University of Groningen

Phone Masking Augmentation for Automatic Recognition of Whispered Speech

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science

Igor Marchenko

(S5754798)

June 11, 2024

Supervised by Assistant Prof. Dr. Shekhar Nayak and Dr. Karthika Vijayan
Second reader: Associate Prof. Dr. Matt Coler

Leeuwarden, the Netherlands

Acknowledgments

I would like to thank my supervisor Dr. Shekhar Nayak and the Sahaj Software team, Karthika Vijayan and Kartikay Shandil, for the thought-provoking discussions we had and for treating me as a colleague. धन्यवाद.

My special thanks go to Zhaofeng Lin from Trinity College Dublin and Heng-Jui Chang from Massachusetts Institute of Technology for their kind assistance with the dataset used in this work. And to everyone who's been along the way: thank you, 谢谢, спасибо, cảm ơn, danke, mulțumesc, 감사합니다, grazie, dank je wel, teşekkür ederim. Good luck to you all.

Lastly, I heartily thank my parents, Irina and Ivan, for their support and the chance to be here. Спасибо!

Contents

Introduction 1

1.1 Whispering as a Means of Communication

The majority of human communicative tasks are resolved through speech (along with nonverbal methods such as facial expressions, gestures, laughter, etc.) with an average intensity range of approximately 55-65 dB. Yet, there are situations where speech must adhere to specific requirements distinct from typical scenarios of communication. For instance, a conversation may occur in environments like libraries or theaters, where loud speech is not appropriate. Additionally, a communication act may require a certain level of privacy and confidentiality, which cannot be achieved with loud voice. In these cases, individuals may resort to whispering – a "secondary aspect" of communication characterized by quiet, voiceless speech where words are articulated without engaging the vocal cords. This type of speech typically has a volume in the range of 20-30 dB (Markides 1986, Lian et al. 2019). For a reference, this volume is close to that of tree leaves rustling.

(Markides 1986): *'Speech levels and speech-to-noise ratios'*

(Lian et al. 2019): *'Whisper to normal speech conversion using sequence-to-sequence mapping model with auditory attention'*

For some speakers, however, whispering is not an auxiliary tool but serves as the only instrument of verbal communication. It is often the main method for individuals with impaired voice production, such as those who have undergone a laryngectomy (Sharifzadeh, McLoughlin, and Ahmadi 2010), or for those advised to rest their voice, for example following surgery or laryngeal damage. Unlike most people who can choose when to whisper and when not to, these individuals may have no choice but to rely on whispering for most of their daily interactions.

(Sharifzadeh et al. 2010): *'Speech rehabilitation methods for laryngectomised patients'*

However, not only is whispering common in human-to-human communication, it can also be an appealing method for human-machine interactions. The ability of digital assistants to recognize whispering would allow users to interact with them in a quieter manner, which is particularly useful in situations where noise levels need to be kept low, such as when someone is sleeping in a room. Moreover, it can make these devices available for people who rely solely on whispering for their communication. Still, despite the occasional use of whispering by all speakers and the exclusive reliance on whispering by some, the prevailing speech recognition systems typically operate under the assumption or necessity of phonated speech and struggle with recognizing whispering, preventing free and inclusive access to technology.

1.2 Technology and Whispered Speech

One of the most dynamically developing areas of artificial intelligence today is automatic speech recognition (hereinafter - ASR), which deals with transcribing human speech into text for various applications, ranging from voice-activated virtual assistants and real-time translation services to automated customer service systems. Over the past half century, significant progress has been made in this field, and modern speech recognition models based on deep neural networks may in some tasks even exceed the abilities of human to recognize speech (Xiong et al. 2017).

(Xiong et al. 2017): Achieving Human Parity in Conversational Speech Recognition.

Despite such significant progress, however, the main goal of ASR — ensuring free communication between human and computer — has not yet been achieved. An obstacle on the way to reaching it, among other technical and philosophical problems, is as well the recognition of whispering, with which even modern sophisticated models cope much worse than with neutral, vocalized speech.

The main reason why modern speech recognition systems still cannot recognize whispered speech as accurately as neutral speech is the lack of adequate data to train corresponding models. All commercial speech recognition systems are trained with huge amounts of speech data (for instance, the state-of-the-art model Whisper[1] of OpenAI (Radford et al. 2023) is trained with approximately 680.000 hours of recordings), however, is it only or predominantly neutral, vocalized speech that was employed for training these models.

(Radford et al. 2023): *'Robust speech recognition via large-scale weak supervision'*

Good quality of recognition requires good amounts of data, but gathering a substantial amount of whispering recordings, not even speaking of Whisper's amounts, is challenging, and mainly due to the risks associated with prolonged whispering. Unlike normal speech, for which it is possible to hire people to make high-quality recordings with which recognition models can be trained, whispering can strain the vocal cords and lead to their damage, making it impossible, due to ethical reasons, to ask individuals to record extensive whisper datasets.

Nevertheless, datasets of whispered speech do exist. However, they are not large enough to train an effective whispered speech recognition model independently. Therefore, to solve this problem, there is a need to find strategies to train models in the absence of sufficient data. For this, various model adaptation techniques can be employed, during which a model originally trained with normal speech is adjusted so as to fit whispered speech parameters. However, even for model adaptation, available whispering data is not sufficient, and thus there is a need to expand these datasets to the extent possible. The main technique to expand the available whisper data is the generation of artificial whispered speech, which has proven to be useful in improving baseline model's accuracies. However, one strategy which has been shown by many works on neutral speech recognition as an effective instrument for expanding data — data augmentation — has not been tested extensively on whispered speech yet.

[1] Despite its name, Whisper does not represent a model trained specifically for recognizing whispered speech.

1.3 Data Augmentation

In conditions of data scarcity, data augmentation has proven to be a saving technique for training effective speech recognition models. The general idea behind it is to create artificial data by applying various transformations to original instances, thereby increasing the volume of the training data. Practically, there is no limit to such transformations: they can consist of any changes to already existing data, depending on the nature of such data. Whereas images, for which these methods originated, are usually subjected to various stretches and rotations, the audio data is subjected to different kinds of signal transformations, i.e., of its speed (Ko et al. 2015), pitch (Shahnawazuddin et al. 2020), or level of noise (Pervaiz et al. 2020).

(Ko et al. 2015): *'Audio augmentation for speech recognition.'*

(Shahnawazuddin et al. 2020): *'Creating speaker independent ASR system through prosody modification based data augmentation'*

(Pervaiz et al. 2020): *'Incorporating noise robustness in speech command recognition by noise augmentation of training data'*

Meanwhile, expanding the training dataset is not the only benefit of data augmentation: in fact, it also helps the model generalize better over data in question. By applying distortions to the original instances and extracting features from these altered versions, the model learns to perform well even under challenging conditions, preparing for real-world scenarios where unseen data may naturally be deformed rather than artificially altered. Thus, the idea of augmenting is also to look a little "further and above" the ideal data.

This way, a special place in the list of data augmentation methods for audio data occupies SpecAugment proposed in (Park et al. 2019), which works by applying time and frequency masking to the input spectrograms. This involves occluding random blocks along the time axis, which is called time masking, and frequency axis, i.e, frequency masking. These occlusions force the model to learn to predict spectral information from the context around these masks, thereby improving its ability to generalize from incomplete or noisy data. However, this approach is indiscriminate and does not account for the linguistic or phonetic importance of different aspects of speech: it has proven to be efficient for speech thanks to its continual nature, but practically, SpecAugment can be applied to any kind of audio, be it human speech or birds singing.

(Park et al. 2019): *'Specaugment: A simple data augmentation method for automatic speech recognition'*

To bring more meaning into the masking process, an extension of SpecAugment called SemanticMasking was introduced in (Wang et al. 2019). Instead of simply masking arbitrary time or frequency regions, in this work it is proposed to mask specific words. This kinds of masking aims to teach the model to handle variations and gaps in a more meaningful way, enhancing its linguistic understanding and generalization capabilities based on semantics of natural language.

(Wang et al. 2019): *'Semantic mask for transformer based end-to-end speech recognition'*

However, when the domain narrows down, for augmentation techniques to be efficient, they must adhere to special needs of this particular domain. As will be discussed in our work, in the context of whispered speech, what make it special are its unique phonetic and acoustic properties. Thus, for whispered speech, I propose another kind of data augmentation based on the mixture of ideas from SpecAugment and Semantic Masking, that would involve masking of groups of specific phonemes which I call Ph[]neMask, where the two brackets [], on the one hand, look like O, but at the same time represent that this O is masked.

1.4 Phone Masking for Whispered Speech

There exists a significant acoustic mismatch between whispered and normal speech. While articulation of phonologically unvoiced consonants is similar to how they are produced in normal speech, voiced consonants and vowels differ markedly as they are produced with no vibrations of vocal folds. Moreover, studies that compared phonetic spaces of normal and whispered speech have shown that there are systematic differences between particular groups of phonemes pronounced in normal and whispered modes of speech. Thus, as a step towards bridging the gap between whispering and normal speech, a more targeted form of masking can be implemented. For instance, by masking specific phonemes that are most different between the two modes of speech and thus pose most difficulties for models trained with normal speech, we can compel the model to specifically learn contextual cues and phonetic patterns that are critical for recognizing these phonemes. Consequently, the model is expected to become adept at recognizing speech in more stressful conditions, which is what whispered speech itself is for ASR models.

If improvement in recognition results is not achieved, I will still be able to evaluate the effectiveness of the augmentation technique that has not yet been tested on whispered speech before. Consequently, this study, regardless of the success of one or another masking method, will make a practical contribution to the general framework of automatic recognition of whispered speech, which is an extremely important aspect of human communication, yet still difficult for machines to understand.

1.5 Thesis Outline

This thesis is organized as follows. Section 2 will provide the necessary background to understand both the achievements of past research in recognizing whispered speech and the experiments performed in this work. Section 3 will describe the methodology of the work, which includes a description of the models that will be trained, the data that will be used to do so, and a description of Ph[]neMask method by which the dataset will be extended. In Section 4, the results of the experiments performed and their description will be provided. Section 5 will provide a discussion of the results obtained. Section 6 will summarize the results of the work and indicate the prospects for future research in the direction of whispered speech recognition.

Background 2

In this chapter a more detailed description of technical aspects relevant to this study will be provided. I will give an overview of phonetics and acoustics of whispered speech that pose challenges for conventional recognition systems, review ASR techniques that emerged and were used on the way to recognizing whispered speech, and provide a description of prior research in adapting normal models to recognize whispered speech. In conclusion, a research question based on gaps in prior studies on whispered speech recognition will be defined.

2.1 Phonetics and Acoustics of Whispered Speech

Whispered speech significantly diverges from normal speech in several key aspects, primarily due to the absence of vocal fold vibration, which leads to distinctive phonetic and acoustic characteristics. These differences affect the spectral envelope and the temporal patterns of speech, making it challenging for models trained on normal speech to accurately recognize whispering. Therefore, before training models to recognize whispered speech, it is crucial to establish these differences through phonetic and acoustic analysis of whispering and only after that find effective ways to incorporate this knowledge into the models.

2.1.1 Acoustic Properties of Whispered Speech

General differences that pose difficulties for systems trained on neutral speech to recognize whispering can be easily noticed at the visual representations of speech signal. Shown below are the waveforms, spectrograms, and spectrums of the phrase *"Each stag surely finds a big fawn"* pronounced in normal and whispered mode, respectively:

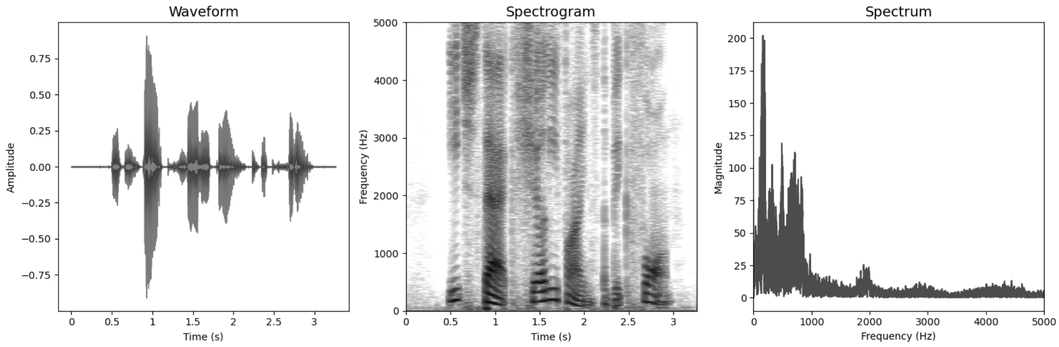

Figure 2.1: "*Each stag surely finds a big fawn*": normal mode

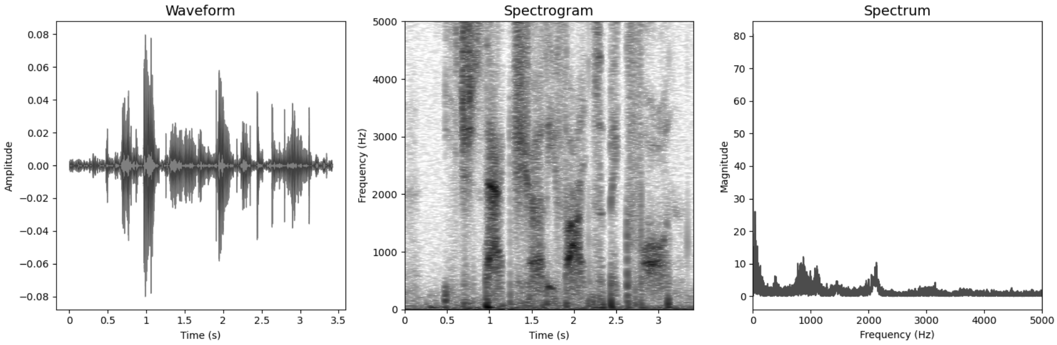

Figure 2.2: "*Each stag surely finds a big fawn*": whisper mode

From these plots, the following features that differentiate whispered from phonated speech can be determined:

1. Absence of f_0. The most important difference is the absence of the fundamental frequency - the frequency at which the vocal folds vibrate when voiced speech sounds are made, which on the spectrogram of normal speech appears as a black band along the bottom of the spectrogram (can be see well at 2.0s).

2. Lower frequencies degradation. Lower frequencies tend to degrade in whispered speech, their structure is not as fine as on the spectrogram of neutral speech as can be seen at the structure of the 500 Hz formant at 0.5s. Higher frequencies, in turn, are still preserved, which can be seen at the 2000 Hz formant at 1.0s. This occurs because the absence of vocal fold vibrations reduces the energy in the lower frequency range, while the turbulence generated in the vocal tract during whispering still maintains the higher frequency components.

3. Formants shift. Experiments in (Kallail and Emanuel 1984) showed that there is a systematic increase in the first three formants of English vowels in whispered speech compared to neutral speech. Our picture also demonstrates this shift well at 1.0s: the third formant is at 2000 Hz for normal speech, and around 100 Hz up in case of whispering.

(Kallail et al. 1984): *‘Formant-frequency differences between isolated whispered and phonated vowel samples produced by adult female subjects.’*

4. More noise. Whispered speech has a higher level of noise compared to that of normal speech, which can be noticed on the spectrogram of whispered speech being "grayer". One possible reason for that can also be attributed to the lack of vocal fold vibrations in whispering, which generate a clear, harmonic structure of voiced speech. Additionally, the turbulent airflow produced during whispering introduces further noise into the signal caused by the partially closed glottis during whispering, which leads to a hissing sound.

5. Longer duration. As better seen on the waveforms, while the phrase pronounced in normal mode took 3 seconds (even less excluding silences), the whispered one took around 3.5s. This extended duration of whispered speech can be attributed to the reduced lower vocal intensity, which then necessitates a slower articulation to maintain clarity and intelligibility. Whispered speech appears to involve a more deliberate and careful pronunciation to compensate for the absence of vocal fold vibration that create prominent and loud sounds, leading to a naturally prolonged utterance.

6. Flatter spectrum. The right plots of spectrums demonstrate that whispered speech has a flatter spectrum compared to normal phonated speech. In regular speech, when the vocal folds vibrate, they create a fundamental frequency and harmonics that result in peaks in the spectral envelope. In contrast, whispering relies solely on turbulent airflow through a partially closed glottis, producing a more uniform distribution of energy across frequencies. This turbulence lacks the periodicity that voiced sounds have, leading to a spectrum with less pronounced peaks and a relatively flatter shape.

Despite these differences, however, the intelligibility of whispering remains maintained. This is mostly due to the preservation of key articulatory movements and the reliance on turbulent airflow to produce sound. Even without vocal folds vibration, the distinct shapes and positions of the tongue and lips continue to form the consonants and vowels that make up intelligible speech, and these articulatory cues are exactly what is enough for recognizing speech sounds. Enough for human - but, turns out, not quite so for machines, as they are still not sufficient for ASR models trained with normal data to recognize whispered speech effectively without prior adaptation.

Nevertheless, even though the fundamental articulatory movements are preserved in whispered speech, studies of phonetics of whispered speech have shown that there are particular groups of sounds that tend to differ from normal speech more than others.

2.1.2 Phonetic Peculiarities of Whispered Sounds

This way, the work (Sharifzadeh, McLoughlin, and Russell 2012) conducted a large-scale comparative study of the vowel spaces in English in whispered and normal speech modes and came to the conclusion that in whisper mode, the greatest shifts in formants occur in central open-mid (up to 24% uprise) and close-mid vowels (up to 52% uprise) (p. 53, keywords: 'significant shifts'), i.e., those where the tongue is positioned halfway between an open and mid position for open-mid vowels (/ɛ/, /ɜ/, /ʌ/, /ɔ/, /æ/, and /ɐ/) and between a close and mid position for close-mid vowels (/e/, /ə/, /o/), rather than in the front-back (like /i/) or open-close vowels (like /a/) that rely heavily on extreme configurations of tongue placement and are thus distinguished easier.

(Sharifzadeh et al. 2012): *'A comprehensive vowel space for whispered speech'*

Regarding consonants, the work (Jovičić and Šarić 2008) conducted a comparative analysis of consonant space in Serbian between whispered and normal speech. This analysis based on place of articulation revealed that the consonants produced at hard palate (/ɲ/, /ʎ/, /ʃ/, /ʨ/, /j/, /ʒ/, /ʤ/, /ʧ/) exhibit the greatest difference in duration between whispered and phonated modes (p. 273, keywords: 'palatal'). This, as the authors note, means that the tongue position at the palatal place of articulation is highly sensitive to vocal cord vibrations, and

(Jovičić et al. 2008): *'Acoustic analysis of consonants in whispered speech'* its absence implies that the articulators need to work harder to create more prominent distinctions between sounds. Same conclusions are found for English in (Osfar 2011), where it is noted that a more careful placement of the tongue on the hard palate is taking place when producing whispered English sounds (p. 60, keywords: 'higher precision movement'), highlighting that these differences are universal and do not depend on the language in question, but characterize whispering speech as a whole.

(Osfar 2011): *'Articulation of whispered alveolar consonants'*

Such seemingly minor peculiarities of particular groups of sounds together make up an overall acoustic picture demonstrated in the previous section that is significantly different from normal. It is these differences that, on the one hand, posit challenges for, but on the other give clues on adapting ASR models that are "used to" neutral speech.

However, to fully appreciate the adaptation methods used for whispered ASR, it is essential to first briefly describe the ASR technology in general that was developed for the recognition of normal speech. These foundational techniques described in the following section, as will be shown later, have been adopted and applied to the specific challenges posed by whispered speech.

2.2 Automatic Speech Recognition

Broadly speaking, automatic speech recognition is a technology that translates spoken language into written text, facilitating smoother communication between humans and machines. A diagram of a typical speech recognition system is shown below:

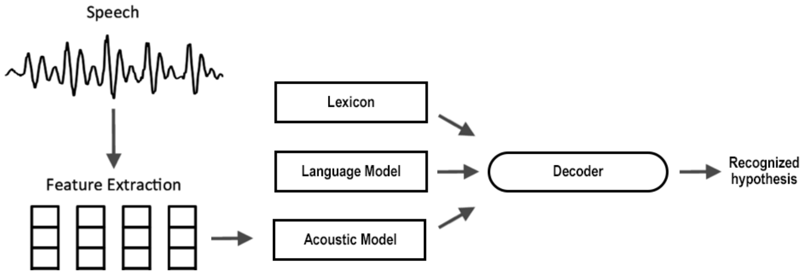

Figure 2.3: *Typical speech recognition pipeline.*

The front-end processing module processes an audio signal to extract relevant acoustic features (usually, it is Mel-Frequency Cepstral Coefficients, or MFCCs, that represent the short-term power spectrum of a sound using a mel scale to mimic human auditory perception) which are then fed into the Acoustic Model which determines the likelihood of an acoustic feature corresponding to a phoneme. The Language Model calculates the probability of a word sequence by analyzing the relationships between words based on the training text. The Decoder constructs a search graph by substituting the word tokens from the Language Model with the corresponding phonetic sequences from the Lexicon and, finally, combines the scores from the Acoustic Model and Language Model to generate the most probable hypothesis.

First speech recognition systems with such architecture appeared in the middle of the 20th century and were based on the ideas of dynamic programming. However, they are out of scope for this research, as by the time whispered speech gained attention in context of automatic recognition, the technology had already shifted towards Hidden Markov Models. It is on these models that the first, and, indeed, most works devoted to whispered speech recognition are based, and thus a short introduction to it is necessary to better understand the adaptation techniques that were adopted from normal ASR to whispered speech recognition.

2.2.1 Hidden Markov Models

The breakthrough in the field of ASR that paved the way to its wider and commercial use occurred in the mid-1980s, when Hidden Markov Models (HMM) were first applied to the task of speech recognition. As noted in (Rabiner 1990), a shift *"from simple pattern recognition methods based on templates and a spectral distance measure to a statistical method for speech processing"* had happened and marked a significant step forward. It was the mainstream approach to speech recognition tasks until about 2010, and the first works in whispered speech recognition were also based on HMMs. In this approach, the speech signal is considered a random pattern that needs to be recognized, and thus the task of speech recognition essentially was a classical pattern classification problem based on the maximum aposteriori probability criterion.

(Rabiner 1990): *'A tutorial on hidden Markov models and selected applications in speech recognition.'*

These systems were significantly superior to those based on the dynamic programming method. However, in case a new speaker was presented to a model, recognition quality dropped substantially, since the feature spaces of each speaker, although coincide at the phoneme level, can differ significantly at the level of phonemes realizations - at sounds themselves.

Many works have been devoted to the solution of this problem, the general goals of which were:

- To remove any channel effects that are speaker-dependent and thus are not significant for recognizing speech.
- To train the existing model with new features without losing useful properties that it learned during initial training.

On the way to achieving these goals, two main methods have emerged: feature normalization and model adaptation.

2.2.1.1 Feature Normalization

Feature normalization aims at distorting the input speech signal or feature vectors extracted from it in order to converge on the average characteristics with the vectors that were initially used to train the model. For this purpose, Vocal Tract Length Normalization (VTLN) (Molau, Kanthak, and Ney 2000) is usually used.

(Molau et al. 2000): *'Efficient vocal tract normalization in automatic speech recognition.'*

2.2.1.2 Model Adaptation

Model adaptation, in turn, refers to shifting and distorting not the data being fed to the model, but the model itself, i.e., the probability density functions of the states, to best fit the new speech data. For this purpose, Maximum Likelihood Linear Regression (MLLR) proved to be an effective algorithm of adaptation (Leggetter and Woodland 1995). From the field of speaker recognition came the method of adapting HMM models using eigenvoices (Kuhn et al. 1998). For adapting models to noise, Vector Taylor Series (VTS) are used (Acero et al. 2000). All these techniques, as will be shown in Section 2.3, have found application in early works on whispered speech recognition.

(Leggetter et al. 1995): *'Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models.'*

(Kuhn et al. 1998): *'Eigenvoices for speaker adaptation.'*

(Acero et al. 2000): *'HMM adaptation using vector taylor series for noisy speech recognition.'*

Nevertheless, despite the undeniable advantages of HMMs, such as effective modeling of temporal variations of the speech signal and the wide range of techniques to adapt models for new data, these systems had a number of significant drawbacks. These include the assumption that sequences of observation vectors are considered statistically independent, which does not hold true for speech. Moreover, speech, as already mentioned, is an extremely dynamic process, but HMMs are somewhat "piecewise constant" in their nature as stated by (Makovkin 2012), i.e., each state has a stationary statistic, regardless of the time being spent in a given state, the emission probability distributions stay the same. These shortcomings are what prevented further development of these models, which prompted researchers to search for alternative approaches to solving the problem of speech recognition. One such approach turned out to be deep neural networks, whose capabilities seemed to many researchers better corresponding to the nature of the speech recognition task.

(Makovkin 2012): *'Hybrid models - Hidden Markov models / Multilayer perceptron - and their application in speech recognition systems. Review.'*

2.2.2 ASR with Deep Neural Networks

Instead of using probabilistic models as it is done in HMMs, deep neural networks (DNNs) function by mimicking the human brain structure, comprising layers of interconnected neurons, each processing and transforming input data so as to find a function that best describes this data. As noted in (Tampel 2015), the first attempts to apply DNNs to speech recognition were made back in the 90s, but were not successful at that time as failed to outperform the then baseline HMM model. It did not, however, reduce the motivation to continue research in this direction, and the power of modern computers allowed neural networks to set a new standard of quality in the ASR realm.

(Tampel 2015): *'Automatic speech recognition - milestones over 50 years.'*

A relevant feature of DNNs is that the inner layers of the neural network can extract features of speech in general, not only of one language. These layers can then be used for a different, but related task, as confirmed, for example, by experiments with adapting ASR models initially trained on one language for recognizing another (Li et al. 2021, p. 6, keywords: ’extending for different domains’), which could even be from another language family. This process of training a pre-trained neural network model on new data is called fine-tuning. The objective is the same as that of model adaptation in HMM: to use the already existing model’s useful features and improve its performance on the new task by adjusting the model’s parameters slightly.

(Li et al. 2021): *’Scaling end-to-end models for large-scale multilingual asr’*

2.2.2.1 Fine-Tuning

The process of DNN fine-tuning involves taking a pre-trained model, which has already learned a variety of general features from a large and diverse dataset, and further training it on a new dataset that is specific to the desired task. By exposing the model to this new data, it can adjust its parameters to better capture the unique characteristics and nuances of the task at hand. Thus, it allows the model to build on its existing knowledge and quickly adapt to new challenges, improving its performance on the specific task while requiring fewer resources, both computational and with regard to data, compared to training a model from scratch.

However, even though fine-tuning requires less data than training a model from scratch, when the data is extremely limited, successful fine-tuning becomes challenging. When it is the case, an effective technique proved to be useful for speech recognition task is data augmentation, which aims at expanding the dataset for a more effective fine-tuning process with artificial data.

2.2.3 Data Augmentation in ASR

Data augmentation involves creating new training examples by transforming existing data. This not only helps in expanding the size of the training dataset but could also potentially improve the model’s robustness and generalization capabilities, making it a valuable strategy alongside fine-tuning. In ASR tasks, there exist two main augmentation strategies: time-domain and frequency-domain augmentations.

Time-domain augmentations involve altering the temporal aspects of an audio signal in some way, such as time-stretching, which involves changing the speed without affecting the pitch, pitch-shifting, or modifying the pitch without changing the speed, and adding noise or reverberation to simulate different recording conditions.

Frequency-domain augmentations, on the other hand, modify the spectral characteristics of the audio. These techniques include frequency masking (randomly hiding parts of the frequency spectrum) and equalization changes that adjust the balance between different frequency components.

A particularly effective and widely used method that combines these two strategies is SpecAugment, which is essentially frequency masking and time masking applied directly to the spectrogram of an audio signal.

2.2.3.1 SpecAugment

Introduced in (Park et al. 2019), SpecAugment is a technique which involves applying random time warping, frequency masking, and time masking directly to the spectrograms of audio signals during training of a neural network model:

(Park et al. 2019): *'Specaugment: A simple data augmentation method for automatic speech recognition'*

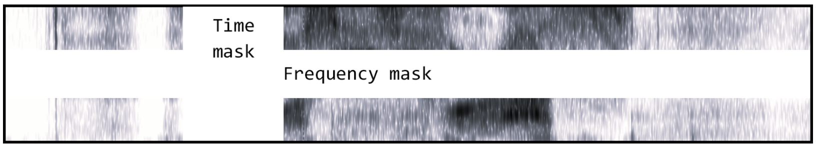

Figure 2.4: *Time- and frequency-masking with SpecAugment*

Time warping shifts the spectral features in time, frequency masking zeroes out random sections of the spectrogram's frequency bands, and time masking zeroes out random sections of the spectrogram's time frames. These augmentations simulate variations in speech patterns and environmental conditions, increasing the volume and diversity of the training data. This leads to enhanced generalization of ASR models by making them less sensitive to variations in the input signal, thereby increasing model's performance in real-world scenarios.

As demonstrated by the authors of the method, on the LibriSpeech[1] dataset, SpecAugment has been shown to reduce WER from 18.1% to 12.9%. On the Switchboard dataset[2], the Word Error Rate (WER)[3] was reduced from 12.9% to 8.5%, which demonstrates the algorithm's effectiveness across different types of speech recognition tasks.

Thus, this technique has been widely adopted in the ASR tasks due to its ability to significantly improve model robustness without the need for additional data. By augmenting the existing instances in a way that mimics realistic variations and noise conditions, SpecAugment helps ASR systems to become more resilient to real-world conditions, which leads to better performance of the model in general.

However, one limitation of SpecAugment is its lack of linguistic awareness. The transformations it applies are purely based on signal characteristics and do not consider the semantic properties of the speech content. To overcome this with bringing more meaning into the masking process, another augmentation technique called Semantic Masking was proposed.

2.2.3.2 Semantic Masking

Inspired by SpecAugment and BERT[4], the paper (Wang et al. 2019) proposes masking the input features corresponding to specific output tokens during training. This way, unlike the random masking in

[1] LibriSpeech is a corpus of approximately 1000 hours of 16kHz read English speech.

[2] Switchboard is a collection of about 2,400 two-sided telephone conversations among 543 speakers from all areas of the United States.

[3] WER is a special metric used in ASR tasks to evaluate model's performance, which in essence is the ratio of errors in a hypothesis to the total words spoken (refer to Section 3.5 for more on WER).

[4] BERT is an NLP model by Google that learns language patterns by masking certain words in text during training and then predicting them, capturing the nuanced relationships between words and their context.

(Wang et al. 2019): *'Semantic mask for transformer based end-to-end speech recognition'*

SpecAugment, this Semantic Masking strategy selectively masks whole segments of features tied to particular words as shown below:

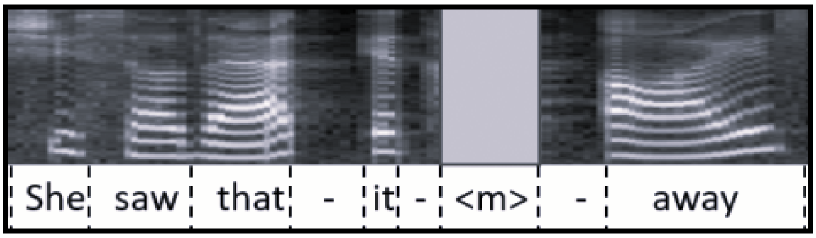

Figure 2.5: *A spectrogram of the phrase "She saw that it moved away" with Semantic Masking applied*

This approach encourages the model to use contextual information to predict the masked tokens, which enhances its language modeling power and robustness to various acoustic distortions.

The authors conducted experiments with Semantic Masking on the LibriSpeech and TedLium2[5] datasets and demonstrated significant improvements in WER with the semantic mask strategy compared to the baseline model: 8.95% vs 7.43% WER on the test-other set of Librispeech, and 10.4% vs 8.5% WER on TedLium2 dataset.

Thus, the Semantic Masking approach significantly enhances the robustness and accuracy of E2E speech recognition models by masking meaningful regions on the spectrogram and employing contextual information to predict them during training.

2.2.4 Summary

Due to the fact that human speech may be very diverse and it is not feasible to train a speech recognition model for each person individually, ASR was faced with the task of bringing the models to some common denominator where they would do a good job at recognizing speech regardless of who speaks and how they speak. At the same time, despite this diversity, the fundamental mechanisms of speech generation do not change from person to person, which makes it possible to preserve useful properties of existing models and adapt them to new conditions, rather than training everything anew. This way, any new data is perceived as a slight "modification" of the data on which the model was previously configured, and therefore it is assumed that in order to teach the model to recognize data from new domain, it is only necessary to slightly modify the model accordingly.

In this context, data augmentation has been particularly useful. By applying simple transformations to the input spectrograms, such as frequency/time or word masking, techniques such as SpecAugment or Semantic Masking are able to expand the dataset in scarcity conditions and enhance the robustness of the model to various variations in speech data.

When the time came to recognize whispered speech, it became clear that it could be viewed in a similar fashion, i.e., as a "modification" of normal speech: the nature of the data remains, but some of its

[5] TedLium2 is an English speech recognition training corpus from TED talks.

acoustic features differ. The following section covers the history of attempts to explain these differences to machine.

2.3 Automatic Recognition of Whispered Speech

If humans can easily recognize whispered speech and machines are already capable of recognizing neutral speech in its variations, is it possible to teach machines to recognize whispered speech as well? It has been several decades since this question first appeared in the minds of researchers, and the interest in the problem of automatic whisper recognition has not waned to this day.

The work that started the road to automatic recognition of whispered speech was a doctoral thesis (Morris 2003). The idea of this research was to see how well a machine would recognize whispering compared to human's abilities with the use of a test called Diagnostic Rhyme Test (DRT). For this purpose, minimal pairs of single syllable words that differed in one phonetic feature (nasality, sibilation, voicing, etc.) were composed and presented to a user for them to determine what they hear between the test and the alternative word. These experiments found that in the case of normal speech, when the role of a user was played by the HMM-based tool called Fast-Talk, it performed worse than when the user was a real human, but it still managed to successfully recognize minimal pairs with all kinds of phonologically distinctive features. In the case of whispered speech, however, the machine failed at telling words in these pairs apart: it was making choice with an accuracy ranging from 38% to 51% (p. 149, keywords: 'ASR DRT scores') depending on the features used for training model. Given the 'choose one out of two' nature of this test, it can be concluded that the machine was making choice randomly.

(Morris 2003): Enhancement and recognition of whispered speech.

At that point, it was realized that the difficulty in training whispered speech recognition models lies largely in the fact that:

- There are significant acoustic differences between whispering and neutral speech.
- There is not as much whispering data for model training as there is normal data available.

Considering these points, the vector of research was directed towards working with models trained with neutral data, which is largely available, and, knowing in what acoustic properties the two modes of speech differ, trying to adapt these models in such a way as to recognize whispered speech. Fortunately, by the time whispered speech recognition gained attention, significant progress had already been made in normalizing features and adapting models for normal speech recognition as discussed in Section 2.2, so for whispered speech, fertile ground was already prepared.

2.3.1 Adapting Models to Recognize Whispered Speech

The fundamental work that highlighted the possibility of adapting models to whispering is (Ito, Takeda, and Itakura 2005), the goal of which was to create a speech recognition system specifically designed to process whispered speech in environments with high levels of noise, such as open offices. They employed HMM to examine various train-test scenarios with neutral and whispered speech, and these experiments revealed a significant decline in model's performance when conducting a mismatched test: the model struggled at recognizing whispering when trained on normal data only with recognition accuracy of approximately 20% (p. 149, keywords: 'recognizing the normal speech').

(Ito et al. 2005): *'Analysis and recognition of whispered speech.'*

However, as the most important outcome of this study, it was demonstrated that ASR systems trained on neutral speech could also be adapted for whispering recognition by incorporating a small dataset of whispered speech to the HMM model and adapting its states using MLLR algorithm (Leggetter and Woodland 1995). The application of such approach resulted in a whispered speech recognition accuracy rate of approximately 70% (p. 149, keywords: 'applying MLLR') in a syllable recognition experiment.

(Leggetter et al. 1995): *'Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models.'*

Subsequent studies have then taken up this idea of adapting existing, trained on normal speech models to recognize whispering by reducing the acoustic mismatch between whispered and normal speech. This way, in the work (Lim 2011) several experiments were conducted with the use of eigenvoices to build whispered speaker-dependent HMMs based on normal data from the same speakers. This approach allowed to achieve recognition accuracy on whispered part of wTIMIT[6] of 66.57% for a speaker-dependent model (p. 111, keywords: 'speaker-dependent').

(Lim 2011): Computational differences between whispered and non-whispered speech.

Nevertheless, (Lim 2011, p. 114, keywords: 'expansion') and, indeed, all studies on adapting models for whispering that I covered starting from the very first work (Morris 2003) discussed above, in results and prospects for future research once again point out the problem of data scarcity and the need to expand the available whispering datasets, which is what most of the current work in the direction of automatic whispered speech recognition is devoted to.

2.3.2 Expanding Whispering Datasets

In order to generate artificial instances for expanding training data, several approaches have been proposed to convert normal speech into whispering employing the knowledge of the acoustics of whispered speech.

[6] wTIMIT is a parallel corpus containing recordings of normal and whispered speech from 50 male and female English speakers from Singapore and the United States. A detailed description of this dataset is provided in Section 3.1.

2.3.2.1 Pseudo-Whisper Generation

This way, in (Ghaffarzadegan, Bořil, and Hansen 2014), the Vector Taylor Series approach coupled with VTLN normalization was probed for transforming neutral speech to pseudo-whispered speech. However, instead of assuming that noisy speech is the sum of neutral speech and noise as it is done in the original paper, in this work it is assumed that neutral speech is the result of whispered speech passed through the channel and corrupted by additive noise. With expanded through VTS dataset, authors report WER decrease from 22% to 18% on the UT-VEII[7] corpus (p. 5, keywords: 'pseudo-whisper samples'). This work was one of the last where HMMs were applied to recognize whispered speech. The following works were centered in the DNN paradigm.

(Ghaffarzadegan et al. 2014): *'Model and feature based compensation for whispered speech recognition.'*

The new approach to artificial whispering generation was attempted with the application of Cycle-Consistent Generation Adversarial Networks (CycleGAN, introduced in (Zhu et al. 2017)) in (Gudepu et al. 2020). In this case, the network comprised two generators, $G_{s \to w}$ and $G_{w \to s}$, and two discriminators, D_s and D_w. $G_{s \to w}$ mapped normal speech to whisper, and $G_{w \to s}$ mapped whisper back to normal speech. Discriminators thus help to ensure that after converting normal speech to whisper and back to normal speech, the original features vector is retained. The so-called adversarial loss is employed to ensure that the generated whisper is indistinguishable from the real whispering sample. Through expanding dataset with the data generated artificially this way, it was possible to reduce WER on whispering set of wTIMIT from 37.1% to 29.4% (p. 2305, keywords: 'comparison').

(Zhu et al. 2017): *'Unpaired image-to-image translation using cycle-consistent adversarial networks.'*

(Gudepu et al. 2020): *'Whisper Augmented End-to-End/Hybrid Speech Recognition System-CycleGAN Approach.'*

The most recent work on whispered speech recognition, (Lin, Patel, and Scharenborg 2023), also made another attempt at converting normal speech into pseudo-whispered speech. The method involved two main steps: first was to removing glottal information, and the second one was to up-shift the formants. Considering absence of f_0 (Section 2.1.2, 'absence of f_0'), a technique called Glottal Inverse Filtering was applied to normal speech to cancel the glottal contributions, resulting in speech with minimal glottal information so as to approximate whispered speech with no vocal folds vibrations. For formants up-shifting (Section 2.1.2, 'formants shift'), Moving Average Filtering was used on the spectral envelope to increase the formant bandwidth and shift the formant frequencies so as to imitate the way formants go up in natural whispered speech. On the dataset augmented with pseudo-whispered speech obtained this way, the authors achieved 38.6% WER on whispered set of wTIMIT.

(Lin et al. 2023): *'Improving Whispered Speech Recognition Performance Using Pseudo-Whispered Based Data Augmentation'*

2.3.2.2 Data Augmentation for Whispered ASR

Regarding data augmentation in whispered speech recognition, research has primarily focused on expanding datasets with artificially generated whispering as shown above. As an addition to this, most of

[7] A dataset of whispered and neutral speech from 37 male and 75 female subjects from (Zhang and Hansen 2009).

these works also rely on traditional augmentation methods, such as SpecAugment, to expand the data further (for instance, (Chang et al. 2021, p. 2, keywords: 'masking') applies SpecAugment for masking out lower frequencies on the spectrograms of whispered speech so as to imitate lower frequencies degradation as described in Section 2.1.2), but have not ventured beyond this technique, although many novel augmentation ideas have emerged recently, such as Semantic Masking.

(Chang et al. 2021): *'End-to-end whispered speech recognition with frequency-weighted approaches and pseudo whisper pre-training.'*

This lack of exploration into other augmentation techniques suggests a potential area for further research in whispered ASR. However, whispered speech required a more nuanced application of these techniques, given its unique acoustic properties as outlined in Section 2.1. This is why I propose another method, Ph[]neMask that is essentially a mixture of SpecAugment and Semantic Masking ideas, where not random frequency/time regions or words, but specific phonemes are masked on spectrograms. These phonemes would be selected based on their significant differences from normal speech, as those that presumably pose the greatest difficulties for ASR models trained with normal speech.

2.3.3 Phone Masking

The idea behind Ph[]neMask is to mask out specific groups of sounds on the spectrogram. Similarly to how random time-masking in SpecAugment works, here timestamps corresponding to specific phonemes are masked out. For example, the spectrogram of the phrase "*This was easy for us*" pronounced in whispered mode without any masking looks as follows:

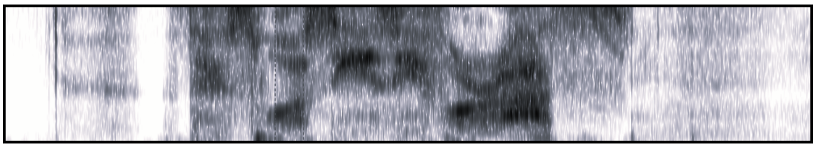

Figure 2.6: *Normal spectrogram of the phrase "This was easy for us" pronounced in whispered mode*

If Ph[]neMask is set, for instance, to mask all close-mid central vowels, the resulting spectrogram would look as follows:

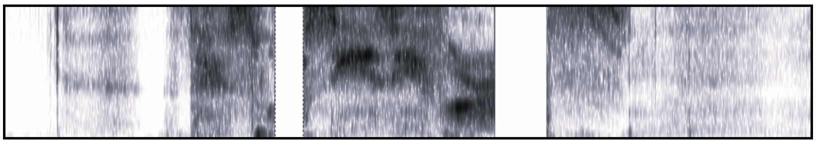

Figure 2.7: *Spectrogram of the phrase "This was easy for us" pronounced in whispered mode with masked out central close-mid vowels*

After applying phone masking, all regions with sounds corresponding to central close-mid phonemes, in this case only the /ə/ is such, have been cut out.

It is known, from the experience of SpecAugment as described in Section 2.2.3.1, that masking specific regions makes model learn cues for these regions from contextual information. By masking out specified groups of vowels that pose special difficulties for the pre-trained model, it is expected that it would encourage model to learn more about them from their surroundings which would result in better speech recognition capabilities.

2.3.4 Summary

The literature review showed that a considerable number of methods of model adaptation have been adopted from normal ASR and proven to be effective for whispered speech recognition, ranging from adaptation of HMM models to whispering with statistical techniques such as MLLR to generating pseudo-whispered speech for expanding existing datasets for fine-tuning DNNs originally trained with normal speech. Yet, there is still a gap in the application of data augmentation methods to whispered speech. However, the success of techniques such as SpecAugment and SemanticMasking in improving neutral speech recognition suggests a promising avenue for whispered speech.

The unique phonetic and acoustic properties of whispered speech outlined in Section 2.2 call for a nuanced application of these techniques. This is why instead of masking bands of frequencies or time regions as it is done in SpecAugment, or entire words as suggested by SemanticMasking, I propose a PH[]NEMASK method for masking specific types of phonemes on the spectrograms of whispered speech that are most different from those in normal speech. It is expected that with this approach, model would learn better to extract contextual cues for masked regions and thus improve at recognizing whispered speech in general.

2.4 Research Questions

In light of the considerations above, the following question arises:

RQ Is it possible to improve the quality of whispered speech recognition by fine-tuning a pre-trained model with a dataset augmented through masking specific phonemes on the spectrograms of whispered speech?

Consequently, this research will explore if PH[]NEMASK, when applied to the unique characteristics of whispered speech, can enhance ASR accuracy. Based on successful applications of SpecAugment's time masking in (Park et al. 2019) and considering systematic phonetic differences between whispered and normal sounds as evidenced by (Sharifzadeh, McLoughlin, and Russell 2012) and (Jovičić and Šarić 2008), I hypothesize that masking palatal consonants and mid vowels, as the most different groups of sounds between normal and whispered speech, can improve the baseline quality of whispered speech recognition.

(Park et al. 2019): *'Specaugment: A simple data augmentation method for automatic speech recognition'*

(Sharifzadeh et al. 2012): *'A comprehensive vowel space for whispered speech'*

(Jovičić et al. 2008): *'Acoustic analysis of consonants in whispered speech'*

The next section will describe the methodology of the experiments which will allow me to confirm or reject this hypothesis.

Methodology 3

In this section, I overview the methodology employed in the present study, describe the details of data, PH[]NEMASK augmentation strategies and the fine-tuning plan implemented for model optimization.

3.1 Data

The dataset employed for experiments in this thesis is wTIMIT, a parallel corpus of whispered and normal speech first presented in the work (Lim 2011).

(Lim 2011): Computational differences between whispered and non-whispered speech.

The process of collecting this data occurred in two phases: the first phase involved 20 Singaporean speakers, and the second phase included 28 North American speakers, resulting in two subsets differing only in accent. All recordings were made in an audiometric booth using an MX-2001 directional condenser microphone, positioned 15cm from the speaker's mouth and slightly tilted to prevent air puffs from hitting the microphone. During whispering, speakers were instructed to move closer to the microphone for a better dynamic range. Each speaker was asked to both whisper and read normally a set of 450 prompts from the phonetically balanced section of the TIMIT corpus, ensuring coverage of common phonetic contexts in spoken English. Considering ethical reasons as described in Section 1.2, prompts were alternately read and whispered in sets of 50 to minimize speaker fatigue.

Initially, this dataset was divided into random train/test sets for training purposes. However, since many speakers uttered the same sentences, it led to train/test overlap. It poses a significant challenge for the DNN-based framework that I employ in this work, in which the training and test sets must be completely different to ensure an accurate evaluation of the model's performance on unseen data. To address this issue, the re-partition of the dataset into train/dev/test sets was done, with each set containing 400, 25, and 25 sentences respectively with no overlaps between the three sets. This division was made in accordance with the work (Chang et al. 2021, p. 3, keywords: 'partition'), and I thank its author, Mr. Heng-Jui Chang from Massachusetts Institute of Technology, for kindly providing the division logic.

(Chang et al. 2021): *'End-to-end whispered speech recognition with frequency-weighted approaches and pseudo whisper pre-training.'*

3.2 Process of Phone Masking

To perform augmentation with phone masking, it is first required to extract exact timestamps for each phoneme in all recordings of the dataset. For this purpose, Montreal Forced Aligner (MFA) was employed. Using pre-trained acoustic models, MFA understands the phonetic characteristics of the input speech. A pronunciation dictionary maps words in the text transcripts to their phonetic representations, helping the aligner understand how the words should sound. Using HMMs, the system then aligns each segment of the audio with the corresponding text at a very fine-grained level, from words down to individual phonemes. The result is a detailed alignment that shows the precise start and end times for each phoneme in the audio file.

With this instrument, for each recording from the whispered part of wTIMIT .TextGrid files with aligned timestamps were extracted that look as follows:

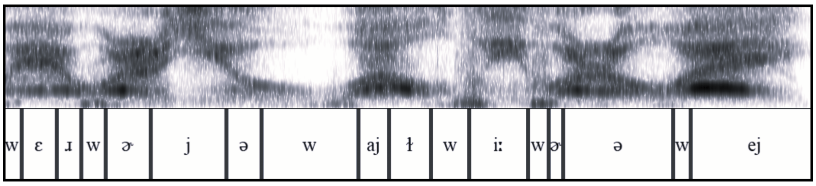

Figure 3.1: *A spectrogram of the phrase "Where were you while we were away?" with forced alignment of phones*

After obtaining alignments, I developed an instrument in Python for masking out phonemes on the spectrograms. A dictionary of phonemes was built, so it is possible to choose the type of phonemes to be masked. After the choice is made, it is also possible to define the probability that a phoneme will be masked out.

3.2.1 Masking Strategies

Following findings on differences in vowels and consonant spaces between whispered and normal speech as described in Section 2.1.1, two strategies will be tested:

- Vowel Space: masking out mid vowels
- Consonant Space: masking out palatal consonants

This way, in the first experiment, I will augment the dataset with instances where mid vowels would be masked: /ɛ/, /ɜ/, /ʌ/, /ɔ/, /æ/, /ɐ/, /e/, /ə/, /o/. The other experiment would involve masking out consonants produced at the hard palate: (/ɲ/, /ʎ/, /ʃ/, /j/, /ʒ/, /ʤ/, /ʧ/).

The model on which the fine-tuning experiments will be conducted is OpenAI's Whisper, currently the most advanced speech recognition model available publicly.

3.3 OpenAI Whisper

Whisper (Radford et al. 2023) is an automatic speech recognition system trained using 680,000 hours of multilingual supervised data sourced from the Internet. The architecture of this model is as follows:

(Radford et al. 2023): *'Robust speech recognition via large-scale weak supervision'*

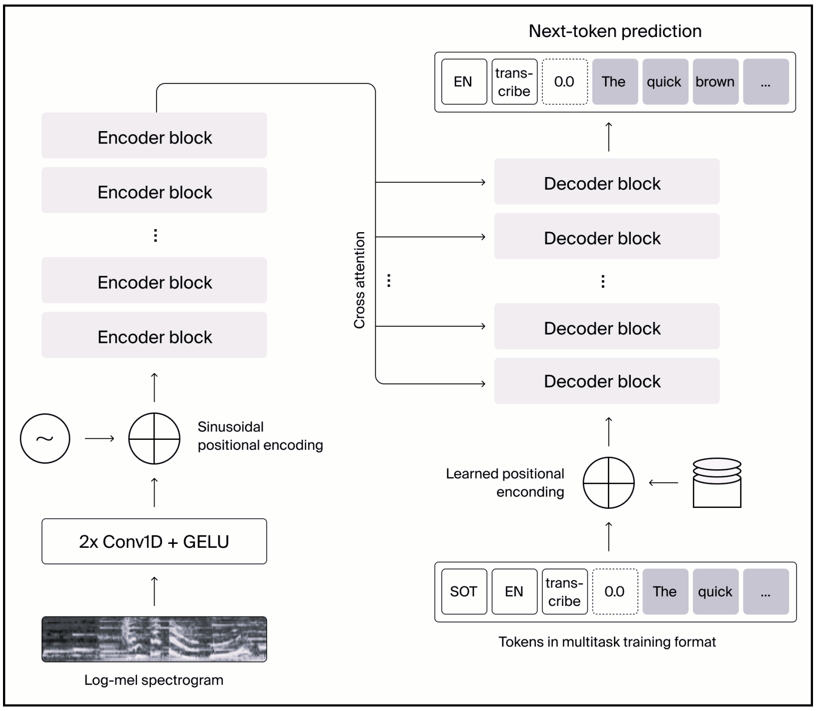

Figure 3.2: *Architecture of Whisper model*

As can be seen, Whisper employs a straightforward end-to-end approach. It processes input audio by dividing it into 30-second segments, converting these segments into a log-Mel spectrogram, and feeding them into an encoder. A decoder is then trained to generate the corresponding text, incorporating special tokens to instruct the model to carry out speech transcription (in fact, language identification and phrase-level timestamping are also possible).

There are five model sizes that Whisper offers: tiny (39M parameters), base (74M), small (244M), medium (769M), and large (1550M). In this work, Whisper-small was chosen for experiments as a model that has a significant number of parameters but does not require heavy computational resources.

3.4 Fine-Tuning Strategies

The fine-tuning process will go as follows. Initially, the base Whisper model will be fine-tuned using the normal portion of the wTIMIT dataset to evaluate the performance on recognizing whispering on a model acclimated to the data within the dataset (such as accents), yet without any adaptation to whispering itself. This model will not be used in subsequent experiments, but will allow for the evaluation of how Whisper handles whispering speech without any prior adaptation.

The subsequent phase will involve fine-tuning the base Whisper model using the whispered portion of the dataset without incorporating any augmented data. This approach will allow for the assessment of the model's performance on whispered speech in its natural form, only when presented whispered speech. The recognition accuracy of this model will serve as a reference point for further experiments.

Finally, datasets augmented with masked data will be constructed and used for fine-tuning the base Whisper model, to evaluate the impact of augmentation using phone masking on the model's performance in recognizing whispered speech.

3.5 Evaluation

To evaluate the quality of the fine-tuned models, Word Error Rate (WER) metric will be used, which quantifies how accurately a model transcribes spoken language by comparing the model's output to a reference transcription. It is computed using the following formula:

$$\text{WER} = \frac{S + D + I}{N} \tag{3.1}$$

where S is the number of substitutions (words incorrectly transcribed), D is the number of deletions (words omitted from the transcription), I is the number of insertions (extra words added to the transcription), and N is the total number of words in the reference transcription. Thus, the lower the WER, the better the performance of the model.

After all WERs are calculated, in order to investigate the statistical significance of their changes across different augmentation strategies compared to the baseline model, the Matched-Pair Sentence-Segment Word Error (MAPSSWE) proposed by (Gillick and Cox 1989) will be applied. As outlined in (Barfuss et al. 2017, p.20, keywords: 'MAPSSWE'), the MAPSSWE test uses aligned reference and model's output strings to identify segments containing misclassified content. These segments are established by finding the regions bounded on both sides by words that both systems get correct.

(Gillick et al. 1989): *'Some statistical issues in the comparison of speech recognition algorithms'*

(Barfuss et al. 2017): *'Robust coherence-based spectral enhancement for speech recognition in adverse real-world environments'*

Let us consider this example (from Jurafsky and Martin 2009, p. 17, keywords: 'MAPSSWE'):

(Jurafsky et al. 2009): Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition

```
         I             II              III                IV
REF:  |it was|the best|of|times it|was the worst|of times|  |it was
      |      |        |  |        |             |        |  |
SYS A:|ITS   |the best|of|times it|IS the worst |of times|OR|it was
      |      |        |  |        |             |        |  |
SYS B:|it was|the best|  |times it|WON the TEST |of times|  |it was
```

In the first region, the system A has made two errors (insertion and deletion), and the system B has zero errors. In the third region, system A and one error (substitution), and system B has two errors. Then, the following variables are defined: N_A^i (the number of errors made on segment i by system A), N_B^i (the number of errors made on segment i by system B), and Z ($N_A^i - N_B^i$, $i = 1, 2, \ldots, n$, where n is the number of segments). With these values, the variance of Z_i's can be calculated:

$$\sigma_z^2 = \frac{1}{n-1} \sum_{i-1}^{n} (Z_i - \mu_z)^2 \tag{3.2}$$

Then, W is the statistic defined as:

$$W = \frac{\hat{\mu_z}}{\sigma_z / \sqrt{n}} \tag{3.3}$$

Thus, with these actions, the test calculates the number of errors in each segment for each system and then tests the null hypothesis that the mean difference in word errors per segment between the two systems is zero.

To conduct this test, I utilize the implementation provided by the National Institute of Standards Scoring Toolkit (NIST 2016).

(NIST 2016): National Institute of Standards and Technology Scoring Toolkit, version 2.4.10

Experiments 4

In this chapter, I present the results of the experiments conducted*. I include plots depicting the training history of the models, as well as plots illustrating the Word Error Rates on the dev set of wTIMIT, and the accuracies of these models on the test set of wTIMIT. Finally, I provide a comprehensive analysis of the WERs obtained on the test set of the wTIMIT dataset to establish statistical significance of the results.

4.1 Preliminary Setup

To start with, a base Whisper-small model without any prior fine-tuning was tested on the test set of wTIMIT, which yielded the following results:

$\mathbf{N_{US}}$	$\mathbf{N_{SG}}$	$\mathbf{W_{US}}$	$\mathbf{W_{SG}}$
4.9	13.16	13.85	30.75

Table 4.1: WER(%) on Whisper-small model without any fine-tuning.

It can be observed that the base model handles normal mode of speech with US accent well, but experiences more difficulties with the normal speech with Singaporean accent. A similar pattern is seen in the case of whispering, with expectedly higher WERs for both accents. In fact, Singaporean accented speech is among the most difficult ones for ASR systems to recognize, as shown by (Winata et al. 2020, p. 4, keywords: 'Singapore'), so such behaviour of Whisper is expected.

(Winata et al. 2020): *'Learning fast adaptation on cross-accented speech recognition'*

Fine-tuning Whisper on the normal part of wTIMIT took around 600 steps before achieving a minima:

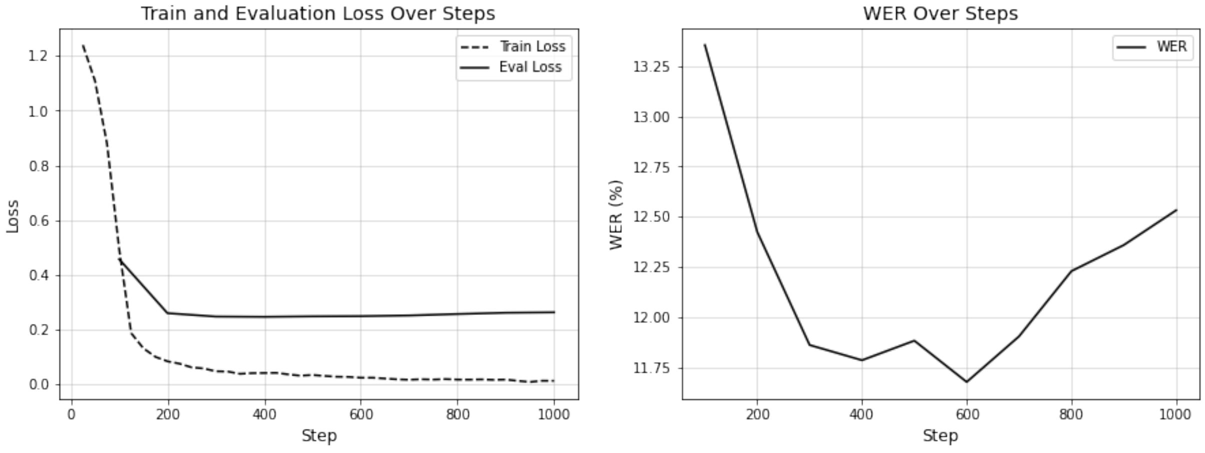

Figure 4.1: *Whisper fine-tuning using normal part of wTIMIT.*

* Available on GitHub: https://github.com/marczenko/phonemask

After approximately 600 steps, the train loss plateaus, suggesting that the model has extracted most of the useful information from the training data and further improvements are minimal[1]. Similar situation is observed with eval loss. Around the 600-steps mark, the eval loss stabilizes and afterwards even shows a slight increase, which indicates the onset of overfitting. The right graph, which tracks the WER, shows a significant decrease up to about 600 steps as well, indicating that the model's predictions are becoming more accurate. However, beyond 600 steps, the WER begins to rise slightly, reinforcing the idea that the minima is reached.

In summary, by around 600 steps, the model has learned enough, as evidenced by the stabilization of evaluation loss. Continuing training beyond degrades the model's performance on evaluation data due to overfitting, as evidenced by WER increase afterwards. This is expected behaviour due to base Whisper model being trained on a substantial amount of English data, so only a small number of fine-tuning steps with the data from the same domain, English in neutral mode in this case, is already enough to reach the performance plateau.

On the test set of wTIMIT, this model yields the following results:

N_{US}	N_{SG}	W_{US}	W_{SG}
5.1	10.9	13.2	26.3

Table 4.2: WER(%) on Whisper-small model fine-tuned with normal part of wTIMIT.

Compared to the model without any prior fine-tuning (Table 4.1), the WER for standard US English has slightly increased, while performance in all other modes has improved. This is also expected since the base Whisper model is already well-acquainted with US English data, leading to overfitting after a certain point when more US data is introduced. However, for Singaporean English data, there is a significant improvement in both normal and whisper modes. This, in my opinion, is a reasonable trade-off between a minor decline in US performance and a substantial gain in Singaporean performance to continue experimenting with Whisper-small[2].

4.2 Fine-Tuning

In this section, experiments with adapting the Whisper model to whispered speech are conducted.

4.2.1 Baseline

The first experiment with fine-tuning is to introduce both the normal and whispered sets of wTIMIT to the model. It would allow us to evaluate the influence of each augmentation strategy afterwards. Below shown the plots of fine-tuning Whisper with the entire training set of wTIMIT:

[1] Several experiments were conducted to select the appropriate hyperparameters, and learning rate of 1e-6, batch size of 64, dropout of 0.1 and weight_decay of 0 have proven to work best.

[2] To confirm whether introducing only Singaporean data to the model would increase the performance on Singaporean accent without degrading it on the US accent, another experiment was conducted with fine-tuning the base model solely with Singaporean data. Not introducing the US into training set significantly degraded the accuracy on US accent:

N_{US}	N_{SG}	W_{US}	W_{SG}
5.9	12.3	15.69	27.14

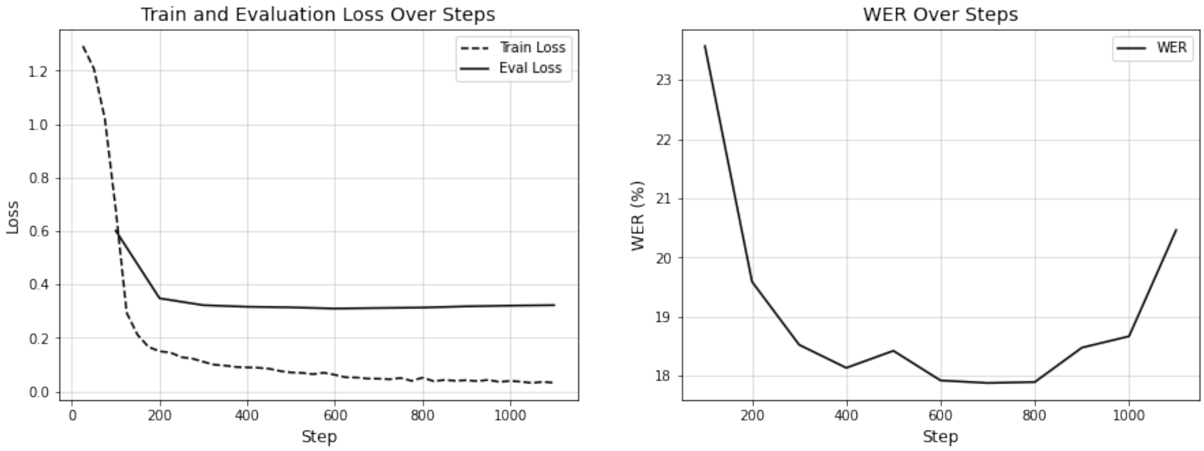

Figure 4.2: *Whisper fine-tuning using whispered and normal parts of wTIMIT with no augmentation.*

In this case, the plateau is achieved at around 600-800 steps, where both evaluation loss and WER on dev set get to their minima. Afterwards, the case of overfitting can be observed as evidenced by both eval loss and WER increase, so no further training improves the quality. The 800-steps checkpoint yields following WERs on the test set:

N_{US}	N_{SG}	W_{US}	W_{SG}
5.2 ↑	11.1 ↑	12.26 ↓	22.76 ↓

Table 4.3: WER(%) on Whisper-small model fine-tuned with normal and whispered parts of wTIMIT.

As can be seen, introducing the whispered part of wTIMIT to the model slightly degrades the quality of normal speech recognition. At the same time, the quality of whispered speech recognition is improved, which is not as prominent in the case of US-accented whispering (7.2% relative decrease), but is especially noticeable on the example of Singaporean accented whispered speech, where the 13.46% decrease was achieved. This suggests that the model adapts to the Singaporean accent quite quickly, while stagnating in the case of the American accent due to the fact that Whisper is trained on a huge amount of US-accented data and therefore bumps against its performance limit in the case of fine-tuning it with more US-accented speech. Still, for whispered speech, improvement for both accents is noted, and further experiments will be compared to these results.

4.2.2 Applying Vanilla SpecAugment

To have a reference point from the world of augmentation techniques before my experiments with Ph[]neMask method, an experiment using vanilla SpecAugment was conducted first. For this, default values that Whisper-small was pre-trained with were used: both time- and frequency-domain random masking with probability of 0.05 were applied to the whole dataset. The training process with this dataset took 800 steps to reach the best performance:

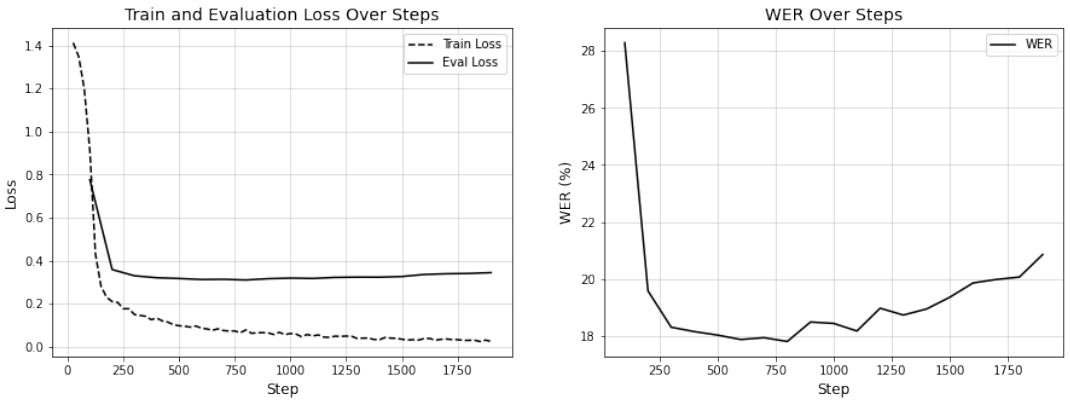

Figure 4.3: *Whisper fine-tuning using wTIMIT dataset with SpecAugment.*

With this model, the following WERs are achieved on the test set:

Table 4.4: WER(%) on Whisper-small model fine-tuned with SpecAugment.

N_{US}	N_{SG}	W_{US}	W_{SG}
5.4 ↑	12 ↑	11.5 ↓	23.5 ↑

For normal speech, a decline in performance can be observed for both accents: 3.7% increase in case of US-accented speech, and 8.1% increase for Singaporean-accented part. For whispered speech, however, a significant decrease is achieved for the US-accented part, with 6.19% relative decrease of WER. Singaporean accented whispering, in turn, shows an increase of 3.25%.

4.2.3 Masking Mid Vowels

In this experiment, the model is fine-tuned with the dataset augmented with audios where all mid vowels are masked out.

Similar situation is observed on the plots of learning history. The model achieves its best results at around 700 steps, and further training does not lead to any improvement, as evidenced by both eval loss history, which goes up after 700 steps, and by history of WER which on the dev set is the lowest at 700 steps checkpoint:

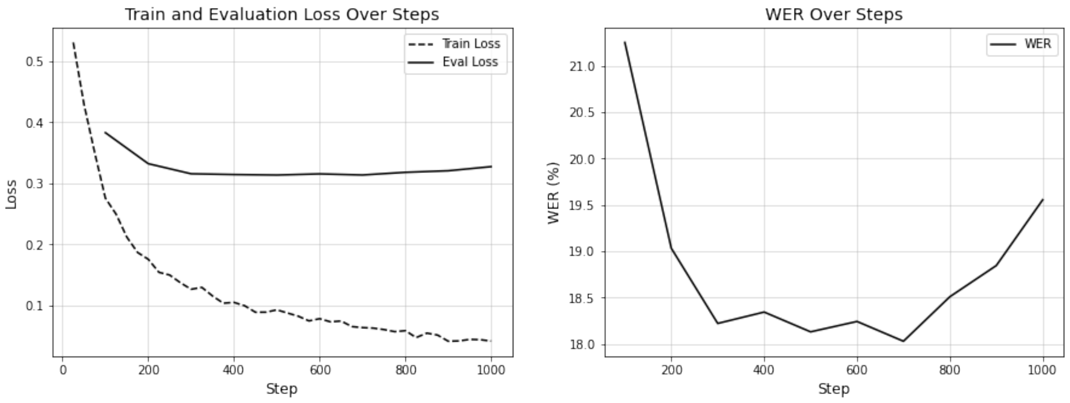

Figure 4.4: *Whisper fine-tuning using dataset augmented with spectrograms with masked mid vowels.*

This model yields the following results on the test set:

$\mathrm{N_{US}}$	$\mathrm{N_{SG}}$	$\mathrm{W_{US}}$	$\mathrm{W_{SG}}$
5.79 ↑	12.03 ↑	12.83 ↑	23.45 ↑

Table 4.5: WER(%) on Whisper-small model fine-tuned using dataset augmented with spectrograms with masked mid vowels.

Compared to the model without augmentation, this strategy shows an increase in WER for all accents and modes. This way, for normal speech, it is 10.96% relative WER increase for the US-accented part, and 8.73% for the SG-accented part. For whispering, it is 4.65% increase for the US-accented speech, and 3% for the SG-accented speech. Thus, masking all mid vowels did not help in improving recognition of whispered speech.

4.2.4 Masking Palate Consonants

In this experiment, the default dataset was extended with spectrograms of whispered speech with masked palatal consonants. This model achieves the best performance after 800 training steps, and further fine-tuning leads to deteriorating performance:

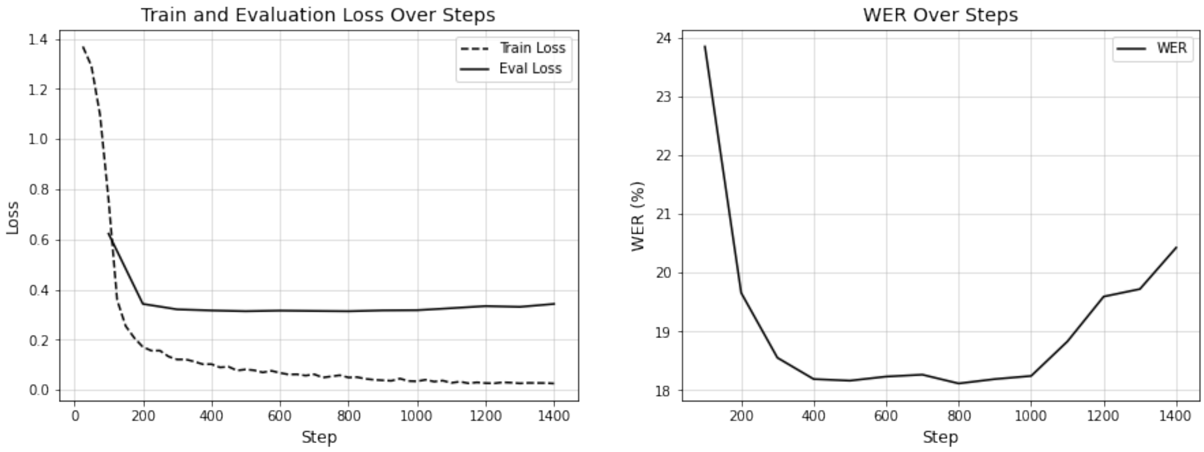

Figure 4.5: *Whisper fine-tuning using dataset augmented with spectrograms with masked palatal consonants.*

The following WERs are achieved with this model:

$\mathrm{N_{US}}$	$\mathrm{N_{SG}}$	$\mathrm{W_{US}}$	$\mathrm{W_{SG}}$
5.77 ↑	11.87 ↑	12.12 ↓	22.96 ↑

Table 4.6: WER(%) on Whisper-small model fine-tuned using dataset augmented with spectrograms with masked palatal consonants.

On the test set, for normal speech in a US accent, the WER increased slightly to 5.77% from the baseline's 5.2%, showing a 10.96% relative increase. In the case of normal speech in a Singaporean accent, the WER also increased, reaching 11.8% compared to the baseline's 11.1% (6.3% relative increase). Thus, palatal-masked dataset slightly worsened the performance for normal speech in both US and Singaporean accents.

For whispered speech, however, the WER improved marginally for the US-accented part, decreasing to 12.12% from the baseline's 12.26%.

This indicates an enhancement in performance for whispered US-accented speech using the palatal consonants masking, but this enhancement is extremely slight (1.14% relative decrease). For whispered speech in a Singaporean accent, the WER increased to 22.9% from the baseline's 22.76%, showing a slight decline in performance of 0.6% relative WER increase.

4.3 Summary

Overall, every fine-tuning strategy showed that the model achieves plateau at around 800 steps and further training led to no improvement. Summary table of the obtained WERs is shown below:

Table 4.7: WER(%) on the test set of wTIMIT, summary

Model	N_us	N_sg	W_us	W_sg
Preliminary Setup				
> Whisper-small (no fine-tuning)	4.9	13.16	13.85	30.75
> normal wTIMIT	5.1	10.9	13.2	26.3
No augmentation				
> normal wTIMIT + whisper wTIMIT	5.2	11.1	12.26	22.76
Augmentation				
> normal wTIMIT + whisper wTIMIT + SpecAugment	5.4	12	11.5	23.5
> normal wTIMIT + whisper wTIMIT + mid_vowels PhoneMask	5.79	12.03	12.83	23.45
> normal wTIMIT + whisper wTIMIT + palatal_cons PhoneMask	5.77	11.87	12.12	22.96

The statistical significance of the obtained results has been checked with MAPSSWE. Null hypothesis is that there is no performance difference between the two systems, significance level is chosen at p < 0.05. Marked with * are the two systems between which a statistically significant difference is established, marked with ~ are the systems with no statistically significant differences.

For the US-accented part the following results are obtained:

Table 4.8: MAPSSWE test for US-accented part of whispered set

	Vowels	Consonants	SpecAugment
No aug.	0.144 ~	0.603 ~	0.033 *
Vowels		0.049 *	0.001 *
Consonants			0.050 *

As can be seen, significant difference is established in the case of SpecAugment, and, as the Table 4.7 reports, this difference lies in the outperformance both the model with no augmentation (11.5% vs 12.26%), and the Ph[]neMask strategies (11.5% vs 12.83% and 12.12% for palatal consonants and mid vowels masking, respectively). Yet, the comparison between palatal consonants masking and SpecAugment is right at the significance threshold, which means that the difference between them is not as prominent. Moreover, the difference between palatal consonants masking and vowel masking is also established, and from the Table 4.7 it is noticeable that the consonants masking strategy outperforms vowel masking (12.12% vs 12.83% WER). Thus, even though no statistical difference between masking consonants and no augmentation strategies is established, I can conclude that masking consonants is a more promising strategy than masking mid vowels.

For the Singaporean-accented part, the following results are obtained:

Table 4.9: MAPSSWE test for SG-accented part of whispered set

	Vowels	**Consonants**	**SpecAugment**
No aug.	0.194 ∼	0.704 ∼	0.187 ∼
Vowels		0.342 ∼	0.928 ∼
Consonants			0.258 ∼

In this case, no significant difference is established across all strategies probed. It means that with the Singaporean accent, none of strategies allowed to achieve any performance change, neither for the better, nor for the worse, although a severe decline in performance can be seen in Table 4.7 compared to no augmentation strategy. In general, Singaporean accented speech turned out to be the most difficult task for Whisper to handle, presumably requiring other approaches than augmentation, which are out of scope for this work, but prospects for which will be covered in the following section which discusses the results of the experiments.

Discussion 5

In this chapter, I will discuss the results obtained and offer explanations for their significance.

5.1 General Observations

Fine-tuning the Whisper-small model with both normal and whispered part of wTIMIT allowed to reach WER on the whispered test set of 12.26% for the US-accented speech, and 22.76% for the Singaporean accented speech, which appears to be the best recognition accuracy for wTIMIT reported so far.

However, neither of the two probed masking strategies have proven to be effective for improving the recognition accuracy for whispered speech, thus, the answer to the research question:

RQ Is it possible to improve the quality of whispered speech recognition by fine-tuning a pre-trained model with a dataset augmented through masking specific phonemes on the spectrograms of whispered speech?

is no with the current experimental setup, and the hypothesis that masking out mid vowels or palatal consonants could lead to a better whispered speech recognition is rejected. I attribute this to the following.

Despite that there is a systematic difference in the formants of mid vowels and duration of palatal consonants between whispered and normal speech as stated by (Sharifzadeh, McLoughlin, and Russell 2012), (Jovičić and Šarić 2008), and (Osfar 2011), masking exactly these phoneme groups may not have sufficiently addressed the broader spectrum of acoustic variations present in whispered speech as discussed in Section 2.1.2. While mid vowels and palatal consonants groups indeed different between normal and whispered speech, the overall spectral characteristics of whispered speech are altered in a more global manner, and by focusing on specific phoneme groups, the two masking strategies may have accidentally neglected other crucial aspects of whispered speech, such as the general up-rise of formants, the overall narrower spectral shape and its noisier nature, all of which are vital for effective speech recognition as shown by number of works on whispered speech, including (Ito, Takeda, and Itakura 2005) and (Lim 2011).

(Sharifzadeh et al. 2012): *'A comprehensive vowel space for whispered speech'*

(Jovičić et al. 2008): *'Acoustic analysis of consonants in whispered speech'*

(Osfar 2011): *'Articulation of whispered alveolar consonants'*

(Ito et al. 2005): *'Analysis and recognition of whispered speech.'*

(Lim 2011): Computational differences between whispered and non-whispered speech.

Still, although no statistical significance was found for the two augmentation methods compared to no augmentation, when compared to each other, masking specifically palatal consonants marginally outperformed masking mid vowels, as confirmed by MAPSSWE with p=0.049, which is a boundary value, but with the threshold chosen is considered significant. I attribute this to the following.

First of all, palatal consonants have clear articulatory boundaries, making it easier for the model to learn about the specific regions where these consonants occur. The articulatory precision of palatal consonants thus both makes it possible to mask out exactly these sounds without affecting the adjacent sounds, and provides a more defined learning target for the model. Mid vowels, on the other hand, have more diffuse and overlapping acoustic features, highly influenced by the sounds on the left and to the right. This less distinctive nature of mid vowels makes it harder for the model to learn specific patterns associated with these regions, leading to less improvement compared to masking consonants. An introspection experiment with pronouncing palatal consonants (for example, the phoneme /j/) highlights difference in the constriction of the tongue with the palate - in whispering, this constriction is significantly more pronounced, supporting the finding of "higher precision movement" by (Osfar 2011). At the same time, an experiment with pronouncing mid vowels would not disclose as noticeable differences, other than that the vocal folds do not vibrate. Thus, consonants contain more pronounced differences between normal speech, while vowels have such subtle differences which are then reflected in as well subtle acoustic peculiarities, that the usual masking in the time domain appears to be insufficient.

(Osfar 2011): *'Articulation of whispered alveolar consonants'*

This is also proved by SpecAugment which has beaten both of the masking strategies, achieving 11.5% WER on the US-accented whispered part of wTIMIT, which is a statistically significant improvement over the baseline. This is also most likely due to the fact that the greatest difficulty for normal models in speech recognition lies in the frequency domain, in which SpecAugment also does masking along with time-domain zeroing. As shown in Section 2.1.1, pretty much every difference between whispered and normal speech is caused by the absence of vocal cord vibrations. One of the most prominent such differences is the degradation of the lower frequencies, which poses particular difficulties for models trained on normal speech, and SpecAugment seemed to slightly overcome this difficulty. Since Ph[]neMask focuses specifically on the time domain, an extension to the frequency domain might be beneficial.

Additionally, it can be observed that the models struggled with recognizing Singaporean-accented speech in both normal and whispered modes, compared to relatively good performance on the US-accented speech. Despite this being out of scope for this research, I would like to also attribute such performance to several factors. Singaporean English, as noted by (Prabhu et al. 2023), has distinct phonetic and prosodic features, including unique vowel and consonant realizations, tonal variations, and influences from other languages such as Malay, Mandarin, and Tamil (p. 124, keywords: 'imprint'). These features deviate significantly from the standard English pronunciation patterns that Whisper is predominantly trained on. Whispered speech further

(Prabhu et al. 2023): *'Accented Speech Recognition With Accent-specific Codebooks'* amplified these differences, making it even more challenging for the model to accurately recognize it. Thus, the scarcity of training data specifically tailored to Singaporean-accented English, all the more so in whispered form, have led to poor generalization and reduced recognition accuracy, and even fine-tuning with wTIMIT data did not help the model to get closer to the recognition accuracy of US-accented speech. As a solution to this problem, it appears that additional fine-tuning of the model on Singapore English data, provided for example by the National Speech Corpus (IMDA n.d.) or The SUSS Corpus of Singapore English (SUSS n.d.), is worth considering.

(IMDA n.d.): National Speech Corpus (NSC), Infocomm Media Development Authority

(SUSS n.d.): The Singapore University of Social Sciences Corpus of Singapore English

Moreover, introducing whispered speech to the model degraded the accuracy of recognizing normal speech, and including more whispered data with masked phones degraded it even further. While the goal of this research was to specifically train the model to recognize whispering, and fine-tuning the base Whisper model with whispered part of wTIMIT allowed for improvement of the recognition accuracy of whispered speech itself, and employing palatal masking and SpecAugment slightly improved it further, the degraded accuracy for normal speech can lead to usability problems of such model. An ideal recognition system should be able to recognize both whispered and normal speech accurately, as a user may want to use the system, for instance, with a loud voice during the day, and with a quiet voice at night. This inconsistency in performance thus could hinder the usability and reliability of such ASR system in real-world scenarios. A potential solution to this problem could be creating a system that first identifies the mode of speech, whether it is voiced or whispered, and then passes it to a specialized model that performs better for the detected mode. This approach would ensure that the system remains versatile and effective in handling both types of speech, providing a better user experience regardless of the communication context.

In conclusion, while the idea of masking mid vowels and palatal consonants on spectrograms is grounded in phonetic research, its practical implementation may not have translated into significant improvement of recognition accuracy for whispered speech due to the complex nature of acoustic changes in whispered speech coupled with possible shortcomings of Whisper fine-tuning process, which will be discussed in limitations of this work.

However, despite statistical significance of the obtained results is not established, comparison of the errors yielded by the models is still relevant as would highlight the weak and strong points of the resulting models.

5.2 Notes on Recognition Errors

In this section, the descriptions of some errors made by the models on the whispered set will be provided that are of special interest for the task. For this, the results from the model with no augmentation, and from the models with augmentations will be compared. Since

Singaporean accented speech poses its unique challenges, the two accents will be described separately.

5.2.1 US-accented speech

The first sentence that is worth analysis is the phrase "The rich should invest in black zircons instead of stylish shoes" pronounced in whispered mode with US accent:

Table 5.1: Recognition of the phrase "The rich should invest in black zircons instead of stylish shoes" by all models

Reference	the rich should invest in black zircons instead of stylish shoes	
Model	**Hypothesis**	**WER**
No Augmentation	the ridge should infest in black circums instead of stylish shoes	0.27
SpecAugment	the rich should invest in black zircons instead of stylish shoes	0
Vowels Augmented	the rich would invest in black circumsents instead of stylish shoes	0.18
Consonants Augmented	the rich should invest in black zircons instead of stylish shoes	0

Interestingly, every augmentation strategy in this case allowed for the correct recognition of phonologically unvoiced consonants, such as [ʧ] in *rich*, and phonologically voiced consonants, such as and [v] in *invest*, whereas the model with no augmentation hypothesized these pairs incorrectly. However, the model with vowels masked augmentation and that with no augmentation made mistakes in the word *zircons*. Both these models recognized [s] in place of phonologically voiced [z], and the former completely misrecognized the word, as *circumsents* does not exist in English and could have been caused by the fusion of *zircons* and the following word.

Similarly, errors such as *ridge* and *infest*, in this case indeed have a phonetic basis, since in whispered speech phonologically voiced and voiceless consonants cannot be distinguished. Nevertheless, it is obvious that these words do not fit the context semantically. The augmented models made no such errors, even though only specific parts of words were masked in them, which suggests that Semantic Masking applied in its original form of masking entire words has the potential to improve the generalization properties of the whispered model even further.

Another case worth attention is the sentence "Each stag surely finds a big fawn":

Table 5.2: Recognition of the phrase "Each stag surely finds a big fawn" by all models

Reference	each stag surely finds a big fawn	
Model	**Hypothesis**	**WER**
No Augmentation	each stack truly finds a big fun	0.42
SpecAugment	each stack surely finds a big fawn	0.14
Vowels Augmented	each stack truly finds a big fun	0.42
Consonants Augmented	each stack surely finds a big fun	0.28

As can be seen, in this case, every strategy failed to correctly recognize the phonologically voiced consonant [g] in the word *stag*, compared to the sentence above where [ʧ]-[ʤ] and [f]-[v] were distinguished successfully. This difficulty with recognizing [g] correctly could be attributed to the fact that in whispered speech, the turbulent air flow is more prominent as described in Section 2.1.1, and thus somewhat of a false aspiration is happening which in normal speech distinguishes plosives along with presence of vocal folds vibrations. Yet, while vowels augmented model and the model with no augmentation confused the word *surely* with the word *truly* due to the affricatization of [t] before [r] into [ʧ], the model with masked consonants, along with SpecAugment, hypothesized the word *surely* correctly. This suggests that masking palatal consonants allowed the model to learn more about these consonants and slightly improved its performance in this case.

Finally, it is SpecAugment that correctly recognized the word *fawn*, while the two other strategies and the model with no augmentation confused it with *fun*. It is exactly the case with mid vowels, and the mid vowels masking strategy still recognized this sound incorrectly, confusing [ɔ] with [ʌ]. It once again highlights that time-masking strategy is insufficient in case of vowels.

5.2.2 SG-accented speech

Singaporean accented English speech poses its own difficulties for the ASR task as described in the previous section. In some cases, it is quite hard to define what exactly led to wrong recognition: whether the whispered speech itself, or its accent peculiarities, and this is why it is worth analyzing it separately from the US-accented speech. Still, some speculation is possible. Let us consider this example:

Table 5.3: Recognition of the phrase "The 5th jar contains big juicy peaches" by all models

Reference	the 5th jar contains big juicy peaches	
Model	**Hypothesis**	**WER**
No Augmentation	the fixture contains big juicy peaches	0.28
SpecAugment	the fixture contains big juicy peaches	0.28
Vowels Augmented	the fixture contains big juicy peaches	0.28
Consonants Augmented	the fixture contains big juicy peaches	0.28

In this case, it is both the Singaporean accent and whispered speech that caused the problem. On the one hand, the Mandarin influence led to [θ] turning into alveolar [s] (also known as th-alveolarization; however, it is also reported that [θ] can often be substituted with [t] in Singaporean English (Moorthy and Deterding 1997, p. 76, keywords: 'replacement') also known as th-stopping, but in that case something like *feature* could have been expected in place of *fixture*), and on the other, voiced [ʤ] of *jar* was recognized as voiceless [ʧ] because of the whispered mode. Unfortunately, no augmentation strategy allowed to improve the recognition of this sentence.

(Moorthy et al. 1997): *'Three or tree? Dental fricatives in the speech of educated Singaporeans'*

A similar situation is observed in the following example, where the phrase "Please, sing just the club theme" is pronounced:

Table 5.4: Recognition of the phrase "Please, sing just the club theme" by all models

Reference	please sing just the club theme	
Model	**Hypothesis**	**WER**
No Augmentation	please sync just the club team	0.33
SpecAugment	please sync just the club team	0.33
Vowels Augmented	please sync just the club team	0.33
Consonants Augmented	please sync just the club team	0.33

In this case, it is also both the problems caused by whispered mode, and the interference of Mandarin and English can be noticed. This way, the word *sing* was in all cases recognized as *sync* for two reasons: first, the speaker did not pronounce the normal [ŋ] as the English nasalization did not occur here, but pronounced the [ng] cluster instead, in which tbe final [g] was then recognized as [k] due to whispering. The other word, *theme*, was recognized as *team* because of th-stopping, differing from the th-alveorization as in the previous case. Similarly, no augmentation strategy helped to overcome any of these problems, highlighting once again that Singaporean accented speech needs other approaches to handle.

5.2.3 Summary

The error analysis has shown that one of the weakest points of all the models is difficulty with recognizing phonologically voiced and unvoiced plosives and sibilants in whispered speech. Since it is difficult to correctly distinguish them in a purely acoustic manner, improved language models to correctly recognize the word out of context and replace it if necessary may prove beneficial. This, on the one hand, would slow down online recognition process, but would significantly enhance the overall transcription quality. Nevertheless, in some cases, augmentation allowed for the correct recognition of palate consonants, as in case of [ʧ] and [ʤ], which indicates that in the frequency domain the data about these consonants are well preserved, and time-domain masking alone was sufficient to improve the accuracy of their recognition.

The analysis of errors in recognizing whispered speech with Singapore accent showed that there is an overlap between the difficulties of whispered speech and the difficulties in recognizing English speech with specific accent features. Moreover, there is inconsistency in some accent features between different speakers: for example, in case of pronouncing /θ/, th-alveolarization may occur for one speaker while th-stopping may occur for another. Such inconsistency further complicates the process of whispered speech recognition, making simple data augmentation not sufficient here.

Lastly, in all cases of US-accented whispering considered, SpecAugment outperformed the masking strategies that have been tested. It implies that masking frequencies for the whispered speech is a more promising strategy than masking only time bands, which I will discuss in the next section that concludes the work and offers prospects for future research.

Conclusions 6

In this chapter, I will present the conclusion of the study, discuss its limitations and offer avenues for future research.

6.1 Study Outcomes

In the present study, I have tested a method of augmenting whispered data by masking certain types of phonemes that are most different from how they are pronounced in normal speech. Supported by SpecAugment's successful experience in masking random time regions to improve the generalization properties of the model, I expected that masking specifically palatal consonants and mid vowels could improve whispered speech recognition performance by forcing the model to learn more detailed contextual cues to the sounds that corresponds to these phoneme types. However, my hypothesis was not confirmed, and the quality of whispered speech recognition, although increased in the case of recognizing the US-accented part of wTIMIT using palatal consonants masking, has not proven to improve significantly. I attributed such results to the observation that masking in the temporal domain is insufficient to efficiently address whispered speech peculiarities, and the success of SpecAugment's masking of both time and frequency bands compared to masking palatal consonants or mid vowels along the entire spectrum demonstrated this. Yet, comparing the two Ph[]neMask strategies against each other, palatal consonants masking proved to be the more promising strategy. This is due to the fact that palatal consonants have a clear articulatory structure, which allows for their unambiguous distinction in the speech stream and mask them more accurately, forcing the model to pay attention to these particular well-defined regions. Mid vowels, on the other hand, are significantly influenced by neighboring sounds, and it is difficult to establish and mask exactly these boundaries. In general, as the error analysis showed, the biggest problem for all the models is separating phonologically unvoiced consonants from their voiced counterparts, and palatal consonants augmentation in some cases allowed to overcome this problem, but not statistically significantly. Mid vowels, on the other hand, for all the complexity of their structure and subtle differences in articulation between them compared to consonants and even to extreme front/back vowels, require more sophisticated approaches with masking not just their temporal boundaries, but certain frequency zones, presumably lowest ones, as being most influenced by the absence of the vocal folds vibration.

6.2 Limitations

The present study has a number of limitations that may have affected the final result.

First, it is possible that some imperfections are present in the dataset. As the author of wTIMIT notes in (Lim 2011, p. 66, keywords: ’quality-control’), although failed records were usually deleted and re-recorded, little number of poor quality records may still have passed through quality control. Indeed, I found about 20 recordings in the dataset from speaker with the identifier 101, for which there were sampling rate problems, and these were deleted prior to experimentation. However, it is admissible that within this dataset there are other audio recordings with similar problems that I did not have the opportunity to identify during experiments.

(Lim 2011): Computational differences between whispered and non-whispered speech.

Limitations also include that the force alignment produced by MFA could be in some cases not completely correct. Although the general trend is that MFA’s forced alignment on whispered speech is surprisingly accurate, there can still be inaccuracies in determining the exact temporal boundaries of the sound corresponding to a phoneme in whispered speech, as MFA acoustic models are trained on neutral data. Moreover, MFA’s English acoustic model is trained predominantly on the US, UK, Nigerian and Indian corpora, which may have led to problems with precise boundaries determining on the Singaporean part of wTIMIT. The combination of these factors may in some cases have resulted in contextual cues, which are important for accurate speech recognition due to the continuity of speech, being removed along with the target sound.

The last important point I would like to make is the reported tendency of Whisper to overfit on small datasets, noticed by several independent developers. Since it is a fresh model, there is no fundamental research confirming or rejecting this yet, but developers’ experiments (Ma et al. 2024, p. 4, keywords: ’performance gap’) and their reviews on GitHub and other platforms frequently report this property of Whisper. The possible reason for this is the huge dataset on which Whisper is pre-trained, which is an absolute record in ASR models, and so for efficient fine-tuning, large amounts of data are also necessary. Future research could explore experiments with different model sizes provided by Whisper, as well as other potential avenues outlined in the final section.

(Ma et al. 2024): *’Extending Whisper with prompt tuning to target-speaker ASR’*

6.3 Prospects for Future Research

Since I tested a method that has not yet been applied to whispered speech in a rather narrow form of masking palatal consonants and mid vowels, the general idea of masking sounds corresponding to specific types of phonemes may still have potential for future research. Thus, as error analysis has shown, masking out other types of consonants, such as plosives, may prove useful. Moreover, combining the three tested strategies (i.e., incorporating in the dataset spectrograms with palatal consonants masked, those with vowels masked, and those with SpecAugment applied) is also a strategy to try.

It is also possible to mask phonemes not over the whole spectrum, but only over a certain region, i.e., perform "block masking" as proposed in (Huang et al. 2022, p. 3, keywords: ’masking strategies’). This would allow, on the one hand, to mask phonemes that differ most significantly from normal speech, but at the same time to concentrate on particular frequency regions - for example, exclusively on low frequencies, as the most strongly differing from normal speech due to the absence of vocal cord vibrations and therefore not having such a pronounced harmonic structure.

(Huang et al. 2022): *’Masked autoencoders that listen’*

The role of whispering in our communication can hardly be overestimated. Consequently, further explorations, including experiments with phone masking method, are essential to enhance our understanding and handling of whispered speech. The horizons for both theoretical and practical advancements in whispered speech recognition remain open and with great promise.

Bibliography

Markides, Andreas (1986). 'Speech levels and speech-to-noise ratios'. In: *British Journal of Audiology* 20.2, pp. 115–120 (cited on page 1).

Lian, Hailun et al. (2019). 'Whisper to normal speech conversion using sequence-to-sequence mapping model with auditory attention'. In: *IEEE Access* 7, pp. 130495–130504 (cited on page 1).

Sharifzadeh, Hamid Reza, Ian Vince McLoughlin, and Farzaneh Ahmadi (2010). 'Speech rehabilitation methods for laryngectomised patients'. In: *Electronic Engineering and Computing Technology*, pp. 597–607 (cited on page 1).

Xiong, W. et al. (2017). *Achieving Human Parity in Conversational Speech Recognition.* (Cited on page 2).

Radford, Alec et al. (2023). 'Robust speech recognition via large-scale weak supervision'. In: *International Conference on Machine Learning*. PMLR, pp. 28492–28518 (cited on pages 2, 20).

Ko, Tom et al. (2015). 'Audio augmentation for speech recognition.' In: *Interspeech*. Vol. 2015, p. 3586 (cited on page 3).

Shahnawazuddin, S et al. (2020). 'Creating speaker independent ASR system through prosody modification based data augmentation'. In: *Pattern Recognition Letters* 131, pp. 213–218 (cited on page 3).

Pervaiz, Ayesha et al. (2020). 'Incorporating noise robustness in speech command recognition by noise augmentation of training data'. In: *Sensors* 20.8, p. 2326 (cited on page 3).

Park, Daniel S et al. (2019). 'Specaugment: A simple data augmentation method for automatic speech recognition'. In: *arXiv preprint arXiv:1904.08779* (cited on pages 3, 12, 18).

Wang, Chengyi et al. (2019). 'Semantic mask for transformer based end-to-end speech recognition'. In: *arXiv preprint arXiv:1912.03010* (cited on pages 3, 12).

Kallail, Ken J and Floyd W Emanuel (1984). 'Formant-frequency differences between isolated whispered and phonated vowel samples produced by adult female subjects.' In: *Journal of Speech, Language, and Hearing Research* 27.2, pp. 245–251 (cited on page 6).

Sharifzadeh, Hamid Reza, Ian V McLoughlin, and Martin J Russell (2012). 'A comprehensive vowel space for whispered speech'. In: *Journal of voice* 26.2, e49–e56 (cited on pages 7, 18, 31).

Jovičić, Slobodan T and Zoran Šarić (2008). 'Acoustic analysis of consonants in whispered speech'. In: *Journal of voice* 22.3, pp. 263–274 (cited on pages 7, 18, 31).

Osfar, Megan J (2011). 'Articulation of whispered alveolar consonants'. PhD thesis. University of Illinois at Urbana-Champaign (cited on pages 8, 31, 32).

Rabiner, Lawrence R (1990). 'A tutorial on hidden Markov models and selected applications in speech recognition.' In: *Readings in Speech Recognition*, pp. 267–296 (cited on page 9).

Molau, Sirko, Stephan Kanthak, and Hermann Ney (2000). 'Efficient vocal tract normalization in automatic speech recognition.' In: *Proc. of the ESSV'00*, pp. 209–216 (cited on page 10).

Leggetter, Christopher J and Philip C Woodland (1995). 'Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models.' In: *Computer speech & language* 9.2, pp. 171–185 (cited on pages 10, 15).

Kuhn, Roland et al. (1998). 'Eigenvoices for speaker adaptation.' In: *Fifth International Conference on Spoken Language Processing* (cited on page 10).

Acero, Alex et al. (2000). 'HMM adaptation using vector taylor series for noisy speech recognition.' In: *INTERSPEECH*, pp. 869–872 (cited on page 10).

Makovkin, K. (2012). 'Hybrid models - Hidden Markov models / Multilayer perceptron - and their application in speech recognition systems. Review.' In: *Speech Technologies* 3, pp. 58–83 (cited on page 10).

Tampel, I. B. (2015). 'Automatic speech recognition - milestones over 50 years.' In: *Scientific and Technical Journal of Information Technologies, Mechanics and Optics* 15.6, pp. 957–968 (cited on page 10).

Li, Bo et al. (2021). 'Scaling end-to-end models for large-scale multilingual asr'. In: *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, pp. 1011–1018 (cited on page 11).

Morris, Robert W (2003). *Enhancement and recognition of whispered speech.* Georgia Institute of Technology (cited on pages 14, 15).

Ito, Taisuke, Kazuya Takeda, and Fumitada Itakura (2005). 'Analysis and recognition of whispered speech.' In: *Speech communication* 45.2, pp. 139–152 (cited on pages 15, 31).

Lim, Boon Pang (2011). *Computational differences between whispered and non-whispered speech.* University of Illinois at Urbana-Champaign (cited on pages 15, 19, 31, 39).

Ghaffarzadegan, Shabnam, Hynek Bořil, and John HL Hansen (2014). 'Model and feature based compensation for whispered speech recognition.' In: *Fifteenth Annual Conference of the International Speech Communication Association* (cited on page 16).

Zhang, Chi and John HL Hansen (2009). 'Advancements in whisper-island detection within normally phonated audio streams.' In: *INTERSPEECH*, pp. 860–863 (cited on page 16).

Zhu, Jun-Yan et al. (2017). 'Unpaired image-to-image translation using cycle-consistent adversarial networks.' In: *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232 (cited on page 16).

Gudepu, Prithvi RR et al. (2020). 'Whisper Augmented End-to-End/Hybrid Speech Recognition System-CycleGAN Approach.' In: *INTERSPEECH*, pp. 2302–2306 (cited on page 16).

Lin, Zhaofeng, Tanvina Patel, and Odette Scharenborg (2023). 'Improving Whispered Speech Recognition Performance Using Pseudo-Whispered Based Data Augmentation'. In: *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, pp. 1–8 (cited on page 16).

Chang, Heng-Jui et al. (2021). 'End-to-end whispered speech recognition with frequency-weighted approaches and pseudo whisper pre-training.' In: *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 186–193 (cited on pages 17, 19).

Gillick, Laurence and Stephen J Cox (1989). 'Some statistical issues in the comparison of speech recognition algorithms'. In: *International Conference on Acoustics, Speech, and Signal Processing,* IEEE, pp. 532–535 (cited on page 22).

Barfuss, Hendrik et al. (2017). 'Robust coherence-based spectral enhancement for speech recognition in adverse real-world environments'. In: *Computer Speech & Language* 46, pp. 388–400 (cited on page 22).

Jurafsky, Daniel and James H Martin (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (cited on page 22).

NIST (2016). *National Institute of Standards and Technology Scoring Toolkit, version 2.4.10.* https://github.com/usnistgov/SCTK (cited on page 23).

Winata, Genta Indra et al. (2020). 'Learning fast adaptation on cross-accented speech recognition'. In: *arXiv preprint arXiv:2003.01901* (cited on page 24).

Prabhu, Darshan et al. (2023). 'Accented Speech Recognition With Accent-specific Codebooks'. In: *arXiv preprint arXiv:2310.15970* (cited on page 32).

IMDA (n.d.). *National Speech Corpus (NSC), Infocomm Media Development Authority*. `https://www.imda.gov.sg/how-we-can-help/national-speech-corpus` (cited on page 33).

SUSS (n.d.). *The Singapore University of Social Sciences Corpus of Singapore English*. `https://susscse.suss.edu.sg/SUSSCSE/` (cited on page 33).

Moorthy, Shanti Marion and David Deterding (1997). ‘Three or tree? Dental fricatives in the speech of educated Singaporeans’. In: *Malay* 1, p. 1 (cited on page 35).

Ma, Hao et al. (2024). ‘Extending Whisper with prompt tuning to target-speaker ASR’. In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 12516–12520 (cited on page 39).

Huang, Po-Yao et al. (2022). ‘Masked autoencoders that listen’. In: *Advances in Neural Information Processing Systems* 35, pp. 28708–28720 (cited on page 40).