of the CTC model, which is also

a loss function calculation method similar to CTC. It incorporates semantic features of text during training, integrates acoustic and linguistic feature information, and optimizes the model structure collectively. Rao et al. proposed the CTC multilevel pre-training mechanism to overcome the challenges associated with RNN-T training. This involved incorporating CTC in the pre-training of the encoding grid. The CTC loss function is incorporated into the pre-training of the coding grid, thereby accelerating the model's convergence. Zhang et al. [35] employ the Transformer in lieu of the RNN structure in RNN-Transducer, enabling the Transducer model to parallelize the computation, enhance its computational efficiency, and achieve a word error rate of 2.4% to 5.6% on the LibriSpeech dataset. The RNN-Transducer model achieved a 5.6% word error rate on the LibriSpeech dataset. Wu et al. proposed integrating dialect, accent, or locale information into the RNN-Transducer model as a signal input to improve the overall recognition accuracy of the model. The RNN-Transducer model is capable of independently modeling context through the use of regularization algorithms. This model is particularly advantageous in online recognition due to its ability to simultaneously consider both acoustic features and linguistic information. However, it should be noted that the model is only capable of considering these two types of information simultaneously. The RNN-Transducer model is capable of considering both acoustic and linguistic information, which is particularly advantageous in online recognition. However, it employs a single objective function for the joint optimization of acoustic and linguistic information, which makes training challenging. In the next phase of this chapter, the precise architectural configuration of the present E2E model will be elucidated through the use of Whisper, the central model in this research, as an illustrative example.

2.2 WHISPER

The Whisper model, as mentioned above, specifically for its prowess in large-scale supervised pre-training and its adaptability across numerous languages. This section provides an in-depth examination of Whisper's architecture, focusing on its attention mechanism and the transformative role of Transformers, before synthesizing these elements into an overview of the complete model architecture.

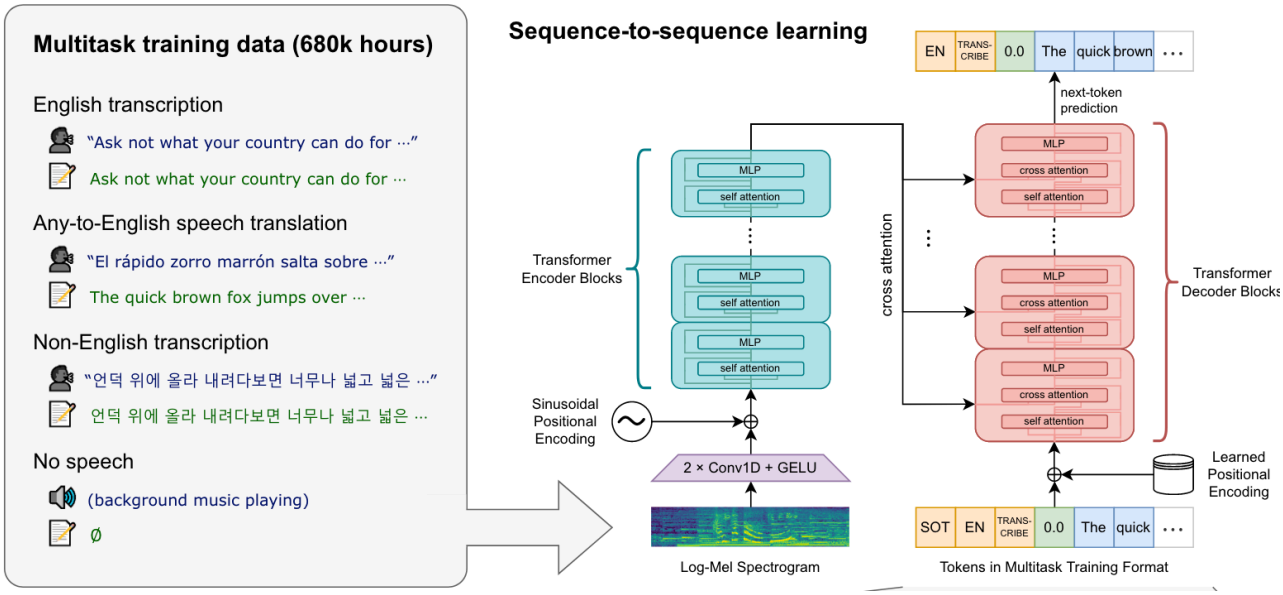

Figure 2: Whisper[4]

2.2.1 Transformer Speech Recognition System

In recent years, encoder-decoder models based on the attention mechanism have made great breakthroughs in sequence-to-sequence tasks such as speech recognition and machine translation [36, 27]. Before the introduction of the Transformer, these models were mainly based on RNN networks, which usually calculated along the order of symbol positions in the input and output sequences. They aligned the position with the recursive steps during computation, generating a series of hidden states h_t based on the previous hidden state h_{t-1} and the input at position t. The dependence of RNN networks on the sequence order leads to the inability to train samples in parallel, but parallelization is essential for improving training efficiency. Recent work has significantly improved computational efficiency through factorization tricks [37] and conditional computation [38], the latter also improving model performance. However, the basic constraint of sequential computation in RNNs still exists. At the same time, the attention mechanism has become an indispensable part of sequence modeling and transformation models in various tasks, as it allows modeling dependencies without considering their distance in the input or output sequence [36, 39]. However, in almost all cases, such attention mechanisms are used in conjunction with RNNs.

In order to overcome the above problems in modeling time series, Vaswani first proposed an end-to-end model based on self-attention, the Transformer [40]. It is a model structure that avoids loops and fully utilizes the self-attention mechanism to model the global dependency relationship between inputs and outputs. Compared with other RNN-based methods, it greatly reduces training costs and has set the best results at the time in the WMT2014 English to French

translation task. Through the multi-head self-attention mechanism (Multi-head Attention, MHA), the Transformer can effectively capture global dependencies within sequences [41], learn direct dependencies between long-distance modeling units [42], and also support model parallel training [40]. This makes the performance of Transformer in speech recognition tasks far exceed that of other traditional HMM speech recognition models and end-to-end models [43, 44].

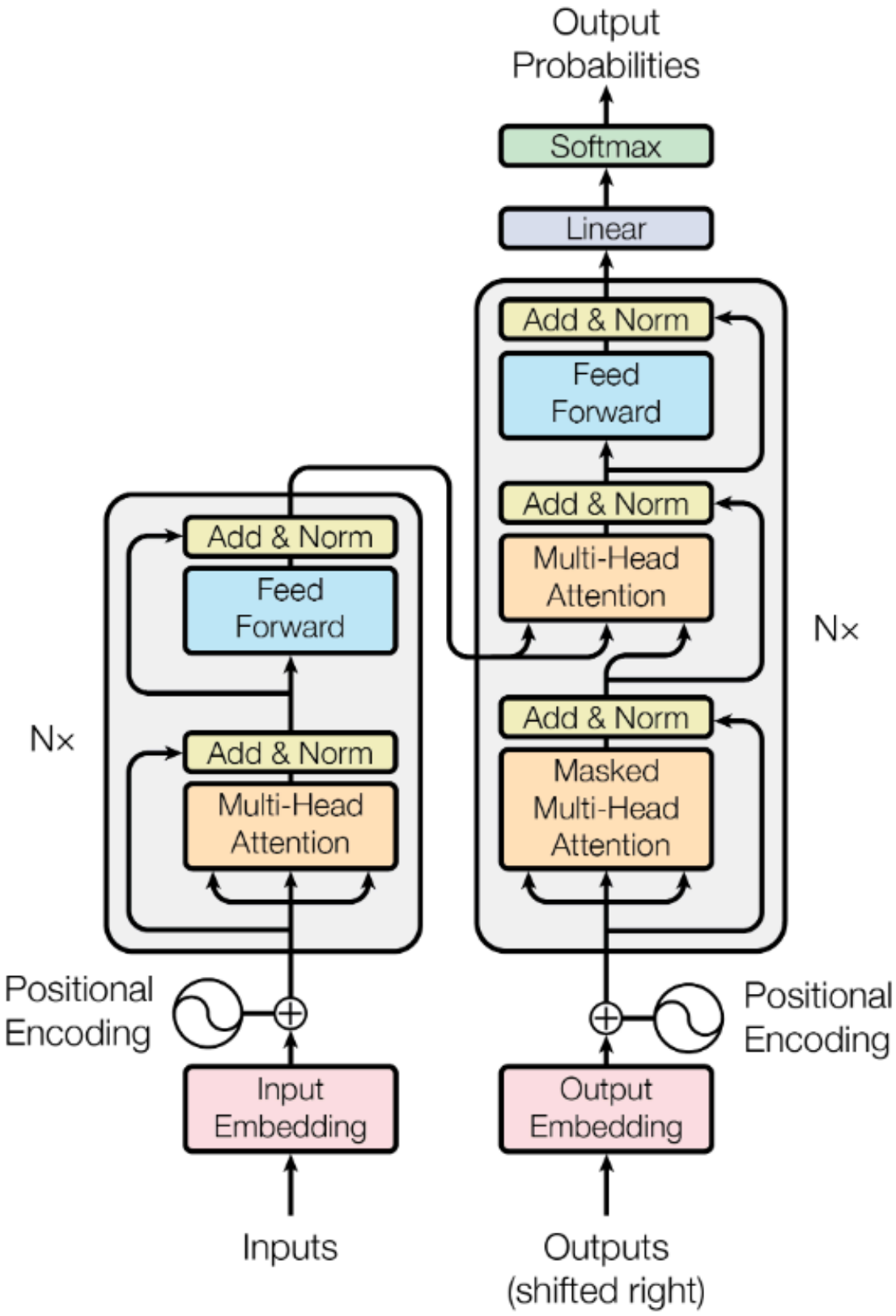

Figure 3: The Transformer - model architecture.[40]

2.2.2 Encoder and Decoder

The model structure of Transformer applied in speech recognition is shown in Figure 3. The encoder maps the input sequence $(x_1,\ldots,x_n)$ to a series of feature representation vectors $h = (h_1,\ldots,h_n)$. The decoder is responsible for generating the output symbol sequence $(y_1,\ldots,y_m)$ one by one through decoding h. Each step of the model is based on an auto regressive approach [45]: when generating the next, the previously generated symbols are used as additional input. Compared with DNN-HMM-based systems, the Transformer speech recognition system does not rely on any conditional independence assumptions and directly estimates the posterior probability $P(C|X)$ based on the chain rule in probability theory.

$$P(C_i \mid c_1,\ldots,c_{i-1},X) = \mathrm{Patt}(C \mid X)$$

$$C = \{c_i \in \mathcal{U} \mid i = 1,\ldots,L\}$$

, where

$$\mathcal{U}$$

is a set of non-repeating characters with L elements, serving as the system's minimum modeling unit set. $P_{att}(C|X)$ is the objective function based on the Attention model.

$$h_t = Encoder(X)$$

$$P(c_i|c_1,\ldots,c_{i-1},X) = Decoder(h_t,c_1,\ldots,c_{i-1})$$

The above equations represent the computation process of the encoder and decoder networks, respectively.

Encoder: The encoder consists of N independent encoder layers, each containing two sub-layers. To reduce the length difference between the audio feature sequence and the label sequence, and to save computational resources, the audio features need to undergo down sampling before being input into the encoder, and then position encoding is added to join the temporal information of the feature sequence. The first sub-layer of the encoder is an MHA mechanism, and the second layer is a position-related fully connected feed-forward network. For these two sub-layers, each is equipped with residual connections (RC) [46] and layer normalization (LN) [47]. Therefore, the output of each sub-layer can be represented as $\mathrm{LayerNorm}(x+\mathrm{Sublayer}(x))$, where $\mathrm{Sublayer}(*)$ represents the function of the sub-layer. The output vector dimension of all sub-layers and the word embedding layer in the model is d_{model}.

Decoder: The decoder is also composed of N independent decoder layers. To compute the input sequence, the decoder first calculates the word vector of each label in the sequence using the word vector layer. In addition to the self-attention sub-layer identical to that in the encoder,

the third MHA sub-layer (referred to as the source-target attention sub-layer in the following text) is responsible for connecting the decoder with the encoder output. Like the encoder, each sub-layer also has the corresponding RC and LN. In addition, a mask is added to the computation of the self-attention sub-layer in the decoder. According to the last equation, to ensure that the decoder output is the next token c_i given the current token sequence $c_1, \ldots, c_{i-1}$ during decoding, a special token "SOS" (indicating the start of the current sentence) is added to the left side of the input sequence, causing the input sequence to shift one position to the right. Also, to ensure that the length of the decoder input and output sequence is consistent for the calculation of the loss function, and to facilitate the determination of the end position during decoding, a special termination token "EOS" (representing the end of the current word) is added to the right side of the given supervised label sequence. Based on the model's output, there is a token offset relative to the input position, and the mask in the decoder ensures that the prediction of the position will only depend on the content of the known output that is less than the position.

2.2.3 Attention mechanism

The key mechanism of attention can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key.The image (the left in Figure 4) describes the Scaled Dot-Product Attention and the Multi-Head Attention mechanisms in the Transformer model.

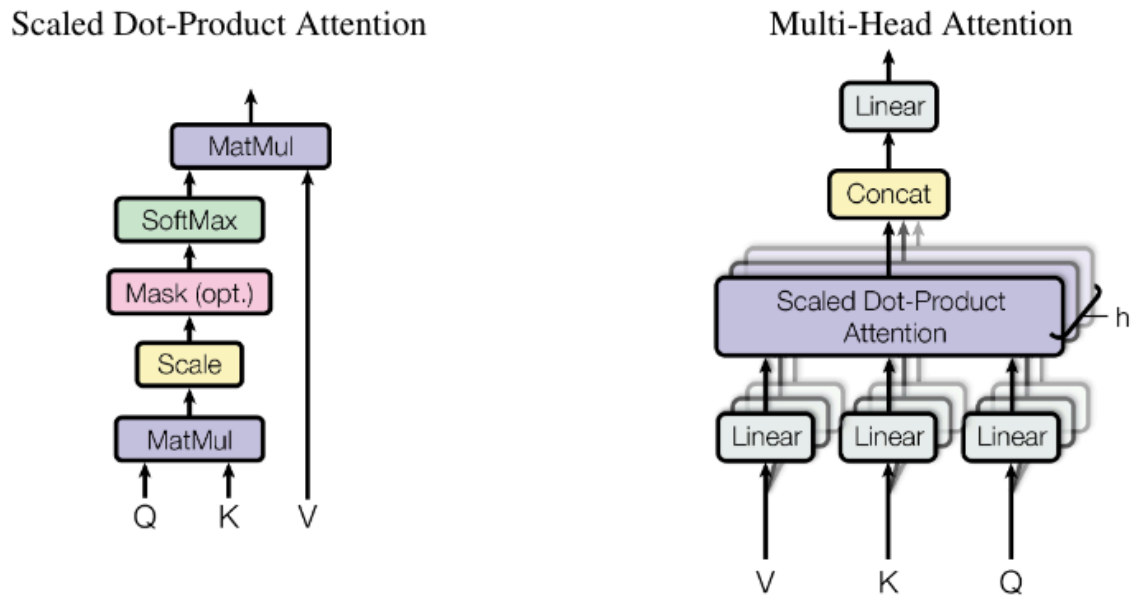

Figure 4: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.[40]

The attention function can be described using a dot-product of the Query and Key vectors, and then dividing each by the square root of the dimension of the keys to mitigate the impact of large dot products which can push the softmax function into regions where it has extremely small gradients.

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

Multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions. This is done by projecting the queries, keys, and values h times with different, learned linear projections to d_k, d_k, and d_v dimensions, respectively. Then the attention function is applied to each of these projected versions of queries, keys, and values, and the results are concatenated and once again projected, resulting in the final values.

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,\ldots,\text{head}_h)W^O$$

$$\text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

2.2.4 Feed-Forward Neural Networks

After the attention layer, a fully connected feed-forward network is applied to each position separately and identically. This consists of two linear transformations with a ReLU activation in between (equations 3-4).

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

2.2.5 Positional Encoding

Since the Transformer contains no recurrence and no convolution, in order for the model to make use of the order of the sequence, "positional encodings" are added to the input embeddings at the bottoms of the encoder and decoder stacks. The positional encodings have the same dimension d_{model} as the embeddings, so that the two can be summed. They use sine and cosine functions of different frequencies:

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}})$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{model}})$$

2.2.6 The Construct of the Whisper Model Architecture

The Whisper model architecture, which synthesizes the critical elements mentioned above, is deliberately structured to optimize outcomes in multilingual and code-switched ASR tasks. The model processing begins with the input speech being resampled to a frequency of 16,000 Hz and subsequently converted into an 80-channel log-Mel spectrogram - the standard feature representation in contemporary ASR systems. The initial spectrogram then passes through two convolutional layers with a GELU activation function to fine-tune the representation. Subsequent to these convolutional layers, sinusoidal position embeddings are added, preparing the spectrogram data for input to the Transformer encoder. The encoder, fortified with the aforementioned pre-activation residual blocks and layer normalization mechanisms, processes the spectrogram through multiple Transformer blocks. Mirroring the specifications of the encoder in terms of scale and block count, the decoder incorporates learned position embeddings and employs tied input-output token representations for efficiency. This strategic design choice fosters an efficient information flow from acoustic input to textual output. Furthermore, the model utilizes a shared lexicon based on a byte-level Byte Pair Encoding (BPE) tokenizer. Initially developed for GPT-2 , the BPE tokenizer has been readapted for multilingual usage to reduce language-specific fragmentation. Figure 1 encapsulates the details of this sophisticated

architecture, illustrating the interaction between each component to provide a holistic picture of Whisper's exceptional performance in transcribing speech irrespective of language-switching scenarios.

2.3 CODE-SWITCHING

Code switching (CS) is a linguistic phenomenon where speakers fluidly alternate between two or more languages or dialects within a single conversation. This practice often reflects the speaker's identity, social context, and communicative goals. In the realm of Automatic Speech Recognition (ASR), CS poses a formidable challenge due to the intricate nature of recognizing and interpreting multilingual speech patterns. The complexity lies in capturing the nuances of various languages, managing language shifts seamlessly, and ensuring accurate transcription and translation of the spoken content.

Myers-Scotton's Matrix Language-Frame (MLF) Model, introduced by Namba (2004) [48], presents a structured framework for analyzing CS in bilingual speech. It emphasizes the notion of a "matrix language," representing the dominant language in a conversation, and an "embedded language," which consists of language elements borrowed from the non-dominant language. By delineating these roles within the communication process, the MLF model helps elucidate the mechanisms underlying CS patterns and language alternation strategies used by speakers.

Despite ongoing advancements in ASR technology, current systems encounter significant hurdles when processing CS. Factors such as imbalances in language ratios, phonetic conflicts arising from diverse linguistic structures, and the scarcity of annotated CS datasets impede the accurate transcription and translation of multilingual speech. Mandarin-English CS, in particular, introduces distinct challenges including discrepancies in tonal systems, variations in word order conventions, and lexical disparities that necessitate tailored adaptations within ASR models to achieve optimal performance.

Recent research endeavors in ASR have focused on leveraging advanced methodologies such as deep learning models, multi-task learning approaches, and hybrid systems that integrate sophisticated language models with acoustic features. These approaches aim to enhance the recognition capabilities of ASR systems when confronted with the complexities and intricacies of CS scenarios, including Mandarin-English code-switching.

In conclusion, a holistic comprehension of CS definitions, conceptual frameworks, and inherent challenges is paramount for the progression of ASR research, especially in relation to Mandarin-English code-switching. Future investigations should center on the development of robust ASR models capable of accommodating the intricate nature of multilingual speech patterns. By

addressing these challenges and refining ASR technology, researchers can facilitate accurate transcription and effective translation services across diverse linguistic contexts, thus heralding advancements in the field of automated speech processing.

3 Methodology and Experiments

This chapter provides a methodological overview of the experiments dedicated to testing the hypothesis. The mixed Chinese and English speech recognition experiments in this paper are mainly derived from TAL's audio dataset of adult mixed Chinese and English lectures.This dataset contains a mixture of Chinese and English speech, with 200 speakers.The audio data is in mainstream 16kHz, 16-bit WAV format, the whole corpus includes training, development and test sets, due to GPU and time constraints, this thesis only uses the development and test sets, detailed information is shown in Table 1.

3.1 Methodology

3.1.1 Datasets

A dataset for speech recognition usually consists of recorded speech and its corresponding transcription. The recorded speech can be in the form of readings or spontaneous conversations, but it usually takes thousands of hours of data to build a practical speech recognition system. There are several publicly available hybrid speech recognition datasets in English and Chinese as follows.

- SEAME [49] is a mixed Mandarin-English conversation dataset from respondents in Singapore and the Malay system, with Mandarin as the dominant voice. It contains 112 hours of speech data from 154 speakers.
- The HKUST Mandarin-English dataset [50] is also a conversational speech dataset for interviews and contains 5 hours of labeled audio data and 15 hours of unlabeled audio data.
- CECOS [51] contains 12 hours of mixed Mandarin-English audio data from 77 speakers.
- OC16-CE80 [52] contains 80 hours of voice recordings from more than 1,400 speakers in China, with Chinese as the primary language and at least one or more English words in each sentence.

3.1.2 Approaches[1]

All datasets used are open source datasets.

[1]https://ai.100tal.com/openData/voice

	train set	test set
number of audios	15,000	5,000
percentage of Mandarin	47.90%	48.08%
percentage of English	52.10%	51.92%

Table 1

3.1.3 Model Selection

In the realm of Whisper experiments, two models were chosen: Whisper-small and Whisper-large-v3. Despite the Whisper-large model demonstrating superior performance over its smaller counterpart, the Whisper-small model was selected for its faster processing capabilities.

3.1.4 Evaluation - Mixture Error Rate

Given the distinct metrics for Mandarin and English speech recognition, the Word Error Rate (WER) is the most common method for evaluating English.The word sequence hypothesised by the ASR system is aligned with a reference transcription, and the number of errors is computed as the sum of substitutions (S), insertions (I), and deletions (D). If there are N total words in the reference transcription, then the word error rate WER is computed as follows [53]:

$$WER = (I + D + S)/N \times 100. \tag{1}$$

While the Character-level Error Analysis, Character Error Rate (CER) [54] is prevalent for Mandarin. To comprehensively assess the recognition performance across both languages, this paper introduces the Mixture Error Rate (MER). This metric involves calculating the WER and CER for the respective English and Mandarin segments and then averaging these values to obtain the MER, then the MER is computed as follows:

Furthermore, it was observed during practical experimentation that the small model exhibited difficulties in differentiating between traditional and simplified Chinese characters, which impacted the CER. In the context of Mandarin, both traditional and simplified scripts convey the same meaning, differing only in their written form. Since the focus of this paper is on speech recognition, the distinction between the two should not affect the MER. Consequently, for the small model's experiments, the author first standardized the recognition results into simplified Chinese before calculating the MER.

3.2 Experiments

To empirically investigate the efficacy of fine-tuning the Whisper model with a Mandarin-English code-switched dataset, we conducted a series of experiments. The objective was to evaluate the impact of this fine-tuning on the model's MER in recognizing mixed-language speech, particularly in intra-sentential code-switching scenarios. The author first implemented Mandarin-English CS speech recognition with Whisper-small and large-v3 to investigate how original Whisper works toward CS dataset as baseline. In order to further investigate a better fine-tuning training strategy, the author also fine-tuned both small and large-v3 models on the dataset. The second experiment aims to improve the models' performances. The details of the motivation are discussed in the next chapter.

3.2.1 Data Preprocessing

For fine-tuning, the author employed a dedicated Mandarin-English code-switched publicly available dataset, carefully curated to reflect the nuances and complexities of intra-sentential code-switching prevalent in Mandarin-dominant contexts infused with English terms. Prior to training, the author preprocessed the dataset to prepare it for ingestion by the Whisper model. This involved encoding audio samples into log-Mel features, which serve as input representations for the model. Additionally, we tokenized the corresponding transcriptions using the Whisper tokenizer, converting them into sequences of label IDs compatible with the model's architecture. Notably, we resampled audio data to ensure consistency in sampling rates across the dataset, a crucial step for model training and evaluation.

3.2.2 Setup

The experiments were conducted on the HPC cluster of the University of Groningen. The GPU used is an Nvidia V100 GPU accelerator card with 32 GB of VRAM available. Testing the small model experiments takes around 8 hours. For Whisper-large, testing took X hours to complete.

3.2.3 Training Procedure

The training procedure for fine-tuning the Whisper model on the Mandarin-English code-switched dataset was meticulously designed to enhance the model's ability to recognize speech in mixed-language contexts. Here's an overview of the steps taken:

We began by initializing the Whisper-small and Whisper-large-v3 models using their respective

configurations provided by the Hugging Face library [2]. As mentioned earlier, audio samples were converted into log-Mel spectrogram features, which are commonly used in speech recognition tasks due to their ability to capture the spectral characteristics of speech. The transcriptions were tokenized using the Whisper tokenizer to align with the model's input requirements.Then we utilized a cross-entropy loss function, which is standard for classification tasks, to measure the discrepancy between the predicted output and the true labels during training. The AdamW optimizer was chosen for its effectiveness in training deep learning models, with a learning rate schedule that decays the learning rate over time to help converge to a good solution. To train with larger batches without exceeding GPU memory limits, we employed gradient accumulation, which involves accumulating gradients from multiple mini-batches before performing a single update to the model's weights. To speed up training and reduce memory usage, we utilized mixed-precision training with NVIDIA's Apex library, which allows for faster computation with half-precision floats while maintaining model accuracy. Dropout and weight decay were applied to the model to prevent overfitting to the training data. We trained the model for a fixed number of epochs, with periodic evaluations on a separate validation set to monitor for overfitting and to adjust hyperparameters as necessary. Model checkpoints were saved at regular intervals, allowing us to resume training or to select the best-performing model based on validation performance. With the optimized settings, we conducted the final training run, ensuring that the model was thoroughly trained and ready for deployment.

3.3 Ethical issues

While the primary goal of the study is to develop an Automatic Speech Recognition (ASR) system that will be beneficial for Mandarin-English code-switching, it is crucial to acknowledge that any technological advancement may come with unforeseen consequences. To address and mitigate potential risks, the research team is committed to communicating the study' s results and implications in a manner that is both accessible and transparent to all stakeholders.

To ensure ethical considerations are met, I will not be collecting any sort of data from human participants. Instead, all the audio recordings used in the study are sourced from open access repositories. These recordings have been validated and vetted by the community, ensuring their authenticity and relevance to the research. The corpus utilized in the study is licensed under the Creative Commons Zero (CC0) license, which permits any form of distribution, adaptation, or modification without the need for attribution or acknowledgment. This open licensing approach fosters a collaborative environment and encourages the free use of the data for research purposes.

[2]https://huggingface.co/blog/en/fine-tune-whisper

In the evaluation of the ASR models developed, I will not involve human participants. The decision is based on the fact that objective metrics, which are more aligned with the field' s standards and requirements, will be employed for assessment. The use of subjective evaluation methods involving human participants is deemed less meaningful in this context, thus eliminating concerns regarding the ethics of participant involvement.

Furthermore, there are no ethical concerns regarding the involvement of human participants or any issues that conflict with the faculty' s ethical guidelines. Should there be a scenario where data from human participants becomes necessary, I will undertake the requisite steps to secure approvals from the ethics committee, ensuring full compliance with ethical standards.

In terms of replicating the research, all the code used in the study will be made publicly available via GitHub [3]. Detailed instructions on how to reproduce the experiments described in the proposal can be found in the Methodology section of the study. The dataset is also publicly accessible for download and use, facilitating other researchers in the field to verify the findings and build upon the work.

It is important to note that while the outcomes should be largely similar, they may not be identical due to certain elements that introduce randomness in the trained models. Variations in the hardware used can also impact the performance of the models. The experiments in this study will be conducted on the university' s high-performance computing cluster, which may offer different computational capabilities compared to other environments. This aspect should be taken into account when attempting to replicate the study, as it could influence the results.

In summary, the study is designed with a strong emphasis on ethical considerations, transparency, and replicability. By utilizing open-source data, avoiding human participant involvement, and sharing all research materials and methodologies, the study aims to contribute to the field of Mandarin-English code-switched ASR in a responsible and collaborative manner.

3.4 Summary

By following this rigorous training procedure, we aimed to fine-tune the Whisper model to achieve the best possible performance on the Mandarin-English code-switched speech recognition task.

[3] https://github.com/LinDeng134340/Mandarin-English-CS-using-Whisper.git

4 Results

This chapter is dedicated to presenting and discussing the findings derived from the comprehensive set of tests that were conducted as part of the research. The primary objective of this section is to provide a detailed account of the outcomes, which are crucial for evaluating the effectiveness of the fine-tuning process on the Whisper models for Mandarin-English code-switched speech recognition.

4.1 Baseline Experiments

The Whisper series models served as the subjects of the experimental procedure. These models were first subjected to baseline experiments to provide an initial performance measure.

An example of one such model is the Whisper-small model. Designed with a compact architecture, it yielded a Match Error Rate (MER) of approximately 66.80%. Even though the number overshoots the preferred target of 50%, it indicates the starting limitations or initial performance capacity of the model in its raw state. It ' s important to note that such performance levels are not unusual for baseline models given they represent the unoptimized state of the model.

Contrastingly, the Whisper-large-v3 model, equipped with a comparatively extensive architectural design, showed an improvement with an MER of 55.14%. Although the figure still surpasses the desired 50% mark, it is intrinsically an enhanced performance when compared with the Whisper-small model. This suggests that larger models have an inherent higher accuracy potential but are still in need of certain optimization strategies to decrease their MER.

4.2 Fine-tuning Phase

To leverage the larger models' inherent accuracy potentials, a fine-tuning phase was devised. This phase was conceived to improve both models' performances with the execution of targeted modifications.

After undergoing the fine-tuning process, the ME-Whisper-small model demonstrated a notable decrease in its MER, which was measured at around 45.52%. This drop in the error rate represents a considerable improvement from its originally established baseline performance. Essentially, this suggests that the application of proper optimization strategies has the potential to greatly enhance a model's performance.

In a similar vein, the ME-Whisper-large-v3 model, which previously outperformed in the baseline performance measurement, managed to further decrease its MER to 39.07% after under-

Experiment	Model	Test MER
Baseline experiment	Whisper-small	66.80%
Baseline experiment	Whisper-large-v3	55.14%
Fine-tune experiment	ME-Whisper-small	45.52%
Fine-tune experiment	ME-Whisper-large-v3	39.07%

Table 2: Results of the baseline compared to each experiment on the test sets.

going fine-tuning. This improvement serves to signify the strength of the fine-tuning process in augmenting model performances. It also illuminates the adaptability of the model concerning the application of adjustments aimed at boosting its accuracy.

4.3 Comparative Analysis - Table 2

Table 2 succinctly puts forth a comprehensive comparative analysis of the models' performances at the baseline level versus the fine-tuning level. The generated data clearly highlights a striking decrease in the MER values for both models. This attests to the fine-tuning phase' s indispensable role in enhancing model performance. Moreover, the ME-Whisper-large-v3 model emerged as the most effective among the tested models after the fine-tuning phase. With the lowest MER, it demonstrates the benefits of coupling larger model architecture with strategic fine-tuning techniques.

5 Discussion

The in-depth scrutiny of the outcomes delineated in table 2 from the antecedent section unmistakably affirms the alignment of the fine-tuning process with the hypothesis expounded in Subsection 1.1. The empirical findings resoundingly establish that fine-tuning the Whisper models using a Mandarin-English code-switched dataset unequivocally engenders a discernible uplift in performance for executing code-switched speech recognition tasks. The essence of the results encapsulates a successful validation of the hypothesis, elucidating the potency and efficacy of orchestrating fine-tuning methodologies to finely calibrate the Whisper models for optimal performance within the realm of Mandarin-English code-switched speech recognition.

5.1 Validation of the Hypothesis

The initial baseline experiments served as the cornerstone for establishing a comparative benchmark between the Whisper-small and Whisper-large-v3 models, with the latter already showcasing a notable performance edge. The hypothesis positing that fine-tuning would serve as a catalyst for bolstering the models' capabilities garnered validation through a meticulous series of experiments. Post fine-tuning, the ME-Whisper-small model exhibited a remarkable reduction in Mixture Error Rate (MER) to 45.52%, marking a substantial enhancement from its baseline MER of 66.80%. Moreover, the ME-Whisper-large-v3 model displayed an even more impressive performance leap, achieving a MER of 39.07% post fine-tuning, down significantly from the initial 55.14%. These outcomes emphatically underscore the pronounced efficacy of fine-tuning methodologies in not only enhancing but also customizing the models to adeptly navigate the distinctive intricacies embedded within code-switched speech patterns.

5.2 Limitations

The limitations of the study, as previously outlined, are significant and warrant an expanded discussion to fully understand their implications on the research and its outcomes.

5.2.1 Extent of the Dataset

The primary limitation stems from the size and diversity of the Mandarin-English code-switched dataset used for fine-tuning the Whisper models. The dataset, while providing a solid foundation for the initial experiments, is somewhat limited in scope. This limitation is twofold:

Size of the Dataset: The number of samples available for training is a critical factor in the performance of machine learning models. A larger dataset typically allows models to learn more complex patterns and generalize better to unseen data. The restricted size of the current

dataset may have restricted the models' ability to capture the full spectrum of Mandarin-English code-switching nuances.

Variety of the Dataset: In addition to size, the diversity of the dataset is equally important. The current dataset may not fully represent the wide range of accents, speaking styles, and linguistic variations present in natural Mandarin-English code-switched speech. This lack of diversity could lead to a model that performs well on the training data but fails to generalize to new, unseen examples.

5.2.2 Temporal Restrictions

The second major limitation is the time constraints faced during the experimental phase. These constraints affected the research in several ways:

Data Collection: Time limitations may have restricted the ability to collect and incorporate a larger and more diverse set of data. A more extensive data collection process could have been beneficial in enhancing the dataset's representativeness.

Experimentation: Time constraints could have limited the number of experiments that could be conducted. More experiments would have allowed for a more thorough exploration of the fine-tuning process and potentially uncovered additional insights into the models' performance.

Hyperparameter Tuning: The process of optimizing hyperparameters is time-consuming but can significantly impact model performance. Limited time may have restricted the ability to perform exhaustive hyperparameter tuning, which could have further optimized the models.

5.2.3 Impact on Model Optimization

The fine-tuning process, while effective within the given constraints, may not have been able to fully optimize the models for the task of Mandarin-English code-switched speech recognition. The models' potential may have been under-realized due to:

Inadequate Exposure: The models may not have been exposed to a wide enough variety of speech patterns to fully adapt to the complexities of code-switching.

Suboptimal Hyperparameter Configuration: The models may not have been tuned to their optimal hyperparameter settings due to limited experimentation time.

5.2.4 Addressing Limitations for Future Research

Acknowledging these limitations is the first step towards addressing them in future research. Possible strategies include:

Expanding the Dataset: Efforts should be made to increase the size and diversity of the Mandarin-English code-switched dataset. This could involve collecting more data, incorporating a wider range of speakers, and including more linguistic variations.

Longitudinal Studies: Conducting studies over a longer period could allow for more comprehensive data collection and experimentation.

Resource Allocation: Allocating more resources, such as computational power and research personnel, could help overcome time constraints and allow for more extensive experimentation and hyperparameter tuning.

Collaboration: Collaborating with other research groups or institutions could provide access to additional datasets and resources, further enhancing the research capabilities.

In conclusion, while the current study has made significant strides in Mandarin-English code-switched speech recognition using Whisper models, it is clear that there is room for improvement. By addressing the limitations related to dataset size, diversity, and temporal restrictions, future research can build upon these findings and continue to push the boundaries of ASR technology in the context of multilingual communication.

5.3 Future Work

In light of the acknowledged limitations, the forthcoming research endeavors will pivot towards a concerted emphasis on broadening the dataset earmarked for fine-tuning purposes. The strategic infusion of a more expansive and diverse compendium of Mandarin-English code-switched speech data is poised to yield a commensurate elevation in the accuracy and resiliency of the models. This dataset expansion not only promises improved performance but also beckons a comprehensive evaluation of the models' efficacy across a wider spectrum of code-switched speech scenarios. Moreover, the infusion of additional time and resources will empower a meticulous and rigorous exploration of hyperparameter tuning, eschewing the models towards their zenith potential. This decisive step is poised to not only redress the extant limitations but also chart a course for the models to seamlessly align with real-world applications and the multifaceted demands of diverse user needs. The ennobled dataset size will invariably facilitate an exhaustive scrutiny of potential biases, paving the way for the formulation of nuanced strategies to mitigate them. The cumulative effect of these strategic interventions heralds the development of a more inclusive, equitable, and efficacious model tailored for Mandarin-English code-switched speech recognition.

6 Conclusion

The study conducted a thorough investigation into the domain of Mandarin-English code-switched speech recognition, leveraging the refinement of Whisper models through the process of fine-tuning. This unique approach aimed to address the intricate challenges posed by the combination of Mandarin and English languages within speech recognition systems. By exploring the realm of self-supervised learning (SSL) techniques tailored to this specific linguistic domain, the study unveiled promising outcomes, emphasizing the successful adaptation of SSL in overcoming data scarcity concerns prevalent in automatic speech recognition (ASR) tasks.

Delving into the fine-tuning methodology as elaborated in the discussions, the study substantiated the initial hypothesis that customizing Whisper models to suit the complexities of code-switched speech would result in substantial performance enhancements. Empirical evidence stemming from the experiments, particularly the noteworthy reductions in Mixture Error Rates (MER) exhibited by the ME-Whisper-small and ME-Whisper-large-v3 models, underscored the efficacy of tailored models in augmenting ASR precision.

However, amidst the successes observed, the study also shed light on critical limitations that warrant attention for future research pursuits. The relatively constrained size of the Mandarin-English code-switched dataset and the temporal constraints imposed on experimental optimization delineated the boundaries within which model refinement could occur. Nonetheless, these limitations function as springboards for improvement and refinement in subsequent investigations.

Looking towards the horizon, the imperative expansion of the dataset for fine-tuning surfaces as a pivotal progression. The augmentation of the dataset with a more extensive assortment of diversified code-switched speech instances not only fortifies model accuracy and resilience but also facilitates a more thorough assessment under varied speech contexts. By allocating additional resources and time towards meticulous hyperparameter optimization, the trajectory towards achieving peak model performance can be steadily advanced.

Moreover, the exploration of cross-lingual datasets and the formulation of severity-dependent models present captivating avenues for prospective research ventures. These novel methodologies hold promising potential in tailoring models to cater to diverse user requirements and in mitigating biases, thereby fostering the development of more equitable and efficient ASR frameworks.

In conclusion, this research represents a groundbreaking endeavor in implementing fine-tuning strategies to Mandarin-English code-switched speech recognition employing Whisper models. The significance of this work lies not only in the technical advancements achieved but also in

its potential to redefine the landscape of ASR technology. By focusing on Mandarin-English code-switching, a common phenomenon in multilingual communities, this study addresses a gap in the current offerings of ASR systems, which often struggle to accurately process speech that fluidly transitions between languages.

Although the findings are preliminary, they exude optimism and offer glimpses into a future where ASR systems not only exhibit heightened accuracy but also demonstrate adaptability to the multifaceted nature of multilingual communication. The improvements in performance metrics such as WER and MER, as a result of fine-tuning, are indicative of the models' enhanced ability to understand and transcribe code-switched speech. This is a promising step towards developing ASR systems that can better serve the needs of multilingual speakers.

The author harbors aspirations that this study will catalyze further exploration and innovation within the domain. By sharing the methodologies, findings, and open-source code, this research aims to inspire other researchers and developers to build upon these initial results. The hope is that through continued collaboration and knowledge sharing, the field will witness rapid advancements in the development of ASR systems that are more responsive to the complexities of real-world, multilingual interactions.

Ultimately, the goal is to contribute to the evolution of ASR systems that are genuinely inclusive and attuned to the diverse linguistic tapestry of our global society. As the world becomes increasingly interconnected and multilingualism becomes the norm, the need for ASR systems that can accurately and efficiently process code-switched speech becomes more critical. This research is a step towards creating technology that respects and accommodates the rich linguistic diversity of our world, ensuring that no voice is left unheard or misunderstood.

Furthermore, the study's emphasis on transparency, replicability, and ethical considerations sets a precedent for how future research in this field should be conducted. By ensuring that the research is accessible, reproducible, and respectful of ethical standards, this work aims to foster a culture of trust and integrity within the ASR community.

In summary, this research is more than just an academic pursuit; it is a commitment to advancing technology in a way that is mindful of the world's linguistic diversity and the need for inclusive communication tools. The author is hopeful that this study will not only pave the way for more sophisticated ASR systems but also inspire a broader conversation about the role of technology in promoting inclusivity and understanding in our interconnected world.

References

[1] B. E. Bullock and A. J. E. Toribio, *The Cambridge handbook of linguistic code-switching.* Cambridge university press, 2009.

[2] C. Myers-Scotton, *Contact linguistics: Bilingual encounters and grammatical outcomes.* OUP Oxford, 2002.

[3] K. Li, J. Li, G. Ye, R. Zhao, and Y. Gong, "Towards code-switching asr for end-to-end ctc models," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6076–6080, IEEE, 2019.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*, pp. 28492–28518, PMLR, 2023.

[5] P. Xie, X. Liu, Z. Chen, K. Chen, and Y. Wang, "Whisper-mce: Whisper model finetuned for better performance with mixed languages," *arXiv preprint arXiv:2310.17953*, 2023.

[6] M. Gales, S. Young, *et al.*, "The application of hidden markov models in speech recognition," *Foundations and Trends® in Signal Processing*, vol. 1, no. 3, pp. 200–203, 2008.

[7] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[8] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[9] J. Ao and T. Ko, "Improving attention-based end-to-end asr by incorporating an n-gram neural network," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 1–5, IEEE, 2021.

[10] N. Seshadri and C.-E. Sundberg, "List viterbi decoding algorithms with applications," *IEEE transactions on communications*, vol. 42, no. 234, pp. 313–323, 1994.

[11] H. Ney, D. Mergel, A. Noll, and A. Paeseler, "A data-driven organization of the dynamic programming beam search for continuous speech recognition," in *ICASSP'87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 12, pp. 833–836, IEEE, 1987.

[12] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[13] A.-r. Mohamed, G. Dahl, G. Hinton, *et al.*, "Deep belief networks for phone recognition," in *Nips workshop on deep learning for speech recognition and related applications*, vol. 1, p. 39, Vancouver, Canada, 2009.

[14] D. Yu and L. Deng, "Deep learning and its applications to signal and information processing [exploratory dsp]," *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 145–154, 2010.

[15] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[16] V. Mendelev, T. Raissi, G. Camporese, and M. Giollo, "Improved robustness to disfluencies in rnn-transducer based speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6878–6882, IEEE, 2021.

[17] W. Wang, Z. Chen, and H. Yang, "Long short-term memory for tibetan speech recognition," in *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, vol. 1, pp. 1059–1063, IEEE, 2020.

[18] S. Zhang, C. Liu, H. Jiang, S. Wei, L. Dai, and Y. Hu, "Feedforward sequential memory networks: A new structure to learn long-term dependency," *arXiv preprint arXiv:1512.08301*, 2015.

[19] K. O'shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[20] S. Zhang and M. Lei, "Acoustic modeling with dfsmn-ctc and joint ctc-ce learning.," in *INTERSPEECH*, pp. 771–775, 2018.

[21] Q. Gao, H. Wu, Y. Sun, and Y. Duan, "An end-to-end speech accent recognition method based on hybrid ctc/attention transformer asr," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7253–7257, IEEE, 2021.

[22] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*, pp. 1764–1772, PMLR, 2014.

[23] J. Yi and J. Tao, "Self-attention based model for punctuation prediction using word and speech embeddings," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7270–7274, IEEE, 2019.

[24] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, pp. 369–376, 2006.

[25] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, pp. 173–182, PMLR, 2016.

[26] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4945–4949, IEEE, 2016.

[27] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4960–4964, IEEE, 2016.

[28] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, "Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm," *arXiv preprint arXiv:1706.02737*, 2017.

[29] N. Moritz, T. Hori, and J. Le, "Streaming automatic speech recognition with the transformer model," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6074–6078, IEEE, 2020.

[30] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *arXiv preprint arXiv:1901.02860*, 2019.

[31] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[32] C.-C. Chiu and C. Raffel, "Monotonic chunkwise attention," *arXiv preprint arXiv:1712.05382*, 2017.

[33] H. Inaguma, M. Mimura, and T. Kawahara, "Enhancing monotonic multihead attention for streaming asr," *arXiv preprint arXiv:2005.09394*, 2020.

[34] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[35] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7829–7833, IEEE, 2020.

[36] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[37] O. Kuchaiev and B. Ginsburg, "Factorization tricks for lstm networks," *arXiv preprint arXiv:1703.10722*, 2017.

[38] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint arXiv:1701.06538*, 2017.

[39] Y. Kim, C. Denton, L. Hoang, and A. M. Rush, "Structured attention networks," *arXiv preprint arXiv:1702.00887*, 2017.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[41] B. Yang, Z. Tu, D. F. Wong, F. Meng, L. S. Chao, and T. Zhang, "Modeling localness for self-attention networks," *arXiv preprint arXiv:1810.10182*, 2018.

[42] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber, *et al.*, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," 2001.

[43] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5884–5888, IEEE, 2018.

[44] S. Zhou, L. Dong, S. Xu, and B. Xu, "A comparison of modeling units in sequence-to-sequence speech recognition with the transformer on mandarin chinese," in *International Conference on Neural Information Processing*, pp. 210–220, Springer, 2018.

[45] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[47] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[48] K. Namba, "An overview of myers-scotton' s matrix language frame model," *Senri International School (SIS) Educational Research Bulletin*, vol. 9, pp. 1–10, 2004.

[49] D.-C. Lyu, T.-P. Tan, E. S. Chng, and H. Li, "Seame: a mandarin-english code-switching speech corpus in south-east asia," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[50] Y. Li, Y. Yu, and P. Fung, "A mandarin-english code-switching corpus.," in *LREC*, pp. 2515–2519, 2012.

[51] H.-P. Shen, C.-H. Wu, Y.-T. Yang, and C.-S. Hsu, "Cecos: A chinese-english code-switching speech database," in *2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)*, pp. 120–123, IEEE, 2011.

[52] D. Wang, Z. Tang, D. Tang, and Q. Chen, "Oc16-ce80: A chinese-english mixlingual database and a speech recognition baseline," in *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, pp. 84–88, IEEE, 2016.

[53] A. Ali and S. Renals, "Word error rate estimation for speech recognition: e-wer," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 20–24, 2018.

[54] I. S. MacKenzie and R. W. Soukoreff, "A character-level error analysis technique for evaluating text entry methods," in *Proceedings of the Second Nordic Conference on Human-Computer Interaction*, NordiCHI '02, (New York, NY, USA), p. 243–246, Association for Computing Machinery, 2002.