UNIVERSITI TEKNOLOGI MARA

COMPARATIVE ANALYSIS OF SPEECH DETECTION MODELS WITH A FOCUS ON THE JAPANESE LANGUAGE

MUHAMMAD ALIFF AIMAN BIN ZOLKIFELI

MSc

March 2025

UNIVERSITI TEKNOLOGI MARA

COMPARATIVE ANALYSIS OF SPEECH DETECTION MODELS WITH A FOCUS ON THE JAPANESE LANGUAGE

MUHAMMAD ALIFF AIMAN BIN ZOLKIFELI

Dissertation submitted in partial fullfillment
of the requirements for the degree of
Master of Computer Science

Faculty of Computer and Mathematical Sciences

March 2025

ABSTRACT

The summary of everything. Master-minimum 200 words. PhD-minimum 250 words. Limit to one(1) page only. Font: Times New Roman - 12pt. Single spacing.

ACKNOWLEDGEMENT

A very special thanks to...

Finally, this dissertation is dedicated to my father and mother for the vision and determination to educate me. This piece of victory is dedicated to both of you. Alhamdulilah.

Must include all supervisors names. Limit one (1) page only. Font Times New Roman 12pt.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE
INTRODUCTION

1.1 Research Background

The way we interact with computer has changed rapidly throughout the history. We start by sending instruction to computer using punch card and nowadays we are able to instruct computer using our own voice. This advancement is made possible by using speech-to-text technology that converting the spoken language into text format (Wei Xu & Gao, 2023). Speech-to-text technologies has already exist in industries such as customer service and medicine. But with the advancement of the machine learning and artificial intelligence, it has made the speech-to-text technology to become more precise and faster (Latif et al., 2020). These advancements have enabled its application across many areas, including transcription services and the development of inclusive tools for individuals with disabilities (Koenecke et al., 2020).

Although the speech-to-text technology has advance rapidly, it still has challenge like accurately transcribing Japanese language. According to Kanno (1996) in "An Introduction to Japanese Linguistics", Japanese language has may words that sound the same but have different meaning and to know which word is being used is based on the current context of the sentence. This is because Japanese language is using syllable-based word formation rather than individual phonemes, it means that the words are created using syllables like "ka", "ki’ or "ku" instead of using a single consonant or vowel. Japanese language also using combination of three script with each has its own set of rule makes it harder to convert from spoken language to text.

With the advancement of machine learning and artificial intelligence, it has significantly improved speech recognition algorithms, enabling them to adapt to language nuances (Xu et al., 2023). This study investigates prominent models that is Whisper from Open AI, wav2vec2 by Facebook, and ALMA-7B fine-tuned model from Google Gemma. These models have their pros and cons when transcribing spoken Japanese into text and typically use learning frameworks and undergo training on extensive datasets to improve accuracy in recognizing speech patterns (Ando & Fujihara, 2021). It is important to compare Japanese speech recognition systems because most research currently focuses on English or European languages, with limited exploration of how well these systems work with Japanese, especially in casual conversations and real-world contexts.

Only a few studies is comparing the Japanese speech recognition systems which created a gap in this area. It is important to examine how well these models can handle language feature like dialects and how well they able to transcribe spoken language based on accuracy and the speed to convert speech to text.The findings will contribute to the development of Japanese speech recognition technology.

1.2 Problem Statement

Current speech-to-text model are trained on standardized language which might not capture the complexities of Japanese language dialect and informal expression (Imaizumi et al., 2022). This has led to the models cannot perform well when transcribing the conversational Japanese especially when informal words or dialects is being used. Despite the advancements of AI, which significantly increase the quality of text-to-speech model (Karita et al., 2021a), there is still lack of comprehensive evaluation between these models performance with Japanese language.

The lack of effective speech-to-text solution that tailored for Japanese language has its implication in industries. Industries that relying on speech-to-text technology such as telecommunication, education, technology, may face a problem because ineffective speech recognition can resulting in problems such as misunderstandings and will diminished the user satisfaction (Sztahó & Fejes, 2023). Additionally, the speech detection technology will not be adopted in industries if it fails to accurately capture the full spectrum of the language, limiting usability and accessibility. Because of that, a study focused on Japanese speech recognition quality is important not only to improve practical outcome but also supports the ongoing advancement in AI field.

1.3 Research Objectives

1. To identify the main challenges and the key requirements for more effective speech-to-text model within the context of Japanese language.

2. To analyze which of the technique is the most efficient to improve the WER(Word Error Rate) and the transcription latency of speech-to-text models that target Japanese language.

3. To evaluate the WER(Word Error Rate) and the transcription latency of different speech-to-text model when transcribing Japanese formal and informal language.

1.4 Research Questions

1. How can we define the challenges and requirement in order to make speech-to-text models for Japanese language to become more effective?

2. What is the best techniques and approaches that can be applied to improve the accuracy and speed of speech-to-text models for transcribing Japanese language?

3. How can we calculate the performance and effectiveness of different speech-to-text model in context of Japanese language?

1.5 Scope of Study

This study will be focusing on examining the effectiveness of large language model(LLM) for Japanese language speech-to-text technology. The key model is Whisper from Open Ai, wav2vec2 from Facebook's fined-tuned XLSR large language model, and the third model is ALMA-7B model that fine-tuned based on the Google's gemma large language models. In this study we will analyze the specific linguistic challenges that is unique in the Japanese language.

Some of the challenges is Japanese language using Syllable-based word formation that created many words that have same sounds but different meaning, which is crucial for the LLM to distinguish which word is being used based on the context of the sentence. On top of that, Japanese language also using three distinct writing system that can be used in one sentence. So it is also important to see how accurate the LLM transcribe the correct character from the speech.

After that we will evaluate each model performance based on how well it transcribing both formal and informal Japanese language. Formal language is the language that is usually being used in professional setting where informal language is used in the daily life conversation. We will also evaluate the models performance based on its performance transcribing the different Japanese dialect.

Aside from the model accuracy, in this study also will be comparing the speed of the model to complete the transcription task. Speed is also one of the important aspect when the model is being used in real-time application where any delay will cause impact on the user experience. By evaluating both the accuracy and processing speed of the model, this study aims to identify which model is high in performance with minimal latency.

1.6 Significance of Study

This study is aim to address the gap of effective speech-to-text solution that focusing on Japanese language. Most of the developed LLM is focusing in English language or a generic transcribe model that is developed for multi-language. In this study we want to highlight the industry implication that caused by the inefficiency in speech to text technology in industries such as telecommunication and technology.

This paper also contribute to the research by identify the current gaps in speech to text LLM, mainly in complex structured language like Japanese. This is achieved by offering a comparative analysis on which model is the best performance that can guide future improvement and innovation. This paper aims to increase the effectiveness of speech to text adoption thus enhance the real world application that rely on efficient transcription.

1.7 Conclusion

In this chapter, we discuss that the advancement in machine learning and artificial intelligence has made the computer can understand us better by improving the Speech To Text model accuracy and speed. However, there is still challenges to transcribe a language that has complex structure like Japanese that include syllable-based formation and the use of multiple writing systems. Because of this, a study to find which implementation and which model is the most performance for handling Japanese language. The finding from this study is very important to answer the question of which model is the best for speech-to-text solution in Japanese language. By identifying the specific linguistic challenges and comparing these models, this study will provide a valuable information that will be able to guide future advancements in speech-to-text technology in Japanese language and ultimately will be able to support its broader application across the industries that rely heavily on precise and efficient transcription.

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

The technology for automatic speech recognition (ASR) has advanced rapidly in these years. Starting from traditional models like GMM and HMM into more sophisticated deep learning approaches such as DNNs, CNNs, RNNs, and Transformer-based architectures. However, to apply these technologies to the Japanese language may pose few challenges due to its complex writing systems, phonetic ambiguities, and dialectal variations. In this chapter we will explore the intricacies of Japanese speech detection, reviews traditional and modern ASR models. The state-of-the-art systems like Whisper, wav2vec 2.0, and ChirpV2 will also be discussed based on their applicability to Japanese. By identifying the key challenges and gaps in existing research, this chapter prepared for a focused analysis of Japanese-specific ASR systems.

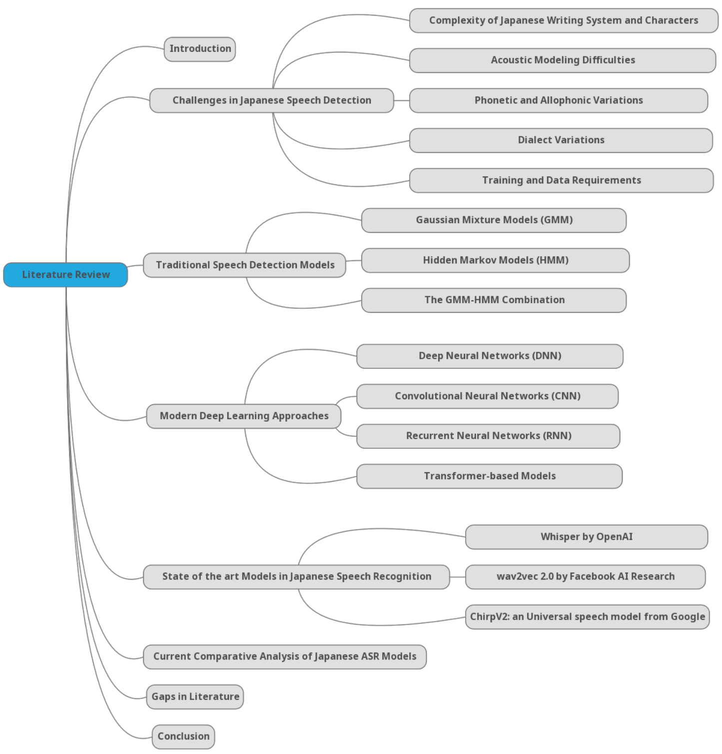

Figure 2.1 Literature Review Mind Map

2.2 Challenges in Japanese Speech Detection

2.2.1 Complexity of Japanese Writing System and Characters

The complexity of Japanese writing system and character can cause challenge in Automatic Speech Recognition(ASR) system especially in end-to-end neural network architectures. Japanese writing system is a combination of multiple character sets, such as the Hiragana, Katakana, Kanji (ideographic characters), Roman letters and various symbols, leading to a considerably larger and more varied character (Rose, 2019). As mentioned by Ito et al. (2016, 2017), the number of possible Japanese character labels can exceed several thousand.

A single character of kanji may have a few ways to pronounce it because each character of kanji has Onyomi (Chinese derived) and Kunyomi (native Japanese) readings, and these readings can change depending on the word context (Curtin, 2020). Because of this ambiguity, the ASR system must be able to model and distinguish numerous acoustic differences in the speech data with same sound. The training model must be able to handle thousands of character and each of the character is potentially linked to multiple context dependent phonetic outcome which require a significant computational resources and large scale training data to ensure adequate coverage (Ito et al., 2016, 2017).

2.2.2 Acoustic Modeling Difficulties

Ito et al. (2017) also discussed that one of the core acoustic modeling challenges in Japanese ASR is from the huge number of characters that must be predicted directly from the acoustic signal. This large number of character can reach thousands of label and can quickly increase the number of parameters in the acoustic models. Because of this, a traditional model like HMM-GMM and more recent end-to-end framework to face challenges when scaling into these high quantity of label sets. The combination of different character sets like Kanji, Hiragana, Katakana, Roman letters, digit and symbol forces the acoustic model to handle a massive output data, which makes the straightforward end-to-end approaches less efficient(Glasmachers, 2017).

The intonation and pitch accent in Japanese language is important to distinguish between words with similar sound but has different meaning. However, to accurately modeling these pattern is challenging because it is not depends on the specific word that being spoken but also on the sentence structure and grammatical context (Yasuda & Toda, 2022). This is different from the languages where the pronunciation can often be predicted based on the nearby sound because Japanese language require a deeper understanding of the overall linguistic context to capture the correct pitch accent.

2.2.3 Dialect Variations

Another notable challenges in Japanese speech detection is dialectical variation that exist in the language. This dialects usually different from the standardized Tokyo based variety that is referred to as standard Japanese (Takahashi et al., 2024). As mentioned by Imaizumi et al. (2020) that simply merging merging dialectal data with standard Japanese data into a single multi-condition training set often proves insufficient. The dialect specific pattern become saturated and leads to inefficient modelling and reduce the recognition accuracy. This issue can be addressed by integrating dialect labels as auxiliary features within the end-to-end ASR modeling framework. By explicitly encoding dialect information into the model, the recognition performance is improved by 19.2% relative error reduction (Imaizumi et al., 2020).

Another approach suggested by Imaizumi et al. (2022) is by employs a multi-task learning framework that able to optimize both dialect identification (DID) and multi-dialect ASR. This is achieveable by jointly training for dialect classification and ASR, the system exploits the strong interdependencies between acoustic-phonetic cues that is important for identification of dialectal features and linguistic structures that is crucial for accurate transcription). Three proposed architectures on how DID and ASR are integrated are as below:

1. DID2ASR: First performs dialect identification, then uses the predicted dialect label to inform ASR decoding.

2. ASR2DID: Reverses the order, first performing ASR and then using the recognized text to identify the dialect.

3. DID+ASR: Conducts dialect and ASR prediction jointly, using distributions over dialect classes rather than a single predicted label.

From these experiments demonstrate that multi-task learning with DID and ASR reduces word error rates and improves dialect classification accuracy (Takahashi et al., 2024).

2.2.4 Phonetic and Allophonic Variations

A study conducted by Nguyen et al. (2002) highlighted the impact of modeling frequent allophones to enhance recognition accuracy in the Japanese ASR systems. They observed that Japanese's predominant Consonant-Vowel (CV) syllabic structure leads to a skewed frequency distribution of certain CV pairs (e.g., /n-o/, /sh-i/, /k-a/), with frequent allophonic variations often misclassified by standard ASR systems that rely on broad phonemic units.

Aside from Phonetic variation, another challenges in Japanese speech detection is that the Japanese allophonic variation. These variation is influenced by factors such as vowel length, nasalization and gemination that often lead to context-dependent changes in phoneme realization. For traditional speech recognition, allophonic variation pose a certain challenges (Halpern, 2008). However, for refined acoustic models that explicitly capture frequent allophones by isolating allophones associated with common Consonant-Vowel (CV) combinations, this challenges can be mitigated. These models dedicate specific mixture densities and Gaussian components to high-frequency variants and allows the system to better differentiate subtle acoustic differences.

2.2.5 Training and Data Requirements

The quality of the Japanese speech detection system is relying heavily on the quality of the available training data as well as the methods and techniques employed during model training. As highlighted by Karita et al. (2021a) , Japanese ASR is often benchmarked using substantial corpora, such as the 581-hour Corpus of Spontaneous Japanese (CSJ). Similarly, Yasuda and Toda (2022) mentioned that large-scale textual resources are essential for pre-training language models like PnG BERT that can capture both character and phoneme-level representations, enabling important aspect such as pitch-accent-aware TTS.

It is important to do pre-training on a large text corpora as example shown by Yasuda and Toda (2022) that uses massive raw text and phoneme transcriptions to build strong contextual representations. While it skips paired speech-text data during pre-training, fine-tuning on high-quality and labeled speech is still necessary to align phoneme and acoustic features effectively. Fine-tuning strategies, such as layer freezing and gradual unfreezing, also can can help adapt pre-trained text models like PnG BERT in the speech recognition tasks (Ardestani et al., 2024). This ensures linguistic knowledge from pre-training is preserved while shifting focus to acoustic features, especially when speech datasets are smaller.

2.3 Traditional Speech Detection Models

2.3.1 Gaussian Mixture Models (GMM)

GMM have been the earliest technology used for developing Japanese speech detection and recognition systems because of their capability in capturing the statistical distribution of speech features very well (Imaishi & Kawabata, 2022). Because of the absence of word boundaries and the nuances of pitch accent in the Japanese language, it is really complicated to understand the context of the spoken words. However, GMM would be useful by employing probabilities to manage and characterize intricate patterns (Sun & Chol, 2020). For example, Povey et al. (2011) were able to use GMM to model phoneme-based acoustic features, and this approach led to a good performance of speech recognition systems.

Imaishi and Kawabata (2022) developed an approach within the EM algorithm that leads to the stabilization of the GMM parameters as well as increasing the discriminative power of the model in cases where there is not much evaluation data available. In other work, Povey et al. (2011) point out that it is possible to represent the distribution of speech features in GMM mode by employing a combination of several Gaussian components. This way the GMM ca n account for the phonetic or speaker variability which is known to be present during word is being pronounce.

Takami and Kawabata (2020) emphasized a different direction which starts with the creation of the Universal Background Model, which is a Gaussian Mixture Model calculated from the collection of a large number of speech samples from every dialect. To develop a model of the characteristics of a given UBM, the UBM is modified through Maximum A Posteriori (MAP) Adaptation. This method adjusts parameters of the UBM such as mean vectors, covariance matrices, and mixture weights depending on the individual's data (Dehak et al., 2009). Studies also have shown that the use of speaker factor space constructed in the GMM and Joint Factor Analysis (JFA) can greatly improves the accuracy and efficiency of GMM systems (Matrouf et al., 2011).

2.3.2 Hidden Markov Models (HMM)

HMM is working quite well with Japanese speech detection because of the incorporation of the acoustic and temporal characteristics of speech, including the difficulties found in the encoding of Japanese speech (Tokuda, 1999). Moreover, HMM is so useful in ASR because they are very efficient in the representation of time varying systems by a succession of discrete time states. A unique segment of the speech signal is represented in each state, and the segment is described using a specific set of acoustic features (Juang & Rabiner, 1991). ASR systems incorporated with HMM are more superior in portraying Japanese speech characteristics' rhythm and tone including essential features like pitch accent and moraic timing which features will enhance the performance of the systems on the phonology aspects of the language (Tokuda, 2000).

ASR systems based on HMMs give quite satisfactory results especially on languages like Japanese because it is a possible to interpolate between a discrete set of states, where each state stands for a segment of the speech signal that has distinct acoustic features like pitch, duration, and phoneme quality (Juang & Rabiner, 1991). To further Increase Japanese ASR project, few other model is used along with HMM which is context-sensitive such as Tri-phone method. Tri-phone method is a phonetic expansion that employs phonetics of the neighboring sounds to the phoneme as context in order to increase the recognition accuracy by taking into account the co-articulation that takes place during fast speech production (Tokuda, 2000). Other models by Gales and Young (2008) were used together with HMM are Maximum

Mutual Information and Minimum Phone Error which are useful for optimizing the parameters of the HMM and improve the recognition performance.

2.3.3 The GMM-HMM Combination

The GMM-HMM model uses GMM for the observation probabilities corresponding to each state of the HMM. Each state of an HMM is assumed to have a library of Gaussian mixtures with which the state's acoustic feature is pooled. Because transition probabilities of each state are determined by the HMM, temporal dependency of speech is well modelled. This combination allows the system to account for some of the variations in speech signals, such as those related to accent and the differences in the pronunciation of words in the Japanese language (Taheri & Taheri, 2006). While HMMs trained with large datasets under maximum likelihood criteria may have limited discriminative power, incorporating GMMs as observation models captures a broader range of acoustic variations. This method works really well for Japanese language, which are sensitive to the duration of phonemes in the context of the language.

Furthermore, the integration of GMM and HMM eliminates the need for applying state-of-the-art feature extraction techniques like Mel-Frequency Cepstral Coefficients (MFCC), hence increasing recognition performance (Sonali Nemade, 2019). This hybrid approach has been successful in speaker-dependent as well as in speaker-independent systems. When fuzzy clustering and the expectation-maximisation algorithm are used, lower error rates are usually obtained by GMM-HMM than the methods used in isolation. For example, in a paper on speech data collected in a noisy environment, it was demonstrated that GMM-HMM provided much improvement in recognition performance over the conventional HMM scheme (Sonali Nemade, 2019; Taheri & Taheri, 2006).

2.4 Modern Deep Learning Approaches

2.4.1 Deep Neural Networks (DNN)

The use of DNN in conjunction with HMM, also known as DNN-HMM has been shown to improve performance in Japanese speech recognition tasks. Seki et al.

(2014) compared syllable-based and phoneme-based DNN-HMM and found that the syllable-based DNN-HMM was better, as its parameter space is less coupled with the context of the syllables. They reported that an 11% relative decrease in the word error rate (WER) for triphone DNN-HMMs over syllable-based DNN-HMMs when used on large databases such as ASJ+JNAS. The multilayered structure of DNNs makes it much suitable for developing models of contextual dependencies for speech signals (Hojo et al., 2018). GMM-HMM models are less effective compared to DNN when the task involves the estimation of posterior probabilities. In particular, pre-training with restricted Boltzmann machines has been quite useful for weight initialization, the vanishing gradient problem, and overall performance (Masato Mimura et al., 2013).

Mu et al. (2020) developed a double-deep neural network for the evaluation of Japanese pronunciation to address the problems of text-to-speech alignment and scoring. The DDNN integrated CNN and RNNs with attention and it is effective for detecting pronunciation mistakes. Lin et al. (2017) noted the importance of addressing the particular problem of the lack of annotated Japanese speech corpora by emphasising the use of transfer learning with DNN. First, pre-training on large universal datasets increases the generalisation ability. Then, fine-tuning on Japanese databases enhances the performance that is critical in low-resource applications. The authors were also able to use CNN and recurrent architectures to attend to the granularity features of the Japanese language.

2.4.2 Convolutional Neural Networks (CNN)

There is difficulties in the visual speech recognition areas and specifically within lipreading because a limitation for the use of CNNs for phoneme recognition tasks was considered to be the number of training datasets (Noda et al., 2014). The research was conducted using elastic net regression on a seven-layer CNN structure and 58% of phoneme recognition accuracy was obtained for Japanese datasets. Building upon this work, Yalta et al. (2019) constructed a functional speech recognition framework inclusive of several types of words spoken intended for tight spots like houses. There are more focused methods for connecting microphones such as incorporating residual connections and batch denormalisation.

Noda et al. (2014) investigated the use of CNNs for solving the problem of creating a Japanese speech acoustic model. CNN used to encode the frequency-time domain images and properly exploit the spatial and temporal aspects. The C-nets employed in this model aided in recognising fine speech traits that i mproved performance in terms of recognition in contrast to the prevalent GMMs and HMMs methods. The combination of CNNs with attention mechanisms has yielded some results in the accurate detection of Japanese speech. This integration has been beneficial in increasing accuracy and interpretability during the detection of long utterances and multi-speaker datasets (Kohei Mukohara et al., 2015).

2.4.3 Recurrent Neural Networks (RNN)

In the work of Takeuchi et al. (2020), a novel design of the RNN is introduced, which enables the processing of input speech while removing noise caused by the room impulse response. This network mitigates the vanishing and exploding gradient problems often seen in RNNs while also keeping the parameter count low, making it very suitable for real-time applications. Yusuke Kida et al. (2016) investigated linear prediction filters based on LSTM. Their method trained an LSTM which did not require direct access to raw information and thus can extract features from distorted signals, as an LSTM estimated linear prediction coefficients.

Kubo (2014) broadened approaches incorporating RNNs into synthesizing speech for Japanese, particularly focusing on improving prosody and intonation. Their work underscored the necessity to consider the sequential modelling features of RNNs units, especially LSTMs, techniques for natural voice synthesis of Japanese language sounds. Takeuchi et al. (2020) took advantage of the RNN-based architectures for the acoustic modelling for Japanese automatic speech recognition system (ASR). They showed that even though GRUs have a simpler gating strategy than LSTMs, they could achieve a similar level of classification accuracy with lower compute requirements. Then, the studies on bidirectional LSTMs (BLSTMs). Imaizumi et al. (2022) revealed that they could utilise the past context and the future context of the signal for better performance of the speech recognition device. Many applications of automatic speech recognition in which the Japanese language is used have demonstrated that BLSTMs are particularly helpful for modelling complex phonological and prosodic structures of the Japanese language.

2.4.4 Transformer-based Models

Taniguchi et al. (2022) propose a series of Transformer-based automatic speech recognition (ASR) models aimed at improving Japanese speech recognition, particularly in the context of simultaneous interpretation. They investigate the possibility of utilizing auxiliary input like the source language text to resolve issues such as disfluencies, hesitations, and self-repairs commonly observed in the interpreter speech which helps to improve the transcription quality (Futami et al., 2020). The models combined audio and text data via multimodal transformer encoders and decoders, which offers a broader scope of recognition by using previously provided source language text for interpreter training programs (Taniguchi et al., 2022).

A wide range of datasets for source text and simultaneous interpretation speech are however not readily available, so the authors use a adapted speech translation corpora from MuST-C and CoVoST 2 while also introducing TED based Japanese texts for evaluation purposes (Taniguchi et al., 2022). With an additional goal of enhancing performance, the authors fine-tune the source language text encoder by using large machine translation corpora which helps in lowering the word error rates during translation of English, Dutch, German and Japanese (Taniguchi et al., 2024). Results consistently demonstrate that incorporating source language text into Transformer-based ASR models significantly improves recognition performance, with the greatest impact observed when auxiliary input is introduced at later stages of the audio encoding and decoding process (Futami et al., 2020).

2.5 State of the art Models in Japanese Speech Recognition

2.5.1 Whisper by OpenAI

Largescale weak supervision has emerged as one of the major approaches in speech recognition as noted by Radford et al. (2023) in their development of whisper model that has been trained on multilingual and multitask audio datasets that has a combined duration of 680,000 hours. This work is a continuation to the self-supervised methods such as Wav2Vec 2.0 (Baevski et al., 2020), which demonstrated learning without supervision from audio without any human-provided labels. However, dataset-specific fine-tuning is often necessary to obtain good performance, whereas with Whisper such reliance is reduced because of the efficacy of weak supervision.

By scaling weak supervision across diverse datasets, Whisper able to bypass the need for dataset-specific adaptation while able offer a robust zero-shot performance across languages and tasks. The authors also mentioned that by using this method, it will ensure the generalization and the robustness of the model while at the same time addressing main limitation in traditional models that is struggle to transcribe unfamiliar audio. This method also resulting in the models to have similar trends with other state of the art model in machine learning where a large, diverse datasets will improve model resilience which is align the with the advancements in computer vision (Kolesnikov et al., 2020) and NLP (Radford et al., 2019).The Whisper model's architecture, a simple encoder-decoder Transformer, reinforces the effectiveness of minimal preprocessing and sequence-to-sequence training, simplifying the transcription pipeline while achieving near-human-level accuracy.

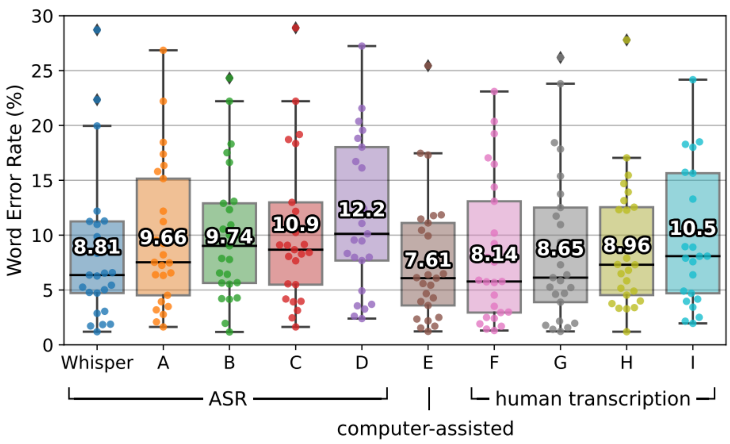

Figure 2.2 Whisper WER

Based on the work of Radford et al. (2023), there is further research that seeks to improve the performance of multilingual models on tasks that involve a japanese language. Bajo et al. (2024) detail their work on adapting OpenAI's Whisper model to enhance its performance in Automatic Speech Recognition (ASR) for the Japanese language. The research draws attention to the dilemma faced in balancing the multilingual being and the accuracy of an English-only product, ReazonSpeech, that seeks to maximize on the Japanese language ASR, which is monolingual in nature. By using a Japanese dataset while utilizing Low-Rank Adaptation (LoRA) and fine-tuning methods, they were able to lower Whisper-Tiny's Cumulative Expenditure Rate (CER) from 32.7% to 14.7%. This fine tuning method showed that smaller multilingual models give more promising result, after being tuned for the desired language outperform their larger baseline models, for example the case of the Whisper-Base model (Bajo et al., 2024).

2.5.2 wav2vec 2.0 by Facebook AI Research

The research conducted by Baevski et al. (2020) proved that self-supervised learning greatly reduces the dependency on large amounts of labeled data in speech recognition. They achieved this by using a technique called wav2vec 2.0. With this method, models are trained over a significant set of unlabeled speech data by masking the raw audio inputs and then treating a contrastive task. Thereafter, a model can be fine tuned using a limited set of labeled data which enables it to perform better when compared to semi-supervised techniques. Furthermore, according to Baevski et al. (2020), while their approach performed well with all the available labeled data by raising the word error rate (WER) to 1.8% for clean data and 3.3% for other data, it went even better with 10 mins of labeled and 53k hours of unlabeled data which had WER rates of 4.8% and 8.2%.

In Japanese speech recognition, self-supervised learning (SSL) has emerged as one of the major tools for tackling the problems of dialectal diversity and low-resourced datasets. Miwa and Kai (2023) showcased what they refer to as successful adaptation of the wav2vec 2.0-based XLSR model to the Corpus of Japanese Dialects (COJADS), a collection of data capturing various dialects from different regions of Japan. They reported significant gains in automatic speech recognition (ASR) metrics for dialectal speech, achieving character error rate (CER) reductions of as much as 8.9% relatively to the models only trained on tagged data.

2.5.3 ChirpV2: an Universal speech model from Google

Zhang et al. (2023) demonstrated a novel technique that scales ASR to more than a hundred languages, this is achieved with the aid of large multilingual datasets with self-supervised learning, they refer to their model as the Universal speech model. The model was pretrained on 12 million hours of unlabelled audio data collection of 300 languages, in addition to 90 thousand hours of multilingual labelled audio data. One of the crucial innovations is BEST-RQ (BERT-based Speech pretraining with Random-projection Quantizer) because it improves the performance of speech representation without complicated quantization modules.

The model also outperformed specialized models including Whisper that have previously been trained with more data. In addition to this, chunk-wise attention is used to solve the performance drop-off problem that USM has with long audio, allowing USM to transcribe long audio. Other language resource enabling techniques such as noisy student training and adapter modules have enhanced USM performance with low resource and unseen languages considerably, as it did with low resource languages ensuring a robust ASR system (Zhang et al., 2023). USM proves the efficacy of self-supervised models in minimizing multilingualism, and far supersedes existing standards for ASR systems.

2.6 Current Comparative Analysis of Japanese ASR Models

A comparative analysis that carried out by Karita et al. (2021b) shows that Conformer-based models perform better than Conformer BLSTM architectures, as they obtained 4.1, 3.2, and 3.5 character error rates for CSJ in eval1, eval2, and eval3 tasks respectively. It is noted that both the BLSTM and Conformer models have character error rates below 7% and the character error rate is lower when using Conformer Itself. Conformer encoders also offer increased accuracy and efficiency, with a throughput of 628.4 utterances processed per second and 430.0 for the BLSTM models. The scope of the work also emphasizes the importance of the analysis of the specific problem of training parameters optimization, noting the importance of the implementation of SpecAugment, exponential moving average (EMA) and variational noise (VN). The SpecAugment technique results in the largest shifts which affect the performance. The integration of the Conformer transducers with the described set of training approaches surpasses all existing solutions in Japanese ASR and open the path for further development (Karita et al., 2021b).

Another comparative analysis in the domain of the Japanese language is presented by Takahashi et al. (2024), focusing on the accuracy of speech recognition for different dialects. The study evaluates three models: Whisper, XLSR, and XLS-R, which are self-supervised learning frameworks. The Whisper model significantly underperformed for any Japanese outside of standard Japanese, recording a 4.1% character error rate (CER) only after it has gone through fine tuning. However, when the accuracy is low when the language identification marker is absent where some instances of cetus being higher than 100%.This marks the weakness of Whisper in terms of its application for wide ranging applications in different dialects of Japanese. However, Whisperer and XLS-R both of which were trained on multilingual speech data show improvement in the recognition of Japanese dialects. These models apply multi-task learning paradigms such as dialect identification (DID) and automated speech recognition (ASR) to increase their efficiency. Multi tasking adds significantly to the dialect accuracy and a three-step efficient training of the models reduce the character error rate (CER) by 3-4% relative to conventional transfer learning. Some dialects, especially those from Kyushu and Chubu, have larger CER than those spoken in Kanto, where there is a greater linguistic affinity (Takahashi et al., 2024).

2.7 Gaps in Literature

The improvement of ASR technologies with the progress of machine learning and AI models is very impressing but there is still a gap in the research on Japanese language ASR systems. Most of the existing work currently is focusing English or other widely used languages, while not much focus on Japanese which has unique features like syllable-based word formations, multiple script systems, and contextual nuances. However, more studies are needed to explore their abilities in handling dialectical variations and informal expressions in the Japanese ASR for modern and advanced models like Whisper, wav2vec 2.0 and Chirp. Furthermore,the issue non-availability of annotated datasets for Japanese and difficulty in balancing model accuracy with real-time processing speed is still persist. In order to narrow this gap, an assessment on the model precision, pace as well as how they treat such elements peculiar to Japanese is important in order to improve applicability of ASR systems in industries that require accurate speech-to-text solutions.

2.8 Conclusion

This chapter highlighted the challenges in Japanese ASR system that is its writing systems, phonetic variations, and dialectal diversity. The evolution of traditional ASR model to the cutting-edge technologies also has been highlighted in this chapter. Despite the advancements of the model and ASR framework, there is still a gap in adapting these models to Japanese-specific contexts, especially in handling informal speech and dialects. There is still work to be done to further refine the accuracy, speed, and dataset availability to advance Japanese ASR. These result can be use as a foundation for developing more effective and inclusive speech-to-text solutions tailored for Japanese language.

CHAPTER THREE
RESEARCH METHODOLOGY

3.1 Introduction

This chapter presents the methodology for evaluating three pre-trained ASR models—OpenAI's Whisper, Meta's wav2vec 2.0, and Google's CHIRP (USM)—on Japanese speech recognition. The study focuses on comparing Word Error Rate (WER) and transcription speed using formal (TED Talks) and informal (dialectal, COJADS) datasets. The chapter details the data collection, preprocessing steps (such as audio conversion and resampling), model selection rationale, and the standardized testing environment. It concludes by discussing key challenges, limitations, and ethical considerations relevant to the evaluation process.

3.2 Research Design

3.3 Data Collection

3.3.1 Dataset Selection

To evaluate the performance of the models, this study will be using two source of data. The first dataset is from Ted Talk Youtube dataset which will be responsible for formal speech with clear pronunciation completes with rich and diverse vocabulary. The second dataset is the Corpus of Japanese Dialects(COJADS) which will be responsible as the input for informal speech which categorized based on regional accent and expression. The COJADS dataset only can be obtain from National Institute for Japanese Language and Linguistics (NINJAL).

3.3.2 Data Pre-processing

The pre-processing steps is the first step in doing the model comparison. This step will be tailored to prepare the datasets for input into the selected speech recognition models. For the TED Talks YouTube dataset, the audio files will be extracted from video recordings and transcribed using Python moviepy library into Waveform Audio File Format (WAV) format.

Listing 3.1: Python code to convert video to WAV format using moviepy

```
def convert_video_to_wav(video_path, output_wav_path):
try:
    video_clip = VideoFileClip(video_path)
    audio_path = output_wav_path\
        .replace(".wav", "_temp_audio.mp3")
    video_clip.audio.write_audiofile(audio_path)
    return audio_path

except Exception as e:
    return None
```

After that, each audio file was converted to the standardized 16 kHz, 16-bit PCM format to ensure the data is compatible with the ASR models. Then the text will be manually transcribed and aligned with the corresponding audio to create accurate transcriptions for evaluation.

For the Corpus of Japanese Dialects, the data is in MP4 format and the audio files will be extracted to wav using the same method as the TED Talks YouTube dataset. Then the audio files will be resampled to resampled to the 16 kHz, 16-bit PCM format using Python pydub library.

Listing 3.2: Python code to resample audio to 16 kHz using pydub

```
def resample_audio(input_audio_path, output_audio_path,\
    target_sample_rate=16000):
try:
    audio = AudioSegment.from_file(input_audio_path)
    audio = audio.set_frame_rate(target_sample_rate)
    audio.export(output_audio_path, format="wav")

    print(f"Resampled audio saved to {output_audio_path}")
except Exception as e:
    print(f"Error during audio resampling: {e}")
```

The reason for resampling the audio into 16 kHz, 16-bit PCM format is because it will align with the standard input requirements for most Automatic Speech Recognition

(ASR) models. This sampling rate has the capability to capture the full frequency range of human speech, making it suitable for speech recognition tasks.

3.4 Model Selection

This study will be evaluating three advanced speech recognition models to determine their performance in handling the challenges of Japanese ASR. the first model is the whisper model by OpenAI is selected for its capabilities in speech-to-text transcription across diverse languages. It is also develop by OpenAI which is known for their cutting-edge research in AI and machine learning. The second model is the wav2vec 2.0 model by Meta/Facebook AI Research is chosen for its self-supervised learning approach, which reduces the dependency on large annotated datasets by leveraging unlabeled speech data. The third model is the CHIRP model by Google USM is selected for its cutting-edge multilingual capabilities, designed to handle over 100 languages with high accuracy efficiently.

3.5 Model Testing and Evaluation

3.5.1 Testing Environment

The model evaluations will be using the Hugging Face platform, which is a widely used and versatile framework for deploying and testing Machine Learning models. Hugging Face provides out of the box integration with the chosen speech recognition models. This will enable more efficient setup and execution of the testing workflows.

To provide better and more consistent evaluations, the testing environments for all models will be standardized. The sample will be prepared using the same preprocessed method for audio files which are sampled at 16 kHz and follow the 16 bit PCM standard. This will make sure all the input data will be compatible with the models Then the test will be conducted in the same environment to provide identical hardware and software configurations, so the results of the models can be compared. Each experiment will be run five times to overcome the variations in the performance of the model and the averages will be use for more accurate result.

3.5.2 Performance Metrics

To evaluate the performance of the selected speech recognition models, these key metrics will be used:

- Word Error Rate: Measures the accuracy of transcriptions by calculating the percentage of words that incorrectly transcribed. This is a critical metric for assessing the overall precision of the models.
- Transcription Latency: Measures the time taken by each model to transcribe audio input, providing insight into their suitability for real-time applications.
- Handling of Formal vs. Informal Language: Evaluates how well the models perform across different linguistic contexts, including formal speech and informal, dialectal speech.

3.5.3 Test Procedure

The following step-by-step procedure was implemented to evaluate the models:

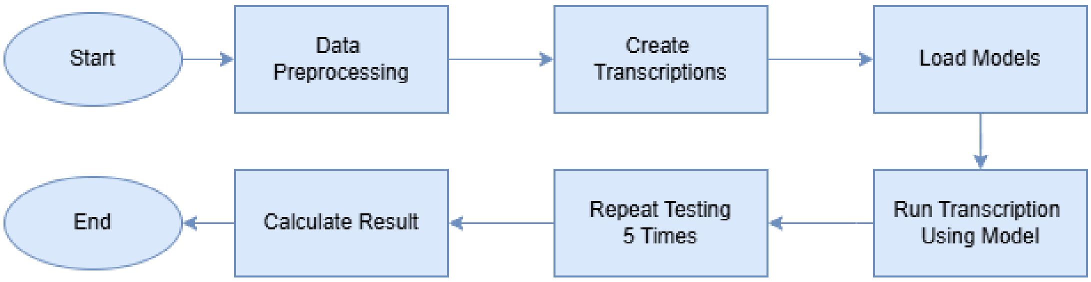

Figure 3.1 Testing procedure

The model evaluation process will be start by pre-processing the audio files to ensure consistency in format. This will make sure that the audio files are compatible with the models. Then the data audio from the TED Talks YouTube dataset will be used to create transcriptions manually. The transcriptions will be aligned with the audio files to ensure accurate evaluation. The COJADS dataset did not need to be transcribed as the transcriptions are already provided.

Then the models will be loaded into the Hugging Face platform and the pre-processed audio data will be inputted into the models one by one. The output transcriptions will be evaluated against the reference transcripts to gather the comparison metrics. The evaluation will be repeated five times for each model to ensure reliability. The average performance will be calculated for each metric to provide a more accurate representation of the models' capabilities.

3.6 Challenges and Limitations

one significant challenge in this study is the limited availability of high-quality, annotated datasets for Japanese speech recognition. While datasets like TED Talks YouTube and the Corpus of COJADS provide valuable resources, they may not fully capture the breadth of linguistic diversity in Japanese, particularly in less-represented dialects and informal speech contexts. Additionally, variations in recording quality and noise levels within the datasets can introduce inconsistencies that may affect the models' performance.

another limitation is the potential biases introduced by the selection of test samples. The TED Talks YouTube dataset predominantly features formal speech, which may not adequately represent everyday conversational Japanese. The Corpus of Japanese Dialects includes informal and regional speech but might not cover all dialects or account for significant inter-speaker variability. These biases could skew the evaluation results, favoring models better suited to the specific characteristics of the datasets.

3.7 Ethical Considerations

The ethical considerations of this study center on data privacy, consent, and fairness in model evaluation. The datasets used in this study contain publicly available speech recordings, ensuring that no private or sensitive information is disclosed. The TED Talks YouTube dataset and the COJADS are widely accessible and do not require individual consent for research purposes. The study will focus on evaluating the performance of speech recognition models rather than analyzing the content of the recordings, minimizing potential privacy concerns.

3.8 Summary

In this chapter, a clear research design was laid out to compare three ASR models on Japanese speech. Formal (TED Talks) and informal (COJADS) datasets were chosen for linguistic variety, and consistent preprocessing steps were described. The reasoning for selecting Whisper, wav2vec 2.0, and CHIRP (USM) was explained, followed by details on how testing will be conducted in a standardized environment to measure WER and transcription latency. Lastly, potential challenges in dataset availability, biases, and ethical considerations around data privacy were addressed.

REFERENCES

Ando, S., & Fujihara, H. (2021). Construction of a large-scale japanese asr corpus on tv recordings. *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6948–6952.

Ardestani, M., et al. (2024). *A study of text summarization with graph attention networks* [Doctoral dissertation, Lethbridge, Alta.: University of Lethbridge, Dept. of Mathematics and . . .].

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.

Bajo, M., Fukukawa, H., Morita, R., & Ogasawara, Y. (2024). Efficient adaptation of multilingual models for japanese asr. *arXiv preprint arXiv:2412.10705*.

Curtin, K. (2020). Japanese kanji power: A workbook for mastering japanese characters.

Dehak, N., Kenny, P., Dehak, R., Glembek, O., Dumouchel, P., Burget, L., Hubeika, V., & Castaldo, F. (2009). Support vector machines and joint factor analysis for speaker verification. *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 4237–4240.

Futami, H., Ueno, S., Mimura, M., Sakai, S., Kawahara, T., et al. (2020). Rescoring hypotheses of automatic speech recognition with bidirectional transformer language model. *Proceedings of the 82nd National Convention of IPSJ*, *2020*(1), 175–176.

Gales, M., & Young, S. (2008). The application of hidden markov models in speech recognition. *Foundations and Trends® in Signal Processing*, *1*(3), 195–304.

Glasmachers, T. (2017, November). Limits of end-to-end learning. In M.-L. Zhang & Y.-K. Noh (Eds.), *Proceedings of the ninth asian conference on machine learning* (pp. 17–32, Vol. 77). PMLR. https://proceedings.mlr.press/v77/glasmachers17a.html

Halpern, J. (2008). The role of phonetics and phonetic databases in japanese speech technology.

Hojo, N., Ijima, Y., & Mizuno, H. (2018). Dnn-based speech synthesis using speaker codes. *IEICE TRANSACTIONS on Information and Systems*, *101*(2), 462–472.

Imaishi, R., & Kawabata, T. (2022). Examination of iterative estimation counts in gaussian mixture model-based speaker recognition. *Journal of the Acoustical Society of Japan*, *78*(11), 650–653.

Imaizumi, R., Masumura, R., Shiota, S., & Kiya, H. (2020). Dialect-aware modeling for end-to-end japanese dialect speech recognition. *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 297–301.

Imaizumi, R., Masumura, R., Shiota, S., & Kiya, H. (2022). End-to-end japanese multi-dialect speech recognition and dialect identification with multi-task learning. *APSIPA Transactions on Signal and Information Processing*, *11*. https://doi.org/10.1561/116.00000045

Ito, H., Hagiwara, A., Ichiki, M., Mishima, T., Sato, S., & Kobayashi, A. (2016). End-to-end neural network modeling for japanese speech recognition. *Journal of the Acoustical Society of America*, *140*, 3116–3116. https://doi.org/10.1121/1.4969755

Ito, H., Hagiwara, A., Ichiki, M., Mishima, T., Sato, S., & Kobayashi, A. (2017). End-to-end speech recognition for languages with ideographic characters. *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1228–1232. https://doi.org/10.1109/APSIPA.2017.8282226

Juang, B. H., & Rabiner, L. R. (1991). Hidden markov models for speech recognition. *Technometrics*, *33*(3), 251–272.

Kanno, K. (1996). An introduction to japanese linguistics. *The Journal of the Association of Teachers of Japanese*, *30*(1), 64–69. Retrieved November 19, 2024, from http://www.jstor.org/stable/489672

Karita, S., Kubo, Y., Bacchiani, M. A. U., & Jones, L. (2021a). A comparative study on neural architectures and training methods for japanese speech recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2*, 2092–2096. https://doi.org/10.21437/INTERSPEECH.2021-775

Karita, S., Kubo, Y., Bacchiani, M. A. U., & Jones, L. (2021b). A comparative study on neural architectures and training methods for japanese speech recognition. *arXiv preprint arXiv:2106.05111*.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America*, *117*, 7684–7689. https://doi.org/10.1073/PNAS.1915768117/SUPPL_FILE/PNAS.1915768117.SAPP.PDF

Kohei Mukohara, S. N., Koichiro Yoshino, et al. (2015). Investigation of dnn and cnn bottleneck features in emotional speech recognition. *Research Report on Speech and Language Processing (SLP)*, *2015*(15), 1–6.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., & Houlsby, N. (2020). Big transfer (bit): General visual representation learning. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, 491–507.

Kubo, Y. (2014). Deep learning for speech recognition (series explanation: Deep learning [part 5]). *Artificial Intelligence*, *29*(1), 62–71.

Latif, S., Qadir, J., Qayyum, A., Usama, M., & Younis, S. (2020). Speech technology for healthcare: Opportunities, challenges, and state of the art. *IEEE Reviews in Biomedical Engineering*, *14*, 342–356.

Lin, S., Tsunakawa, T., Nishida, M., & Nishimura, M. (2017). Dnn-based feature transformation for speech recognition using throat microphone. *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 596–599.

Masato Mimura, T. K., et al. (2013). Application of dnn-hmm to japanese lecture speech recognition using csj and investigation of speaker adaptation. *Research Report on Speech and Language Processing (SLP)*, *2013*(9), 1–6.

Matrouf, D., Verdet, F., Rouvier, M., Bonastre, J.-F., & Linarès, G. (2011). Modeling nuisance variabilities with factor analysis for gmm-based audio pattern classification. *Computer Speech & Language*, *25*(3), 481–498.

Miwa, S., & Kai, A. (2023). Dialect speech recognition modeling using corpus of japanese dialects and self-supervised learning-based model xlsr. *Proc. INTERSPEECH 2023*, 4928–4932.

Mu, D., Sun, W., Xu, G., & Li, W. (2020). Japanese pronunciation evaluation based on ddnn. *IEEE Access*, *8*, 218644–218657.

Nguyen, L., Guo, X., & Makhoul, J. (2002). Modeling frequent allophones in japanese speech recognition. *Interspeech*. https://api.semanticscholar.org/CorpusID:7693997

Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., Ogata, T., et al. (2014). Lipreading using convolutional neural network. *Interspeech*, *1*, 3.

Povey, D., Burget, L., Agarwal, M., Akyazi, P., Kai, F., Ghoshal, A., Glembek, O., Goel, N., Karafiát, M., Rastrow, A., et al. (2011). The subspace gaussian mixture model—a structured model for speech recognition. *Computer Speech & Language*, *25*(2), 404–439.

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

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Rose, H. (2019). Unique challenges of learning to write in the japanese writing system. *L2 writing beyond English*, *66*.

Seki, H., Yamamoto, K., & Nakagawa, S. (2014). Comparison of syllable-based and phoneme-based dnn-hmm in japanese speech recognition. *2014 International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA)*, 249–254.

Sonali Nemade, R. D. P., Yogesh Kumar Sharma. (2019). To improve voice recognition system using gmm and hmm classification models. *International Journal of Innovative Technology and Exploring Engineering (2019) 8(11) 2724-2726*.

Sun, R. H., & Chol, R. J. (2020). Subspace gaussian mixture based language modeling for large vocabulary continuous speech recognition. *Speech Communication*, *117*, 21–27.

Sztahó, D., & Fejes, A. (2023). Effects of language mismatch in automatic forensic voice comparison using deep learning embeddings. *Journal of Forensic Sciences*, *68*, 871–883. https://doi.org/10.1111/1556-4029.15250

Taheri, A., & Taheri, M. (2006). Fuzzy hmm and gmm models for speech recognition. *2006 2nd International Conference on Information & Communication Technologies*, *1*, 1242–1245.

Takahashi, N., Miwa, S., Kamiya, Y., Toyama, T., Nahar, R., & Kai, A. (2024). Comparison of large pre-trained models and adaptation methods for japanese dialects asr. *2024 IEEE 13th Global Conference on Consumer Electronics (GCCE)*, 811–814.

Takami, J., & Kawabata, T. (2020). Speaker recognition performance metric for small-scale voice dialogue systems based on adaptation speed and convergence accuracy of gaussian mixture models. *Journal of the Acoustical Society of Japan*, *76*(5), 254–261.

Takeuchi, D., Yatabe, K., Koizumi, Y., Oikawa, Y., & Harada, N. (2020). Real-time speech enhancement using equilibriated rnn. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 851–855.

Taniguchi, S., Kato, T., Tamura, A., & Yasuda, K. (2022). Transformer-based automatic speech recognition with auxiliary input of source language text toward transcribing simultaneous interpretation. *INTERSPEECH*, 2813–2817.

Taniguchi, S., Kato, T., Tamura, A., Yasuda, K., et al. (2024). Pre-training of transformer-based asr for simultaneous interpretation with auxiliary input of source language text using large machine translation corpus. *Proceedings of the 86th National Convention of IPSJ*, *2024*(1), 397–398.

Tokuda, K. (1999). Application of hidden markov models to speech synthesis. *IEICE Technical Report, SP99-61*, 48–54.

Tokuda, K. (2000). Fundamentals of speech synthesis using hmm. *IEICE Technical Report, SP2000-74*, 43.

Wei Xu, L. G., Marvin J. Dainoff, & Gao, Z. (2023). Transitioning to human interaction with ai systems: New challenges and opportunities for hci professionals to enable human-centered ai. *International Journal of Human–Computer Interaction*, *39*(3), 494–518. https://doi.org/10.1080/10447318.2022.2041900

Xu, C., Ye, R., Dong, Q., Zhao, C., Ko, T., Wang, M., Xiao, T., & Zhu, J. (2023). Recent advances in direct speech-to-text translation. *arXiv preprint arXiv:2306.11646*.

Yalta, N., Watanabe, S., Hori, T., Nakadai, K., & Ogata, T. (2019). Cnn-based multichannel end-to-end speech recognition for everyday home environments. *2019 27th European Signal Processing Conference (EUSIPCO)*, 1–5.

Yasuda, Y., & Toda, T. (2022). Investigation of japanese png bert language model in text-to-speech synthesis for pitch accent language. *IEEE Journal of Selected Topics in Signal Processing*, *16*(6), 1319–1328. https://doi.org/10.1109/JSTSP.2022.3190672

Yusuke Kida, T. T., et al. (2016). Reverberant speech recognition based on linear predictive filter estimation using lstm. *Research Report on Speech and Language Processing (SLP)*, *2016*(25), 1–6.

Zhang, Y., Han, W., Qin, J., Wang, Y., Bapna, A., Chen, Z., Chen, N., Li, B., Axelrod, V., Wang, G., et al. (2023). Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*.

APPENDICES

APPENDIX A
THE DATA

This is the data.

APPENDIX B
THE CODING

This is the C code.

```
#include <stdio.h>
int main() {
   printf("Hello World\n");
return 1;
}
```