Improving whispered speech recognition using pseudo-whispered based data augmentation

Improving whispered speech recognition using pseudo-whispered based data augmentation

Thesis

for the purpose of obtaining the degree of Master of Science
in Electrical Engineering
at Delft University of Technology
to be defended publicly on Tuesday 29, August 2023 at 10 o'clock

by

Zhaofeng LIN

Faculty of Electrical Engineering, Mathematics & Computer Science,
Delft University of Technology, Delft, Netherlands
born in Dongguan, China

Thesis committee:

Dr. O. Scharenborg,	Delft University of Technology
Dr. J. Dauwels,	Delft University of Technology
Dr. T.B. Patel,	Delft University of Technology

So therefore I dedicate myself, to my art, my sleep, my dreams, my labors, my suffrances, my loneliness, my unique madness, my endless absorption and hunger because I cannot dedicate myself to any fellow being.

Jack Kerouac

Contents

Abstract

Whispering, characterized by its soft, breathy, and hushed qualities, serves as a distinct form of speech commonly employed for private communication and can also occur in cases of pathological speech. The acoustic characteristics of whispered speech differ substantially from normally phonated speech and the scarcity of adequate training data leads to low automatic speech recognition (ASR) performance. This project aims to build an ASR system that can recognize both normal and whispered speech and discover which acoustic characteristics of whispered speech have an impact on whispered speech recognition. In my study, I use signal processing techniques that transform the spectral characteristics of normal speech to those of pseudo-whispered speech, called pseudo-whispered-based data augmentation. I enhance an End-to-End ASR system by incorporating pseudo-whispered speech and state-of-the-art (SOTA) data augmentation methods, speed perturbation and SpecAugment, yielding an 18.2% relative reduction in word error rate compared to the strongest baseline. Results for the accented speaker groups in the wTIMIT database show the best results for US English. Further investigation uncovers that the lack of pitch in whispered speech has the largest impact on the performance of whispered speech ASR.

Index Terms: Whispered speech, pseudo-whisper, end-to-end speech recognition, signal processing

1

Introduction

1.1 Motivation

Whispering is a unique form of human speech characterized by distinct acoustic properties compared to normally phonated speech. It serves the purpose of engaging in private and subtle conversations or avoiding disturbances in environments like libraries or meetings. In addition to its common applications, whispered speech also occurs in pathological speech contexts: speech from individuals who face vocal system challenges such as diseases affecting the vocal folds or post-larynx surgery [1, 2] often is whisper-like. Moreover, whispering has been found to be beneficial in reducing or avoiding stuttering [3].

Automatic Speech Recognition (ASR) systems have become ubiquitous in our daily lives, particularly when integrated into virtual assistants on devices like smartphones and home devices (e.g., Apple Siri and Amazon Alexa). While ASR is primarily designed for normal voice interactions, there are situations that necessitate whispered speech to ensure privacy and subtlety during conversations, especially in sensitive or confidential matters.

In recent years, End-to-End (E2E) ASR approaches such as Connectionist temporal classification (CTC) [4], Attention-based encoder-decoder [5], Hybrid CTC/Attention [6] and Conformer models [7] have achieved state-of-the-art performance and become a trend in ASR community. There are several major advantages of E2E models over traditional hybrid models [8]. One of the greatest advantages is that E2E models use a single objective function consistent with the ASR objective to optimize the whole network, while traditional hybrid models optimize individual components separately, which cannot guarantee the global optimum.

However, ASR models are predominantly trained and employed for normally phonated speech, and they struggle to perform well on whispered speech [9, 10]. Whispered speech exhibits distinctive characteristics compared to phonated speech, primarily due to the absence of vocal fold vibrations and glottal source excitation. Additionally, whispered speech displays unique features such as an upward shift in formant frequencies of vowels [9, 11–14], wider formant bandwidths for whispered vowels [9, 11, 12], a flatter spectrum for voiced consonants [9], and lower energy [15]. These distinct acoustic characteristics pose difficulties for ASR systems trained on normally phonated speech to recognise whispered speech, creating a mismatched train/test scenario. Moreover, the limited availability of whispered speech data for training often leads to unsatisfactory performance when ASR systems are trained specifically on whispered speech.

To address these challenges, various methods have been proposed in the literature to improve whispered speech recognition. One of the first research focusing on the automatic recognition of whispered speech is started by Ito *et al.* [9] in 2005. They first introduced a specific ASR system for whispered speech, designed for communication over mobile phones in noisy environments. Their research explored different scenarios involving mismatched training and testing, showing severe degradation in ASR when using mismatched data (trained on normal speech and tested on whispered speech). And they also showed that ASR systems that are trained on normal speech can be adapted to recognize whispered speech by using a small amount of whispered speech data, which can improve the accuracy of whispered speech recognition to around 66%.

After that, several approaches have been proposed in the literature to improve whispered speech ASR. These methods include using Teager Energy Cepstral Coefficients instead of traditional Mel-frequency cepstral coefficients (MFCCs) [16], showing large improvements for Serbian whispered speech. [12] proposed a method to generate pseudo-whispered speech segments using denoising autoencoders showing considerable performance improvements on their own dataset. In recent years, Chang *et al.* [17] showed that a system trained with a frequency-weighted SpecAugment, a frequency-divided Convolutional Neural Network extractor, a layer-wise transfer learning approach, and pre-training outperformed their baseline with about 44% relative improvement in character error rate on whispered speech from wTIMIT [18]. Gudepu *et al.* [19] generated whispered speech from normal speech using Generative Adversarial Networks-based voice conversion (VC) techniques for training data augmentation obtaining the current best results on wTIMIT: 29.4% word error rate (WER). Other methods used multimodal data including articulatory cues from motion data [2, 20] and visual information [21]. Although these techniques show that it is possible to improve the recognition of whispered speech, there is still a performance gap with normal speech.

Even though we manage to improve the recognition performance of whispered speech, it is still worth discovering which acoustic characteristics of whispered speech have an impact on whispered speech recognition and how they lead to a performance gap between normal and whispered speech recognition.

1.2 Research Questions

This thesis addresses two primary objectives:

1. Dealing with the data scarcity problem by generating artificial whispered speech to augment the training data for improved E2E whispered speech ASR;

2. Understanding the (detrimental) effects of the specific acoustic characteristics of whispered speech on whispered speech recognition performance.

Regarding the second objective, using signal processing techniques would be a suitable approach as it allows for specific modification of one or multiple acoustic features of speech. And signal processing techniques offer the advantage of not requiring extra data for training. To that end, signal processing methods are considered to convert normal speech to pseudo-whispered speech in two independent steps. These steps allow us to create “intermediate forms” of normal-to-whispered speech, which in turn allow us to investigate the effect of the specific acoustic characteristics of whispered speech on whispered speech ASR.

Specifically, the research questions are:

- **RQ1:** Can we improve ASR systems performance for whispered speech by generating artificial whispered speech data through signal processing techniques as additional training data?
- **RQ2:** Which and to what extent do acoustic characteristics of whispered speech impact whispered speech recognition performance?

1.3 Outline

In this thesis, Chapter 2 provides detailed background information for this work. Chapter 3 introduces the datasets, proposed method, and experimental setup. Chapter 4 presents the experimental results. Chapter 5 discusses the results, answers the RQs, talks about limitations and future research, and gives conclusions.

2

Background

In this chapter, I provide detailed background information for this thesis. First, the introduction of whispered speech and the acoustic comparison between normal and whispered speech is given in Section 2.1. Then, in Section 2.2, basic signal processing methods are introduced, including the source-filter model and linear predictive coding. ASR-related knowledge, i.e. traditional ASR, E2E ASR, and dictionary are explained in Section 2.3, evaluation metrics in Section 2.4, and data augmentation techniques in Section 2.5.

2.1 Whispered speech

Whispered speech has characteristics that differ from normally phonated speech. Normally phonated speech is produced by modulation of the airflow from the lungs by the vibrations of the vocal folds, while there is no vibration of the vocal folds in whispered speech. Due to the absence of vocal fold vibrations, whispering lacks the fundamental frequency of the voice and much prosodic information. Furthermore, whispered speech has significantly lower energy compared to normally phonated speech [15], and the slope of the spectrum is much flatter than in normal speech [22]. In addition, in real-world environments where background noise is present, the signal-to-noise ratio (SNR) of whispered speech is low. Therefore, processing and recognition of whispered speech are expected to be more difficult than normal speech.

2.1.1 Acoustic analysis

To understand the acoustic differences between normal and whispered speech, a detailed comparison is given below.

Figures 2.1 and 2.2 present the time-domain waveforms and time-frequency domain spectrograms of the same sentence *"I gave them several choices and let them set the priorities."* produced by female speaker 101 in a normal and whispered voice in wTIMIT corpus [18]. The amplitude of whispered speech signal is smaller than that of normal speech signal. The spectrogram of whispered speech appears to have less clear articulatory features and a different acoustic profile compared to normal speech. These differences are primarily due to the altered vocal production mechanism and reduced vocal fold activity during whispering. In whispered speech, voicing harmonics are significantly attenuated or absent, resulting in a lack of distinct harmonic patterns (red box in Figure 2.2) and the formant structure is often less pronounced and less distinct compared to normal speech.

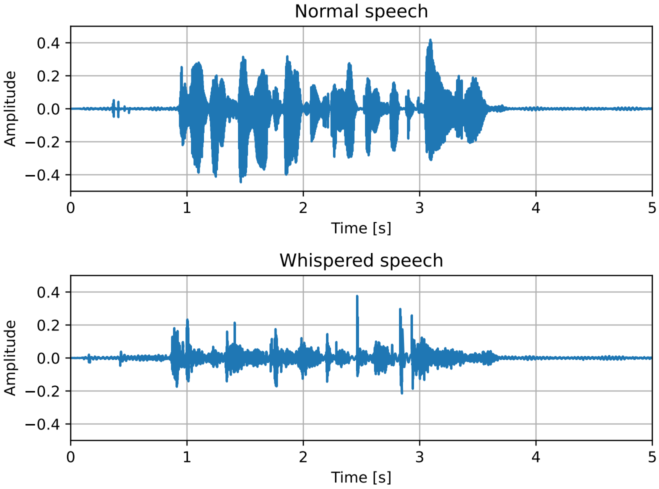

Figure 2.1: Waveforms of the same sentence produced by the same speaker in a normal and whispered voice in wTIMIT corpus.

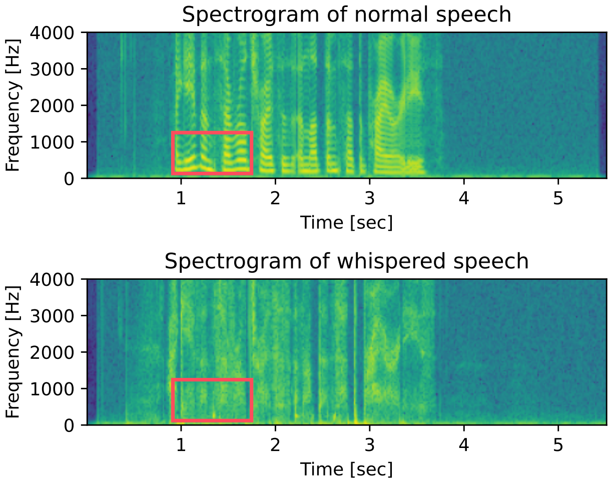

Figure 2.2: Spectrograms of the same sentence produced by the same speaker in a normal and whispered voice in wTIMIT corpus. The red box shows an example of harmonic patterns that is present in normal speech but absent in whispered speech.

Compared to normally phonated speech, whispered speech has:

No fundamental frequency (F0)

The Fundamental Frequency (F0) is the physical measure of the rate at which the vocal folds vibrate during speech production. Pitch is the perceptual attribute of a sound that allows us to differentiate between low and high sounds. In the context of human speech, it refers to the perceived "highness" or "lowness" of a person's voice. Pitch is the perceptual outcome of the physical phenomenon of F0.

Figure 2.3 shows the pitch information of the same sentence *"I gave them several choices and let them set the priorities."* produced by female speaker 101 in a normal and whispered voice in wTIMIT corpus, extracted by the WORLD vocoder [23]. Since voicing is absent when whispering, whispered speech signals theoretically have no F0/pitch. Because of the F0 estimation algorithm in the WORLD vocoder, the result may not match completely with the theory. But we can still say in whispered speech, pitch is more or less absent.

An upward shift in format frequencies of vowels

The format frequencies of vowels in normal speech are generally higher than that in whispered speech [9, 11–14]. Table 2.1 shows the first and second formant frequencies, F1 and F2, estimated from a typical spectrum of each of the five vowels in the Serbian language for normal and whispered speech of the same speaker with the same context. The mean shift for the first formants for all four vowels (/i/, /e/, /a/, and /o/) is 264 Hz and for second formants is 153 Hz in [11]. The whispered vowel /u/, in comparison to the other four vowels, shows a completely different behaviour.

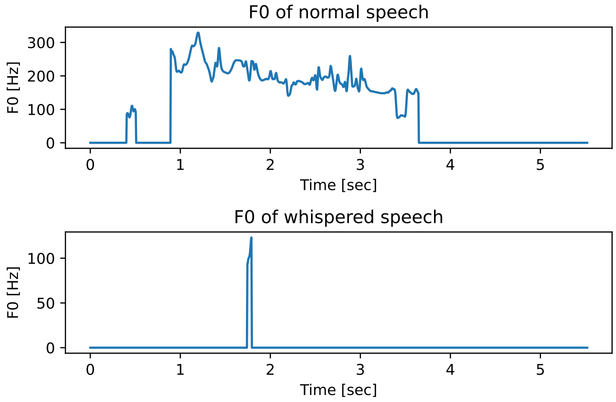

Figure 2.3: Pitch information of the same sentence produced by the same female speaker in normal and whispered voice.

Table 2.1: Mean values of first and second formant frequencies (F1 and F2) of normal and whispered vowels in the Serbian language for male and female speakers derived from [11].

Frequencies		/i/		/e/		/a/		/o/		/u/	
[Hz]		normal	whisp.	normal	whisp.	normal	whisp.	normal	whisp.	normal	whisp.
F1	male	341	492	534	812	708	955	503	724	337	196
	female	357	535	588	868	826	1113	513	842	365	226
F2	male	2140	2306	1830	1756	1192	1433	912	1187	688	685
	female	2573	2683	2141	2265	1403	1586	1067	1267	831	742

Wider formant bandwidths for whispered vowels

The formant bandwidths of whispered vowels are wider than that of normal vowels [9, 11, 12]. Considering broader phone classes (unvoiced/voiced phones), it also shows a similar pattern that unvoiced formants exhibit broader bandwidths than the voiced ones [12]. Table 2.2 shows the formant bandwidths estimation of whispered vowels, in comparison with voiced vowels. Results are obtained as mean values of bandwidth for all speakers and the last column is obtained as mean over all five vowels. The bandwidths for all formants in whispered vowels are wider than those in normal vowels.

Table 2.2: Mean formant bandwidths of first and second formants (F1 and F2) for voiced and whispered vowels in the Serbian language, and mean over all five vowels, derived from [11].

Bandwidths	Vowels										Mean	
[Hz]	/i/		/e/		/a/		/o/		/u/		bandwidths	
Formants	normal	whisp.	normal	whisp.	normal	whisp.	normal	whisp.	normal	whisp.	normal	whisp.
F1	45.8	160.1	54.5	102.5	63.0	116.5	68.1	104.1	59.2	111.6	58.1	119.0
F2	48.7	104.0	44.6	131.6	67.4	133.6	576	136.0	90.7	105.2	61.8	122.1

2.2 Signal processing methods

In this section, I introduce some signal processing methods which are the basis of the methodology chapter. Specifically, the algorithm in methodology is based on the Source-Filter theory and Linear Predictive Coding.

2.2.1 Source-filter model of speech production

In the source-filter theory, speech signals are composed of a combination of a sound source, such as the vocal folds, a linear acoustic filter, known as the vocal tract, and a differentiator (simulating lip radiation effects) [24]. Written in an equation, a time-domain speech signal is produced by a sound source modulated by a linear filter, which can be stated as

$$\begin{aligned} s(n) &= u(n) * h(n) \\ &= u(n) * v(n) * l(n) \end{aligned} \tag{2.1}$$

where $s(n)$ denotes the time-domain speech signal, $u(n)$ denotes the sound source representing the excitation, $h(n)$ denotes the linear transfer function, $v(n)$ denotes the vocal tract filter, $l(n)$ denotes the lip radiation differentiator and $*$ is convolution.

Source

A speech signal can be classified as voiced or unvoiced depending on whether the vocal folds are oscillating or not. In the voiced speech, the source originates from the vibrations of the vocal folds, generating a glottal flow waveform with fundamental frequency $F0$ which is often referred to as the pitch. Unvoiced speech is a non-periodic, noise-like signal, caused by air passing through a narrow constriction of the vocal tract. Hence the source of unvoiced speech is often modelled as white noise. In general, normal speech consists of voiced vowels interspersed with voiced and unvoiced consonants, whereas real whispers are unvoiced [25].

Filter

In the source-filter theory, the source signal is then passed through and modulated by an acoustic filter. The shape of the vocal tract and consequently the positions of the articulators (i.e., jaw, tongue, velum, lips, mouth, teeth, and hard palate) provide a crucial factor to determine the characteristics of this filter [26]. The source-filter model can be stated in the z-domain as

$$S(z) = \underbrace{E(z)\cdot G(z)}_{U(z)} \cdot \underbrace{V(z)\cdot L(z)}_{H(z)} \tag{2.2}$$

where $S(z)$ is the speech spectrum, $E(z)$ is the excitation, $G(z)$ is the glottal contribution, $V(z)$ is the vocal tract filter, and $L(z)$ is the lip radiation. $E(z)$ and $G(z)$ are often combined and called the glottal excitation. $V(z)$ and $L(z)$ are often combined and called the transfer function.

Figure 2.4 shows a diagram of the source-filter model of speech production derived from [27] as described in Eq. 2.2.

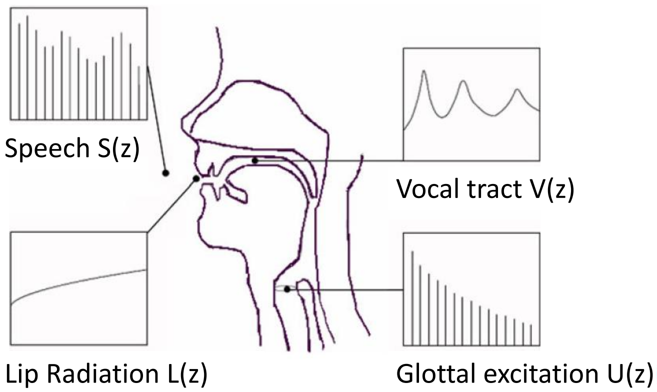

Figure 2.4: Source-filter model of speech production derived from [27]. The speech signal is described as a cascade of three processes: (i) glottal excitation, (ii) vocal tract filtering, and (iii) lip radiation.

2.2.2 Linear Predictive Coding

Linear Predictive Coding (LPC) or Linear Predictive (LP) analysis is a signal processing method to model human voice production based on the source-filter model. As mentioned in Eq. 2.1, time-domain speech signals are produced by a sound source $u(n)$ going through a linear filter $h(n)$. LPC is a method to estimate vocal tract filter $h(n)$ and residual $u(n)$ from a speech signal $s(n)$.

In the theory of LPC, the source signal $u(n)$ is either an impulse train for voiced speech or random white noise for unvoiced speech and the filter $h(n)$ is a p-th order all-pole filter. Figure 2.5 shows the speech production model under the hypothesis of the source-filter theory with LPC.

The reason why I introduce LPC here is that LPC is the basis of the method used for removing glottal information, which is introduced in section 3.2.1.

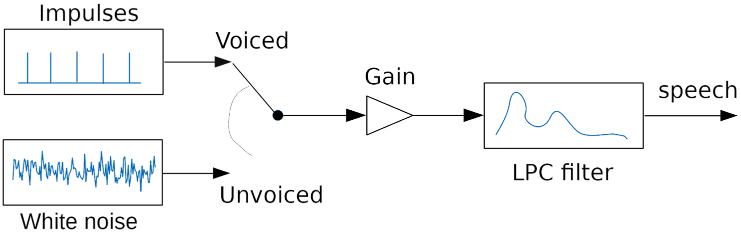

Figure 2.5: Speech production model with LPC[1]. The source signal $u(n)$ is either an impulse train for voiced speech or random white noise for unvoiced speech. Then the source signal is modulated by the filter $h(n)$ and the output is the speech signal $s(n)$.

[1] Source: `https://jmvalin.ca/demo/lpcnet/`.

2.3 Automatic Speech Recognition

This section first briefly gives an overview of ASR systems and introduces the dictionary in ASR. Then several E2E ASR models are introduced. After that, the evaluation metrics and two SOTA data augmentation methods, speech perturbation and SpecAugment, are presented.

2.3.1 Overview

Traditional ASR

An ASR system produces the most likely word sequence given a speech signal. The problem of speech recognition is defined as the conversion of spoken utterances into textual sentences by a machine. In the traditional ASR framework, the Bayesian decision rule is employed to find the most probable text sequence, $\hat{Y}$, given the observation sequence, X:

$$\hat{Y} = \arg\max_{\mathrm{Y}} p(Y \mid X) = \arg\max_{\mathrm{Y}} p(X \mid Y)p(Y) \tag{2.3}$$

where $p(X \mid Y)$ is calculated by the acoustic model and $P(Y)$ is modeled by the language model.

Figure 2.6 illustrates a framework of the traditional ASR system. First, acoustic features are extracted from the input speech using signal processing methods. Notably, log-mel filterbanks are used as the acoustic features in this project, which are also commonly used in deep neural network-based ASR systems. The log-mel filterbanks are calculated by mapping the spectrogram of a signal from the frequency (Hz) scale to the Mel scale. Then the acoustic model is trained to model the probability of a phoneme sequence based on the given acoustic features; the language model is trained on the text data to produce the probability of a certain word sequence; and the lexicon is given as a handcrafted pronunciation dictionary that maps phonemes to words. The most likely word sequence (transcripts) are searched by the multiplication of the output probabilities computed from the acoustic model and the language model as in Eq. 2.3.

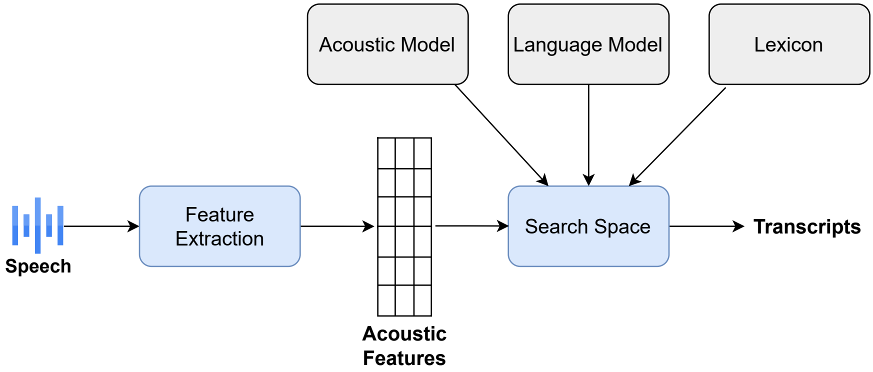

Figure 2.6: The framework of traditional ASR system.

End-to-End ASR

E2E ASR systems are designed to map the input acoustic feature sequence $X = (x_1, ..., x_T)$, i.e. log-mel filterbanks, to the output transcription $Y = (y_1, ..., y_U)$. However, the length of the input sequence is equal to or longer than, i.e. $T \geq U$. This characteristic poses a challenge when computing the loss function during training. Within E2E architectures, two major approaches tackle the task of aligning input acoustic features and output label sequences while computing the loss function between these sequences of variable lengths. These methods encompass the Connectionist Temporal Classification (CTC) [4] and Attention-based Encoder-Decoder [5].

2.3.2 Dictionary

Most E2E ASR models use character-based vocabularies: characters, subwords, or words. In this thesis, I introduce the character-level dictionary and Byte Pair Encoding (BPE) token dictionary.

Character-level dictionary

The character-level dictionary holds a straightforward structure. In English, words are constructed from a fixed set of characters. Consequently, an English character-level dictionary primarily comprises the 26 letters of the alphabet along with a selection of special tokens, such as unknown tokens <unk>, word delimiters <space>, and more.

BPE token dictionary

Byte Pair Encoding (BPE) [28] is originally a simple lossless data compression algorithm in which the most common pairs of consecutive bytes of data are replaced with a byte that does not occur within that data. It was initially adopted as a tokenizer for neural machine translation [29] and has since been widely adopted for many Natural Language Processing tasks as it is simple yet offers a good balance between character and word representations as well as being deterministic, which makes it computationally inexpensive. And now BPE is a popular subword tokenization method in ASR tasks for building a robust vocabulary.

BPE tokenization algorithm first initializes the symbol vocabulary with the character vocabulary and represents each word as a sequence of characters. Then it iteratively counts all symbol pairs and replaces each occurrence of the most frequent pair (‘A’, ‘B’) with a new symbol ‘AB’. It continues to count and merge, creating new longer character strings, until the vocabulary size reaches the desired limit N, an adjustable hyperparameter of the algorithm.

Table 2.3 shows an example of a character-level dictionary and a BPE token dictionary (vocabulary size $N = 100$) generated from the combined dataset TIMIT + wTIMIT.

2.3.3 CTC-based End-to-End Models

The Connectionist Temporal Classification (CTC) approach, introduced by Graves *et al.* [4], employs an intermediate label representation denoted as $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_T)$. This representation allows for label repetitions and the presence of a special blank label ($-$) signifying emission without labels, i.e., $\pi_t \in \{1, \cdots, K\} \cup \{-\}$. CTC trains the model to maximize $P(Y \mid X)$, the probability distribution over all possible label sequences $\Phi\left(Y'\right)$:

Table 2.3: Example of a character-level dictionary (left), and BPE token dictionary (right) generated from the combined dataset TIMIT + wTIMIT.

Character	BPE
<unk>	<unk>
<space>	’
’	a
a	age
b	al
c	an
d	ate
e	b
f	c
g	ce
h	ch
i	d
...	...

$$P(Y \mid X) = \sum_{\boldsymbol{\pi} \in \Phi(Y')} P(\boldsymbol{\pi} \mid X) \tag{2.4}$$

where Y' is a modified label sequence of Y, which is made by inserting the blank symbols between each label and the beginning and the end for allowing blanks in the output, i.e., $Y = (c, a, t)$, $Y' = (-, c, -, a, -, t, -)$.

CTC is generally applied on top of Recurrent Neural Networks (RNNs). Each RNN output unit is interpreted as the probability of observing the corresponding label at a particular time. The probability of label sequence $P(\boldsymbol{\pi} \mid X)$ is modelled as being conditionally independent by the product of the network outputs:

$$P(\boldsymbol{\pi} \mid X) \approx \prod_{t=1}^{T} P(\pi_t \mid X) = \prod_{t=1}^{T} q_t(\pi_t) \tag{2.5}$$

where $q_t(\pi_t)$ denotes the softmax activation of π_t label in RNN output layer q at time t.

The CTC loss to be minimized is defined as the negative log-likelihood of the ground truth character sequence Y^*, i.e.

$$\mathcal{L}_{\text{CTC}} \triangleq -\ln P(Y^* \mid X) \tag{2.6}$$

2.3.4 Attention-based Encoder-Decoder Models

Another branch of the E2E system is the Attention-based Encoder-Decoder architecture [5]. Unlike the CTC approach, the attention model does not make any conditional independence assumptions, and directly estimates the posterior, $P(Y \mid X)$ on the basis of a probabilistic chain rule, as follows:

$$P(Y \mid X) = \prod_u P(y_u \mid y_{1:u-1}, X) \tag{2.7}$$

$$\mathbf{h} = \text{ Encoder } (X) \tag{2.8}$$

$$y_u \sim \text{AttentionDecoder}(\mathbf{h}, y_{1:u-1}). \tag{2.9}$$

The framework consists of two RNNs: Encoder and AttentionDecoder so that it is able to learn two different lengths of sequences based on the cross-entropy criterion. Encoder transforms X, to high-level representation $\mathbf{h} = (h_1, \cdots, h_L)$ in Eq. 2.8, then AttentionDecoder produces the probability distribution over characters, y_u, conditioned on h and all the characters seen previously $y_{1:u-1}$ in Eq. 2.9. L is the number of the frames of Encoder output, and $L < T$. Here, a special start-of-sentence(sos)/end-of-sentence(eos) token is added to the target set, so that the decoder completes the generation of the hypothesis when (eos) is emitted. The loss function of the attention model is computed from Eq. 2.7 as:

$$\mathcal{L}_{\text{Attention}} \triangleq -\ln P(Y^* \mid X) = -\sum_u \ln P\left(y_u^* \mid y_{1:u-1}^*, X\right) \tag{2.10}$$

where $P\left(y_u^* \mid y_{1:u-1}^*, X\right)$ is the ground truth of the previous characters.

2.3.5 Hybrid CTC/Attention Models

The biggest advantage of CTC is that it preserves the monotonic relationship between acoustic frames and output labels, which fits the ASR tasks. However, there are some drawbacks of CTC. The CTC assumes that the outputs are conditionally independent of each other, which is a wrong assumption to make when dealing with speech. Another drawback is that the alignments are many-to-one mappings, where multiple inputs can be aligned to at most one output.

One of the advantages of using Attention-based models is not requiring conditional independence assumptions compared to CTC-based models. However, the attention is too flexible to satisfy monotonic alignment constraints in speech recognition tasks.

Since CTC and Attention models both have their advantages and limitations, it can be useful to integrate them, preserve their advantages and avoid their limitations. The Hybrid CTC/Attention model utilizes both benefits of CTC and Attention during the training and decoding steps in ASR.

The training method of the Hybrid CTC/Attention model uses a CTC objective function as an auxiliary task to train the Attention model encoder within the multiobjective learning (MOL) framework. Figure 2.7 illustrates the overall architecture of the framework, where the same BLSTM is shared with the CTC and Attention encoder networks. The multiobjective learning loss function is represented as follows by using both CTC in Eq. 2.6 and Attention model in Eq. 2.10:

$$\mathcal{L}_{\text{MOL}} = \lambda \mathcal{L}_{\text{CTC}} + (1-\lambda)\mathcal{L}_{\text{Attention}} \tag{2.11}$$

with a tunable parameter $\lambda : 0 \leq \lambda \leq 1$.

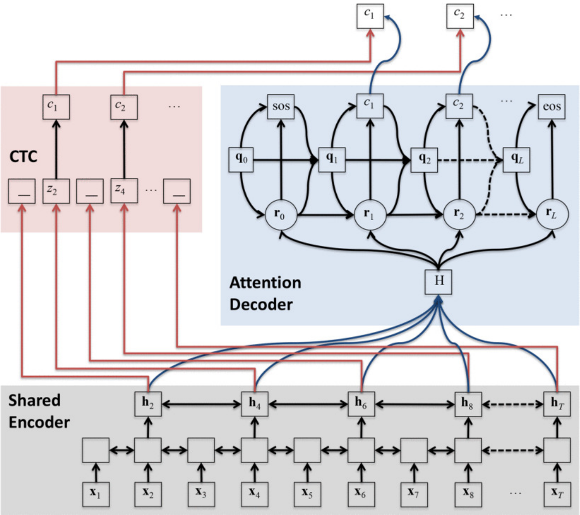

Figure 2.7: Hybrid CTC/attention based end-to-end framework from [6]. The shared encoder is trained by both CTC and attention model objectives simultaneously. The shared encoder transforms the input sequence x into high-level features **h**, the location-based attention decoder generates the character sequence y.

2.3.6 Conformer-based Hybrid CTC/Attention Models

In the previously mentioned Hybrid CTC/Attention Models, the encoders are all RNNs. In recent years, Gulati *et al.* [7] proposed a novel architecture with a combination of self-attention and convolution in ASR models, which is named Conformer. It has achieved the SOTA performance in many widely-used datasets in the field of ASR [30]. Conformer combines convolution neural networks and transformers to model both local and global dependencies of an audio sequence in a parameter-efficient way.

The Conformer model [30] consists of a Conformer encoder proposed in [7] and a Transformer decoder. It predicts a target sequence Y of characters or BPE tokens from an input sequence of 80-dimensional log-mel filterbank features with/without 3-dimensional pitch features. X is first sub-sampled in a convolutional layer by a factor of 4 and then fed into the encoder and decoder to compute the cross-entropy loss. The encoder output is also used to compute a CTC loss for joint CTC/attention training and decoding.

2.4 Evaluation metrics

Word Error Rate (WER) is chosen as the metric to evaluate the performance of the ASR systems in this project. The WER is computed as:

$$WER = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+C} \tag{2.12}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, C is the number of correct words, and N is the number of words in the reference.

2.5 Data augmentation

Data augmentation has been proposed as a method to generate additional training data for ASR. It proves to be an effective method for dealing with data sparsity and improving the performance of normal speech recognition [31, 32] as well as abnormal speech recognition [17, 19, 33]. Here I introduce two widely-used data augmentation methods: speed perturbation [31] and SpecAugment [32].

2.5.1 Speed perturbation

Speed perturbation is a widely-used data augmentation method for speech [31]. It changes the speed of the audio signal, producing a faster or slower version of the original signal with a speed factor. When the speed factor is larger than 1.0, the speech signal is accelerated; when the speed factor is smaller than 1.0, the speech signal is decelerated.

2.5.2 SpecAugment

SpecAugment is another popular way for speech data augmentation [32]. There are three augmentation policies in SpecAugment:

- **Time Warping**: This policy is to warp the spectrogram in the time axis randomly. Unlike speed perturbation, this method does not increase or reduce the duration but squeezes and stretches the spectrogram locally.

- **Frequency Masking**: Here, some consecutive Mel frequency channels are randomly masked.

- **Time Masking**: This method is similar to frequency masking but randomly masks consecutive time steps of a spectrogram.

3

Methodology

In this chapter, I introduce the datasets, proposed method, and experimental setup. In section 3.1, the three datasets used in this project are introduced in detail. Section 3.2 presents the proposed approach to convert from normal to pseudo-whispered speech, Pseudo-whispered speech conversion. The experimental setup including the data pre-processing, feature extraction, ASR systems, and data augmentation specifics are given in section 3.3.

3.1 Datasets

In this section, all datasets used in this thesis are introduced. The Whispered TIMIT [18] and TIMIT [34] datasets are used to train ASR systems for baseline experiments. And Whispered TIMIT, TIMIT and LibriSpeech [35] datasets are used to generate pseudo-whispered speech data.

3.1.1 Whispered TIMIT

The Whispered TIMIT (wTIMIT) corpus [18] consists of 450 phonetically balanced sentences in both normal (wTIMIT-n) and whispered (wTIMIT-w) speech from speakers from two accent groups: US and Singaporean English with 28 (12 male and 16 female) and 20 (12 male and 8 female) speakers, respectively.

wTIMIT was originally partitioned into a training and test set [18]. To prevent overfitting the E2E models, a re-partitioning of wTIMIT into training, development, and test sets is needed. Preliminary experiments performed in [17] showed that a partitioning of the training and test data where there was no speaker overlap in the training and test set, degraded performance by approximately 10% relatively compared to a partitioning of the training and test data where the same speaker could occur in both. This relatively small difference in performance was attributed to the pitch being mostly absent in whispered speech. Also given the fact that partitioning by speakers can lead to less data that can be used as training data, [17] suggested that prohibiting speaker overlap between the training and test sets is unnecessary. Following [17], wTIMIT was re-partitioned into a training, development, and test set allowing speaker overlap. Each data set consisted of 400/25/25 sentences, respectively, split from the 450 sentences.

Table 3.1: Subsets of wTIMIT.

		Accent	Hours	Utterances
Train	Normal	US	15.23	10738
		SG	10.51	7999
	Whisper	US	15.54	10707
		SG	10.64	8000
Dev	Normal	US	0.95	670
		SG	0.65	500
	Whisper	US	0.97	668
		SG	0.67	500
Test	Normal	US	0.94	674
		SG	0.65	500
	Whisper	US	0.96	672
		SG	0.66	500

For the sake of simplicity, the test sets in wTIMIT and their abbreviations are listed below:

- **N**$_{\text{US}}$: Normal speech with US accent
- **N**$_{\text{SG}}$: Normal speech with SG accent
- **W**$_{\text{US}}$: Whispered speech with US accent
- **W**$_{\text{SG}}$: Whispered speech with SG accent

3.1.2 TIMIT

The TIMIT corpus [34] consists of read speech and was designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition systems. There are a total of 6300 utterances, each consisting of 10 sentences spoken by each of the 630 speakers (438 male and 192 female) from 8 major dialect regions across the United States. The prompts for the 6300 utterances consist of 2 dialect sentences (SA), 450 phonetically compact sentences (SX), and 1890 phonetically-diverse sentences (SI). All 450 prompts of wTIMIT were obtained from the SX section of the TIMIT corpus, which makes TIMIT a good option as additional training data for normal speech and to generate pseudo-whispered speech. Table 3.2 shows the subsets of TIMIT used in this project and their duration and number of utterances.

Table 3.2: Subsets of TIMIT.

	Hours	Utterances
Train	3.15	3696
Dev	0.34	400
Test	0.16	192

3.1.3 LibriSpeech

LibriSpeech [35] is a corpus of approximately 1000 hours of 16kHz read English speech. The data is derived from read audiobooks from the LibriVox project and has been carefully segmented and aligned. Table 3.3 shows the three training subsets with a total size of 960 hours and their duration. The first two train-clean sets are on average of higher recording quality and with accents closer to US English than the third training set. In this project, I only used the train-clean-100 subset to generate pseudo-whispered speech and augment the training data.

Table 3.3: Subsets of the training set of LibriSpeech.

	hours	per-spk minutes	female speakers	male speakers	total speakers
train-clean-100	100.6	25	125	126	251
train-clean-360	363.6	25	439	482	921
train-other-500	496.7	30	564	602	1166

3.2 Pseudo-whispered speech conversion

The proposed method to convert normal to pseudo-whispered speech is based on that of Cotescu *et al.* [36] who proposed a handcrafted digital signal processing recipe that converts normal speech into whispered speech in three steps by making acoustic modifications to the normal speech: 1) remove the glottal contribution using spectral subtraction; 2) shift the first formant using frequency warping; 3) increase the formant bandwidth using moving average filtering. The WORLD vocoder [23] is used to extract features for re-synthesizing high-quality speech.

In step 1, instead of using spectral subtraction as in [36], I implemented a glottal cancellation method, which does not require parameters for modelling glottal flow but removes the glottal information directly from a given normal speech signal. Moreover, preliminary experiments using the method from [36] showed that moving average filtering not only widens the formant bandwidth but also up-shifts the formant frequencies. Hence, the proposed method for pseudo-whispered speech conversion is implemented in 2 steps, which is also shown in Figure 3.1:

- **Step 1: Removing glottal information**
  Removing the glottal source using GFM-IAIF-based Glottal Inverse Filtering (see section 3.2.1),
- **Step 2: Changing formant information**
  Increasing the formant bandwidth and up-shifting the formant frequencies using moving average filtering (see section 3.2.2).

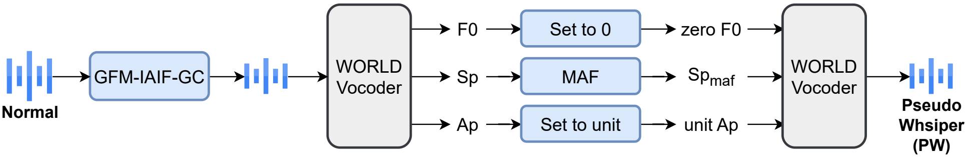

Figure 3.1: The proposed pipeline for pseudo-whispered speech conversion, where GFM-IAIF-GC is GFM-IAIF-based glottal cancellation and MAF is moving average filtering. Input is normal speech and the output is pseudo-whispered speech (PW).

3.2.1 Step 1: Removing glottal information

As introduced in section 2.2.1 Eq.2.2, a speech signal is composed of an excitation E, vocal tract filter V, lip radiation filter L, and glottal contribution G, which can be written in the frequency domain as $S(z) = E(z) \cdot G(z) \cdot V(z) \cdot L(z)$. Glottal Inverse Filtering (GIF) estimates the source of voiced speech, specifically the glottal volume velocity waveform. Iterative Adaptive Inverse Filtering (IAIF) [37] is one of the most widely used algorithms for GIF. IAIF successively models the vocal tract filter $V(z)$, lip radiation $L(z)$, and glottis (glottal contribution) $G(z)$ using linear prediction (LP) analysis, then removes their effect by inverse filtering. After two iterations, it ultimately removes $V(z)$ and $L(z)$ to leave an estimate of the glottal flow $g(n)$, where n is the discrete-time index.

The Glottal Flow Model-based Iterative Adaptive Inverse Filtering (GFM-IAIF) [38] is an improved version of IAIF. It constrains glottal flow by a 3^{rd} order spectral model $G(z) = \left\{ \left(1 - az^{-1}\right)\left(1 - a^{*}z^{-1}\right)\left(1 - bz^{-1}\right) \right\}^{-1}$. GFM-IAIF performs competitively for normal phonations [39], which makes it suitable for our case: cancelling the glottal contribution of normally phonated speech. The GFM-IAIF method comprises four essential steps for an accurate estimation:

In the initial step, referred to as gross glottis estimation, the spectral tilt contribution of the glottis is eliminated from the speech signal, laying the groundwork for the vocal tract (VT) estimation. The second step, VT gross estimation, involves the deconvolution of the gross glottis and lip radiation filters from the original signal. The VT autoregressive coefficients are then estimated using high-order LPC. For the third step, fine estimation of the glottis, the signal undergoes the removal of lip radiation and the estimated VT contributions (thus eliminating all VT formants). A 3^{rd} order LPC is utilized to ensure the final glottis filter aligns with the glottal flow model. Lastly, the fourth step, fine VT estimation, yields the final estimate of the vocal tract transfer function.

In the proposed method, I employ GFM-IAIF to extract glottal flow, after which the effect of the glottis is cancelled by inverse filtering, and the output is a speech signal without glottal contribution. This process is called GFM-IAIF-based glottal cancellation (GFM-IAIF-GC) and is shown in Figure 3.2, where the input is a normal speech signal and the output is a speech signal without glottal contribution.

Subsequently, $F0$, the spectral envelope (Sp), and the aperiodic spectral envelope (Ap) are extracted from the speech signal without glottal contribution using the WORLD vocoder. To ensure that the pitch is removed entirely, the $F0$ is set to zero and Ap values are set to all units.

3.2.2 Step 2: Changing formant information

To increase the formant bandwidth and up-shift the formant frequencies, I employ moving average filtering on the spectral envelop Sp extracted by the WORLD vocoder with a 400 Hz-wide triangular window across all frequency axes and get a new spectrogram Sp_{maf}.

The three adapted features, *zero* $F0$, Sp_{maf} and *unit* Ap are passed to the WORLD vocoder for re-synthesising the pseudo-whispered speech (PW). Figure 3.3 shows an example of the conversion results, it shows a zoomed-in spectrogram of normal speech (left panel), whispered speech (middle panel) and pseudo-whispered speech (right panel).

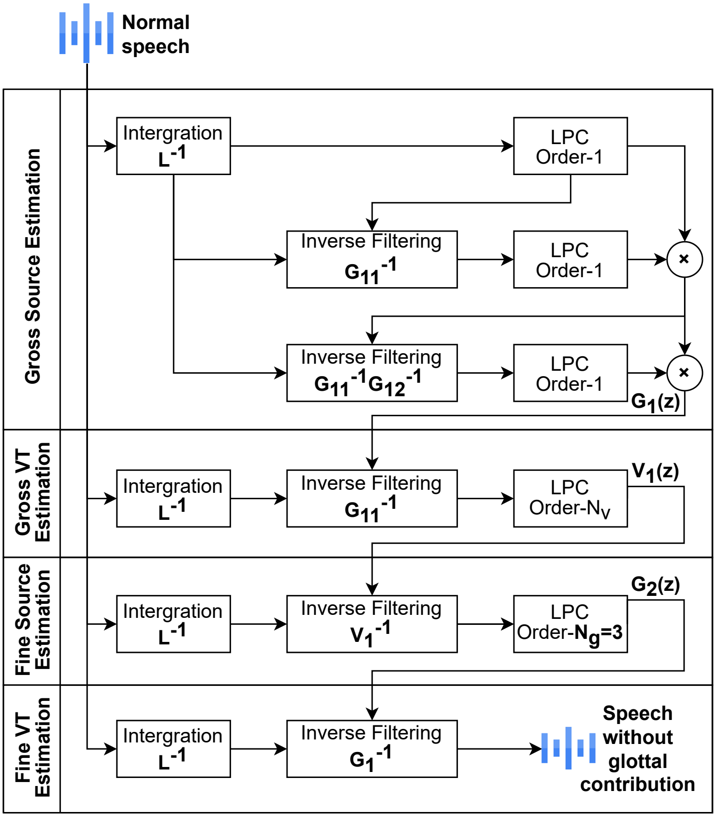

Figure 3.2: GFM-IAIF-based glottal cancellation derived from [38]. The input is a normal speech signal and the output is a speech signal without glottal contribution.

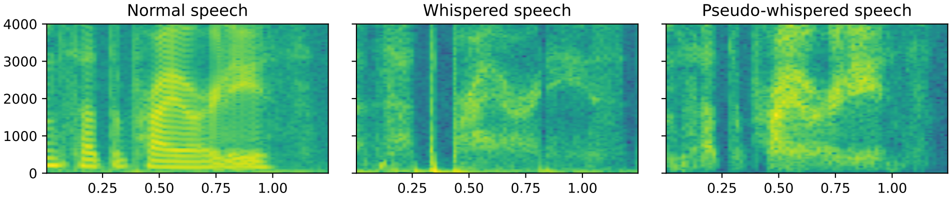

Figure 3.3: Spectrogram of normal (left panel), whispered (middle panel), and pseudo-whispered speech (right panel) of the words “the priorities” from the same utterance as in Figure 2.2.

3.3 Experimental Setup

All E2E models were trained with the ESPNet toolkit [40]. The sampling rate of speech data in the wTIMIT dataset is 44.1 kHz, while the sampling rates in TIMIT and LibriSpeech are 16 kHz. It is necessary to keep the sample rate consistent among all datasets, hence all speech data in wTIMIT were downsampled from 44.1 kHz to 16 kHz before the Feature Extraction stage simply using SoX [1]. The front-end features are 80 dimensional log-mel filterbank features with 3-dimensional pitch features used for network training.

3.3.1 Baseline models

First, I trained a strong baseline by investigating the

1. **Training data**: TIMIT plus the normal speech from wTIMIT (TM+wTM-n) vs. TIMIT plus both normal and whispered speech from wTIMIT (TM+wTM-wn);
2. **Data augmentation**: none vs. speed perturbation (SP) [31] at 90% and 110% of the original rate of the training data and SpecAugment (SpecAug) [32] which was used with a maximum width of each time and frequency mask of $T = 20$, $F = 10$, respectively;
3. **Dictionary types and sizes**: six models used a character-level dictionary, and the remaining two used a BPE token dictionary;
4. **Model architectures**: I compared two architectures: the Hybrid CTC/Attention Model (Hybrid-CTC) [6] and Conformer-based Hybrid CTC/Attention Model (Conformer) [30] (from the LibriSpeech recipe from the ESPNet framework).

In total, eight models were trained. The models were evaluated on the TIMIT and wTIMIT-n and wTIMIT-w datasets. Performance was measured in terms of Word Error Rate (WER) for both accent groups separately.

3.3.2 Pseudo-whisper data augmentation

The second experiment investigated the effect of adding pseudo-whispered (PW) data to the training data on whispered speech ASR performance. To that end, the pseudo-whispered speech was created from TIMIT, wTIMIT-n, and LibriSpeech-100h (referred to as Libri-100) and each was successively added to the training data: first only PW speech from TIMIT (PW(TM)), then the PW speech from wTIMIT-n (PW(wTM-n)) was added, and finally also the PW speech from LibriSpeech-100h (PW(Libri100)). Each set of training data was used to train both the Hybrid-CTC and Conformer architectures, yielding six models. The effect of the pseudo-whisper data augmentation on the two accent groups was also analysed.

3.3.3 Acoustic characteristics of whispered speech

In the final experiment, the effect of the specific acoustic characteristics of whispered speech on whispered speech ASR performance was investigated by comparing the recognition results on speech in which either the glottal information was removed or in which the formant bandwidth had been widened and the formant frequencies had been shifted.

To that end, I individually applied each of the two steps of the proposed pseudo-whispered speech conversion method on the normal test set in wTIMIT and synthesized

[1] https://sox.sourceforge.net/

the modified speech. Figure 3.4 shows the pipelines of generating speech without glottal contribution (referred to as NG) and speech with widened formant bandwidth and shifted formant frequencies (referred to as WB). The pipelines are subsets of the full pipeline in Figure 3.1.

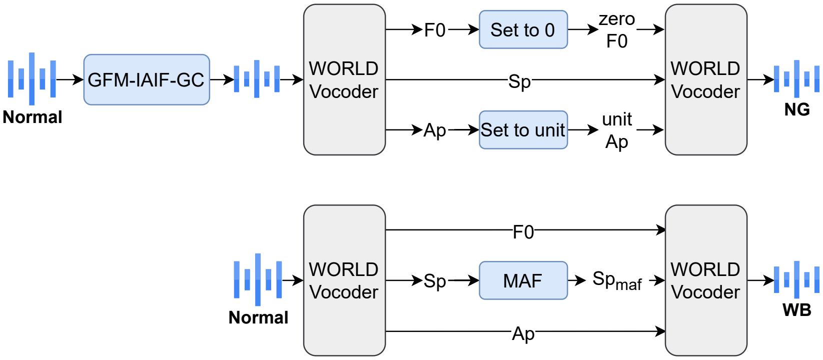

Figure 3.4: The pipeline for generating speech with only glottal cancellation (top panel) and with only a widened formant bandwidth and shifted formant frequencies (bottom panel).

The modified speech was subsequently tested using three models: the Hybrid-CTC architecture trained on only normal speech (row TM+wTM-n in Table 4.1); trained on normal and whispered speech (row TM+wTM-wn in Table 4.1); and trained on normal, whispered, and pseudo-whispered speech (TM+wTM-wn+PW(TM+wTM-n)). SP and SpecAug were not applied to all three models in this final experiment.

4

Results

In this chapter, we show the experimental results. Section 4.1 presents the results of the baseline experiments. Section 4.2 gives the results of adding pseudo-whispered speech data. Section 4.3 shows the results of RQ2: Impact of acoustic characteristics.

4.1 Baseline Experiments

Before diving into the analysis of the proposed pseudo-whisper data augmentation and its impact on recognition performance, it is crucial to establish a strong baseline model using the available data. Therefore, we first evaluate the performance of a model trained solely on normal speech. By incorporating actual whispered speech in the training data and employing SOTA data augmentation techniques, we achieved a relatively strong baseline.

Table 4.1 presents the results of the baseline models on the TIMIT and wTIMIT test sets for the two accent groups separately and averaged over both accent groups. The first model was trained on only normal speech (TM+wTM-n) and gave a WER of 40% (Hybrid-CTC) and 52% (Conformer) on the TM test set, while the performance on wTIMIT-n averaged over both accent groups ($\mathbf{N_{Avg}}$) showed a fairly large WER drop of 10-20%. The performance on whispered speech is much lower than that on normal speech, with WERs of over 100%. Initially, training a model on normal speech alone led to a significant performance gap of nearly 50% between normal and whispered speech recognition.

To bridge this performance gap between normal and whispered speech recognition and try to improve recognition accuracy, we incorporated real whispered speech data into the training set (TM+wTM-wn). This addition improved recognition performance for whispered speech substantially, and reduced the gap with performance on normal speech to less than 10% for both architectures, but at the cost of a slight increase in WER for normal speech. The improvement clearly indicates the benefits of utilizing matched training and test sets in enhancing overall ASR performance for whispered speech.

Furthermore, we investigated the effectiveness of the SOTA data augmentation techniques, including speed perturbation and SpecAugment. While the Hybrid-CTC model did not demonstrate notable improvements with these techniques, the Conformer model demonstrated a promising improvement of more than 25% for both average normal and whispered speech. This suggests that as the size of the training data increases, the Conformer models outperform the Hybrid-CTC models.

Finally, we investigated the use of the BPE token dictionary as an alternative to the character-level dictionary. Leveraging BPE enables both models to further improve the performance, achieving WERs of 37.9% and 44.4% for normal and whispered speech for the Conformer model. This final baseline model is selected as the baseline model to which the models trained using additional pseudo-whispered speech will be compared.

Table 4.1: Details of baseline ASR systems and their performances in WER (%) on TIMIT and wTIMIT

Details						**TM**	**wTIMIT**					
Training	**Augmentation**	**Hours**	**Architecture**	**Token**	**#Token**	**Test**	N$_{\mathrm{US}}$	N$_{\mathrm{SG}}$	**W**$_{\mathrm{US}}$	**W**$_{\mathrm{SG}}$	N$_{\mathrm{Avg}}$	**W**$_{\mathrm{Avg}}$
TM+wTM-n	None	28.89	Hybrid-CTC	Char	29	40.6	45.4	59.1	99.4	105.2	51.2	101.9
			Conformer	Char	29	52.9	73.4	82.7	102.5	109.6	77.4	105.5
TM+wTM-wn	None	55.07	Hybrid-CTC	Char	29	41.2	51.5	62.8	55.3	74.4	56.3	63.5
			Conformer	Char	29	44.7	78.0	86.4	81.4	92.0	81.6	85.9
TM+wTM-wn	SP + SpecAug	166.33	Hybrid-CTC	Char	29	42.3	52.0	67.3	57.0	77.7	58.5	65.9
			Conformer	Char	29	38.3	49.6	58.3	53.2	68.3	53.3	59.7
TM+wTM-wn	SP + SpecAug	166.33	Hybrid-CTC	BPE	100	44.6	41.2	55.8	44.1	65.9	47.4	53.4
			Conformer	BPE	100	**34.1**	**34.9**	**41.8**	**37.7**	**53.5**	**37.9**	**44.4**

4.2 RQ1: Effect of pseudo-whispered speech

This section focuses on **RQ1: Can we improve ASR systems performance for whispered speech by generating artificial whispered speech data through signal processing techniques as additional training data?**

In the following experiments, we subsequently added pseudo-whispered speech data of TIMIT, wTIMIT, and LibriSpeech-100h to the baseline training data (trained on TM and wTM-wn with SP + SpecAug). Table 4.2 presents the results of the pseudo-whispered speech experiments. For ease of comparison, the results of the baseline Hybrid-CTC and Conformer models are added (identical to those reported in Table 4.1).

Table 4.2: WER (%) on the TIMIT and wTIMIT test sets when using pseudo-whispered training data generated from TIMIT, wTIMIT-n, and LibriSpeech-100h. Relative improvement (%) of the proposed method compared to the baseline is also reported. Results of the chosen baseline Hybrid-CTC and Conformer models are added (identical to those reported in Table 4.1).

Details				**TM**	**wTIMIT**						**Relative Imp.**	
Training Data	**Hours**	**Architecture**	**#Token**	**Test**	N$_{\mathrm{US}}$	N$_{\mathrm{SG}}$	**W**$_{\mathrm{US}}$	**W**$_{\mathrm{SG}}$	N$_{\mathrm{Avg}}$	**W**$_{\mathrm{Avg}}$	N$_{\mathrm{Avg}}$	**W**$_{\mathrm{Avg}}$
Baseline	166.3	Hybrid-CTC	100	44.6	41.2	55.8	44.1	65.9	47.4	53.4	-	-
		Conformer	100	34.1	34.9	41.8	37.7	53.5	37.9	44.4	-	-
+PW(TM)	175.8	Hybrid-CTC	100	46.5	40.1	52.4	41.2	59.9	45.4	49.2	4.2	7.9
		Conformer	100	36.4	32.1	**35.2**	33.4	**40.1**	33.4	**36.3**	11.9	**18.2**
+PW(TM+wTM-n)	253.6	Hybrid-CTC	100	43.4	36.1	44.3	38.0	51.8	39.6	43.9	**16.5**	17.8
		Conformer	100	34.6	32.4	36.6	33.6	42.1	34.2	37.2	9.8	16.2
+PW(TM+wTM-n+Libri100)	557.6	Hybrid-CTC	300	16.9	35.5	55.0	39.2	63.4	43.8	49.5	7.6	7.3
		Conformer	300	**11.0**	**26.8**	38.6	**30.7**	49.2	**31.8**	38.6	16.1	13.1

4.2.1 Pseudo-whisper data augmentation

In this subsection, we only focus on the results of the TM Test, the average results of normal (N$_{\mathrm{Avg}}$) and whispered speech (**W**$_{\mathrm{Avg}}$), and the relative improvement the proposed method brought.

Adding Pseudo-whispered speech of TIMIT

In Table 4.2, rows **+ PW(TW)** present the results obtained by incorporating only 3 hours of pseudo-whispered speech from the TIMIT training set (with speed perturbation yielding 9 hours), and the main findings are:

- On the TM Test set, there is a slight degradation in performance for both the Hybrid-CTC and Conformer models, with an increase in WERs of approximately 2%.
- The Hybrid-CTC model shows small improvements in performance for normal and whispered speech compared with the baseline, with relative WER improvements of 4.2% and 7.9%, respectively.
- The Conformer model achieves larger relative improvements in performance for normal speech and whispered speech, with relative improvements of 11.8% and 18.2%, respectively, which is comparatively higher than the Hybrid-CTC model.
- The Conformer model performs better than the Hybrid-CTC model on all test sets.

Overall, adding only 3 hours of pseudo-whispered data from TIMIT improved the average WER of whispered speech compared to the baseline for both models, with the largest relative improvement for the Conformer model (18.2%). Interestingly, adding pseudo-whispered speech also improved the WER on the normal wTIMIT speech was reduced for both models.

Adding pseudo-whispered speech of TIMIT and wTIMIT-n

In Table 4.2, rows **+ PW(TM+wTM-n)** show the results of adding pseudo-whispered speech of wTIMIT-n on top of the training data used in the previous experiment (TM + wTM-wn + PW(TM)), and the main findings are:

- On the TM Test set, there is a slight improvement compared to the previous section. The Hybrid-CTC model performs better than the baseline model on this test set. However, the Conformer model still has a higher WER on the TM Test.
- For the Hybrid-CTC model, the WERs of both normal speech and whispered speech decrease significantly, with a relative improvement of over 15% for each.
- On the other hand, the Conformer model does not achieve further enhancement compared to only adding PW(TM), although its performance is better than the baseline model in all cases.
- Whatsoever, the Conformer model still performs better than the Hybrid-CTC model on all test sets.

Overall, adding the pseudo-whispered speech of the wTIMIT-n training set further improved recognition performance for the Hybrid-CTC model but performance for the Conformer model deteriorated for the normal and whispered speech. Recognition performance on the TM test set was again similar to the baseline models.

Adding pseudo-whispered speech of TIMIT, wTIMIT-n and LibriSpeech-100h

In Table 4.2, rows **+ PW(TM+wTM-n+Libri100)** show the results of further adding pseudo-whispered speech of LibriSpeech-100h, and the main findings are:

- In the TM Test set, there is a massive improvement compared to the previous sections: the WER of TM Test drops to 16.9% (Hybrid-CTC) and 11.0% (Conformer), achieving the best performance on TIMIT.
- Both the Hybrid-CTC and Conformer models show improvements in normal speech, with a relative improvement of 7.6 and 16.1%. The Conformer model achieves the best performance on wTIMIT normal speech, with a WER of 31.8%.
- On the other hand, both models do not enhance the performances on wTIMIT whispered speech. However, they are worse than the previous models (**+ PW(TM)** and **+ PW(TM+wTM-n)**) without adding pseudo-whispered speech of LibriSpeech-100h.

Overall, further adding the pseudo-whispered speech generated from LibriSpeech-100h gives the best recognition performance for normal wTIMIT speech, but it deteriorated the performance for the whispered speech. The best whispered speech results are obtained with the Conformer model trained with (only) the pseudo-whispered TIMIT speech added.

4.2.2 Comparing different speaker groups

In this subsection, I focus on the recognition performance on normal and whispered speech for the two accent groups in the wTIMIT test set in Table 4.2, namely, $\mathbf{N}_{\text{US}}$, $\mathbf{N}_{\text{SG}}$, $\mathbf{W}_{\text{US}}$, and $\mathbf{W}_{\text{SG}}$.

Comparing the recognition performance on normal and whispered speech for the two accent groups in the wTIMIT test set showed that Singaporean English normal and whispered speech is consistently worse recognised than US English. This performance gap is the largest for whispered speech.

Adding pseudo-whispered speech always improves the recognition performance of normal and whispered US and Singaporean English, even if the pseudo-whispered speech was based on US English only (PW(TM)). In fact, adding only the US English pseudo-whispered speech from TIMIT gives the best result for $\mathbf{N}_{\text{SG}}$ and $\mathbf{W}_{\text{SG}}$ and reduces the performance gap with US English to 3.1% for normal speech and 6.7% for whispered speech for the Conformer, i.e., the smallest performance gap for whispered speech for the Conformer.

Interestingly, adding pseudo-whispered speech from Singaporean English does not further improve recognition performance for $\mathbf{N}_{\text{SG}}$ and $\mathbf{W}_{\text{SG}}$ for the Conformer, although it does further improve performance for the Hybrid-CTC model, giving the best results for normal and whispered Singaporean English for the Hybrid-CTC model.

When adding the pseudo-whispered US English from LibriSpeech (PW(Libri-100)) as additional training data, the performance on whispered US English ($\mathbf{W}_{\text{US}}$) improves to 30.7%, the best result, but it adversely affects the performance on $\mathbf{W}_{\text{SG}}$, widening the gap between the US and Singaporean English to 18.5% for the Conformer model and even more for the Hybrid-CTC model. Thus, adding a large amount of pseudo-whispered speech based on US English negatively impacts the recognition of Singaporean English normal and whispered speech.

4.3 RQ2: Impact of acoustic characteristics

This section focuses on **RQ2: Which and to what extent do acoustic characteristics of whispered speech impact whispered speech recognition performance?**

Table 4.3 presents the results of the experiments on normal speech, real whispered speech, pseudo-whispered (PW) speech and the intermediate forms of whispered speech (see section 3.3), i.e., normal speech without glottal contributions (NG) and normal speech with widened formant bandwidth and shifted formant frequencies (WB). Note that to create PW, steps used to generate NG and WB are applied, or we can roughly say $\mathrm{PW} = \mathrm{NG} + \mathrm{WB}$.

Table 4.3: WERs (%) of different test groups when the model is trained on normal speech (row TM+wTM-n in Table 4.1); normal and whispered speech (row TM+wTM-wn in Table 4.1); and normal, whispered, and pseudo-whispered speech (TM+wTM-wn+PW(TM+wTM-n)).

Training data	**Normal**	**Whisper**	**PW**	**NG**	**WB**
TM + wTM-n	51.2	101.9	79.7	78.0	56.5
TM + wTM-wn	56.3	63.5	65.5	65.2	59.2
TM + wTM-wn + PW(TM+wTM-n)	55.9	61.3	59.2	59.4	62.3

When the model is trained on only normal speech (TM + wTM-n), the gap between Normal and NG (>25%) is larger than the one between Normal and WB (5%). This indicates that performance is worse for speech without glottal contribution and that the widened formant bandwidth and shifted formant frequencies in whispered speech are less detrimental to recognition performance. Combining NG and WB into pseudo-whispered speech only shows a small deterioration compared to NG. This indicates that the effect of both removing glottal information and widening the formant bandwidth and shifting the formant frequencies is not entirely additive.

Not surprisingly, adding real whispered speech from wTIMIT-n into training data (TM + wTM-wn) greatly improves the recognition performance of real whispered speech. Recognition performance of pseudo-whispered speech and NG speech also greatly improves, to the level of that of real whispered speech. Performance on WB speech slightly deteriorates. This again indicates that the glottal information is the most important acoustic information to explain the whispered speech recognition performance.

Adding pseudo-whispered speech (TM + wTM-wn + PW(TM+wTM-n)) improves the recognition performance of real and pseudo-whispered speech, indicating that the pseudo-whispered speech is close enough to real whispered speech for real whispered speech to benefit from the added data. The recognition performance of PW is actually better than that of real whispered speech which shows the benefit of adding matched training data. Speech without glottal information is now actually better recognised than WB speech, which shows that adding pseudo-whispered speech is most beneficial for NG speech and that the benefit for WB speech is less great.

5

Discussions and Conclusions

5.1 Discussions

For the **RQ1: Can we improve ASR systems performance for whispered speech by generating artificial whispered speech data through signal processing techniques as additional training data?**

The proposed pseudo-whisper data augmentation improves ASR performance compared to the baseline Hybrid-CTC and Conformer models. All three experiments, i.e. **+PW(TM)**, **+PW(TM+wTM-n)** and **+PW(TM+wTM-n+Libri100)**, have shown an improvement for whispered speech recognition compared to the baseline. For the Hyrbid-CTC model, we achieve an improvement from 53.4% to 43.9% on WER, with a relative improvement of 17.8%. For the Conformer model, we achieve an improvement from 44.4% to the best result on whispered speech: 36.3% WER, with a relative improvement of 18.2%. This WER reduction underscores the efficacy of utilizing pseudo-whispered speech as a means of augmentation. And we also discover that accents in the training data may have an impact on the recognition performance of different speaker groups. In the wTIMIT dataset, there are two different accents: North American and Singaporean English. When we add the pseudo-whisper of LibriSpeech-100h as additional training data, the amount of US-accented speech dominates quite more in the training data. The addition of it improves the performance of US English normal and whispered speech but has a negative impact on the performance of Singaporean English normal and whispered speech. So when comparing the performance between models trained on different training data, results of further adding pseudo-whispered speech of LibriSpeech-100h are worse in the average WER over US and SG accents.

Comparing our results to the SOTA on wTIMIT shows that our WER on whispered speech is higher than in [19]; however, [19] does not report which accent group from wTIMIT they use in their evaluation. Assuming that they only used the US English part of wTIMIT, considering that they used large amounts of US English data from LibriSpeech and their internally recorded dataset for training, our results are very close to theirs (29.4% vs. our 30.7%) on the US English whispered speech, but using far less data. Comparing our results to those of Chang *et al.* [17]: both their approach and ours showed a relative improvement (their 44.4% in CER; our 18.2% in WER); however they only report phone and character error rates, making a direct comparison impossible. Agrawal *et al.* [10] also achieved a WER of 8.86 % on the wTIMIT whisper test set; however they used speech enhancement techniques like Denoising Autoencoders, Variational Autoencoders to map speech from the whisper domain into the normal domain and recognised the speech through an ASR model trained on normal speech instead of building an ASR system trained on whispered speech.

For the **RQ2: Which and to what extent do acoustic characteristics of whispered speech impact whispered speech recognition performance?**

For the model only trained on normal speech, the performance on speech without glottal information (NG) is worse than speech with widened formant bandwidth and shifted formant frequencies (WB). And when real whispered speech data is added to training data, the performance on WB speech deteriorates. They both prove that the lack of glottal information in whispered speech has the largest impact on whispered speech recognition.

I believe this research question brings more meaning to future research on whispered speech recognition or related tasks. The results of the final experiment suggest that further investigation on feature modelling for whispered ASR is needed, which may emphasize the importance of glottal information and then improve the performance of whispered ASR. Also now that we know the glottal information in speech plays a more important role in whispered speech recognition and we are able to generate the speech without glottal information, I would like to bring back the case in pathological speech contexts. For patients with impaired larynx (vocal folds) who can only produce whispery voices, we can try using the NG speech to enhance the ASR systems specially designed for them.

5.2 Future Research

Based on what this thesis has done and its limitations, I can provide some ideas for future works.

In this work, I did not try training on the full 960 hours of LibriSpeech and generate pseudo-whispered speech from it. It is a widely-used dataset and other related works on whispered speech recognition [17, 19] also used LibriSpeech-960h. For now, our results on whispered speech are slightly higher than those in [19]. It is worth trying to add it to the training set so that we can compare our proposed method with theirs in a more direct and fair way. Also, the related works, e.g. [17, 19], both used different evaluation metrics. Especially this work used the same wTIMIT re-partition as [17], but it did not train or report at the phone or character level. In future, conducting more experiments reporting on phone and character error rates can help us compare with their work.

This thesis proposed a signal-processing-based normal-to-whisper conversion method to create pseudo-whispered speech. When I listen to synthetic pseudo-whispered speech, the quality is still not as good as real whispered speech. Although this thesis does not aim at high naturalness or intelligibility, it is worth seeing how "good" the pseudo-whispered speech is by measuring the naturalness, intelligibility and even speaker identity as in [36]. And further, discover if the quality of the synthetic speech data will influence the performance.

Based on the findings in RQ2, it shows a large impact of glottal information in whispered speech. Many other acoustic features have also been used to train the ASR systems, or even modelling from the raw waveform. Glottal information or pitch is more related to the front-end acoustic features, so I think it is worth trying new acoustic features.

5.3 Conclusions

This thesis project aims to deal with the data scarcity problem of whispered speech by generating artificial whispered speech to augment the training data for improved E2E whispered speech ASR and understand what acoustic characteristics of whispered speech have the largest effect on whispered speech recognition performance. The proposed signal-processing-based normal-to-whisper conversion method was used to create pseudo-whispered speech from three databases, i.e. TIMIT, wTIMIT-n, and LibriSpeech-100h. Utilizing the proposed pseudo-whispered-based data augmentation and the SOTA data augmentation methods, i.e. speech perturbation and SpecAugment, the best model reduced the WER of whispered speech from 44.4% to 36.3%, with an 18.2% relative reduction, leaving only a small WER gap with normal speech. This work also showed results for the individual speaker groups in the wTIMIT database giving the best results for US English. The final experiment shows the lack of glottal information in whispered speech has the largest impact on whispered speech ASR performance.

Bibliography

References

[1] Nancy Pearl Solomon, Gerald N McCall, Michael W Trosset, and William C Gray. Laryngeal configuration and constriction during two types of whispering. *Journal of Speech, Language, and Hearing Research*, 32(1):161–174, 1989.

[2] Beiming Cao, Myungjong Kim, Ted Mau, and Jun Wang. Recognizing Whispered Speech Produced by an Individual with Surgically Reconstructed Larynx Using Articulatory Movement Data. In *Proc. 7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT 2016)*, pages 80–86, 2016.

[3] Roger J. Ingham, Anne K. Bothe, Erin Jang, Lauren Yates, John Cotton, and Irene Seybold. Measurement of speech effort during fluency-inducing conditions in adults who do and do not stutter. *Journal of Speech, Language, and Hearing Research*, 52(5):1286–1301, 2009.

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

[5] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. *Advances in neural information processing systems*, 28, 2015.

[6] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253, 2017.

[7] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

[8] Jinyu Li et al. Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, 11(1), 2022.

[9] Taisuke Ito, Kazuya Takeda, and Fumitada Itakura. Analysis and recognition of whispered speech. *Speech Communication*, 45(2):139–152, 2005.

[10] Vikas Agrawal, Shashi Kumar, and Shakti P. Rath. Whisper Speech Enhancement Using Joint Variational Autoencoder for Improved Speech Recognition. In *Proc. Interspeech 2021*, pages 2706–2710, 2021.

[11] Siobodan T Jovičić. Formant feature differences between whispered and voiced sustained vowels. *Acta Acustica united with Acustica*, 84(4):739–743, 1998.

[12] Shabnam Ghaffarzadegan, Hynek Boşil, and John H. L. Hansen. Generative modeling of pseudo-target domain adaptation samples for whispered speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5024–5028, 2015.

[13] Ken J. Kallail and Floyd W. Emanuel. An acoustic comparison of isolated whispered and phonated vowel samples produced by adult male subjects. *Journal of Phonetics*, 12(2):175–186, 1984.

[14] Hamid Reza Sharifzadeh, Ian V McLoughlin, and Martin J Russell. A comprehensive vowel space for whispered speech. *Journal of voice*, 26(2):e49–e56, 2012.

[15] Slobodan T. Jovičić and Zoran Šarić. Acoustic analysis of consonants in whispered speech. *Journal of Voice*, 22(3):263–274, 2008.

[16] Đorđe T. Grozdić and Slobodan T. Jovičić. Whispered speech recognition using deep denoising autoencoder and inverse filtering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2313–2322, 2017.

[17] Heng-Jui Chang, Alexander H. Liu, Hung-yi Lee, and Lin-shan Lee. End-to-end whispered speech recognition with frequency-weighted approaches and pseudo whisper pre-training. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 186–193, 2021.

[18] Boon Pang Lim. *Computational differences between whispered and non-whispered speech.* University of Illinois at Urbana-Champaign, 2011.

[19] Prithvi RR Gudepu, Gowtham P Vadisetti, Abhishek Niranjan, Kinnera Saranu, Raghava Sarma, M Ali Basha Shaik, and Periyasamy Paramasivam. Whisper augmented end-to-end/hybrid speech recognition system—cyclegan approac. *Proc. Interspeech 2020*, pages 2302–2306, 2020.

[20] Szu-Chen Jou, T. Schultz, and A. Waibel. Whispery speech recognition using adapted articulatory features. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages I/1009–I/1012 Vol. 1, 2005.

[21] Stavros Petridis, Jie Shen, Doruk Cetin, and Maja Pantic. Visual-only recognition of normal, whispered and silent speech. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6219–6223. IEEE, 2018.

[22] Chi Zhang and John H. L. Hansen. Analysis and classification of speech mode: whispered through shouted. In *Proc. Interspeech 2007*, pages 2289–2292, 2007.

[23] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99(7):1877–1884, 2016.

[24] Gunnar Fant. *Acoustic Theory of Speech Production: With Calculations based on X-Ray Studies of Russian Articulations.* De Gruyter Mouton, Berlin, Boston, 1971.

[25] Olivier Perrotin and Ian V McLoughlin. Glottal flow synthesis for whisper-to-speech conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:889–900, 2020.

[26] Isao Tokuda. The source–filter theory of speech, 11 2021.

[27] Paavo Alku. Glottal inverse filtering analysis of human voice production—a review of estimation and parameterization methods of the glottal excitation and their applications. *Sadhana*, 36:623–650, 2011.

[28] Philip Gage. A new algorithm for data compression. *C Users Journal*, 12(2):23–38, 1994.

[29] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

[30] Pengcheng Guo, Florian Boyer, Xuankai Chang, Tomoki Hayashi, Yosuke Higuchi, Hirofumi Inaguma, Naoyuki Kamo, Chenda Li, Daniel Garcia-Romero, Jiatong Shi, et al. Recent developments on espnet toolkit boosted by conformer. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5874–5878. IEEE, 2021.

[31] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. Audio augmentation for speech recognition. In *Sixteenth annual conference of the international speech communication association*, 2015.

[32] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

[33] Mengzhe Geng, Xurong Xie, Shansong Liu, Jianwei Yu, Shoukang Hu, Xunying Liu, and Helen Meng. Investigation of Data Augmentation Techniques for Disordered Speech Recognition. In *Proc. Interspeech 2020*, pages 696–700, 2020.

[34] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93:27403, 1993.

[35] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

[36] Marius Cotescu, Thomas Drugman, Goeric Huybrechts, Jaime Lorenzo-Trueba, and Alexis Moinet. Voice conversion for whispered speech synthesis. *IEEE Signal Processing Letters*, 27:186–190, 2019.

[37] Paavo Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech communication*, 11(2-3):109–118, 1992.

[38] Olivier Perrotin and Ian McLoughlin. A spectral glottal flow model for source-filter separation of speech. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7160–7164, 2019.

[39] Parham Mokhtari, Brad Story, Paavo Alku, and Hiroshi Ando. Estimation of the glottal flow from speech pressure signals: Evaluation of three variants of iterative adaptive inverse filtering using computational physical modelling of voice production. *Speech Communication*, 104:24–38, 2018.

[40] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. ESPnet: End-to-end speech processing toolkit. In *Proceedings of Interspeech*, pages 2207–2211, 2018.