UNIVERSITY OF SOUTHAMPTON

FACULTY OF PHYSICAL SCIENCES & ENGINEERING

School of Electronics and Computer Science

Improving Automatic Speech Recognition Transcription Through Signal Processing

by

Afnan Shah

Thesis for the degree of Doctor of Philosophy

Supervisor: Prof. Mike Wald

June_2017

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL SCIENCES & ENGINEERING

School of electronics and computer science

Thesis for the degree of Doctor of Philosophy

IMPROVING AUTOMATIC SPEECH RECOGNITION TRANSCRIPTION THROUGH SIGNAL PROCESSING

Afnan Shah

Automatic speech recognition (ASR) in the educational environment could be a solution to address the problem of gaining access to the spoken words of a lecture for many students who find lectures hard to understand, such as those whose mother tongue is not English or who have a hearing impairment. In such an environment, it is difficult for ASR to provide transcripts with Word Error Rates (WER) less than 25% for the wide range of speakers. Reducing the WER reduces the time and therefore cost of correcting errors in the transcripts.

To deal with the variation of acoustic features between speakers, ASR systems implement automatic vocal tract normalisation (VTN) that warps the formants (resonant frequencies) of the speaker to better match the formants of the speakers in the training set. The ASR also implements automatic dynamic time warping (DTW) to deal with variation in the speaker's rate of speaking, by aligning the time series of the new spoken words with the time series of the matching spoken words of the training set.

This research investigates whether the ASR's automatic estimation of VTN and DTW can be enhanced through pre-processing the recording by manually warping the formants and speaking rate of the recordings using sound processing libraries (*Rubber Band* and *SoundTouch*) before transcribing the pre-processed recordings using ASR.

An initial experiment, performed with the recordings of two male and two female speakers, showed that pre-processing the recording could improve the WER by an average of 39.5% for male speakers and 36.2% for female speakers. However the selection of the best warp factors was achieved through an iterative 'trial and error' approach that involved many hours calculating the word error rate for each warp factor setting.

Finding a more efficient approach for selecting the warp factors for pre-processing was then investigated.

The second experiment investigated the development of a modification function using, as its training set, the best warp factors from the 'trial and error' approach to estimate the modification percentage required to improve the WER of a recording. A modification function was found that on average improved the WER by 16% for female speakers and 7% for male speakers.

Table of Contents

List of Tables

List of Figure

List of Appendix Tables and Figures

DECLARATION OF AUTHORSHIP

I, Afnan Shah

declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

Improving Automatic Speech Recognition transcription through signal processing

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

Signed: Afnan Shah

Date: 28/05/2017

Acknowledgements

First of all, I would like to thank my advisor Prof. Mike Wald with the core of my heart. Prof. Mike has provided me continuous support and motivation during my Ph.D. study and research. I have benefitted a lot from his immense knowledge, patience and enthusiasm. Prof. Mike has continuously guided me through the research phase as well as during the writing phase. I cannot imagine having a better advisor and mentor for my Ph.D.

I would also like to thank my lab mates at University of Southampton. They have been involved in stimulating discussions and have spent sleepless nights with me in order to meet different deadlines. Apart from that, they have been a source of great fun during the four years of my Ph.D.

Last but not the least, I would like to thank my mother Sabah Sabagh and my sisters Maryam Halawani and Asmaa Halawani, who experienced all of the ups and downs of my research, for the moral and spiritual support that they have been providing me throughout my life. I must express my gratitude to them, for their continued support and encouragement.

Definitions and Abbreviations

The following table shows the abbreviations used in this work.

Abbreviation	**Meaning**
AC	Accuracy rate
ASR	Automatic speech recognition
D	The number of deletions
DF	Deep Formants – a formant extraction tool
DTW	Dynamic time warping
$\boldsymbol{F_1}$	The first formant
$\boldsymbol{F_2}$	The second formant
$\boldsymbol{F_3}$	The third formant
Ffmpeg	An open source library/tool that can record, convert and split audio and video into equal chunks
I	The number of insertions
N	The total number of transcript words
WER	Word error rate
OR	The values extracted of the original speaker recordings
ORT	The values extracted of the original speaker recording transcripts
PDA	Pitch detection algorithm
PR	The processed values extracted after modifying the speaker recordings
PRT	The processed values extracted after modifying the speaker recording transcripts
Praat	Free computer software package for the scientific analysis of speech in phonetics
RB	Rubber Band sound processing library
S	The number of substitutions
SEM	Standard error mean
SPSS	Statistical Package for the Social Sciences
SR	Speech recognition
ST	SoundTouch sound processing library
VTN	Vocal track normalization

Chapter 1: Introduction

This chapter introduces the research motivation followed by the research aim. The research objectives are specified then and the challenges involved are explained. Section 1.4 addresses the importance of the contribution. The research structure is finally presented.

1.1: Motivation

Students who have difficulty interpreting speech or taking notes, such as those deaf or hard of hearing, dyslexic, or non-native English, have a challenge with the accessibility of classroom lectures. Giving these students access to the spoken words presented within lecturers' speech and multimedia, gives them an opportunity to increase their engagement in the classroom and prevents their disability from being a barrier in the way of their academic success.

A number of methods are available to assist such students to access the spoken words of lectures, such as stenographers, sign language interpreters, lip-reading, and third-party note-takers. The use of lip-reading is insufficient for hearing-impaired students because even the best lip readers can miss important words, since only 30-40% of speech is visible on the lips (Berke, 2009). There is a shortage of fully-trained stenographers and they are costly to hire (Osterrath, et al., 2008). Third party note-takers are usually inexperienced in that lecture discipline and there is no guarantee of the notes' quality.

The previous methods listed are costly, time-consuming, and the students might be not able to afford to pay for them in order to get accurate transcripts or captions. Also, some of these methods, such as interpreters, can cause hearing-impaired students to develop low self-esteem or isolation because they are another barrier to the communication and interaction between hearing-impaired students and lecturers, and between hearing-impaired students and other students (Kersting, 1997).

To provide text transcripts of lecture recordings, automatic speech recognition (ASR) can be used but it can only reach an 85% accuracy rate on average, for 40% of speakers (Wald, 2007). Baeker et al. (2007) found that ASR transcripts had a low rate of accuracy, e.g. 55-60%. Hirschberg, et al. (2000) concluded that 60% of lecturers' transcriptions would not reach the required accuracy rate, hence ASR performs better with some speakers rather than others of same gender and accent. This presented an interesting question about does modifying the speech of speakers that have low accuracy rate to match the speech of speakers that have high accuracy rate helps the ASR performance.

Although the student might be able to understand the gist of the lecture, the low accuracy rate affects the disabled student's ability to gain full understanding of the lectures and to search the lecturers' transcripts (Amento, et al., 2002). Papadopoulos and Pearson (2009) found that the quality of transcriptions of lecture recordings did not improve significantly with training. Papadopoulos and Pearson (2011) also developed a tool that reduce editing time by 44% for 88% accurate transcripts that students found usable. The use of a human intermediary can help to improve the ASR accuracy in classrooms. Editing the ASR transcripts with human intervention may require the work of human editors be checked to ensure that the transcripts are consistent and achieve the correct level of quality (Melanie & Cialdini, 1998).

Although ASR systems that produce transcripts with fewer errors will require less editing time and effort to correct and are accepted by academics, these systems do not provide transcriptions that support disabled students, even after training those systems with the academic's voice (Pearson & Papadopoulos, 2011). The accuracy of the ASR is one of its biggest problems. It might produce low quality transcripts, which are not useful for educational purposes. Also, the ASR cannot be widely adopted (Huang & Deng, 2004) given the challenges of system error correction. Many research studies have been carried out aimed at improving the accuracy of ASR transcripts and reducing the WER (Boulain, et al., 2007) (Bell, 2007) (Daher, et al., 2005) (Baecker, et al., 2006). These use two techniques: enhance the ASR system by training the acoustic and the language models provided in the commercial ASR systems, and reducing the WER (Padrnanabhan & Mangu, 2001) in the ASR transcripts by editing their errors either automatically or using human intervention. Both techniques are costly and time-consuming when provided with a reasonable amount of training data. This research therefore focuses on improving ASR performance and thus reducing the transcript errors, in order to simplify the automatic correcting tools or those based on the human intervention to improve the transcripts sufficiently to support disabled students. Since higher accuracy means less editing time (Pearson & Papadopoulos, 2011), any error reduction could be effective.

1.2: Research Challenges

This work investigates the need to provide a method for improving access to lecturers' speech at low cost by providing more accurate lecture transcripts via ASR. This work attempts to reduce the need for training the system with the lecturer's voice, as it is time-consuming to provide a reasonable amount of training data. Pearson & Papadopoulos (2011) pointed out that the quality of their system's transcripts did not improve significantly, even after extensive training of the system.

A way needs to be found to make speaker independent recognition systems more accurate as it's impossible to train such systems especially when these systems failed to get a reasonable accuracy rate while recognizing some speaker voices. Also, improving the accuracy of the speaker independent recognition might be leading to minimize their transcript errors, therefore reducing the cost of correcting these errors, since paying human editors to correct errors is expensive.

1.3: Research aim

Much research has investigated improving ASR, either by improving the system's techniques and methods, or by correcting the system's transcripts.

The aim of this research is to find a way that students can improve the accuracy of speech recognition of lecturers. By making speaker independent recognition more accurate, students will not have to ensure that the lecturer trains a speaker dependent recognition to their voice, or pay somebody to correct errors.

This work investigates whether it is possible to use speech processing library to transform the speech characteristics of people whose ASR systems normally achieved low accuracy rate while recognizing their voice, so that it more closely resembles the characteristics of the speech of people who achieve normally high accuracy.

A literature search did not reveal any previous studies investigating whether accuracy could be improved by processing the audio files before transcribing these via the ASR system.

1.3.1: Research question and hypotheses

The aim of this investigation is to find the answers to the following research question:

How can speech recognition accuracy be improved by signal processing audio files before transcribing these via an ASR system?

This research question has been explored in two ways: the iterative approach, and finding a more efficient approach for selecting the modification factor. The iterative approach was used to answer the following sub research question *(SQ1) What is the impact on ASR transcription* *accuracy of pre-processing recordings before transcribing them*? This sub-question has been converted into hypotheses for the pilot phase. The WER improvement and frequency warping were used for forming corresponding hypotheses, each addressing the overall success. The frequency warping dimensions of processing recordings was also used for forming the hypothesis relating to WER improvement. The hypotheses are shown below for testing the relationship between processing the recordings before transcribing them and WER improvement of ASR.

H1: pre-processing the recording by adjusting pitch and speech rate leads to adjustment of the formants in these recordings.

H2: adjustment of the formants in recordings by pre-processing through adjusting pitch and speech rate before transcribing them with ASR can lead to significant improvement in WER.

Figure 1.3.1-1 shows the hypotheses of SQ1 for testing the relation between processing the recordings and warping the formants, and the overall success of processing the recordings before transcribing them.

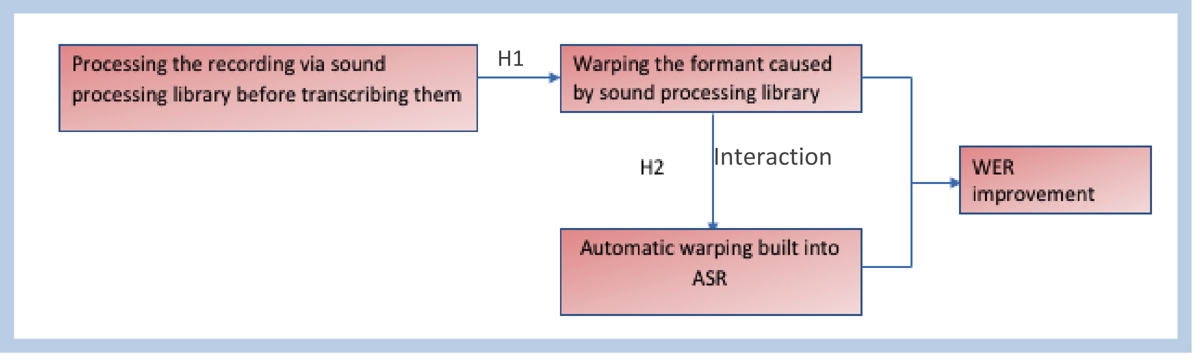

Processing the recording via sound processing library before transcribing them
H1
Warping the formant caused by sound processing library
H2
Interaction
WER improvement
Automatic warping built into ASR

Figure 1.3.1-1 Conceptual relationship between processing recordings and automatic warping built into ASR with overall success factors

The other way, finding an appropriate efficient approach by limiting the time taken to find the optimal modification for each recording, has been explored by answering the following sub research questions:

(SQ2) What is the minimum recording length and transcription that can be used to represent the average WER of the whole recordings for evaluating the transcript?

(SQ3) To what extent is the ASR transcription accuracy improved with the use of a missing value estimation algorithm to estimate the modification factor for processing the recording of a new speaker?

The following figure shows the report structure based on the research sub-question and hypotheses

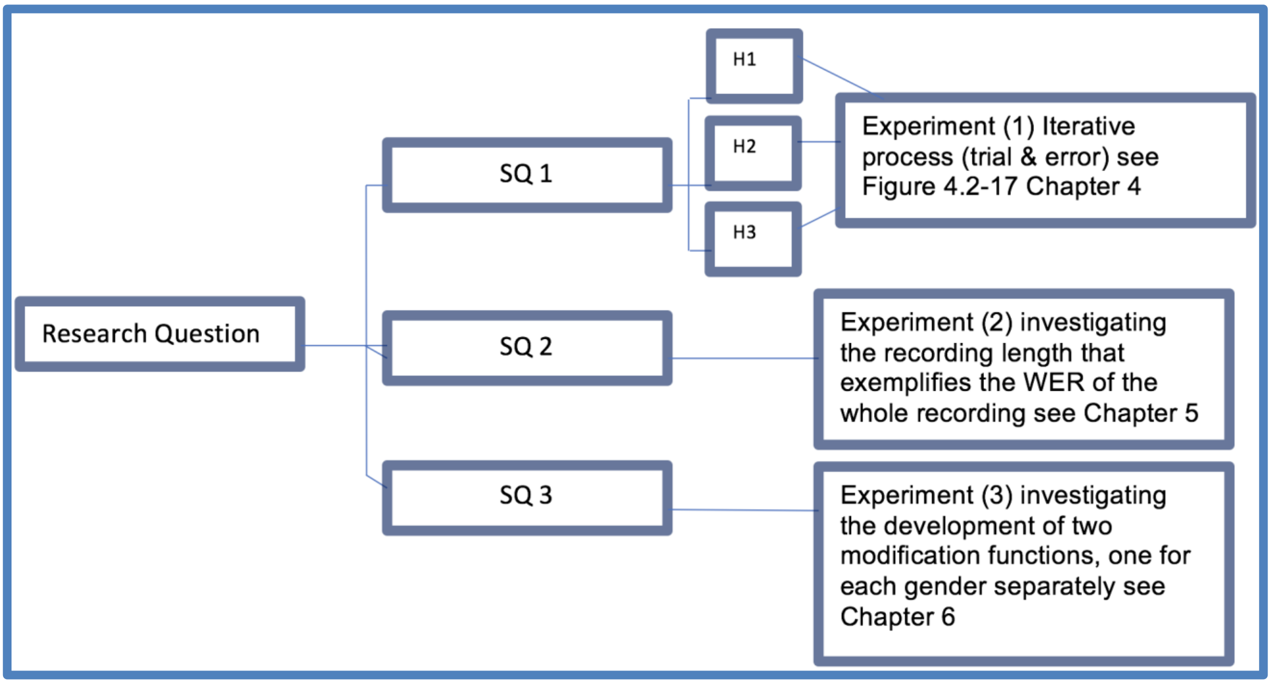

Research Question
SQ 1
SQ 2
SQ 3
H1
H2
H3
Experiment (1) Iterative process (trial & error) see Figure 4.2-17 Chapter 4
Experiment (2) investigating the recording length that exemplifies the WER of the whole recording see Chapter 5
Experiment (3) investigating the development of two modification functions, one for each gender separately see Chapter 6

Figure 1.3.1-2 Research sub-questions and hypotheses

1.4: Research contribution

To enhance ASR performance, previous research has investigated vocal tract normalization (VTN) in a speaker-independent ASR system. The method allows the ASR to automatically warp the input signal of a new speaker to be similar to the average signal of all other speakers in the system's training set, to cope with speaker variability (Guoping, et al., 2012) and to reduce the amount of the training data required (Jakovljević, et al., 2008). Other research enhanced ASR accuracy by correcting the system's transcripts, using either editors to correct the errors or automatic correction.

The contribution described in this work is a new approach to improving the accuracy of ASR transcription of lecture recordings through pre-processing the recordings by warping the formants and speech rate before transcribing these recordings using ASR. This modification of speakers' recordings is carried out with sound processing libraries that adjust formants through adjusting pitch and adjusting speech rate. Hence, this research suggests that since the warping factor of ASR was estimated automatically, warping the signal manually before warping it automatically through ASR improved the estimated value of the warping factor, which in turn improved the performance of the system.

The recordings were modified using a sound processing library in two ways: the best modification factor for each recording was iteratively chosen as the one which gave the minimum WER among the modified versions of that recording output from ASR, while the second approach developed a modification function by using the best warp factors resulting from the iterative procedure as a training set for estimating a modification function for each gender separately.

The results of the iterative procedure showed that, stretching the fundamental frequency and decreasing SR for recordings of males before transcribing them improved the WER by 39.5% while compressing both fundamental frequency and SR for recordings of females improved the WER by 36.2%.

These results encourage application to more samples to enhance the reliability of the results, and to establish a computationally-efficient algorithm for pre-processing recordings using a suitable warping factor.

Secondly, this research also contributes an efficient approach to selecting the adjustments. The recordings of males and females were modified, based on a modification function developed for each gender separately, which resulted in a significant improvement in the WER by 7% and 16% respectively.

1.5: Thesis Structure

Chapter 2 addresses the background literature on ASR systems and performance improvements, considers methods already implemented in ASR to improve the system performance, and the challenges of these methods. It also discusses the impact that improving ASR performance is having on increasing student accessibility, and in understanding and searching the lecturers' transcripts.

Chapter 3 describes the research methodology used to review and validate the pre-processing recordings procedure. It also discusses the methodology for a more efficient approach to selecting the warp factors. This research methodology combines aspects of multiple experimentation and a case study approach.

Chapter 4 investigates and answers '***SQ1***'. The experiment of chapter 4 describes the procedure for pre-processing speakers' recordings, whose components and sub-components encompass iteratively modifies to the speaker's formants to find the 're-mapping' that produces the lowest WER. This chapter explains an experimental design methodology for evaluating the pre-processing method and discusses the findings of their validation and review. The experiment implemented in chapter 4 took a random sample of four speaker recordings. Identifying the impact of processing the speakers' recordings on ASR performance through the sample data initiated this research. The 'trial and error' approach was used in processing the recordings to find the optimal modification for each speaker's recording. Each speaker's original and modified recordings were chunked into 20 audio files to calculate means and standard deviations for each speaker. To analyse the sample data, the *NIST* measurement tool was used to measure the WER of the original and modified recordings' transcripts. The *ANOVA* test was used to compare the mean WER on original recordings' transcripts (ORT) with mean WER on processed recordings' transcripts (PRT). Accordingly, the total average WER difference between the original and modified transcripts was calculated to get the improvement ranges for the transcript errors.

The results showed that warping formants through altering pitch and SR of recordings improved the performance of ASR by an average of 39.5% for male speakers and an average of 36.2 for female speakers, which in turn reduced the time required to edit these transcripts to correct any remaining errors. The result also showed that recordings of males were improved by increasing their pitch, while female recordings were improved by decreasing their pitch.

Chapter 5 shows the statistical analysis results of the pre-processing procedure and the findings of the pilot study and the user validation for the transcripts that resulted from the pre-processing procedure. It also discusses the findings of the user validation for the transcripts that resulted from the pre-processing procedure.

The ANOVA test showed a statistically significant difference between the mean WER in the PRTs and the mean WER in the ORTs with $p<0.001$, also the results of chapter 5 showed that 95% confidence that the mean WER of all speakers improved by between 15% and 20%.

Additionally, chapter 5 investigates and answers ***'SQ2'*** by implementing experiment 2 to investigate the chunk of the speaker's recordings that exemplifies the WER of whole recording. This experiment has been implemented by splitting the first and second minute of audio file for each speaker into 4 different segments so that each segment contained 15 seconds to investigate whether the average WER of the first or second minute of the recordings could be used to represent the average WER of the whole recordings. It is conjectured that recognition can take a short while to adapt to the speaker and also the first few words spoken have no previous words/context to assist with that recognition. The reason behind this is that ASR automatically looks for a sequence of three words in real time to adapt its statistical models to new voices, and the first spoken words are less accurately observed because they are not preceded by any content that can be used by the statistical language models (Machlica, et al., 2009).

A paired t-test was used to compare the average WER of the difference between the first minutes' WER and the second minutes' WER, of both original and processed recordings, with the average WER of the difference between the whole original recordings' WER and the whole processed recordings' WER.

The result of this investigation showed that the average WER of the second minute of the speaker recordings could be used to represent the average WER of the whole recordings (Chapter 5).

The second part of chapter 5 experiment investigates whether it was possible to pick the transcript that has the lowest WER from the PR transcript which discussed in details in that chapter. The reason for this is that if the results showed that the best matching transcripts could be distinguished without any correcting procedure having been performed on them, then time would be saved in finding the best pre-processing for each speaker. This investigated whether pre-processing provides a noticeable improvement in transcript accuracy. The transcripts of the audio segments were presented while listening to the original audio segments for each speaker, without any correcting procedure being carried out on those transcripts.

The result showed that the improvement in the transcript errors was noticeable when the transcript errors reduced by more than 2 and the WER improvement was more than 11% (Chapter 5).

Chapter 6 investigates and answers ***'SQ3'***. This chapter explains an experimental design methodology for investigating and establishing an efficient way to determine the optimum pre-processing frequency warp factor and discusses the findings of that experiment. It reports the statistical analysis of the results.

Firstly, chapter 6 investigates whether the estimation algorithm helps in finding the best warp settings quicker than the manual process. A cross-validation test was used to split the dataset of formants detected in the original and modified recordings into training data sets to build the estimation algorithm and validation data sets to test that algorithm.

The result showed that using an estimation algorithm for selecting the warp factors helped reduce the time taken to find the best warp factor for the new speaker, rather than the 'trial and error' approach that involved many hours to calculate the WER for each warp factor setting.

Additionally, chapter 6 investigates whether an estimation function for the modification values for the new speakers improves the efficiency of creating accurate transcripts through ASR, similar to the manual process. The paired t-test was used for this investigation.

The results showed that an estimation function based gender improves the efficiency of creating accurate transcripts significantly by 16% and 7% for female and male speakers respectively.

Chapter 7 summarises the contributions and the possible impact on research. Firstly, investigates whether the ASR's automatic estimation of VTN and DTW can be enhanced through pre-processing the recording by manually warping the formants and speaking rate of the recordings by using sound processing libraries (rubberband and soundTouch) before transcribing the pre-processed recordings using ASR. The selection of the best warp factor for each speaker's recording was achieved through an inefficient iterative 'trial and error' approach that involved calculating the word error rate for each warp factor setting. The ASR improved by an average of 39.5% with increasing the pitch and decreasing the speech rate of the male speakers resulting in stretching the formants and compressing the time series. For female speakers, the system improved by an average of 36.2% with decreasing both pitch and speech rate, resulting in compressing both the formants and time series. Secondly, original 320 recordings and their best modified versions were analysed statistically where the results showed with 95% confidence that the mean WER of all speakers improved by between 15% and 20% with p <0.01. The second approach used, as a training set, both the original recordings and the best modified version of those recordings resulting from the ‘trial and error’ approach to investigate a more efficient approach for selecting the warp factors for pre-processing the speakers’ recordings. The formants of the original and modified recordings for each speaker were extracted and then the mean and median of those formants were used as training sets, and each function that was defined by these training sets for the male and female separately was used to compute the new speaker modification mean and median formant values. The results showed that modifying the male recordings based on the estimated values by the male modification function, and modifying the female recordings based on the values estimated by the female modification function improved the WER significantly by an average of 7% and an average of 16% with p <0.01 for male and female speakers respectively. Finally, the important directions of future work are identified.

Chapter 2: Literature review

In the educational environment, automatic speech recognition (ASR) could be a solution to the problem of gaining access to the spoken words of a lecture. Specifically, ASR can be used to provide transcripts of lecture recordings to assist students who have difficulty interpreting speech data and find note-taking difficult, such as those with hearing impairments, dyslexic students, and foreign language speakers. Papadopoulos and Pearson (2011), supported by the findings of Wald (2007), found that disabled students get the full benefit of a lecture's transcript when its accuracy rate is 85% or more. However, ASR can only reach this level of accuracy for 40% of the teaching staff (Wald, 2007). Thus, 60% of lecture transcriptions would not achieve the required accuracy to enable a student to understand the gist of the lectures. Gaur et al. (2016) described the quality of ASR systems as still too low, and improvements are thus needed to make them accurate enough to use instead of manually converting the speech to text. In order to use ASR transcription, the WER has to be less than 30%, where editing a greater WER takes longer than producing the transcription from scratch (Gaur, et al., 2016). Gaur et al. (2016) pointed out that the limitations in WER arising from the wide range of speakers, the variation of acoustic features between speakers, the lack of training data, and the need to train the ASR system on each new voice.

To cope with both the perceptual problems, such as the variation of loudness and speed levels among speakers, and the wide range of speakers, by reducing the need to train the system on each new voice, ASR implements automatic vocal tract normalization (VTN) that warps the formants (resonant frequencies) of the speaker to better match the formants of the speakers in the training set (Acero & Stern, 1991) (Emori & Shinoda, 2001). While this automatic warping method still presents a challenge to ASR systems (Madhavi, et al., 2016), an investigation suggested that manually estimating these warp factors can produce lower WERs than automatically estimating them (Cohen, et al., 1995).

Section 2.1 presents the definition of ASR and points out the common problems that complicate speech recognition in processing the speech signal, and also highlights the use of ASR in the classroom for students who have difficulty in interpreting speech data. This is followed by a review of basic concepts, which are: formant frequencies (section 2.2), formant frequency estimation (section 2.3), sound processing libraries (section 2.4), and speaker normalization techniques (section 2.5). Section 2.6 presents a comparison between the formant modification caused by speaker normalization techniques and that caused by pitch modification. Section 2.7 reviews some of the speaker normalization methods that have been implemented in ASR.

2.1: ASR and the system's common problems

Ventre, et al. (1996) defined ASR as a process that maps an acoustic speech signal to text. Babu & M.Bennett (2011) defined ASR as a system to process the speech into text using a language model and an acoustic model. Gouvêa & Evandro (1998) pointed out that an ASR system uses acoustic modelling techniques, based on multiple modelling, and auxiliary acoustic features to process the acoustic features of the speech signal. Ariki & Shigemori (2003) indicated that speaker-independent ASR systems include acoustic modelling based on techniques for adapting the system to a new speaker's acoustic features. This results in good performance, achieved by matching the model to the task with adequate training data (Benzeghiba, et al., 2007). Other techniques are based on pronunciation modelling such as Hidden Model Sequences and use of representation frameworks such as FST, and larger and diverse training corpora. Picheny (2015) represented the basic structure of a speech recognizer, shown in Figure 2.1-3 below. The main components involved are a feature extractor, a speech engine, a language model, and an acoustic model. The feature extractor takes the speech signal as its input and extracts important features, which are then used by the acoustic model to describe how different words are realized as sequences. The different words or strings of words are then assigned probabilities by the language model. Bahl, Jelinek, & Mercer (1983) defined the language model as a technique for using statistics to assign a probability to a sentence, which represents a probability of how words might be displayed in the sentence. It also contains the vocabulary that represents a group of words, and a written representation of each word. Schwartz (1985) defined the acoustic model as a technique using statistical representation to train the ASR system on the sounds that each word makes, and it produces acoustic-phonetics for processing the acoustic aspects of speech sounds corresponding to a particular phoneme. Picheny (2015) pointed out the difficulty for ASR systems in recognizing words with similar sounds, such as "Austin" and "Boston". For that reason, speech recognition implements probability techniques as a part of the system's language model to find the word that has the higher probability in the recognized sentence. For example, "Austin" has a higher probability in the phrase "*Austin, Texas*" compared with "Boston" in this phrase. As mentioned by Picheny (2015), the speech engine of an ASR can choose the most suitable sequence of words by combining the inputs from the feature extractor and the acoustic and the language model; this was supported by Padmanabhan & Picheny (2002) who pointed out that the speech engine's output still needs to be adapted. These adaptation techniques, mentioned by Padmanabhan & Picheny (2002), consist of an acoustic model mechanism that adapts the model, based on limited amounts of test data from the speakers, and vocal tract length normalization that maps a new speaker to a matching speaker in the training set.

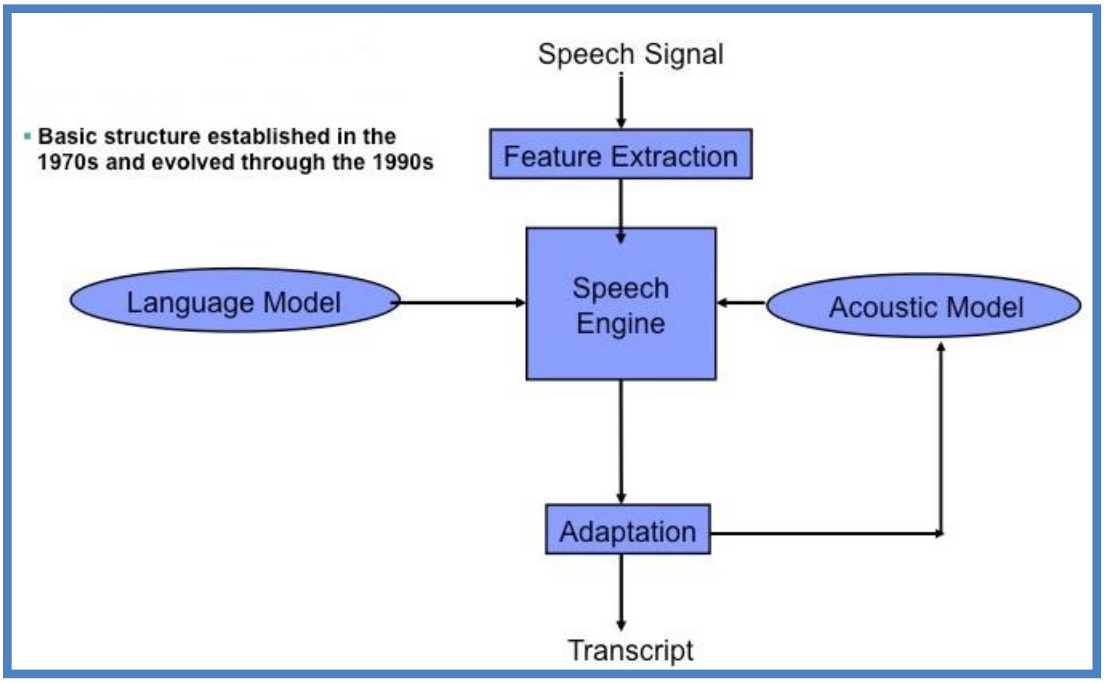


Figure 2.1-3 Traditional speech recognition system

There are some common problems that complicate speech recognition, such as the wide range of speakers, the variation of acoustic features between them, the lack of training data, and the need to train the ASR system with each new voice. Sommer (1994) found that the spectral and temporal properties of speech signals that distinguish phonetic categories can be substantially altered by factors such as phonetic context, stress, vocal tract size and shape, and speaking rate. Due to factors such as variation in speaking rate and source characteristics, challenges such as variability in acoustic realization of phonetic items arise (Bradlow, et al., 1996). Related research therefore considers the speaker variability as a perceptual problem which affects spoken-word recognition (Guoping, et al., 2012). Sommers & Barcroft (2006) found that the variation in the speaking style and speaking rate have significant effects on spoken word recognition while variations in the fundamental frequency were found to have no significant effect. The research also highlighted that varying the speaking style, fundamental frequency and speaking rate at the same time would reduce spoken-word recognition performance to a greater extent than by varying them one at a time. Benzeghiba (2007) divided the effects on speech recognition into three classes: modification by physiological or behavioural factors, intentional modification of voice to transmit important information, and alteration of word pronunciation.

Additionally, ASR faces a problem caused by the intrinsic variation of speech derived from the following points.

- Shrawankar & Thakare (2013) indicated that modelling the complex shape of vocal organs due to physiological differences is a difficult problem that affects ASR greatly. For example, one speaker's vocal folds vibrate more irregularly than another's; if one person has weaker articulatory muscles they may speak more slowly or less clearly (UCL, 2016).
- Jousse (2009) found that spontaneous speech could lead to conveying less information when processed by ASR.
- Russell (2010) indicated that children's speech could be difficult to recognize by ASR because of incorrect pronunciation due to less experience, and their use of language, especially a difference in vocabulary from adults.
- Benzeghiba (2007) found that long term habits, such as smoking and singing, have not been studied well.

So far, researchers have been attempting to improve the performance of ASR systems to match human performance levels in the classroom for students who have difficulty interpreting speech. However, ASR still produces low accuracy rates, requiring extensive editing, for the following reasons:

- The features used for ASR training may not contain all the useful acoustic or language information for recognition.
- There is a huge mismatch between the acoustic modelling assumptions of the system and the acoustic features of recognition.
- The modelling approaches may be too sensitive to speech variability.

Therefore, this research investigates whether modifying the speakers' acoustic features would enhance the ASR in the above cases, so that an ASR chooses the acoustic model that best matches this new speaker feature. Editing may still be required if the accuracy does not improve sufficiently.

2.1.1: ASR for students who have difficulty interpreting speech data

This section discusses the use of ASR in the lecture room to improve the access to information for students who have difficulty interpreting speech data and find note-taking difficult, such as those with hearing impairments, dyslexic students, and foreign language speakers. ASR in the educational environment could be one solution to the problem of gaining access to the spoken words of a lecture. Bain, et al. (2005) reported that instructors were trained in IBM Liberated Learning Projects to use IBM ViaScribe, a specialised speech recognition tool for note-taking in classrooms. Note-taking is made easy by displaying the classroom speech in the form of words on a big screen in the classroom or on handheld devices that students may have. The converted text may also be edited by the teachers and uploaded to the internet. In order to generate a multimedia script, ViaScribe can also incorporate PowerPoint slides into the speech and text, which can result in increasing the students' attention. Wald (2007) reported that, compared to a standard benchmark, nearly 40% of users achieved 85% accuracy. However, to achieve 95% accuracy, one hour of editing time is required for each hour of lecture, while to achieve 65% accuracy, the editing time for a lecture is much longer. Even so, the total time and cost required for ViaScribe is still less compared to a transcription created by typing. The research also indicated that to create good transcripts, it is currently essential to use human intervention in correcting transcript errors. The research method attempts to speed the correction procedure by highlighting the transcript's errors and thus ensure that the editor notices the errors and makes the corrections. An experiment was performed using two people to edit the transcript to see if this would decrease the correction time (Wald, 2007); one person would type the corrections while the other spotted any further errors simultaneously, and amend using two screens. A single editor was capable of correcting 24% of the errors in the transcript, while two editors corrected 44% of the errors. Thus concluding that more editors would help to minimize errors and improve speech recognition accuracy, because as soon as one person spots an error they do the correction while the other person's role would be to move a cursor to a different error. It was also noted that real-time editing was feasible without specialist skills and training. Although Wald (2007) succeeded in speeding up the editing procedure to correct errors in real time, the estimated efficiency of corrections was not determined. Iglesias, et al. (2014) discussed a Spanish educational project that uses real time captioning through Dragon Spanish speech recognition, with the aim of inclusive education. Their research evaluated the provision of real time captioning/transcription during 50 minute lectures for hearing-impaired students at a university. Students of a university's regular undergraduate course and those at a school for hearing-impaired children were satisfied with the results when the WER of the generated captions and transcripts was less than 10%, which was rarely obtained using ASR in lecture rooms. Since the level of the lectures at university is usually high and lectures are not easy to understand, even changing the key word of a caption's sentence makes that sentence difficult to understand.

Wald (2010) developed Synote that used speech recognition for advanced learning for both hearing impaired and non-impaired people. The system synchronises audio or video recordings with a visual transcript in the form of slides and notes. The system is highly valued by all students, including those who are deaf and use lip-reading or require a sign language interpreter, as well as those who are dyslexic and foreign language speakers. Of 100 students surveyed on their use of Synote, more than 80% were satisfied with its ease of use. The students also appreciated the design of synchronised slides, audio, video and transcripts. Further, 97% wanted all their lectures to be available on Synote, such was its benefit for all who want to learn with multimedia applications such as PowerPoint. Synote uses human editors to edit each generated video's captions, only one person being able to correct a caption at a time. Synote then uses a matching algorithm to compare all the different corrections to each caption by finding agreement among the corrections, before accepting the edit as 'correct'.

Other research has studied how to generate accurate transcripts of lectures by ASR, where disabled students receive the full benefit of these transcripts (Pearson & Papadopoulos, 2011). They developed a Semantic and Syntactic Transcription Analysing Tool (SSTAT) based on natural language processing and human interface design techniques. This tool helps produce post-lecture materials with minimal time and effort by the staff and students, that makes full access to the lecturer sources. The tool uses three functions: analysing text and identifying erroneous syntactic and semantic transcription, classifying errors, and removing lexical inconsistencies such as false start or hesitations. The lectures are recorded and the audio files are processed by the ASR tool. The transcripts then pass through the SSTAT tool for analysis. After analysis, lexical errors are removed. SSTAT produces a text-based document which identifies all kinds of error and serves as the basis for targeting the ASR training process. Finally, the academics can correct the transcript errors and train the software by dictation. This research shows that HCI technology can be combined with Natural Language Processing research, and both domains could be used in removing different kinds of error, leading to improved accuracy. However, after spotting and highlighting all the transcript errors, the editing procedure still requires the use of human editors, there is still a need to implement a quality control classifier that prevents editors from submitting poor quality transcripts by checking each editor correction to find whether this correction is the most suitable edit for the highlighted error (Lee & Glass, 2011).

The use of ASR in the lecture room saves the time and effort of transcribing the lecture recordings manually and benefits the students when the system produces a transcript with 85% accuracy. However, as highlighted by Wald (2008), more than 60% of users did not achieve 85% accuracy in the lecture room, whereas editing WER of greater than 15% Wald (2008) or greater than 30% Gaur et al. (2016) took more time and effort than manually transcribing the audio files. Pearson & Papadopoulos (2011) found that to obtain high accuracy, extensive training of the system is necessary by the academics delivering the lectures. Where Pearson & Papadopoulos

(2009) research showed that an untrained system can only achieve 65% accuracy, this method only improved by 4.3% after training for 100 minutes (consisting on average of 15,000 words, based on Wald (2008) which demonstrated that speakers talk an average 150 words per minute).

There is a need to investigate a method that could improve the performance of ASR systems while decreasing the WER to less than 30%. For such a system to be acceptable in the lecture room, it could employ correcting methods to edit the transcripts with a WER less than 30%, such as those presented by Wald (2009) or Pearson & Papadopoulos (2011).

2.2: Formant frequencies

Acoustic analysis describes the changes in the speech production system, which consists of the respiratory, laryngeal, and vocal tract (oral and pharyngeal resonating cavities) subsystems (Stathopoulos, et al., 2014). The frequency formants (F1, F2, F3 and F4) are the spectral peaks of the sound spectrum in the frequency domain, caused by resonances in the vocal tract (Sommers, et al., 1994). Figure 2.2-4, from Romenesko, et al. (2016), shows the relation between acoustic features that include the formants and anatomical structures of speech.

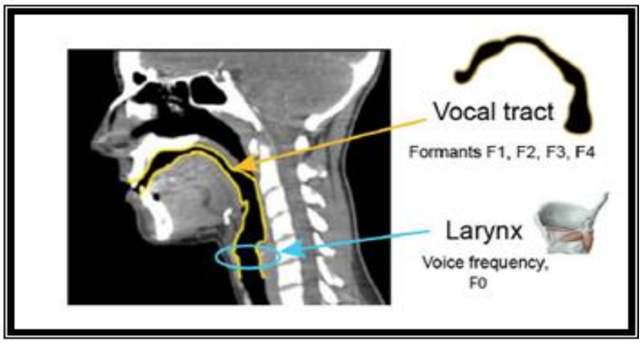


Figure 2.2-4 The vocal tract (resonator) and the larynx (sounds source)

The commonly used definition for the formants is a range of frequencies of sound in which there is an absolute or relative maximum in the sound spectrum (American Standards Association, and Acoustical Society of America, 1960). Romenesko, et al. (2016) defined the formant frequencies as the resonant frequencies of the vocal tract for a particular vowel production. A vowel is defined as a spoken sound produced by a comparatively open configuration of the vocal tract and with vibration of the vocal cords but without the audible friction (Chauhan, 2013). Vowels are classified into different categories based on the position of the tongue and lips and whether or not the air is released through the nose when they are spoken (Chauhan, 2013).

Rabiner & Juang (1993) indicated that, although vowels have extremely low relevance for recognition of written text, most speech recognition systems depend on vowel identification for high performance because vowels are more simply and reliably recognized, since all vowels differ spectrally (Beigi, 2011). Also, vowels contain most of the periodic parts of speech and they possess more information about the resonance of the vocal tract, namely the fundamental frequency and the formants (Beigi, 2011), so they provide all the information needed about the speakers and their accent (UCL, 2016). Peterson & Barney (1952) measured the formants of vowel utterances from a number of speakers, and discovered that the measurements of formants of dissimilar vowels tend to cluster, although with some overlap among clusters. They also discovered that vowel tokens of the same category uttered by different speakers had widely differing formant frequencies, while vowel tokens of different vowel categories uttered by different speakers had identical formant frequencies. Gouvêa (1998) showed that the centroids of clusters of common vowels could be represented in a "vowel triangle", as in Figure 2.2-5.

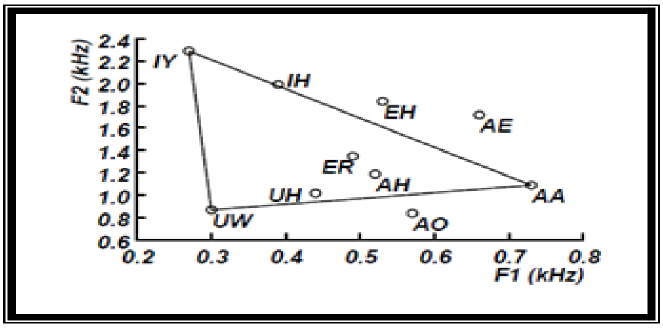


Figure 2.2-5 The "vowel triangle" is the centroids of clusters of common vowels (Gouvêa, 1998)

From previous research, it can be concluded that formants are resonances of the vocal tract and can be specified by the peaks of the spectral envelope, because the formant frequencies correspond to high energy regions of the sound spectrum.

2.3: Formant frequency estimation

Although the formants have a straightforward definition and a clear physical concept, the estimation of formant frequencies for a speech segment still counts as a fundamental problem in speech signal processing (Dissen & Keshet, 2016). Since formants are related to the position of the elements of the vocal tract that specify its features as a cavity resonance, and the changes of this position do not occur suddenly because of the mechanical limitations of the vocal tract, this therefore imposes a continuity constraint on the estimation of formants (Gouvêa, 1998).

The formants are estimated using linear predictive coder (LPC) methods for obtaining an estimate of vocal tract impulse response (Sahoo, et al., 2013). Although LPC analysis has been widely used to find the spectrum envelope (Fong, et al., 2013), LPC might lead to spurious peaks, while its envelope peaks tend to be drawn away from their true values in high pitched speech (Gouvêa, 1998). Therefore, researchers have investigated the use of other techniques with LPC to improve the automatic estimation of formants (Weenink, 2015) (Alku, et al., 2013), for example, using a tracking algorithm with LPC that predicts a range of possible estimation values and rejects values out of this range (Gouvêa, 1998), or using machine learning techniques that are trained on an annotated corpus of read speech (Dissen & Keshet, 2016). A more detailed study of Formant estimation is beyond the scope of this work.

In this work, *Praat*, a tool for general purpose speech analysis (Boersm & Weenink, 2002), has been used to extract the formants of the original and modified speakers' recordings, because *Praat* is the most popular tool used in phonetic research (Dissen & Keshet, 2016). *Audacity* (Audacity Team, 2012) has also been used, which plots the spectrum to give an idea of the frequency range of the original and modified recordings. This was useful in comparing the frequency range of the original recordings with that of the modified recordings. *DeepFormants*, an automated formant tracking and estimation tool, was used by Dissen and Keshet (2016) to estimate the formants of the original and modified recordings. This tool is publicly available at https://github.com/MLSpeech/DeepFormants.

The formants extracted using *Praat* were compared with those obtained using *DeepFormants* (Dissen & Keshet, 2016), to find which of them achieved better results when compared to the manually estimated reference from *Audacity*. The formants were extracted by each tool separately and used to modify the recordings, based on the estimated functions, before transcribing these modified recordings via ASR, thus identifying which tool gave the lower WER. This enabled the evaluation of *Praat*. The results of this experiment are discussed in Chapter 6.

2.4: Sound processing (time and pitch modification)

Sound processing is the intentional alteration of sound by applying an audio effect such as modifying the pitch and tempo. There are numerous sound processing libraries that enable an audio file's pitch and tempo to be modified, such as the Rubber Band library (Quey, 2012) and the SoundTouch library (Parviainen, 2001); both are free and open source. SoundTouch Audio

Processing Library[1] is the tool most commonly used in recent research to manipulate audio signals (Chang, et al., 2011) (Anguera, et al., 2014) (Six, et al., 2014) (Dias, et al., 2016) (Szczypiorski & Zydecki, 2017). It claims to produce high quality processed sounds, where the resulting processed signal keeps the same pitch and most of the original speech characteristics when modifying the speech rate, while the resulting signal keeps the same speech rate when modifying the pitch. This enables the production of a wide variety of processed audio files, each file having been processed by modifying its pitch and speech rate separately, in order to explore the effects of modifying each characteristic of the speech, before transcribing these recordings by ASR. These is why SoundTouch is used in this work. The Rubber Band audio processing library has also been used here to check whether processing the recordings with other sound processing tools would have the same impact on ASR performance. Table 1 shows the basic principles of these libraries.

Principle	**Definition**
Pitch modification	Aims to modify the pitch of the signal, without modifying the signal's duration, in a time-varying manner
Tempo (time stretch)	Slow down or speed up a given signal, without modifying the signal's spectral content, in particular without affecting the sound pitch
Playback rate	Modifies both tempo and pitch together

Table 1 Basic principles of sound processing libraries

2.5: Review of speaker normalization techniques

2.5.1: Speaker normalization

Speaker normalization maps the spectra of two speakers (Wang, et al., 2008) to decrease systematic variations of the acoustic features between speakers (Welling, et al., 2002). The warping functions are an attempt to map one speaker's spectra to another's. Thus, in the context of speaker normalization, a warp function can be defined as a function mapping two spectra (Gouvêa, 1998). This warp function is controlled by a scalar value called the warping factor (Mimura & Kawahara, 2011).

Ban, et al. (2014) pointed out that, because of the acoustic mismatch caused by differences of vocal tract lengths among speakers, the warping factor has to be estimated to represent the speaker characteristics, and then used to scale the frequency axis. This estimate is either based on

[1] http://www.surina.net/soundtouch

maximizing the likelihood or directly on speaker-specific acoustic features. Although the range of warping factors ranges between 0.88 and 1.2 (Mimura & Kawahara, 2011), estimation of the warping factor has been a systematic problem for normalization techniques. The reasons for this are that slight changes in the warping factor in the wrong direction lead to significant decreases in performance, as well as being computationally expensive to find for each speaker (Madhavi, et al., 2016).

Gouvêa (1998) defined the warping factor as one that could be estimated from the speaker's formants. This factor could be used to map the frequency spectrum of a standard speaker to that of a new speaker in a warping function without any constraints to make it a flexible framework. In other words, to map the spectrum ω to the spectrum ω' uses the warp function f with the warp factor α so that $f(\omega) = \alpha\omega$.

The effect of the warp function will be to compress or expand the spectrum ω, depending on whether α is greater or less than 1 (Ban, et al., 2014). Figure 2.5.1-6 shows the effects of the warp function. The typical warping function has two steps: find the segment with α = 1 that requires no warping, and then perform alignment for all possible warping factors α, for 0.8≤ α ≤1.2, to find the most likely value that aligns the frequency spectrum of ω with the frequency spectrum of ω'. Thus, when the warp factor α = 1.0, no warping is required, a warp factor of α > 1.0 corresponds to the warp function compressing the spectrum, while α < 1.0 corresponds to the warp function stretching the spectrum. In the example represented in Figure 2.5.1-6 **Error! Reference source not found.**, a linear warp function has been used.

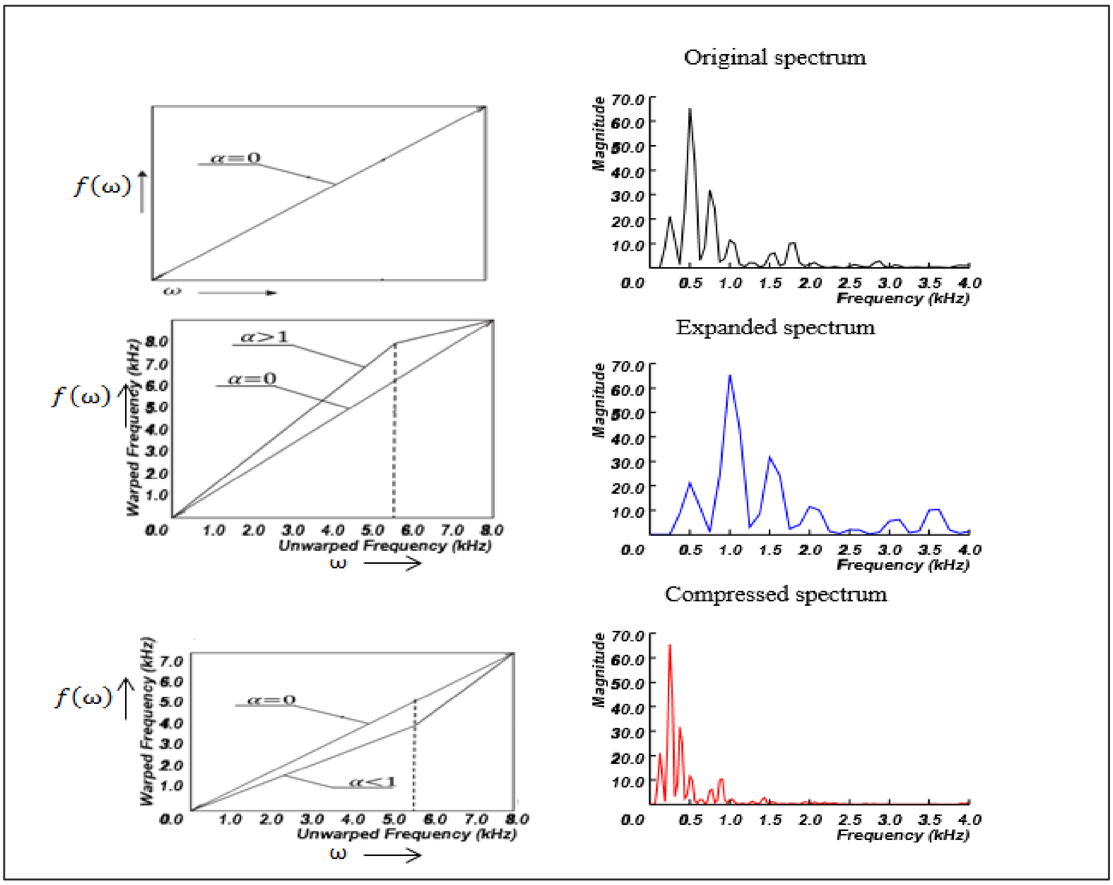


Figure 2.5.1-6 The impact of a linear warping function on the spectrum, compressing the spectrum when the warp factor > 1, and expanding it when < 1

These normalization techniques are mainly focused on compressing and stretching the speakers' formants to closely match those of the ASR training set, so that the performance of ASR can be improved. Accordingly, this work has been investigating the impact on the performance of ASR of compressing and extending the speakers' formants, with sound processing libraries which modify these formants by adjusting pitch and speech rate. The reason for this is to identify whether further modification of the formants with sound processing libraries, before transcribing these words, could interact with the automatic normalization method built into the ASR system, and thus improve WER.

However, in the implementation of normalization techniques there are two important issues: (i) the selection of the warping function, and (ii) the choice of the warping factor. The common choices for the warping function are: a linear function, curves based on speech perception studies such as the bilinear transform or transforms based on the mel scale, and curves based on speech production models (Gouvêa, 1998). The following sections describe some recent research.

2.5.2: Selection of the warping function

The easiest function to apply is simple linear frequency warping (LabMaster, 2013), as shown in Figure 2.5.1-6. Two commonly used warp functions are piece-wise linear warping (Uebel & Woodland, 1999), and bilinear warping (Acero & Stern, 1991). Another complex function used is the Sine-Log All-Pass Transforms (SLAPT) (McDonough, et al., 2001).

The bilinear warping function depends only on the warping factor α and is a nonlinear function based on the mel scale (Ban, et al., 2014). The mel scale was arises from psychophysical studies, which found that human perception of frequency intervals does not follow a linear scale (Gouvêa, 1998). This contrasts with the piece-wise linear function, which depends on the warping factor α and other factors, such as the fixed value ω (to control the bandwidth mismatching problem due to differences of vocal tract lengths of each speaker), and values b and c calculated from ω (Ban, et al., 2014). Figure 2.5.2-7 shows the piece-wise warping function for compressing and stretching the spectrum in the light of the warp factor α, where α > 1.0 corresponds to compressing the spectrum, while α < 1.0 corresponds to stretching it, and α = 1 corresponds to no warping.

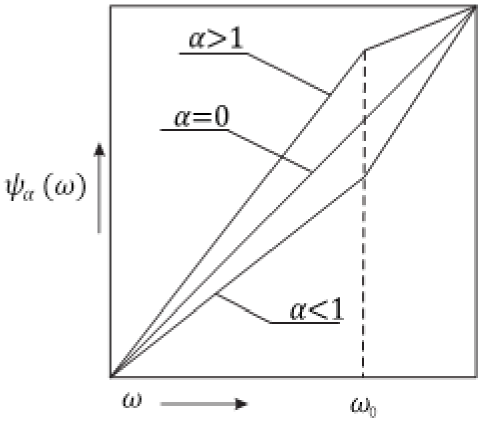


Figure 2.5.2-7 Piece-wise warping function according to Ben, et al. (2014)

The same is true in the bilinear warping function, shown in Figure 2.5.2-8, except that α = 0.42 in the bilinear function corresponds to the mel scale warping.

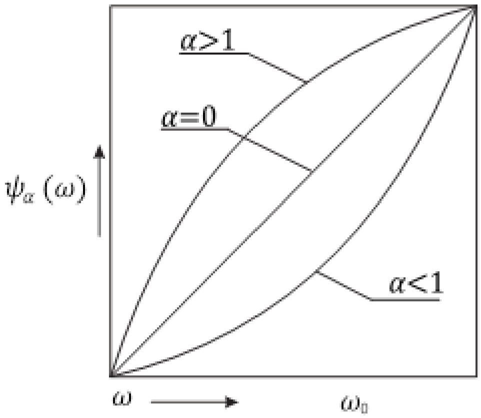


Figure 2.5.2-8 Bilinear warping function according to Ben, et al. (2014)

Obviously, the use of acoustic features that are relevant to speech perception, such as formants, should be central in the selection of the warping function for ASR systems which use exactly these clues to perform normalization.

2.5.3: Selection of the warp factor

The warping factor was selected based on either speaker-specific acoustic features or maximization of likelihood.

The warp factors employed in normalization methods are usually based on mapping the subglottal resonances (SGRs) and the third formant frequency (F3) of a speaker (Xuemin Chi, 2007) (Wang, et al., 2008). The reason for using SGRs is that they connect to the speaker's formants in particular ways while remaining phonetically invariant. The warp factor estimate is also based on F3, rather than the other formants, because it correlates with vocal-tract length (Arsikere, et al., 2013).

Faria & Gelbart (2005) measured pitch (F0), a perceptual property that enables the ordering of sounds in a frequency range (Morandi, 2012), to estimate a joint distribution of warp factors. They measured F0 because, if the recognition system is already performing pitch extraction for other purposes, then VTN can be performed at low cost using the F0-based warp factors. The results showed that pitch-based warp factor estimation reduces the word error rate by 2% per utterance and by 1.1% per speaker. However, other studies based on F3-based warp factors showed greater improvement in ASR performance, which resulted in a relative decrease in WER of more than 15% (Arsikere, et al., 2013).

Garau (2005) proposed an iterative procedure for selecting the warp factor by maximizing the likelihood of the hypothesis transcription at the output of the decoder. This study addressed the application of Vocal Tract Length Normalization (VTLN) to multi-party conversations in a meeting environment. The experiments were performed on NIST Spring 2004 Meeting evaluation data. VTLN normalization uses speaker-specific warping of the frequency axis by employing a scalar warp factor. The warp factor is estimated using maximum likelihood (ML). First, a Brent's method search based on quadratic interpolation is used to estimate the warping factor ∝. Then the likelihood of the normalized acoustic observation $X^{\propto}$ for a given transcription W and an acoustic model λ is maximized by using the equation below:

$$\alpha = arg \max_{\alpha}\big(Pr(X^{\alpha}|\lambda, W)\big).$$

VTN was applied to both training and testing. The following five step iterative procedure was applied.

1. Estimate warping factors using an un-normalized model, followed by normalized computation of features using these estimated warping factors.
2. Single pass retraining performed, starting from un-normalized models and a few Baum-Welch passes.
3. The previous pass acoustic models are used to estimate warping factors and compute normalized features.
4. Starting from the normalized models of the previous pass, a training pass is performed as in step 2.
5. Steps 3 and 4 are repeated until the word error rate (WER) on the data set has stabilized.

This procedure converges the warping factors to between 0.8 and 1.2, where the distribution of warping factors for females decreases to less than 1, and increases for males to greater than 1.

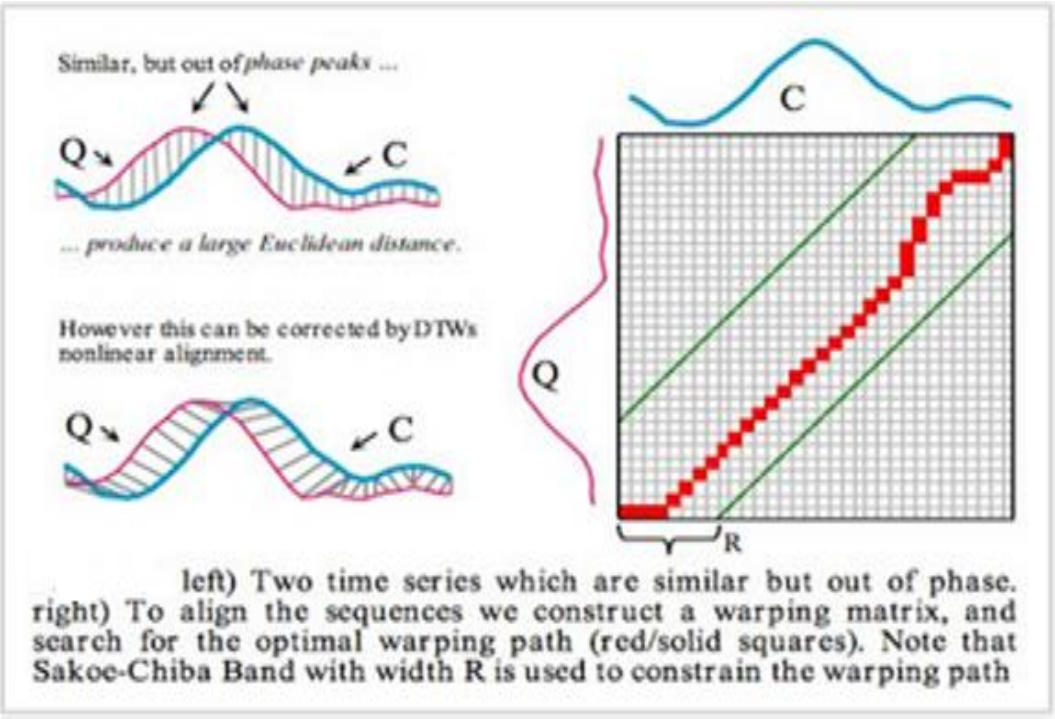
With 12 different speakers in the training set and 15 speakers in the test set, each set comprising 2 different meetings of 10 minutes, this research showed that using ML in VTN to find the best warp factor for the new speaker based on their acoustic features, improved the WER by more than 15%.

Gouvêa (1998) considers everything involved in the process of recognizing speech is needed to guarantee that the warping factor selected is optimal for the decoder being employed. In his research, the warp factor for a particular speaker is chosen from multiple warp factors as the one that maximizes the *posteriori* log-likelihood, computed using a Gaussian mixture model that statistically represents the standard speaker in the ASR training set. The reason for computing and using the Gaussian mixture model is that it is computationally expensive to compute a Gaussian mixture model for each new speaker that the system has to recognize. Gouvêa (1998) believed that the speaker's acoustic features relevant to speech perception, such as formant frequencies, have to be the basic elements while selecting that warp factor for that speaker, as human perception of speech uses the formants as clues to its normalization. Madhavi, Sharma, & Patil (2016) found that, to minimize the speaker differences in ASR, the spectra of the speakers have to be normalized using a scaling factor based on speaker formants such as F3 and F0.

It is clear from this review of the major approaches to speaker normalization and examples, that speaker-specific acoustic features are the major parameters in normalizing the spectra of the speakers. Hence, this work focuses on the selection of the warping factor based on speaker formants, and the iterative procedure in selecting the optimal warp factor for each new speaker discussed in Chapter 4.

2.5.4: Dynamic time warping

Meinard (2007) defined dynamic time warping (DTW) as an algorithm that aligns two time series by constructing a warping metric and warping both time series iteratively with the metric until an optimal alignment between the two sequences is found.


Similar, but out of phase peaks ...
Q
C
... produce a large Euclidean distance.
However this can be corrected by DTWs nonlinear alignment.
Q
C
C
Q
R

left) Two time series which are similar but out of phase. right) To align the sequences we construct a warping matrix, and search for the optimal warping path (red/solid squares). Note that Sakoe-Chiba Band with width R is used to constrain the warping path

Figure 2.5.4-9 Example of dynamic time warping algorithm (Lorica, 2012)

Figure 2.5.4-9 shows Lorica's (2012) implementation of DTW which aligns two time series while remaining inside the warping metric, in order to minimize the cost of the iterative warping, by choosing the minimum distance between the vectors of the two series, to find the optimal warping path and the best warp factor between the two time series.

2.6: Formant modification by speaker normalization and pitch modification

2.6.1: Modification caused by speaker normalization related to gender

Gouvêa & Stern (1997) found that formants from females are usually higher than those from males. The reason is that females usually have smaller vocal tracts than males. Hence using the same the normalization techniques result in compressing the females' formants, while expanding the males' formants. Figure 2.6.1-10 shows counts of warping factors for a linear warping function for a set of speakers. Gouvêa (1998) noted that the clusters were separated well by gender.

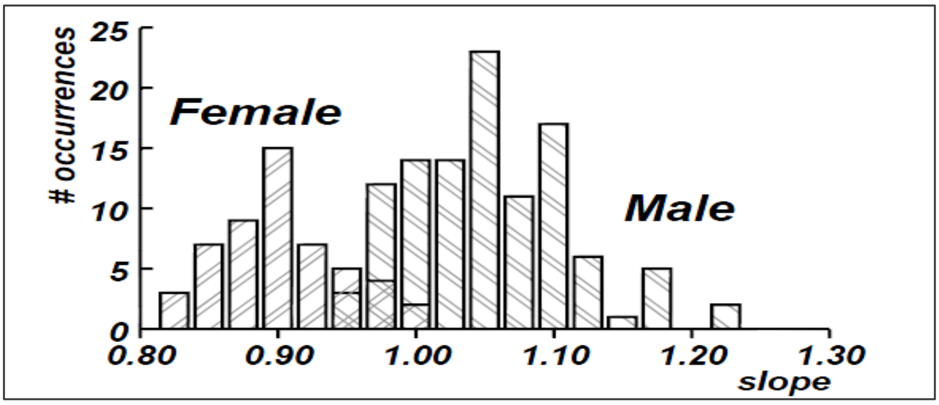


Figure 2.6.1-10 Distribution of warp factors indicating gender (Gouvêa, 1998)

2.6.2: Modification caused by modifying pitch of the signal

Morandi (2012) attempted to identify the change in the spectral content of the human voice caused by pitch modification, and found that modifying the pitch led to modifying the frequency spectrum of the speakers' voices. He also indicated that increasing the pitch resulted in expanding the speakers' formants, while decreasing the pitch resulted in compressing the speakers' formants. Figure 2.6.2-11 shows the frequency spectrum compression caused by decreasing the pitch and the expansion caused by increasing the pitch.

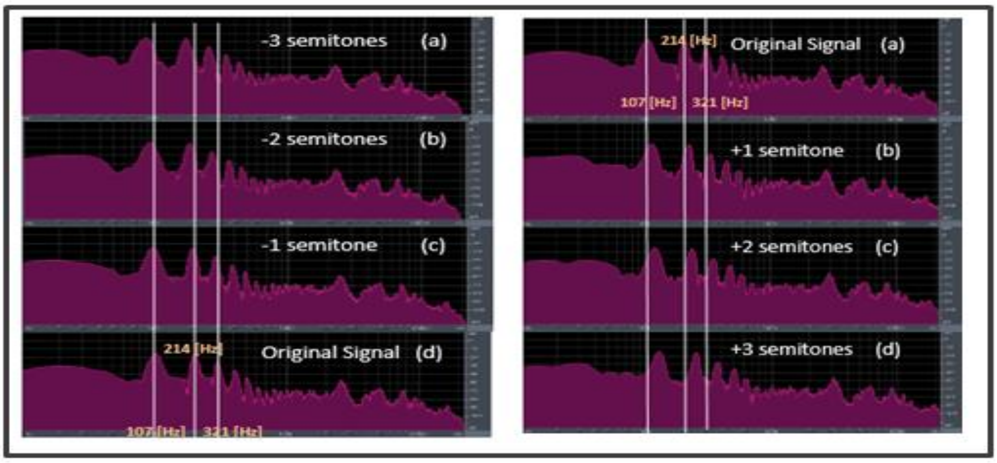


Figure 2.6.2-11 Compression and expansion of the spectrum caused by modifying the pitch (Morandi, 2012)

Obviously, speaker normalization and sound processing with pitch modification both resulted in compressing and extending the speakers' formants. The difference between these two techniques is that the speaker normalizations separated the speakers based on the speaker's gender.

2.7: ASR and speaker normalization methods

To improve the ASR performance and to cope with the problem of inter-speaker variability, the vocal tract length normalization method is frequently employed to transform acoustic features for ASR. Lee & Rose (1996) defined the VTN of ASR as a speaker normalization method that automatically warps the new speaker's acoustic features to the average acoustic features of the speakers in the training set. This method is implemented in ASR to deal with the wide variation of acoustic features among speakers, and so improve the WER of speaker-independent ASR.

Xuemin Chi (2007) and Wang (2008) pointed out that the warp factors employed in the VTN method are usually based on mapping the subglottal resonances (SGRs) and the third formant frequency (F3) of a speaker. Also, the warp factors are usually computed by maximum likelihood (ML) estimation (Emori & Shinoda, 2001). The reason for using SGRs is that they connect to the speaker's formants in particular ways while remaining phonetically invariant, while the reason for using F3 is that it correlates with vocal-tract length (Arsikere, et al., 2013). The vocal tract length varies from one speaker to another and is difficult to measure, but researchers found that F3 can be the best formant to estimate the vocal tract length and it can also be used explicitly in the vocal tract normalization method (Zhan & Westphal, 1997). The reason for this is that F1 and F2 overlap acoustically between vowel classes, which does not make much of a difference when identifying vowels, while the vowel-dependence of F3 is not powerful compared with F1 and F2 (Umesh, 2011). Warp factor estimates, based on ML-VTN, was reducing the WER by between 22% and 9%. The reason for this variation in improvement was because a large number of speakers and less data per speaker made the speaker normalization less effective, whereas a large amount of data per speaker caused the Hidden Markov models in ASR to be better tuned (Gouvêa, 1998).

Research by Liu (2006) and Faria & Gelbart (2005) measured F0 values in order to estimate a joint distribution of warp factors. F0 was used because, if the recognition system is already performing pitch extraction for other purposes, then VTN can be performed at low cost using the F0-based warp factors. Besides, it was shown by Faria & Gelbart (2005) that pitch-based warp factor estimation reduces the word error rate by 2% per utterance and 1.1% per speaker.

Meinard (2007) implemented a DTW algorithm, which aligns the time series (sequence of vectors) of the new spoken phrase with the time series generated by the Hidden Markov model in ASR, to cope with the problem of the variation in speech rate. Similarly, Salvador & Chan (2007) implemented DTW with ASR, with the alignment of the time series remaining inside the warping metric, in order to minimize the cost of iterative warping by choosing the minimum distance between the vectors of the two series.

2.8: Summary

This chapter has reviewed basic concepts and the major approaches to speaker normalization, with representative examples of their implementation. The methods discussed do not take into account whether the acoustic features that are relevant to speech perception, such as formant frequencies, were modified before they reached the ASR system, and the impact of this case on the performance of ASR.

Speaker normalization research has mostly focused on improving the procedure of processing the speech signal ***inside*** ASR, to reduce the speaker differences by warping the new speaker's formants to a matching speaker in the training set of ASR (Madhavi, et al., 2016). Nevertheless, research has not explored the impact of modifying the speech signal ***before*** it is processed by ASR, which has been the motivation for undertaking the present study. In the light of the limited studies into modifying the speakers' recordings before transcribing them, the present inquiry aims to explore the impact on the system's WER of modifying the speakers' recording through sound processing libraries, such as *Rubber Band* and *SoundTouch*, before transcribing these recordings.

This research focuses on two normalization methods: vocal tract normalization (VTN) and dynamic time warping (DTW). The selection of a suitable warping factor for these two methods still presents a challenge in speech recognition. Therefore, this work investigates whether normalizing the formants of the speech signal before normalizing these formants in the ASR system would enhance the selection of the warping factor within the system, and so improve the performance of the system. This work also investigates the effort needed to improve the ASR performance in this way, and the extent of the resulting system improvement.

Gaur, et al. (2016) indicated that starting with the ASR transcripts is worse and takes longer to edit than producing the transcription from scratch unless the Word Error Rate is < 30%. They implemented a system using TED-independent speech recognition, where the system generated transcripts with WERs between 15% and 45%. Their results showed that the participants were easily able to edit the transcripts with WER < 15%, while editing high WER turned out to be time-consuming and it was better to transcribe recordings with high WER manually from scratch. Therefore, this work investigates improving the performance of ASR which results in decreasing the WER of the transcripts generated, in order to either eliminate the editing procedure, if accuracy increased to 85%, or reducing the effort of other editing methods by starting with better quality transcripts to correct.

Chapter 3: Research Methodology

This chapter describes the research objective and a methodology that can be used to achieve this objective.

The objective of this thesis is to reduce the WER in ASR transcripts by investigating whether pre-processing the recording before processing it through ASR would improve the performance. Reducing errors in the generated transcript would eliminate the editing procedure or simplify the processes of other editing methods which involve human intervention to correct transcript errors.

Quantitative research is an investigation to analyse phenomena through statistical, mathematical or computational techniques (Given, 2008). Qualitative research is an exploratory research which used to understand the situation and help decide the most suitable design for the research, data collection method and selection of subjects (Shields & Rangarajan, 2013). It is conducted when a problem has not been clearly defined. It relies on techniques, such as:

- Secondary research that reviews available research methods or data
- Qualitative approaches, such as informal meetings with consumers, employees, and competitors
- Formal approach techniques through in-depth interviews, case studies and pilot studies

Additionally, case study research can be categorized as qualitative and/or quantitative method (Cluett & Bluff, 2006). Soy (1997) defines a case study, based on that presented by Yin (1984), as an experimental inquiry method that investigates a phenomenon with its context, to be used when the boundaries between phenomenon and context are not clearly understood.

Case study methods are applied when an in-depth investigation is needed for a single individual, group, or event. These provide a methodical way to collect data, analyse information and report results (Soy, 1997).

Another method that can be used to obtain knowledge about a problem is Trial and Error, which is one of mankind's problem-solving methods (Radnitzky, Bartley, & Popper, 1987), where the agent performs multiple attempts until success is achieved, or the agent stops trying. Also, trial and error method is used for tuning, or repairing, and may not give the optimal solution. The method is given different names in different fields. Of the two basic approaches to problem solving, trial and error is one, and insight and theory the other (Bei, Chen, & Zhang, 2013). A variation is known as *guided empiricism* which takes guidance from theory to choose a starting point. The method has been studied from its natural computational viewpoint and is commonly used in situations where there is little knowledge about the problem. The features of trial and error are:

- A solution-oriented method that does not investigate why a solution works.
- The solution provided is specifically for the problem at hand and is not generalized for other problems.
- Generally, does not find all solutions but instead gives a solution, which may not be the best solution.
- To proceed with this method, little knowledge or no knowledge about the solution is required.

Table 2 shows common types of research design and the differences between them.

Applied research	**Objective**	**Methods**
Quantitative research (Given, 2008)	- analyses the data with the help of statistics	- statistical techniques - mathematical techniques - computational techniques
Qualitative research or Exploratory research (Shields & Rangarajan, 2013).	- analyse preliminary information - define problems - suggest hypotheses	- reviewing secondary data - reviewing available literature - case studies - pilot studies - interviews - questionnaires
Case study research	- in-depth investigate a single event or person - observational	- case studies

Table 2 Types of research design

The main quantitative research approaches are descriptive, correlational, quasi-experimental, and experimental or causal research (Grand Canyon University, 2016), which differ in the degree of control on the variables in the experiment. Descriptive research describes data and features of the population or phenomenon being investigated and addresses the question What?, but does not answer questions about Why, Who, When and How (Trochim, 2010). Causal research explores the effect of one variable on another through experimentation and simulation to understand the processes (Lawrence, 2009). In other words, it manipulates or controls one or more variables to determine their impacts on another variable. Correlational research investigates the relationship between variables using the statistical analyses, without looking at the causes and effects. This approach is mostly observational in terms of data collection. Quasi-Experimental (also referred to as Causal-Comparative) seeks to establish cause/effect of the relationship between two or more variables, where the researcher does not manipulate the independent variable. Table 3 shows the basic purpose of the different types of quantitative research and the differences between them.

Applied research	**Objective**	**Methods**	**Control of variables**
Descriptive research (Shields & Rangarajan, 2013).	- describe "what is" - observe characteristics of the phenomenon from various aspects	- averages and frequencies - statistical conclusion	- variables not controlled
Causal research (Experimental or true experimentation) (Lawrence, 2009)	- determine the cause and effect relationships - determine which variable causes a certain behaviour - hold the variable that causes the change and measure the changes in the other variable	- experimentation - simulation - trial and error, which could be considered a simple experiment	- all variables are controlled except the independent variable, which is manipulated
Correlational	- investigate the relationships among variables	- descriptive correlation designs - model-testing designs - predictive designs	- variables not controlled
Quasi-Experimental (Causal-Comparative)	- determine the cause and effect relationships between two or more variables - doesn't manipulate the independent variable	- pre-tests and post-tests - interrupted time series regression	- all variables are controlled and the independent variable isn't manipulated

Table 3 Four main types of quantitative research

As the investigation addressed the impact of modifying speakers' recordings on the performance of ASR, an iterative trial and error approach had to specifically find the best modification factor from number of modification factors. The modification factors were selected by changing the pitch of the recordings' in 0.1 increments between –3 semitones and +3 semitones, and the speech rate in 10% increments between –50% and 50%. The pitch and speech rate were modified separately in order to investigate the impact on the formant frequencies of the speech, and thus the effects of those on the performance of the ASR. It was expected to be applied as an efficient method to eliminate the time required to find the best modification factor among all possible factors, and handled by academics and students exclusively without the assistance of external consultants.

This thesis explores the possible effects of modifying speakers' recordings on the performance of Automatic Speech Recognition (ASR). This involved (i) an exhaustive study of manually warping the formants and speaking rate in the recordings by using sound processing libraries such as *Rubber Band* (Quey, 2012) and *SoundTouch* (Parviainen, 2001), before transcribing those recordings using ASR; (ii) an exploration of the impact of formant modification on the Word Error Rate (WER) outcomes; (iii) a comparison between these modification outcomes and the original outcomes. To evaluate the empirical evidence for assertions about the advantages of processing speakers' recordings before transcribing those recordings, a statistically representative sample of speakers' recordings would ideally be selected. To select such a sample, a set that includes speakers' recordings, their correct transcripts, for male and female speakers, would first have to be constructed. The only way to collect correct transcripts of the recordings would be by manual transcription, and this is a costly task to undertake for multiple recordings. However, if a sampling set could be constructed, data could be collected from any online resource that allows access to web recordings and their manual transcripts, such as *Synote* (Wald, 2010).

In addition, the best warp factor for the formants may interact with the automatic vocal tract normalisation (VTN) method built into the ASR system but it may be theoretically possible to still to improve the WER. Because of this, and the formants and average F_3 used to estimate the warp factor are unknown, a 'trial and error' approach would be needed to obtain data for analysis. To draw statistical inferences using this approach, a large number of recordings would need to be sampled and modified, and the formants and the WER outcomes of the original and processed versions of those recordings extracted.

Moreover, a case study approach has been used here because the literature review showed that processing the recording before transcribing those recordings is a gap in the research on improving the accuracy of ASR transcription. It was not clear that the considerable sampling and data collection problems that would entail could be surmounted. In addition, many of the data to be focused on were related to warp processes.

Secondly, regression analysis was used to explore an efficient approach that could be implemented to reduce the time taken to manually select the warp factors.

However, survey research techniques are more suited for measuring characteristics of individuals and organizations rather than studying processes. Hence this study needed to identify and explore the critical success factors.

This chapter describes research design and methods, data collection, and data analysis procedures for the entire study.

3.1: Data collection

The study aimed at speakers in lecture rooms. For the data collection, 16 recordings were selected randomly from *Synote* website, because *Synote* provides the correct transcripts synchronized with its recordings (Wald, 2010). The correct manual transcripts of each speaker recording are required to be able to use any WER measurement tool to compare this transcript with the outcome transcripts of ASR for the original and processed versions of this recording. The collected recordings were pre-processed by modifying the pitch and then transcribing them and calculating the word error rate iteratively for each warp factor setting, until the best WER was found. The second pre-processing step involved acoustic feature extraction that included extracting the pitch (F_0) of the original and pre-processed recordings to obtain the modifications as percentages.

Information was gathered by “brute-force” to measure everything available, in the hope that the informative acoustic features could be isolated. This method was used because the formants of the speakers who originally trained the ASR are unknown, so which acoustic features are likely to be the most informative in estimating the best warp factor cannot be guessed.

3.2: Research design and methods

The data used to implement the research consists of recordings of males and females, and was collected from the Synote website that provides the audio files of recorded lectures with their transcripts. First, the current research design uses a case study to define the problem of measuring the various modifications to recordings, and suggest hypotheses based on this problem.

Secondly, trial and error was used for processing the recordings to select the best warp factors since this can generate all possible modifications, and there are a finite number of modifications that may be tested using the sound processing library.

Finding all possible warp factors would require noting each that is found, and then choosing the best warp factor based on predefined criteria. In situations where only one warp factor exists, all the possible warp factors would be the same and that solution would be the best one.

The ‘trial and error’ approach used here is discussed in chapter 4, see Figure 4.2-17 to determine the effect of recording modification on ASR performance, and to determine the effect of recording modification on the acoustic features such as formants of these recordings. Statistical methods were used to compare the measurements of WER and formants of processed recordings against the measurements of the original/unmodified recordings, to determine whether or not the processed recordings had a statistically significant impact upon the improvement.

Figure 3.2-12 shows the WER outcomes of transcribing the original/unmodified recordings, and Figure 3.2-13 shows WER outcomes of the best modifications for the speakers, where each best modification was found by modifying that speaker’s recording iteratively, which discussed in chapter 4.

Figure 3.2-13 shows the research design and demonstrates that the recordings were modified by altering fundamental frequency or speech rate, or both, before transcription using ASR. Modifying the fundamental frequency of the original recordings resulted in warping the formants by stretching or compressing them. The modification of the speech rate resulted in warping the time series of the spoken phrases of the original recordings. After pre-processing the recording by modifying formants and time series, the pre-processed recording was transcribed by the ASR, and the WER measured and compared with the WER of the original transcripts.

Firstly, trial and error approach was used to iteratively modify each recording until an optimal modification for that speaker was found, the use of trial and error approach has been investigated and evaluated in chapter 4. This approach extracted the most relevant information from each iteration and for better analysis of the data. ***Appendix A*** shows the WER results of three speakers and the impact of the recording modification on altering the formant frequencies of their speech, while ***Appendix B*** shows the results for 16 speakers, WER outcomes of transcribing each speaker’s both modified and unmodified recordings.

Secondly, an experiment was implemented to determine the difference between WER outcomes of the transcripts generated by transcribing the original recordings and those generated by transcribing the modified recordings. This first investigated the length of the recordings that needed to be transcribed so that there was no difference between the original and modified transcripts. Secondly, it identified the rate at which differences between the original and modified transcripts were just noticeable, which helps in eliminating the correction procedure. The both investigations have been explored in chapter 5.

Finally, the use of the estimation algorithm was experimented with to investigate an efficient method that could be used to estimate the optimal modification values for new speakers. Figure 3.2-14 shows the methodology of the current work, which has been discussed in chapter 6.

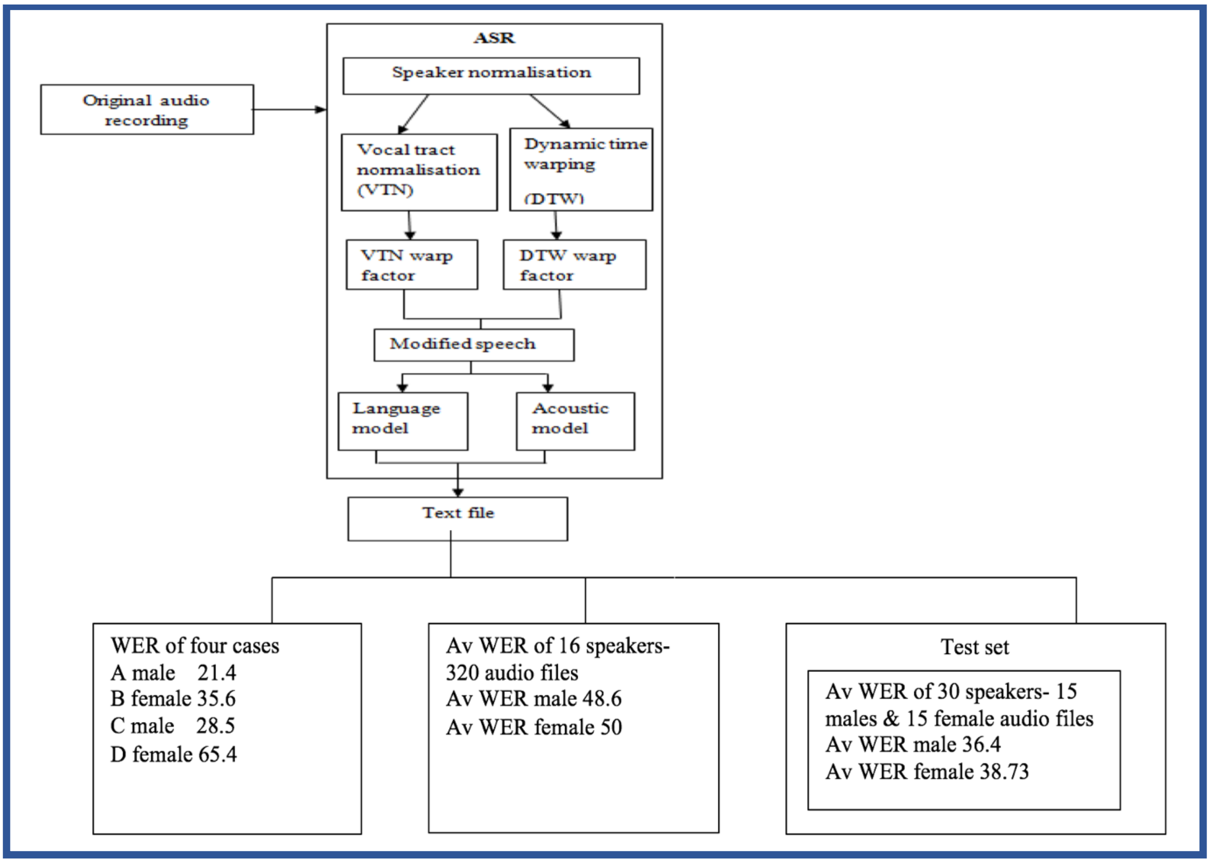


Figure 3.2-12 Processing recording by ASR

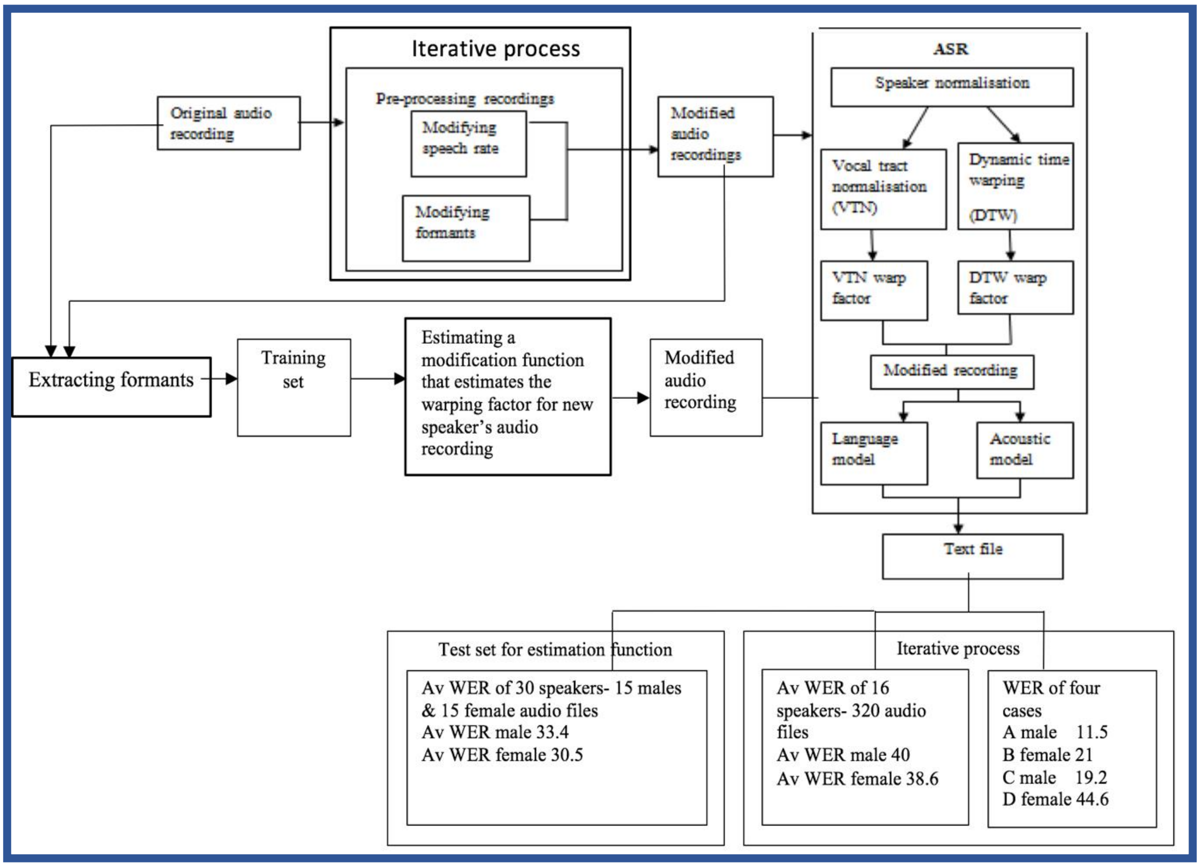


Figure 3.2-13 Pre-processing the recording before processing by ASR

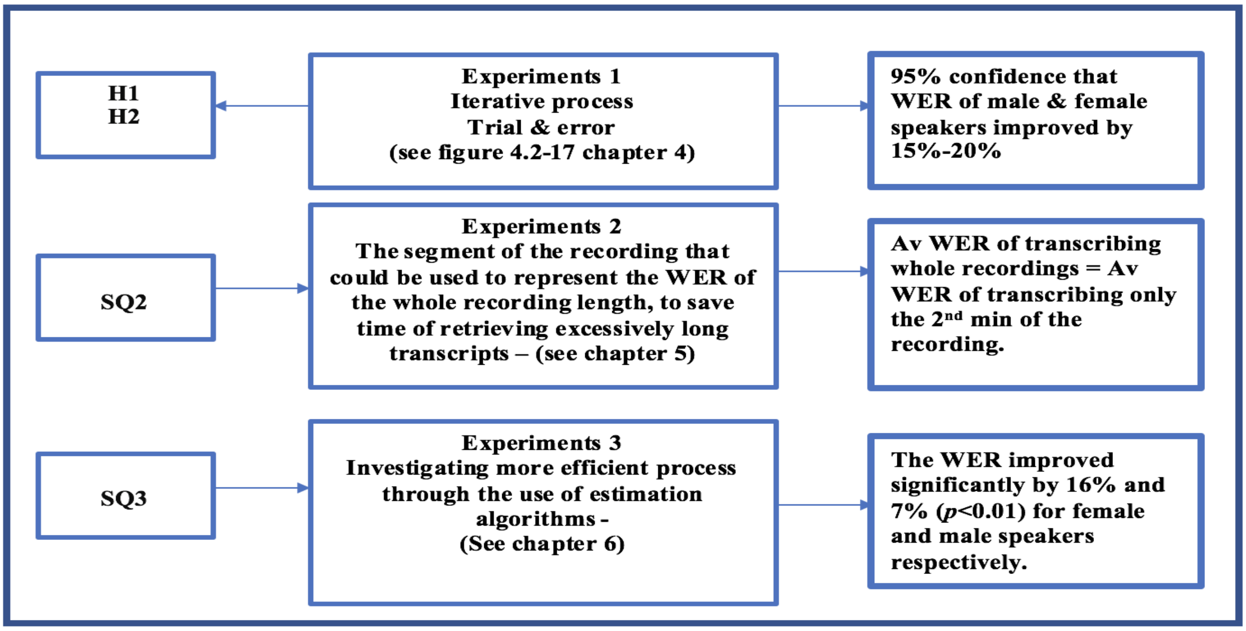

H1
H2
Experiments 1
Iterative process
Trial & error
(see figure 4.2-17 chapter 4)
95% confidence that WER of male & female speakers improved by 15%-20%
SQ2
Experiments 2
The segment of the recording that could be used to represent the WER of the whole recording length, to save time of retrieving excessively long transcripts – (see chapter 5)
Av WER of transcribing whole recordings = Av WER of transcribing only the 2nd min of the recording.
SQ3
Experiments 3
Investigating more efficient process through the use of estimation algorithms -
(See chapter 6)
The WER improved significantly by 16% and 7% (p<0.01) for female and male speakers respectively.

Figure 3.2-14 Research methodology

3.3: Categories chosen for case study

To define whether modifying speakers' recordings improve the efficiency of ASR transcripts and the improvement range caused by this modification, this study focuses on gender. Gender has been the target of most research on ASR performance. This study addresses frequency warping and the range of WER improvement that resulted by using sound processing libraries for recordings of both males and females in order to identify the situations in which pre-processing recordings improves the efficiency of creating accurate ASR transcripts.

3.4: Iterative process and Formulation of hypotheses

The pitch and speech rate of the recordings of males and females were separately and iteratively modified with different percentages to determine the effects. Each of the two was altered at a time to investigate the impact of modifying each factor on ASR performance and to investigate the causes of the variation in the system performance.

First, the pitch of the recordings of males was extracted using the '*Pda*' tool and modified iteratively by changing the pitch by 0.1 from –3 semitones to +3 semitones. It has been found that increasing or decreasing the pitch by more than 3 semitones results in distorting the speech. After each modification, the modified recording was transcribed with the ASR and the generated transcript was compared with the correct transcript by using the NIST tool matcher to compute the WER. Finally, the transcript generated by the modified recording with the lowest WER was selected as the optimal transcript and its modified recording as the optimal modification.

Secondly, the speech rate of the recordings was modified iteratively by changing the speech rate by 10% from –50% to +50%. After each modification, the modified recording was transcribed with the ASR and the generated transcript was compared with the correct transcript by using the NIST tool to determine the optimal speech rate modification.

Thirdly, each recording was modified by changing both the speech rate and pitch, based on the optimal modification factors extracted for that speaker. The best modification for the recordings was chosen that gave the lowest WER, whether resulting from modifying the pitch or speech rate or both. Thus, the optimal modification can be defined as the modified version that gives the lowest WER compared with other modified versions and the original version of the recording.

The sub-research question for the impact of processing the recordings by the trial and error approach is *What is the impact on transcription accuracy of pre-processing recordings before transcribing them on ASR?* Section 3.7: briefly addresses the method used to answer each sub-question with a brief review of the results. The WER improvement and frequency warping were used for forming the corresponding hypotheses, each addressing the overall success. The frequency warping dimensions of processing recordings was also used for forming the hypothesis relating to WER improvement. The hypotheses for testing the relation of processing the recording before transcribing them to WER improvement of ASR is shown below.

H1: pre-processing the recording by adjusting pitch and speech rate leads to adjustment of the formants in these recordings

H2: adjustment of the formants in recordings by pre-processing through adjusting pitch and speech rate before transcribing them with ASR can lead to significant improvement in WER.

3.5: Slight differences

To identify slight differences between the transcriptions of unprocessed speech and the transcriptions of processed speech, the processed transcripts were displayed while listening to the original audio segments for each speaker.

3.6: Data analysis

SPSS – Statistical Package for the Social Sciences (IBM Crop, Released 2013) and Matlab (MATLAB and Statistics Toolbox , Release 2012b) were used for the statistical analysis. The data from the original recordings and the modified recordings were all coded and were entered in to the computer. The data included: original WERs, modified WERs, original pitch, modified pitch, original formant, modified formant, original speech rate, modified speech rate values. The coding of variables in qualitative research is critical for better explanation of the results. SPSS was used to complete the required analysis.

3.7: Research question

The research question stands as: *How can speech recognition accuracy be improved by signal processing audio files before transcribing these via an ASR system?* This has been explored by answering the sub-questions presented below and through validation experiments and user validation.

Part I: Exploring the impact on ASR performance of modifying speaker recordings before transcribing them

SRQ 1 *What is the impact on ASR transcription accuracy rate of pre-processing recordings before transcribing them?*

Approach: This sub-question has been converted into two hypotheses; H1: pre-processing the recording by adjusting pitch and speech rate leads to adjustment of the formants in these recordings

H2: adjustment of the formants in recordings by pre-processing through adjusting pitch and speech rate before transcribing them with ASR can lead to significant improvement in WER. Secondly, this sub-question was analysed by experiment. This experiment took a random sample of four recordings. Identifying the impact of processing the speakers' recordings on ASR performance through the sample data initiated this research. The 'trial and error' approach was used in processing the recordings to find the optimal modification for each speaker's recording. Each speaker's original and modified recordings were chunked into 20 audio files to calculate means and standard deviations for each speaker. To analyse the sample data, the *NIST* measurement tool was used to measure the WER of the original and modified recordings' transcripts. The *ANOVA* test was used to compare the mean WER on original recordings' transcripts (ORT) with mean WER on processed recordings' transcripts (PRT). Accordingly, the total average WER difference between the original and modified transcripts was calculated to get the improvement ranges for the transcript errors.

The results showed that warping formants through altering pitch and SR of recordings improved the performance of ASR by an average of 39.5% for male speakers and an average of 36.2 for female speakers, which in turn reduced the time required to edit these transcripts to correct any remaining errors. The result also showed that recordings of males were improved by increasing their pitch, while female recordings were improved by decreasing their pitch. The ANOVA test showed a statistically significant difference between the mean WER in the PRTs and the mean WER in the ORTs (Chapter 4).

(SQ2) *What is the minimum recording length and transcription that can be used to represent the average WER of the whole recordings for evaluating the transcript?*

Using only the first or the second minute of the recording, what is the probability that its average WER is going to be different from the average WER for the whole recording?

(It is conjectured that recognition can take a short while to adapt to the speaker and also the first few words spoken have no previous words/context to assist with that recognition.)

Approach: Investigate whether the average WER of the first or the second minute of the speaker recordings could be used to represent the average WER of the whole recordings. A paired t-test was used to compare the average WER of the difference between the first minutes' WER and the second minutes' WER, of both original and processed recordings, with the average WER of the difference between the whole original recordings' WER and the whole processed recordings' WER.

The result of this investigation showed that the average WER of the second minute of the speaker recordings could be used to represent the average WER of the whole recordings (Chapter 5).

Part IV: Exploring an efficient approach to pre-processing the speakers' recordings

Firstly, investigated whether pre-processing provide a noticeable improvement in transcript accuracy

Approach: The transcripts of the audio segments were presented while listening to the original audio segments for each speaker, without any correcting procedure being carried out on those transcripts.

The result showed that the improvement in the transcript errors was noticeable when the transcript errors reduced by more than 2 and the WER improvement was more than 11% (Chapter 5).

***(SQ3)** To what extent can the use of missing value estimation algorithm to estimate the modification factor for processing a new speaker recording improve ASR transcription accuracy rate?*

Approach: Investigated whether an estimation function for the modification values for the new speakers improves the efficiency of creating accurate transcripts through ASR, similar to the manual process. The paired t-test was used for this investigation.

The results showed that an estimation function based gender improves the efficiency of creating accurate transcripts significantly where the WER improves by 16% and 7% for female and male speakers respectively (Chapter 6).

Chapter 4: Improving the performance of ASR

This chapter investigates a framework to evaluate the effect of transforming speech, by using two different sound processing libraries *Rubber Band* and *SoundTouch*, to decrease the WER of modified recordings' transcripts.

4.1: Automatic speech recognition and pre-processing recordings

To reduce the variation of acoustic features in speech signals that ASR systems need to deal with, they typically employ speaker normalisation based on frequency warping and speech rate warping methods.

Figure 4.1-15 shows the processing procedure for the speech signal by an ASR system. ASR uses vocal tract normalisation and dynamic time warping methods to warp the formants and time series respectively, of the speech signal provided. The warp factor of VTN warps the formant of the original recording to the average formants in the training set, while the warp factor of DTW aligns the time series of the spoken phrases of the new speech signal with the time series of the matching spoken phrase in ASR. These methods change the acoustic features of the speakers' recording to be as close as possible to those in the ASR, to assist the language and acoustic models process the speech and thus provide an accurate text transcript.

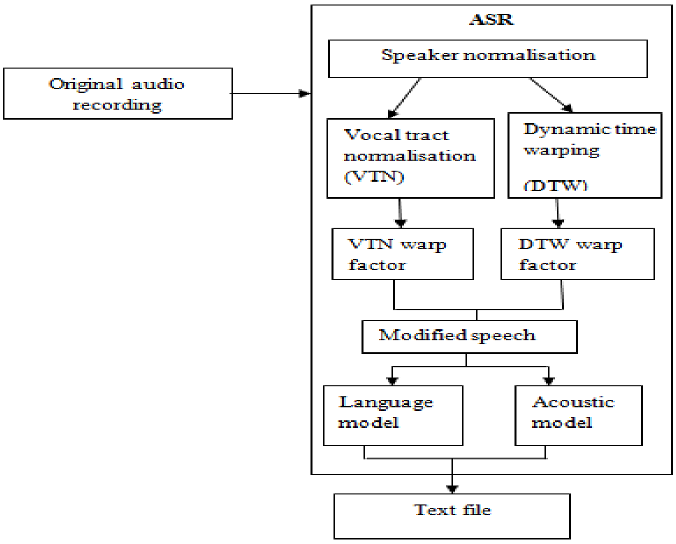


Figure 4.1-15 Processing recording by ASR

The research presented here explores whether modifying the acoustic features (formants and speaking rate) of the speakers' recordings, before transcribing them, would enhance the performance of the ASR.

This research modified the speech signal of recordings by different percentages before transcribing each modified recording via ASR. As shown in Figure 4.1-16, the recordings were modified by altering the fundamental frequency or the speech rate, or both, before transcribing using ASR. Modifying the fundamental frequency of the original recordings resulted in warping the formants by stretching or compressing them. Modification of the speech rate resulted in warping the time series of the spoken phrases. After this pre-processing, the recording was transcribed by the ASR and the WER measured.

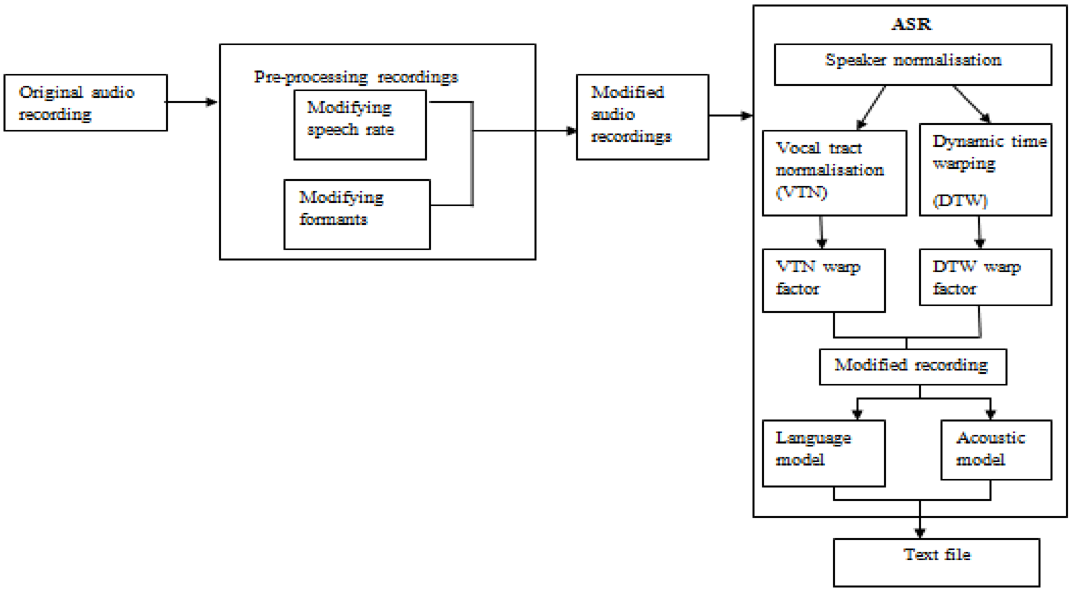


Figure 4.1-16 Pre-processing the recording before processing by ASR

An iterative approach to changing the speaker's formants and time series was used to find the 'mapping' that produced the lowest WER.

The aim of this modification was to improve accuracy by pre-processing the original recording before transcribing using an ASR system.

4.2: Evaluation Framework

This section presents an overview of the evaluation system used here and discusses each of its components. Figure 4.2-17 shows the block diagram of framework used for evaluation. It begins with modifying the speaker recordings by using two different sound processing libraries: *Rubber Band* and *SoundTouch*. Both the original and modified recordings were transcribed via ASR and the WER of the transcripts measured. The WER of the two transcripts was compared to measure the impact caused by altering the recordings. The selection of the best warp factors was achieved through an inefficient 'trial and error' approach that involved calculating the word error rate for each warp factor setting.

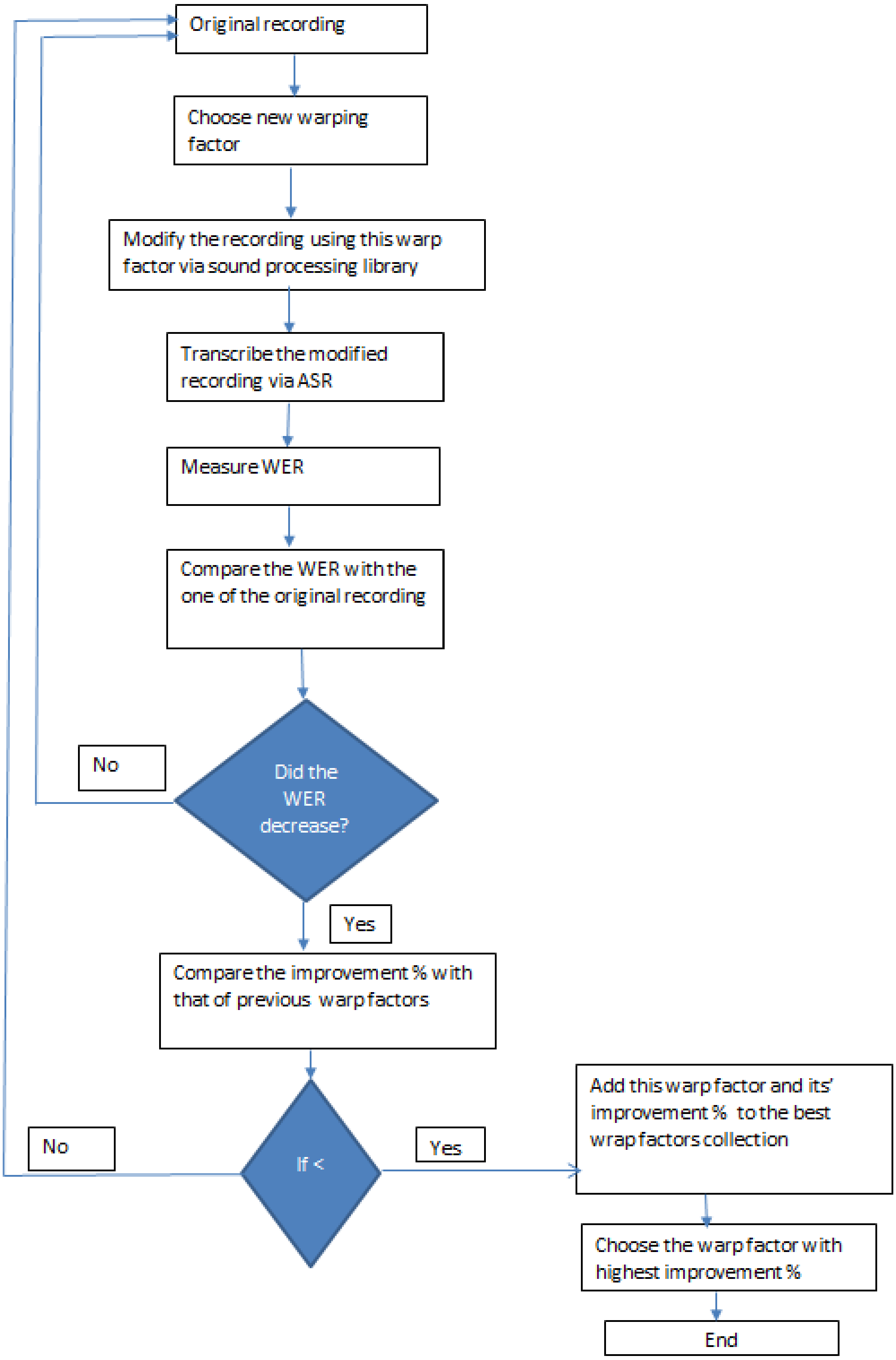

Original recording
Choose new warping factor
Modify the recording using this warp factor via sound processing library
Transcribe the modified recording via ASR
Measure WER
Compare the WER with the one of the original recording
Did the WER decrease?
No
Yes
Compare the improvement % with that of previous warp factors
If <
No
Yes
Add this warp factor and its' improvement % to the best wrap factors collection
Choose the warp factor with highest improvement %
End

Figure 4.2-17 Evaluation framework block diagram

4.3: ASR performance evaluation

The performance of ASR is specified by measuring the “Word Error Rate” (WER) of the generated transcripts. WER is computed by using the equation $\boldsymbol{WER} = \frac{\boldsymbol{S+D+I}}{\boldsymbol{N}}$, where S is the number of substitutions, *D* is the number of deletions, *I* is the number of insertions, and *N* is the total number of transcript words. The accuracy is computed as $\boldsymbol{1 - WER}$**.**

4.4: Experimental database

To investigate pre-processing recordings, *Synote* has been mainly used as a source here. *Synote* is a freely available web-based application that provides access to multimedia web sources and uses speech recognition to automatically synchronise audio or video recordings of this multimedia with its transcripts. The recordings were partitioned into development testing and evaluation testing data. The development testing recordings consisted of 22979 words uttered by 16 speakers, of which 5 were female and 11 male, while each recording lasted from 5 min to one hour. The variety of durations in the data helps with the investigation of the effect of modifying the speech rate in different recordings on ASR performance. *Synote* implements no control over the number of words contained in the recordings and can differ depending on the each speaker’s speaking style. The language model was not taken into account in this work since the main focus was on the acoustic model of ASR and the modification of speaker’s formants. Gouvêa (1998) stated that when the language model is taken into account in the recognition process, it becomes difficult to notice the performance of different implementations of the vocal tract normalization.

Each of the 16 speakers’ recordings were chunked into 20 audio segments to calculate the mean WER and standard deviation for each speaker, and the total of 320 processed audio segments were statistically compared with their original audio segments by *T*-test. The *Synote* database covers a wide range of speakers and speaking styles, and its database includes recordings that taken in different environments and encompassing foreign accents. The classification by Gouvêa (1998) is supported by Garofolo, et al. (1997), who classified different audio segments into a number of focus conditions. Table 4 shows some of these focus conditions that have been considered here when gathering the recordings from *Synote*.

Condition	Dialect	Mode	Background
Dictate speech (F0)	native	Planned	Clean
Spontaneous Speech (F2)	native	Spontaneous	Clean
Degraded Acoustics (F3)	native	(any Mode)	Speech or Noise
Nonnative Speakers (F4)	Non-native	Planned	Clean

Table 4 condition definitions

4.5: The impact of transforming recordings on their transcription

This section discusses the experimental setup for this evaluation. It first describes the sound processing libraries used for pitch and speech rate modification. Then the results of the speech recognition under these modifications are presented. The object is to compare the effects of the two sound-processing libraries on the transcript's error rate.

4.5.1: Experimental data and parameters

The experiment was based on modifying the original speakers' recordings by altering either or both frequency and speech rate of these recordings. Both ***Rubber Band*** (Quey, 2012) and ***SoundTouch*** (Parviainen, 2001) are libraries for shifting-pitch and stretching or compressing speech rate, and were used in this experiment to modify the original speakers' recordings. Modifying the pitch by using these sound processing libraries resulted in modifying the formants of an audio stream dynamically. Also, the '***pda'*** (Taylor, et al., 1999) tool, which is a detection algorithm that produces a fundamental frequency contour from a waveform file, was used to track the average fundamental formants of the original and modified recordings.

The experiment was undertaken on lecture audio files for two male speakers, each containing 1200 words, and for two female speakers, each containing 900 words. The duration of each of the original male recordings was 11:49 min and the duration of each of the original female recordings was 4:59 minutes. The experiment involved:

- Using a Dragon ASR server 11 system to transcribe four audio files of four different speakers.
- Modifying the frequency and speech rate of the speakers' recordings by different percentages and finding the settings that gave the lowest word error rate (WER) after transcription by ASR.
- The WER was calculated by ((Number of wrong words) / (Total of words)) * 100.
- The formant frequency was modified through increasing or decreasing pitch (F0) by

between -5 and 5 semitones via the '*SoundTouch*' and '*Rubber Band* ' libraries before ASR transcription.

- Warping the time series of the spoken phrases of the original recordings through changing speech rate (SR) percentages by between -50% and 50% via the '*Soundtouch*' and '*Rubber Band* ' libraries before ASR transcription.

4.5.2: Results

In this experiment four different original recordings for four different speakers were pre-processed by increasing or decreasing either pitch or speech rate, or both pitch and speech rate. The pre-processed recordings were then processed via a Dragon speech recognition system to generate the transcripts. Finally, the different generated transcripts were measured to find the lowest WER. The detailed results for the pre-processing of the recordings of one of the speakers will be presented as an illustration of the type of effect the pre-processing had on all four speaker recordings and the summary results and analysis then presented for all four speaker recordings.

4.5.2.1: Modifying the fundamental frequency of a recording of male speaker 'A' by using '*SoundTouch*' library

Changing F_0 percentage	Pitch (AV freq Hz)	WER	AR
–11.50%	105.1	89.2	10.8
–10%	106.9	82.1	17.9
–8.40%	108.9	71.2	28.8
–6%	111.7	70.3	29.7
0	118.8	21.4	78.6
6%	126.2	17.2	82.8
7.00%	127.1	14.7	85.3
9%	129.5	16.1	83.9
13.20%	134.5	21.1	78.9
18.40%	140.6	28.6	71.4

Table 5 shows the summary results for WER, resulting from modifying pitch of male 'A' speech by increasing and decreasing the fundamental frequencies of the speech, using '*SoundTouch*' library

Table 5 shows WER and accuracy rate (AR) of generating transcripts via a Dragon ASR system for each pre-processed recording of speaker A. Each pre-processed recording of 'A' was altered by modifying the fundamental frequencies of his speech. The changes in the original pitch percentage were between –11.5% and 18.4%. The modification percentages result in an average fundamental frequency between 105.1Hz and 140.6 Hz. The results of WER of the pre-processed recording's transcripts were found to be between 89.2% (Worst value) and 14.7% (best values) and therefore the AR between 10.8%, and 85% respectively

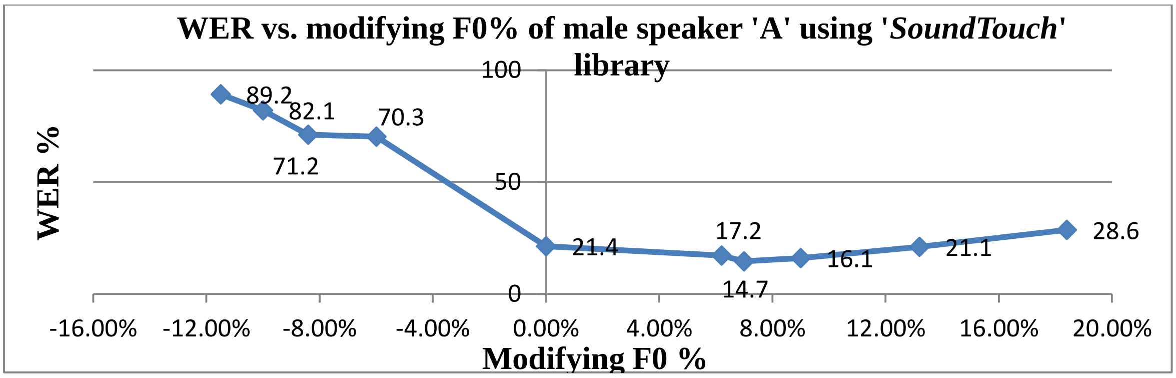


Figure 4.5.2.1-18 WER percentages vs. modifying the F_0 of the original male speaker 'A' via '*SoundTouch*' library

Figure 4.5.2.1-18 plots the results which are summarised below:

- Modifying F0% by 7% reduced WER to a minimum of 14.7%, which was a reduction of 31.3% from the original 21.4% WER.
- Modifying F0% further increased the WER.
- A negative change in F0 % increased WER

4.5.2.2: Modifying fundamental frequency of male speaker 'A' recording by using '*Rubber Band'* library

Changing F_0 percentage	Pitch (AV freq Hz)	WER	AR
–11.4%	105.3	90.5	9.5
–9.3%	107.8	83.5	16.5
–7.9%	109.4	62.1	37.9
–6.1%	111.6	49.6	50.4
0	118.8	21.4	78.6
9.3%	129.9	13.2	86.8
15.3%	137	13.3	86.7
21.3%	144.2	18.5	81.5
27.2%	151.1	28.8	71.2

Table 6 shows the summary results for WER, resulting from modifying pitch of male 'A' speech by increasing and decreasing F_0

Table 6 shows that modifiying F_0 of the original speaker's recordings by between –11.4% and 27.2% resulted in different pre-processed recordings of different average frequency values between 105.3 Hz and 151.1 Hz. Transcribing the pre-processed recodings by the Dragon speech recognition system produced different WER values between 90.5%(worst value) and 13.2% (best value) and therefore corresponding accuracy rate values between 9.5% and 86.8%.

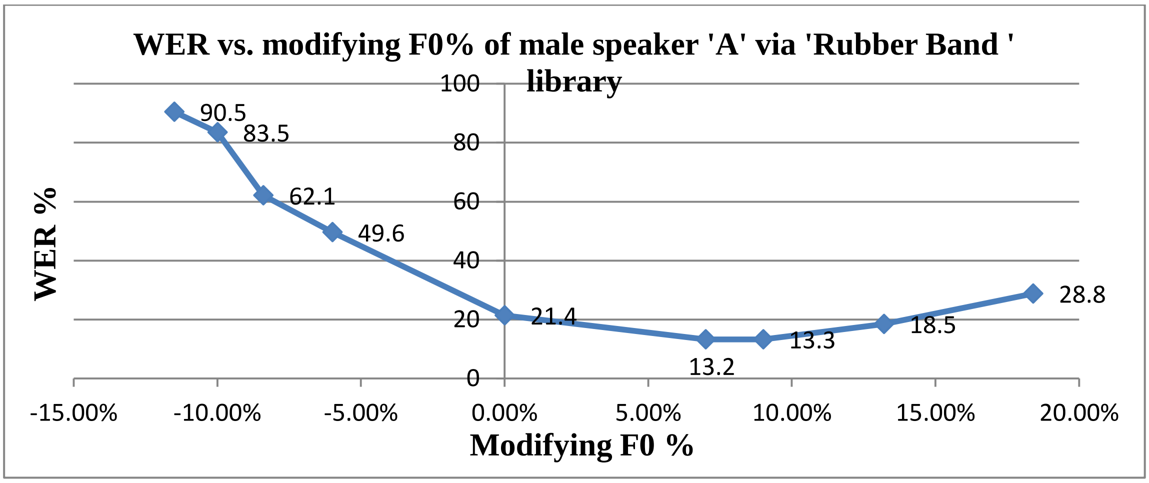


Figure 4.5.2.2-19 WER percentages vs. modifying F_0 of the original speaker 'A' recording via '*Rubber Band* ' library

Figure 4.5.2.2-19 plots the results which are summarised below:

- The '*Rubber Band* ' library showed higher frequency values (mean %) compared to those from the '*SoundTouch*' library.
- Increasing the F_0 of A's speech by 9.3% leads to decreasing the WER to 13.2% (best value), an improvement of 38.3%
- The WER increased with further increasing of F_0

4.5.2.3: Modifying speech rate of a recording of male speaker 'A'

Speech rate (percentage)	Pitch (AV freq Hz)	WER	AR	duration
–50	119	38.8	61.2	23:39
–40	118.5	27.8	72.2	19:42
–30	115.8	23.8	76.2	16:53
–20	114.2	20.6	79.4	14:46
–15	114	20	80	13:54
–13.8	114.3	19.3	80	13:43
–10	115.4	18.8	81.2	13:08
–1.5	118.3	19	81	12:00
0	118.8	21.4	78.6	11:49
1.5	118.6	19.2	80.8	11:39
10	119	19.4	80.6	10:45
13.8	118.9	21.7	78.3	10:23
15	118.9	21.9	78.1	10:16
20	118.4	23.2	76.8	09:51
30	118.9	31.5	68.5	09:05
40	118.9	36.2	63.8	08:26
50	119	44.1	55.9	07:53

Table 7 shows the summary results for WER, resulting from modifying SR of male 'A' speech by increasing and decreasing the original SR

Table 7 shows the effects of modifying the speech rate of the audio file of speaker 'A' by between -50% and 50%. The changes on the audio file were made only on the speech rate while trying to keep the F0 stable. However, it was noticed that modifying SR was still causing between –4% and 1% changes on F0%. The WER was showing changes between the worst value 44.1%, with increasing the SR by 50%, and the best value 18.8%, with decreasing the SR by 10%. Decreasing the SR by 10% caused an improvement on WER by 12.1%.

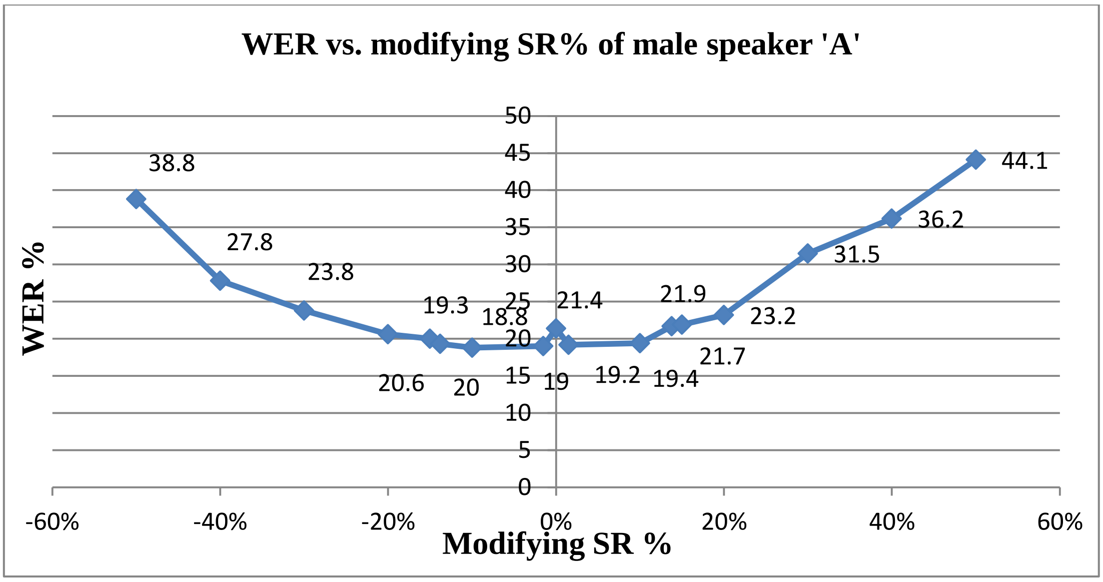


Figure 4.5.2.3-20 shows WER % vs. modifying SR% of male speaker 'A'.

Figure 4.5.2.3-20 plots the results which are summarised below:

- Changing the SR by – 1.5% could improve the WER from 21.4% to 19% which is an 11% improvement, even though that decreased SR % caused a decreasing on F0% by 0.4%.
- The results shown in Figure 4.5.2.1-18 and Figure 4.5.2.2-19 showed that with any decreasing of the male fundamental frequency the WER % always increased.
- However, the result in Figure 4.5.2.3-20 shows a decreased SR% could decrease the WER even if F0% decreased.
- The WER reached the best value of 18.8% after decreasing the SR by 10% even though F0% was decreasing by 2.9%.
- On the other hand, the figure shows increasing the SR by 1.5% caused decreasing of the F0% by 0.2% and an improvement on the WER by 10.3%.

4.5.2.4: Modifying fundamental frequency and speech rate of male speaker ‘A’

Speech rate percentage	Changing F_0 percentage	Av Pitch (Hz)	WER	AR
–20	6%	125.2	13.8	86.2
–20	7%	131.6	16.7	83.3
–20	9%	139.7	22.5	77.5
–15	6%	123.6	15.6	84.4
–15	7%	130	15.2	84.8
–15	9%	137.6	20.8	79.2
–13.8	6%	123.6	17.7	82.3
–13.8	7%	129.7	16.9	83.1
–13.8	9%	137.1	21	79
–10	6%	123.8	13.7	86.3
–10	7%	128.7	14.5	85.4
–10	9%	136	20.5	79.5
–5	6%	124.2	18.7	81.3
–5	7%	128.8	15.8	84.2
–5	9%	134.7	22.2	77.8
–1.5	6%	125.4	11.5	88.5
–1.5	7%	129.3	15.3	84.7
–1.5	9%	134.1	21	79
0	0	118.8	21.4	78.6
1.5	6%	125.9	16.6	83.4
1.5	7%	129.7	15.8	84.2
1.5	9%	134.7	20	79.5
5	6%	127.4	15.7	84.3
5	7%	131.4	16.3	83.7
5	9%	135.3	21.7	78.3
10	6%	128.7	14.5	85.5
10	7%	132.9	15.4	84.6
10	9%	136.7	20.4	79.6
13.8	6%	129.6	17.4	82.6
13.8	7%	134.5	17.3	82.7
13.8	9%	138.9	23.6	76.4
15	6%	129.5	19.4	80.6
15	7%	134.9	17.8	82.2
15	9%	139.1	23.1	76.9

Table 8 shows the summary results for WER, resulting modifying SR for different increases in F0 of male ‘A’ speech

Table 8 shows the WER and AR of speech recognition resulting from modifying both fundamental frequencies and speech rate of the recording of speaker 'A'. The changes in the original pitch were between 6% and 9%. The modification percentages have resulted in an average F_0 of between 123.6 and 139.1 Hz. The results of WER of pre-processed audio files were found to be between 13.8% (best) and 23.1% (worst) and therefore the AR was between 86.2% and 76.9%, respectively.

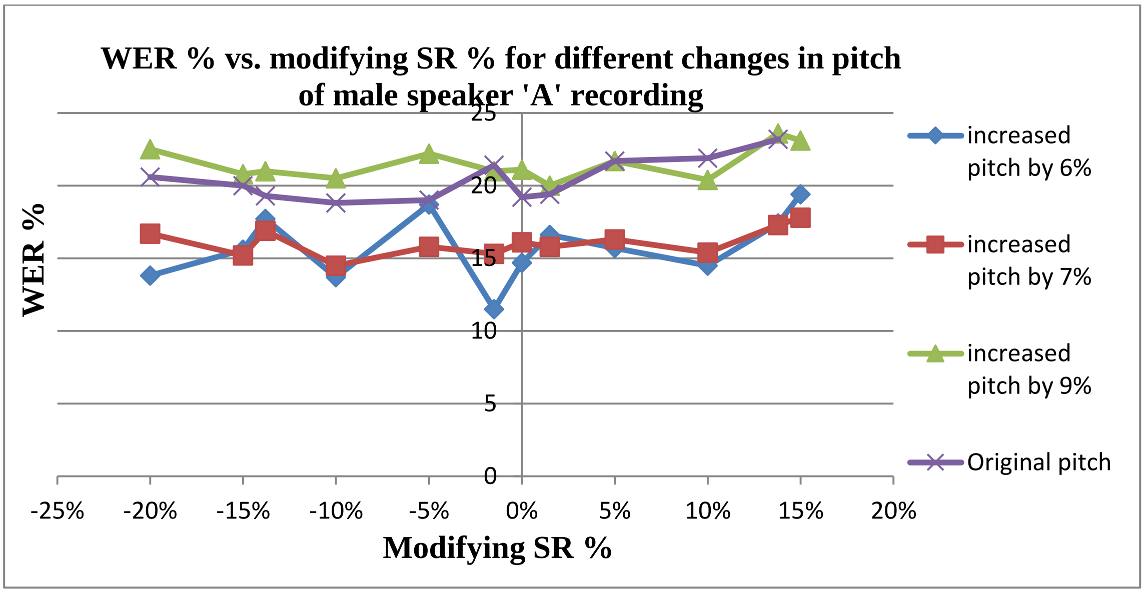


Figure 4.5.2.4-21 plots WER percentages vs. the modifications of SR percentages for different changes in pitch percentages of speaker 'A' male

Figure 4.5.2.4-21 plots the results which are summarised below:

- Modifying SR by –1.5% and increasing the fundamental frequency by 6% improved the WER to 11.5(best value), a change of 46.2%
- The resulting effects on WER were sensitive to the modification of fundamental frequency and speech rate; the figure shows that increasing the SR by 10% and increasing the F0% by between 6% and 9% caused an improvement on the WER of between 32.2% and 4.7%, while keeping the same values of the fundamental frequency and increasing SR by 15% leads to increasing the WER again by 8.4%.
- The WER was increasing and decreasing significantly by between 15.7% and 23.1% with decreasing speech rate by between 5% and 15% and increasing the pitch by between 6% and 9 %.
- Increasing the SR caused a slight further increasing in the F0% of between 0.2% and 1.7%.

4.5.2.5: Spectral frequency (a spectrum of frequencies over a continuous range) warping, resulting from modifying fundamental frequency of speaker ‘A’ recording

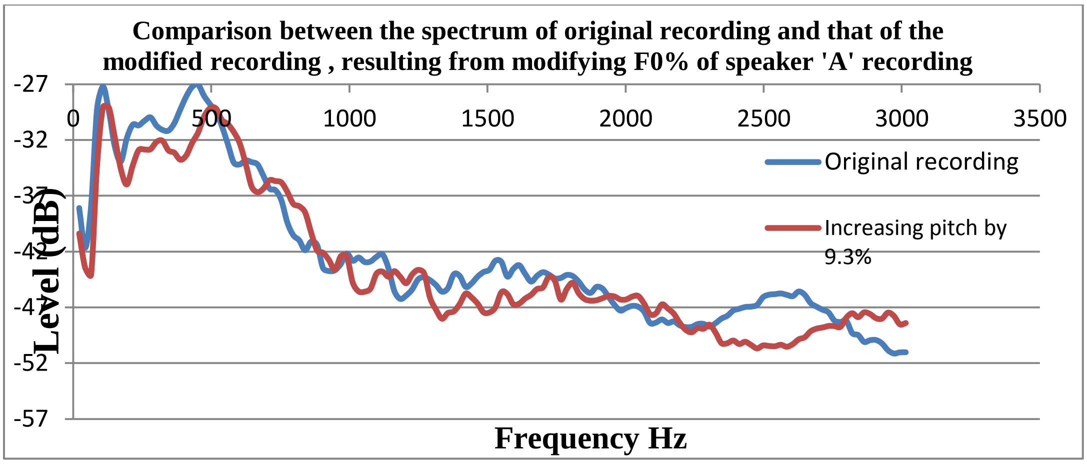


Figure 4.5.2.5-22 Comparison between the spectrums of speaker ‘A’ recording and that of the modified pitch recording. The formants and peak positions offset result from the modified F0%

Figure 4.5.2.5-22 shows the spectral frequency warping caused by increasing the fundamental frequency of speaker ‘A’ recording by 9.3%, which gives the best improvement in the WER. It can be seen that this shifts the formant resonance peaks to the right (i.e. raises their frequencies).

4.5.2.6: Spectral frequency warping resulting from modifying SR and fundamental frequency % of speaker 'A' recording

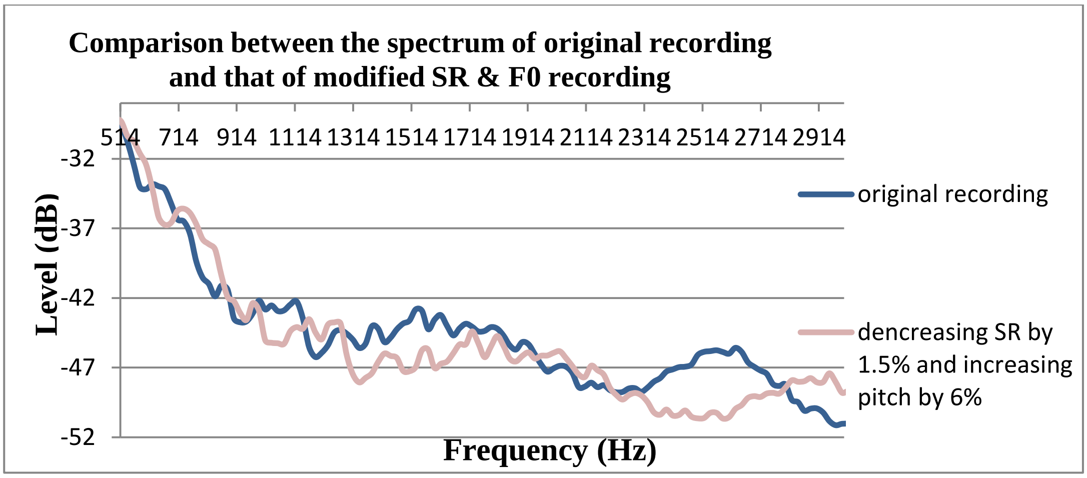


Figure 4.5.2.6-23 Comparison between the spectrums of original 'A' recording and that of modified speech rate and fundamental frequency recording

Figure 4.5.2.6-23 shows how modifying the recording of male speaker 'A' by increasing the pitch by 6% and decreasing SR by 1.5 % resulted in expansions of the 'frequency spectrum (i.e. shifts formant resonance peaks to the right) to give the best WER.

4.5.3: Results discussion

The pitch-shifting libraries (soundTouch and rubber Band) increased and decreased the pitch by between 1 and 12 semitones. However, the *Rubber Band* library showed more accurate results compared with *SoundTouch* after modifying the pitch using the same interval. For example, increasing the average pitch of 118.8 Hz by 2 semitones gives 133.3 Hz, while using the *Rubber Band* library gives an average pitch of 129.9 Hz and using the *SoundTouch* library gives an average pitch of 127.1 Hz. Therefore, using different pitch-shifting libraries produces different modifications to the average pitch of the speakers.

Table 9 below shows the best improvement resulting from modifying either fundamental frequency or speech rate or both simultaneously.

	Pitch		Speech rate		Pitch and speech rate	
speaker	Mod%	Improv%	Mod%	Improv%	Mod%	Improv%
A male	9.3	38.3	–10	12	9.3, –10	24.3
C male	12.8	25.3	–15	16.5	12.8, –15	25.5
B female	–8.6	24.2	–15	26.7	–8.6, –15	39.6
D female	–10.2	27.3	–15	13.8	–10.2, –15	28.3

Table 9 results of best modification% of pitch and speech rate resulting in best improvement% in WER

This table shows that for speaker ‘A’, increasing the pitch by 9.3% improved the WER by 38.3%, while for speaker ‘C’ increasing the pitch by 12.8% improved the WER by 25.3%. On the other hand, for female speakers decreasing the pitch for speaker ‘B’ by 8.6% improved the WER by 24.2%, while for speaker ‘D’ decreasing the pitch by 10.2% improved the WER by 27.3%.

The table also shows that, for males, decreasing the SR for speaker ‘A’ by 10% improved the WER by 12%, while for speaker ‘C’ decreasing SR by 15 improved the WER by 16.5%. On the other hand, for speaker ‘B’ decreasing the SR by 15% improved the WER by 26.7% while for speaker ‘D’ decreasing SR by 15% improved the WER by 13.8%.

Table 10 below shows the best improvement resulting from modifying both fundamental frequency and speech rate together and the improvements when these same values were used independently.

	Pitch		Speech rate		Pitch and speech rate	
speakers	Mod%	Improv%	Mod%	Improv%	Mod%	Improv%
A male	6	19.6	–1.5	11.2	6, –1.5	46.3
C male	9.1	21.1	–13.8	9	9.1, –13.8	32.6
B female	–8.6	24.2	–10	26.2	–8.6, –10	40.6
D female	–10.2	27.3	–20	7.9	–10.2, –20	31.8

Table 10 Modification of pitch and speech rate and improvement in WER

This table shows that modifying both speech rate and pitch improved the WER for male speakers ‘A’ and ‘C’ by 46.3% and 32.6% respectively, while for female speakers ‘B’ and ‘D’ it improved the WER by 40.6% and 31.8% respectively.

One reason for this improvement appears to be that increasing the pitch for male speakers led to increasing the formants. For example, the F_3 of speaker ‘A’ was 2562.45 Hz and increased by 11.8%, resulting in an F_3 of 2863.9 Hz, while the F_3 of speaker ‘C’ was 2153.32 Hz (less than the F_3 of speaker ‘A’) required a smaller increment of 10.7%, resulting in an F_3 of 2411.72 Hz.

On other hand, decreasing the pitch for female speakers led to reducing the formants. For example, the F_3 of speaker ‘B’ was 3768 Hz and decreased by 5.5%, resulting in an F_3 of 3559 Hz, while the F_3 of speaker ‘D’ was 3897 Hz (higher than the F_3 of speaker ‘B’) required a greater decrement of 11.1%, resulting in an F_3 of 3466 Hz.

Another reason for this improvement appears to be that modifying the SR led to warping of the time series of the spoken phrases of the original speakers’ recordings.

4.5.4: The modification of F3 resulting from modifying F0

The resulting modifications of F3 caused by modifying F0 are shown in Table 11, with the word error rate (WER) of the original recordings and that of the modified recordings, and the F3 of the original recording and the modified F3 that resulted from the optimum pre-processing.

Speakers	Original F3	WER% of original recording	Processed F3	WER% of modified recording	Improvement %
Speaker 'A' male	2562.45 Hz	21.4	2863.92Hz	13.2	38.3%
Speaker 'C' male	2153.32Hz	28.5	2411.72 Hz	20.3	28.8%
Speaker 'B' female	3768 Hz	35.5	3559 Hz	23.8	33.5%
Speaker 'D' female	3897 Hz	65.4	3466 Hz	47.5	27.3%

Table 11 shows results of original and processed speaker's F3

As shown in Table 11, F3 of speaker 'A' equal to '2562.45' Hz was increased by 11.8%, resulting in an F3 of 2863.92Hz, leading to a decrease in the WER by 38.3%. While an F3 of speaker 'C' equal to '2153.32' Hz (less than F3 of speaker 'A') required a smaller increment of 10.7%, resulting in an F3 of 2411.72Hz, decreasing the WER by 25.3%.

On the other hand, as shown in the table above, an F3 of speaker 'B' equal to '3768' Hz was decreased by 5.5%, resulting in an F3 of 3559 Hz, which led to a decrease in the WER of 24.2%. While, an F3 of speaker 'D' equal to '3897' Hz (higher than F3 of speaker 'B') required a greater decrement of 11.1%, resulting in an F3 of 3466 Hz, which led to decreasing the WER by 27.3%.

4.6: Reliability

This section investigates the best factors that were found through the ‘trial and error’ approach, and assesses their reliability. It does this in two ways: the reliability of automatically calculating the WER is compared with doing this calculation manually, the reliability of speech recognition in the WER resulting from transcribing the modified recording. These are now elaborated.

4.6.1: Reliability of automated WER calculation compared with manual calculation

The reliability of the calculation of the WER was computed by comparing the results of calculating the WERs manually with the results of calculating those WERs automatically. The manual calculation of the WER was calculated from

$$\frac{Number\ of\ wrong\ words}{Total\ words}*100,$$

while the automatic calculation of WER was measured by using the NIST SCLITE tool (NIST Scoring Toolkit Version 0.1, 1996). The inter-rater reliability agreements were determined using the intra-class correlation coefficient (ICC) of 0.7 (p<0.01).

4.6.2: Test-Retest reliability of WER from transcribing the modified recording

The ASR system was tested and re-tested by transcribing and re-transcribing 50 randomly selected recordings, which were processed with *SoundTouch* and *Rubber Band* sound-processing libraries. In recordings processed by *SoundTouch* (ST), the internal consistency was found to be satisfactory, as confirmed by a Cronbach’s alpha coefficient of 0.978. A high degree of stability was found for the WER of the processed recordings in the transcribe/re-transcribe, with an intra-class correlation coefficient (ICC) = 0.98 (p<0.01; 95% CI = [0.97, 0.98]). Spearman’s correlation coefficient was r = 0.964 (p<0.01), indicating high agreement between the transcribe/re-transcribe WER of recordings processed with ST.

In recordings processed by *Rubber Band* (RB), Cronbach’s alpha coefficient of 0.979 confirmed that the internal consistency was satisfactory, with an ICC = 0.96 (p<0.01; 95% CI=[0.94, 0.97]). Spearman’s correlation coefficient was r = 0.927 (p<0.01), indicating high agreement between the transcribe/re-transcribe WER of recordings processed with RB.

4.7: Comparison between sound-processing libraries

The results of the experiment discussed in Sub-section 4.5.2 showed that using sound-processing libraries such as *Rubber Band* (Quey, 2012) and *SoundTouch* (Parviainen, 2001) to modify the original recordings led to an improvement in the transcripts' WER. This section addresses whether the sound-processing library that was used produced different percentage improvements in the transcripts' WER.

The aim is to answer the following question: Does the WER improvement differ when using different sound-processing libraries? Thus the purpose of this experiment is to compare the impact on WER of *SoundTouch* library's modification and of *Rubber Band* library's modification. The experiment was based on modifying the original speakers' recordings by using both *Rubber Band* (Quey, 2012) and *SoundTouch* (Parviainen, 2001) for shifting-pitch and stretching or compressing speech rate.

The data was taken from 200 lecture audio files with 10 different speakers. The best recordings' modification that resulted best WER for each speaker was taken, as discussed in the previous chapter.

4.7.1: Results

To test whether formant modification that resulted from processing the recordings through *SoundTouch* (ST) differed from those by *Rubber Band* (RB), a chi-squared test was used with alpha = 0.05 as criterion of significance. According to the chi-squared test, there was no significant difference between the mean formants of recordings processed by ST and those processed by RB, (N = 200), $p = 0.15$, so it can be concluded that both sound-processing libraries have changed the speakers' recordings equally.

A Pearson correlation coefficient revealed significant correlation between the WER of the recording processed by RB and that processed by ST, r (200) = 0.96, p <0.01.

A paired-samples t-test revealed that the mean WER of the recording processed by ST (M = 41.73, SD = 18.56) was not significantly different from the mean WER of the recording processed by RB (M= 41.5, SD = 17.57), t (200) = -0.616, p = 0.54. The results showed with 95% confidence that the true difference between these means is CI = [–0.97, 0.51].

4.8: Summary

This chapter has described the implementation of a method of pre-processing speakers' recordings, before transcribing those recordings, by modifying the fundamental frequency and speech rate through the use of pitch shifting and time stretching libraries.

The pre-processed recordings two males and two females were analysed. The main challenge in reducing the WER in ASR transcripts is to find an efficient way to determine the optimum pre-processing frequency and speech rate warp factors.

Shifting the pitch of the speakers' recordings led to warping of the formants of the original speakers' recordings, resulting in an improvement in WER by an average of 20.4% and 25.8% for male and female speakers respectively.

The modification of SR led to warping of the time series, resulting in an improvement in WER by an average of 10.1% and 17.1% for male and female respectively.

Modifying both fundamental frequency and speech rate led to significant improvement in the ASR system WER. The ASR improved by an average of 39.5% with increasing the pitch and decreasing the speech rate of the male speakers, resulting in stretching the formants and compressing the time series. For female speakers, the system improved by an average of 36.2% with decreasing both pitch and speech rate, resulting in compressing both the formants and time series.

Chapter 5: Statistical analysis of processing speakers' recordings

Initial studies in Chapter 4 showed that pre-processing the recordings (PR), through the use of pitch shifting and time stretching libraries before transcribing those recordings, led to improvement in the WER of the ASR system compared with the original recordings (OR). This chapter reports on an experiment conducted to investigate whether this difference was significant. The question addressed is: Does the result suggest that ASR produces transcripts with a better (lower) WER on PR than on OR?

Hypotheses

Null Hypothesis:

There is no difference in the mean WER for PR transcripts and OR transcripts.

Alternative Hypothesis:

- The mean of WER for OR transcripts should be greater than for PR transcripts.
- The mean of WER for OR transcripts should be smaller than for PR transcripts.

5.1: Results

In this experiment, 320 different original recordings with 16 different speakers were pre-processed by increasing or decreasing either pitch or speech rate, or both. The results were then processed through a Dragon ASR server 11 system to generate the transcripts. Finally, the different generated transcripts were measured to find the lowest WER.

ANOVA

Source of Variation	*Sum of Squares*	*Degrees of freedom*	*Mean Square*	*F*	*Probability*	*F crit*
Samples (recordings)	222265.268	319	696.756	15.559	0.00	1.203
WER	12314.836	1	12314.836	275.001	0.00	3.871
Error	14285.134	319	44.781			
Total	248865.237	639				

Table 12 Variance and measures of WER for transcripts of original and processed recordings

There was a statistically significant difference between the WER for the PR and OR transcripts as indicated by the data because F value = 275.001 is much greater than F critical = 3.871, while the probability value < 0.01 which is less than the alpha significance level (0.05), so the null hypothesis is rejected. The PR transcripts were associated with lower WER than the OR transcripts (mean WER of OR = 48.92; mean WER of PR = 40.15).

	Number of samples	Mean	Stand . Dev.	Stand. Error	Confidence Level (95%)	95% confidence interval for mean		minimum	maximum
						Lower bound	Upper bound		
WER on OR transcripts	320	48.92	20.36	1.14	2.24	46.68	51.16	2.6	98.3
WER on PR transcripts	320	40.15	18.09	1.01	1.99	38.16	42.14	1.7	90.0

Table 13 Descriptive statistics of WER for transcripts of original and processed recordings

Table 13 shows the mean WER in OR transcripts of a sample of 320 processed recordings is 48.92, with S.D. of 20.36 and standard error of the sample at 1.14. At the 95% confidence level, the true mean WER of OR transcripts is between 46.68 and 51.16.

Table 13 also shows the mean WER in PR transcripts of a sample of 320 processed recordings is 40.15, the S.D. is 18.09, while the standard error is 1.01. At the 95% confidence level, the true mean WER of PR transcripts is between 38.16 and 42.14.

The mean WER of PR transcripts can be seen to be much lower than the mean WER of OR transcripts. Processing the speaker recordings appears to have led to a better match of the formants of the speakers in the training set when the speech recognition system was built.

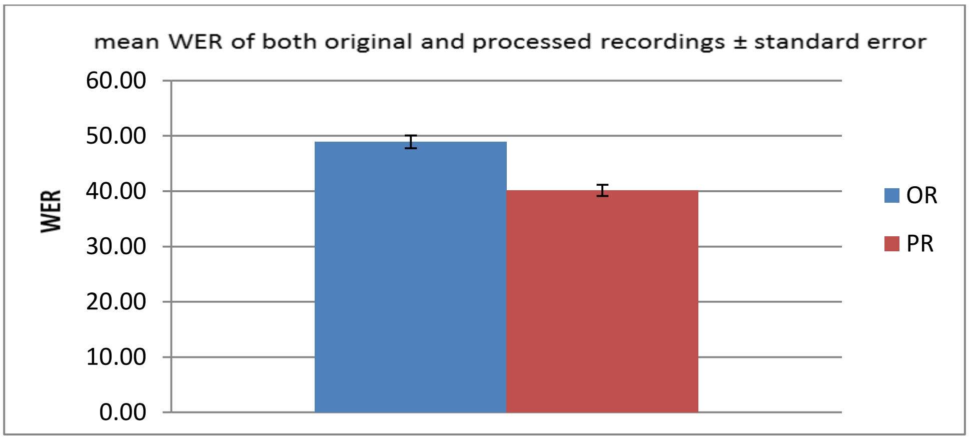


Figure 5.1-24 WER (mean± standard error of the mean) of OR and PR transcripts

In Figure 5.1-24 , the OR plots the mean WER = 48.92 ± 1.14 of n = 320 samples of original recoding transcripts; the PR plots mean WER = 40.15 ± 1.011 of n = 320 samples of processed recoding transcripts.

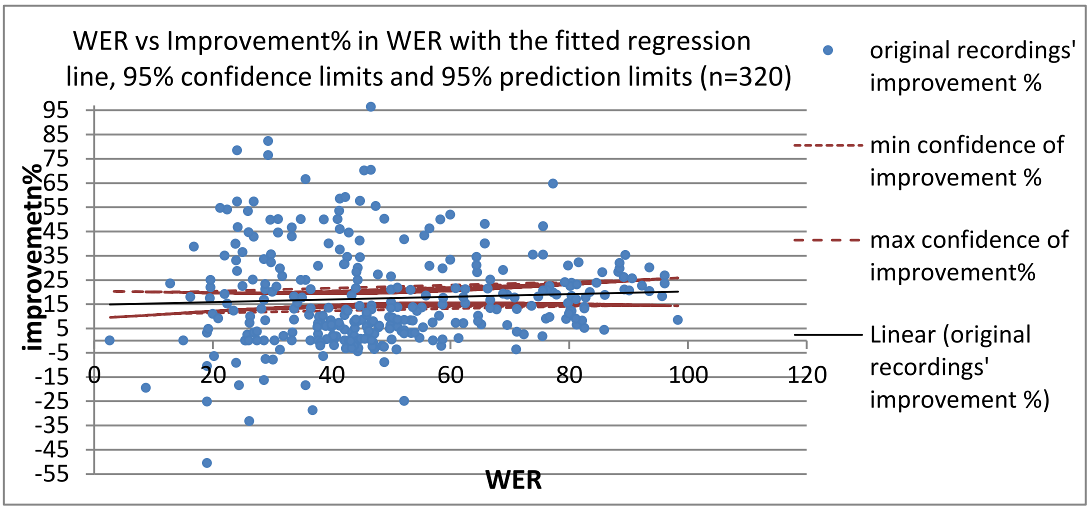


Figure 5.1-25 Correlation between the WER of OR and the improvement % for those WERs that resulted from the optimum processing. The red lines represent 95% confidence interval.

Figure 5.1-25 plots the results, which are summarised below:

- 95% confidence interval for the true mean improvement % in WER falls within the mean range of 15% to 20%
- the interpretation of data plots consistently shows with 95% confidence that the mean improvement range in WER of all speakers is between 15% and 20%

5.2: Discussion

Pre-processing recordings before transcribing them, by modifying the fundamental frequency and speech rate (SR) through the use of pitch shifting and time stretching libraries, led to significant decreasing the mean WER from 48.92 ± 1.14 to 40.15 ± 1.01, which was tested using a paired T-test that resulted a significance value $p < 0.001$

The results of this method were analysed for 320 pre-processed recordings. The main challenge in reducing the WER in ASR transcripts is to establish an efficient way to determine the optimum pre-processing frequency and speech rate warp factors.

Shifting the pitch and SR led to warping of the formants of the original recordings and resulted in improving the ASR system WER. The results of the above investigation showed 95% confidence that the mean WER of all speakers improved by between 15% and 20%.

5.3: Statistical significance of differences in ASR performance

Pre-processing recordings before transcribing them was evaluated here by recognition accuracy (AC), observed using a standard corpus of spoken words. AC was obtained by comparing the word-string outputs for recordings produced by the recognizer, before and after processing, to the word string that was actually spoken.

Pallet, et al. (1990) defined the computation of the word error rate, dependent on a standard string matching program, as the percentage of errors that included substitution, insertion and deletion of words. To evaluate the statistical significance of WER differences, this work applied NIST, an automated benchmark scoring tool, to compare a pair of transcript results. The reference transcript, which contained the word string actually uttered, was compared with the hypothesis transcript, which contained the word string output produced from the ASR. The hypothesis transcript was also compared with the transcript produced from the system for the processed recordings. To interpret the NIST results objectively, it is important to know whether any apparent difference in WER of hypothesis transcripts and transcripts of processed recordings is statistically significant. The t-test was used to determine the statistical significance of recognition results. An improvement was considered significant if the relative difference in WER between two results was greater than 11%.

The WER improvement for each speaker is dependent on the choice of the factor used to warp the formants of the recordings with the sound processing libraries. The modification of each speaker's recording was determined by a 'trial and error' approach. Initially, each warp factor was investigated separately to analyse its influence on speech recognition performance. The determination of the best warp factor for each speaker started by investigating a number of possible factors, selecting those that resulted in an improvement of speech recognition performance, and determining which gave the best improvement. The WER improved by between 15% and 20% by processing.

5.4: The shortest length of recording that represents the WER of the full recording

This section addresses the time taken to find the best pre-processing for the speaker's recording. Two major factors were investigated:

- Finding the shortest length of recording that represents the WER of the full recording, to save the time taken by participants to listen to the full audio recording.
- Investigating whether participants were able to detect the transcript that had the lowest WER.

The experiment chose to take only the first and second minute of original and processed speaker recordings, splitting each minute into 4 segments of 15 seconds. Each segment was then aligned with its original and processed transcripts. The reasons for splitting the recordings into short segments was to reduce the time taken and to avoid possible problems with retrieving excessively long transcripts.

The first question to be answered was whether the average WER of the first or second minute of the recordings could be used to represent the average WER of the whole recordings. The reason behind this is that ASR automatically looks for a sequence of three words in real time to adapt its statistical models to new voices, and the first spoken words are less accurately observed because they are not preceded by any content that can be used by the statistical language models (Machlica, et al., 2009). Thus the question becomes: if we pick the first or the second minute of the recording, what is the probability that the average WER is going to be different from the average WER of the whole recording?

Hypotheses:

H1 There is no difference in the mean WER of recordings compared with the mean WER of the first minute of those recordings.

H2 There is no difference in the mean WER of recordings compared with the mean WER of the second minute of those recordings.

5.4.1: Results

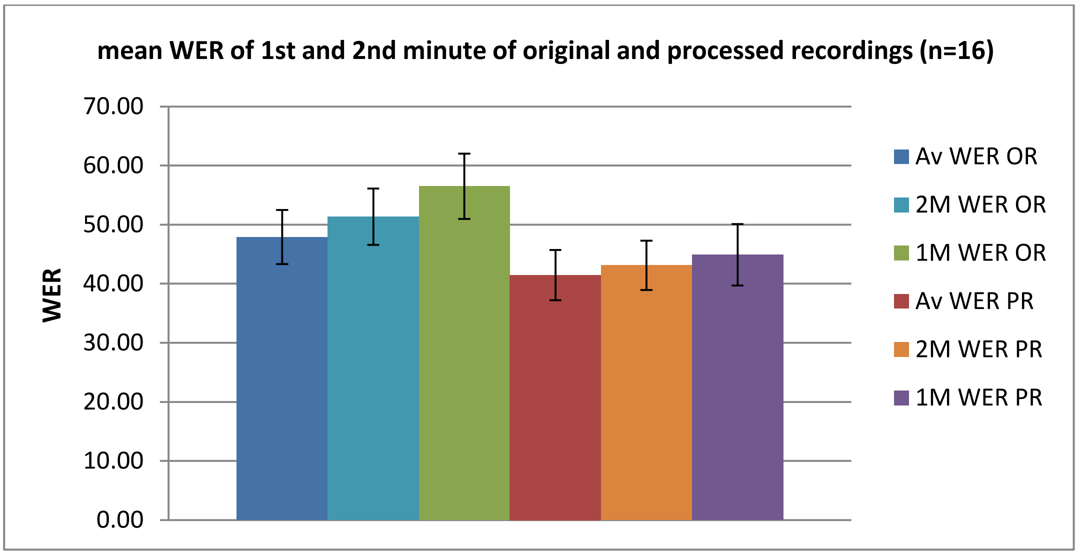


Figure 5.4.1-26 WER mean ± SEM of OR and PR for average and both first and second minutes of those recordings

Figure 5.4.1-26 plots the results for the 16 samples, which are summarised below:

- 'Av WER OR' plots the mean WER = 47.9 ± 4.56 of original recording transcripts of different speakers. The Av 'WER PR' plots mean WER = 41.46 ± 4.23 of processed recording transcripts.

- '1M WER OR' plots the mean WER = 56.49 ± 5.55 of the transcripts of the first minute of each original recording, while the '1M WER PR' plots the mean WER = 44.89 ± 5.21 of transcripts of the first minute of each processed recording.

- '2M WER OR' plots the mean WER = 51.34 ± 4.76 of the transcripts of the second minute of each original recording, while the '2M WER PR' plots the mean WER = 43.12 ± 4.17 of the second minutes of each processed recording.

A paired t-test was used to compare the average WER of the difference between the first minutes' WER of both original and processed recordings, with the average WER of the difference between the whole original recordings' WER and the whole processed recordings' WER. The results of the paired t-test were t(16) = 3.07, df = 14, with *p* = 0.01, which therefore indicates a 1% chance there is no real difference and since *p* is less than the alpha significance level (0.05), so hypothesis H1 is rejected (average WER ≠ average WER of 1st min).

A paired t-test was used to compare the average WER of the difference between the second minutes' WER of both original and processed recordings, with the average WER of the difference between the whole original recordings' WER and the whole processed recordings' WER. The results of the paired t-test are t(16) = 0.95, df = 14, with *p* = 0.36 which indicates a 36% chance there is no real difference, and since *p* is greater than the alpha significance level (0.05), so hypothesis H2 is accepted (average WER = average WER of 2nd min).

5.4.2: Investigating the improvement in PR transcripts

The second step was to investigate whether it was possible to pick the transcript that has the lowest WER from the PR transcripts. First, the second minutes of two different speakers recordings were each split into 4 segments of 15 seconds. Each original segment was then aligned with its transcripts of original and processed audio for that segment. Thirdly, the transcripts of the audio segments were presented while listening to the original audio segments for each speaker and judge (i) which transcript had more correct words, (ii) which transcript would be more effective for understanding the gist of the spoken words. If the results showed that the best matching transcripts could be distinguished without any correcting procedure having been performed on them, then time would be saved in finding the best pre-processing for each speaker.

5.4.2.1: Results

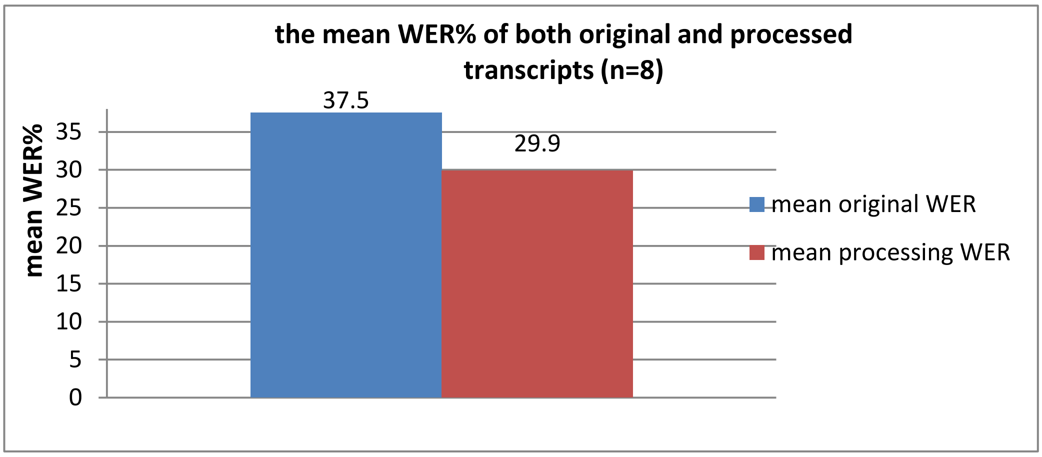


Figure 5.4.2.1-27 Mean WER of both original recording and processed transcripts

Figure 5.4.2.1-27 plots the results for eight different original and processed transcripts, which are summarised below:

- ‘mean original WER’ plots the mean WER = 37.5 of the original recording transcripts, ‘mean processing WER’ plots the mean WER = 29.9 of the processed recording transcripts.

A paired t-test was used to check the effectiveness of processing the speaker recordings before transcribing those recordings in reducing the WER. The results gave a *probability* < 0.02, with processed recording transcripts associated with lower WER than original recording transcripts (mean WER of OR = 37.5; mean WER of PR = 29.9).

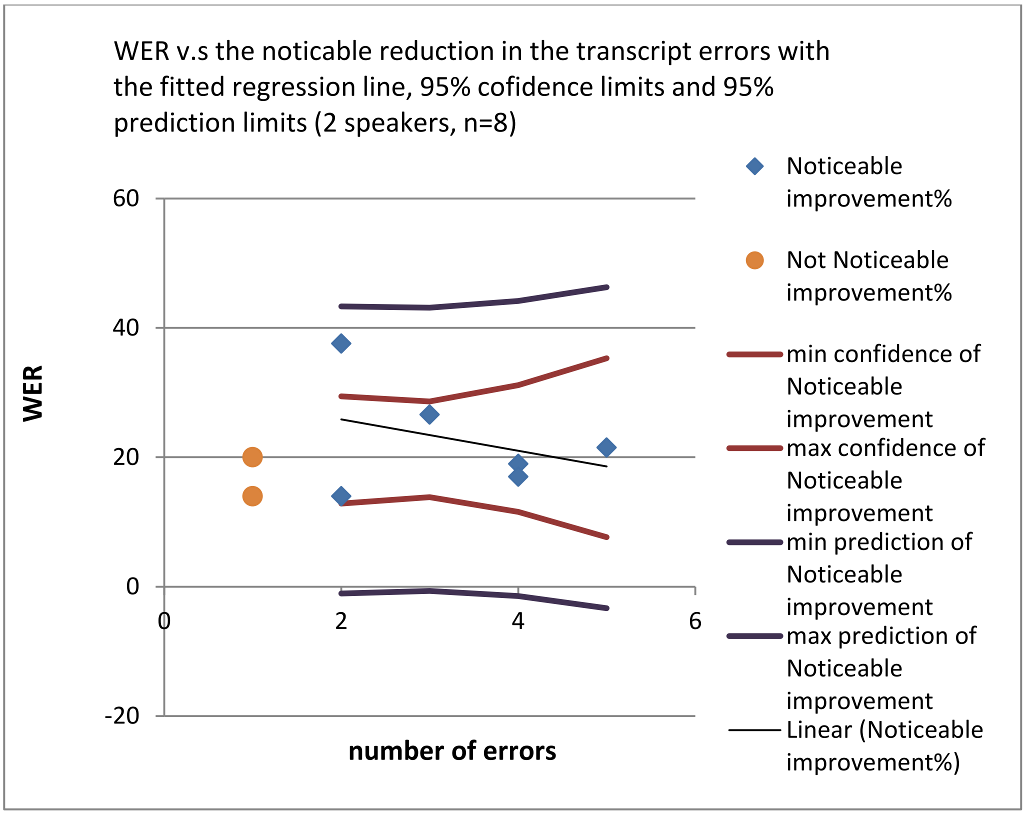


Figure 5.4.2.1-28 Correlation (black line) between both Noticeable and Not Noticeable improvement in the original transcripts and the reduction in those transcript errors that resulted from the optimum processing.

The interpretation of the results plotted in Figure 5.4.2.1-28 are summarised below:

- 95% confidence interval for the noticeable improvement% in the transcript WER lies between the range of the red lines.
- The interpretation of data plots consistently shows 95% confidence that the mean Noticeable improvement% range is between 11.11 and 31.31.
- 95% of the WER Noticeable improvement% to be found for a certain reduction in the transcript errors will be within the prediction interval range around the linear regression line (black).
- Error reduction is Noticeable when the errors reduced are more than 2 and also mean WER improvement% range is between 11.11 and 31.31.

5.4.2.2: Discussion

Transcribing pre-processed recordings through ASR resulted in a significant noticeable reduction in transcript errors, for which a paired t-test gave a significant $p < 0.02$. The reduction in the transcript errors was Noticeable when the transcript errors reduced by more than 2 and the WER improvement was more than 11.11%.

Chapter 6: Formant extraction and modification

This chapter uses the best modifications for a set of recordings found in Chapters 4 and 5 as a training set to find a function that could be used to estimate a new recording modification. The following sections explain the procedure to find the modification function and discusses the results of implementing this function on new recordings.

6.1: Introduction

Modifying speakers' recordings before transcribing those recordings via ASR resulted in improving the WER of the system by an average 17% as shown in Chapter 5, although the iteration procedure for this modification affected its performance measured by computational cost and reliability. Two approaches have been used to investigate the possibility of reducing the computational cost of pre-processing recordings with the *SoundTouch* or *Rubber Band* libraries. First, measuring the noticeability of modifying recordings' improvement, which was discussed in Chapter 5, demonstrated an improvement in ASR transcripts of modified recordings of 11% decrease in the WER. Secondly, using the best warp factors for a set of speakers, which was defined in Chapters 4 and 5, as a training set for investigating a function that estimates the modification required to produce an improvement for recordings of new speakers.

Since the optimal warp factors have been determined previously, this chapter studies the issues related to the combination of these factors as a training set. The speakers' original formants and the best modified formants of these speakers' are considered as points on a plane, and a line is fitted to these points in an attempt to find a function for modifying recordings. The experiment reported here uses the *Synote* database.

6.2: Formant-based recording modification

Formants are the basic cue in computing the warp factor of most normalization methods, which are used for mapping between two spectra. The results of the experiments in Chapter 4 indicated that modifying the pitch of speakers' recordings led to warping the formants of these recordings, before also warping these formants through the ASR system, which resulted in improvement in the system performance. Motivated by these two findings, the formants of the original and modified recordings of each speaker were extracted to compute the warp factors which are then used as a training set in these experiments.

To modify recordings based on speech formants, the distribution of the formant frequencies were collected and implemented separately to analyse the benefits of each parameter, in order to find the one that achieves the desired result.

6.3: Pilot experiment to choose an appropriate formant extraction tool

This experiment compares the formants extracted using *Praat*, a tool for general purpose speech analysis and the most popular tool used in phonetic research, with those obtained using *DeepFormants* (Dissen & Keshet, 2016), a tool to estimate formant frequencies. The reason for this experiment is to find which of these tools achieves the better results for formant extraction when compared to the manually estimated reference. An appropriate sample size for the pilot study should be 10% of the projected sample of a larger study (Connelly, 2008). However, Hertzog (2008) cautions that such studies are generally influenced by many factors and hence this issue is not easy to resolve. A different number of participants have been suggested:

- 10% of the sample size of project was suggested by Treece (1982).
- 10 to 30 participants were suggested by Isaac and Michael (1995)
- 10 to 30 participants were suggested by Hill (1998)
- 12 participants were suggested in the medical field by Belle(2002) and Julious (2005)

For a sample size of 300, it can be concluded that the study sample should be a minimum of 10 and be as much as 30.

Therefore, 10 recordings were used, consisting of 6 male and 4 female speakers, for the pilot experiment. The results of the pilot were then used to modify a different 30 speaker recordings, consisting of 15 male and 15 female speakers. Finally, those 30 recordings were transcribed with

ASR and the ASR performance improvement was calculated. Some of the recordings that were used in this experiment were gathered from the *Synote* website. Other recordings were collected from the *TED* website, a media organization that posts talks on scientific, cultural, and academic topics online, with their transcripts, for free distribution. Formants were extracted by both *Praat* and *DeepFormants* with a maximum formant value of 5.5 kHz and a window length of 100 ms. Each speaker's audio file was chunked into sub-recordings of 100 ms length and processed automatically by *Praat* and *DeepFormants* to extract the formants. Then the mean and median for each formant was calculated for each speaker by computing the mean and median for each formant of all the speaker's sub-recordings.

6.3.1: Formant extraction by Praat and DeepFormants

This section shows the formant extraction values from *Praat* and *DeepFormants* for 10 speaker recordings, and the impact on the WER of ASR when modifying the other 30 speaker recordings based on the estimated formant modification values from those extracted formants.

Table 14 shows the formant values in (Hz) of each speaker's recording extracted by *DeepFormants* and *Praat*. They produced different results, and these also differed from those observed by the manual extraction. These findings matched the results of Dissen & Keshet (2016), who also evaluated *DeepFormants* by comparing the formant values extracted with those extracted by Praat. A paired sample T-test showed that there was no significant difference between the F3 values extracted by either tool compared to those from the manual extraction. A further T-test showed that there was no significant difference between the F2 and F3 values extracted by Praat and DeepFormants, while the values of F1 differed significantly between the tools, with $p < 0.01$.

		Praat						DeepFormants					
		Original			Processed			Original			Processed		
Speaker		F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3
S1	Mean	646.59	1508.29	2351.41	688.77	1635.68	2499.37	441.70	1377.75	2255.37	475.82	1466.44	2385.79
	Median	612.54	1534.44	2340.55	666.30	1660.70	2485.46	438.89	1421.77	2325.75	492.13	1548.61	2512.93
S2	Mean	793.42	1703.79	2581.93	812.17	1760.95	2646.40	468.02	1599.77	2524.06	473.27	1634.45	2564.97
	Median	910.50	1759.52	2616.70	849.89	1839.87	2727.91	468.02	1571.06	2525.88	474.15	1598.91	2557.82
S3	Mean	673.78	1644.34	2219.44	643.35	1589.21	2205.21	460.65	1428.06	2230.59	408.94	1263.18	2084.83
	Median	672.15	1664.71	2209.53	631.17	1620.43	2211.15	478.89	1687.59	2479.78	437.87	1515.84	2371.84
S4	Mean	632.31	1547.75	2184.68	599.02	1504.46	2222.76	510.07	1668.24	2493.67	467.17	1550.12	2387.07
	Median	624.84	1592.39	2171.94	573.81	1544.67	2274.69	482.03	1676.43	2512.94	458.39	1560.28	2408.55
S5	Mean	801.76	1751.99	2676.33	755.30	1730.87	2671.64	423.61	1615.97	2581.37	421.83	1610.19	2570.94
	Median	629.21	1674.52	2539.78	597.97	1669.05	2611.82	425.83	1612.20	2558.95	425.48	1613.74	2557.94
S6	Mean	485.98	1498.77	2410.71	514.86	1557.56	2471.49	415.89	1533.67	2492.41	443.38	1589.53	2554.40
	Median	477.64	1495.10	2405.85	519.32	1587.22	2507.81	409.52	1506.51	2487.19	435.49	1578.91	2529.59
S7	Mean	577.53	1505.52	2401.48	602.86	1510.02	2384.32	452.46	1509.19	2484.49	473.64	1549.67	2531.51
	Median	536.47	1496.84	2397.48	566.79	1508.79	2382.09	437.99	1490.33	2477.83	458.24	1544.43	2525.64
S8	Mean	614.90	1548.66	2388.01	594.82	1536.19	2361.59	485.05	1647.81	2602.17	460.14	1609.41	2518.20
	Median	551.32	1566.19	2479.34	523.79	1606.89	2463.74	473.30	1727.65	2691.86	449.12	1671.64	2584.72
S9	Mean	638.21	1691.64	2545.45	646.66	1695.16	2547.06	425.93	1587.13	2530.02	415.10	1551.49	2491.18
	Median	549.08	1719.22	2689.98	552.29	1719.72	2686.68	422.54	1640.20	2596.65	420.82	1630.90	2604.67
S10	Mean	566.24	1565.64	2485.23	574.78	1630.83	2561.87	401.09	1459.11	2380.73	413.67	1492.97	2438.06
	Median	502.84	1566.24	2477.99	536.74	1652.13	2598.40	398.27	1536.94	2435.86	411.61	1572.17	2508.69

Table 14 Formant values extracted by DeepFormants and Praat

The other 30 speakers were used as a test set and were modified based on values extracted by each tool separately, to find which of *Praat* or *DeepFormants* gave a lower WER when transcribing these recordings with ASR. Table 15 shows the mean and median of the first three formants for each speaker extracted by *Praat* and *DeepFormants,* and the estimated values of the modified mean and median of the first three formants for each speaker. These were estimated using linear regression by defining the median of the original and modified values for the first three formants of the first 10 speaker recordings on two different graphs, one for male speakers and the other for female speakers. On each graph, the median of the original values was plotted against the median of the modified values, and the straight line that best fits the points defines a function that was used to modify the other 30 speaker recordings. Further graphs for males and females were constructed for the mean of the original and modified values for the same formants for these 10 speakers. The resulting straight line functions were used to estimate the modified values for the mean of the original first three formants from the other 30 recordings of males and females. This approach was implemented for both tools.

Table 15 also shows the original WER for each speaker and the WER that resulted from transcribing the modified recording based on the mean and on the mean of formants extracted by each of *DeepFormants* and *Praat*. Most of the WERs from transcribing the modified recordings were better than the WERs from the original recordings. A Wilcoxon test showed that even with an improvement of 6.5% in WER, the difference between the original WER and the modified WER for male speakers was not significant, with $p = 0.2$, when using both *Praat* and *DeepFormants* to modify these male recordings. On the other hand, the Wilcoxon test shows that the WER was significantly improved by 16% for female speakers, with $p < 0.01$, when modified using *DeepFormants*.

Speaker	**Mean formants *DF***	**Mean formants *Praat***	**Median formants *DF***	**Median formants *Praat***	**Mean linear regression estimate *DF***	**Median linear regression estimate *DF***	**Mean linear regression estimate *Praat***	**Median linear regression estimate *Praat***	**Original WER**	**WER of mean *DF***	**WER of median *DF***	**WER of mean *Praat***	**WER of median *Praat***
11 (M)	1470.83	1460.13	1465.07	1482.71	1507.18	1554.508	1521.198	1556.847	67.0	64.7	60.4	57.8	58.8
12 (M)	1483.30	1496.20	1613.23	1604.48	1515.23	1609.402	1550.123	1668.217	24.5	21.7	32.9	24.5	25.2
13 (M)	1506.85	1487.42	1552.48	1580.87	1530.45	1586.894	1543.082	1646.624	44.6	44.0	42.9	41.7	49.4
14 (M)	1514.61	1530.46	1582.76	1606.19	1535.46	1598.113	1577.596	1669.781	44.3	41.2	39.7	44.3	40.5
15 (M)	1479.79	1484.01	1498.93	1480.35	1512.96	1567.054	1540.348	1554.688	51.6	47.2	48.4	50.9	50.3
16 (M)	1502.32	1489.66	1547.95	1480.21	1527.52	1585.215	1544.878	1554.560	37.7	46.4	40.6	47.8	48.6
17 (M)	1513.84	1511.64	1600.19	1587.90	1534.96	1604.570	1562.504	1653.053	34.1	33.3	29.4	31.7	35.7
18 (M)	1429.47	1470.99	1509.69	1521.65	1480.46	1571.040	1529.907	1592.461	46.8	31.9	32.6	29.8	29.1
19 (M)	1505.95	1482.95	1533.17	1515.86	1529.86	1579.739	1539.498	1587.164	24.5	4.5	17.3	9.8	18.8
20 (M)	1592.26	1574.18	1688.13	1653.68	1585.62	1637.152	1612.655	1713.216	24.7	32.0	33.3	30.0	32.0
21 (M)	1462.41	1424.78	1560.33	1497.69	1501.74	1589.802	1492.851	1570.547	32.9	32.9	36.9	35.6	35.6
22 (M)	1484.02	1481.81	1522.67	1505.91	1515.70	1575.849	1538.583	1578.065	42.0	35.3	38.7	43.3	39.3
23 (M)	1502.96	1524.91	1565.80	1577.37	1527.93	1591.829	1573.145	1643.423	21.3	23.3	23.3	23.3	22.0
24 (M)	1529.32	1542.25	1589.69	1601.75	1544.96	1600.680	1587.050	1665.721	7.8	8.6	8.6	11.7	9.4
25 (M)	1496.91	1538.59	1559.41	1590.11	1524.02	1589.461	1584.115	1655.075	36.8	34.5	32.2	29.8	34.5
26 (F)	1630.48	1463.93	1758.54	1534.68	1595.60	1628.643	1439.987	1535.941	29.5	23.0	19.4	24.5	30.9
27 (F)	1595.14	1419.76	1687.73	1444.30	1549.83	1596.779	1389.775	1458.063	31.7	23.0	20.1	23.0	42.4
28 (F)	1581.87	1490.84	1627.36	1556.75	1532.61	1569.612	1470.586	1554.961	42.1	43.0	43.8	41.3	38.8
29 (F)	1625.54	1570.03	1782.26	1789.95	1589.28	1639.317	1560.617	1755.910	25.8	23.3	28.9	22.0	21.4
30 (F)	1640.37	1503.13	1809.33	1711.78	1608.53	1651.499	1484.558	1688.551	41.4	26.3	22.4	29.6	27.6
31 (F)	1585.79	1501.56	1671.36	1618.02	1537.70	1589.412	1482.774	1607.758	60.2	40.7	35.6	39.8	39.0
32 (F)	1607.84	1500.01	1709.42	1593.30	1566.31	1606.539	1481.011	1586.457	42.8	29.9	26.2	26.7	32.1
33 (F)	1597.84	1479.55	1670.69	1574.27	1553.34	1589.111	1457.750	1570.058	35.7	31.2	33.1	32.5	29.2
34 (F)	1519.16	1444.93	1587.24	1474.24	1451.23	1551.558	1418.391	1483.863	25.5	35.0	36.5	35.0	25.5
35 (F)	1554.13	1463.06	1646.73	1546.89	1496.62	1578.329	1439.003	1546.465	21.0	20.2	21.0	15.1	16.8
36 (F)	1631.09	1535.89	1722.89	1544.41	1596.49	1612.601	1521.803	1544.328	28.4	23.4	19.9	22.7	22.0
37 (F)	1582.62	1472.42	1620.38	1501.45	1533.59	1566.471	1449.644	1507.309	53.6	42.9	42.9	44.6	80.4
38 (F)	1569.57	1397.51	1600.19	1437.17	1516.65	1557.386	1364.479	1451.919	35.9	33.3	34.6	34.0	55.6
39 (F)	1605.80	1388.03	1653.29	1346.30	1563.67	1581.281	1353.701	1373.617	60.3	26.5	29.4	33.1	83.8
40 (F)	1581.34	1571.68	1586.52	1582.08	1531.93	1551.234	1562.493	1576.788	47.0	44.0	44.0	45.2	41.0

Table 15 WER from modifying formants based on the mean and median of the first three formants extracted

6.3.2: Relationship between modification% and improvement%

In this section the modification% vs. improvement% for male and female speakers were each plotted separately.

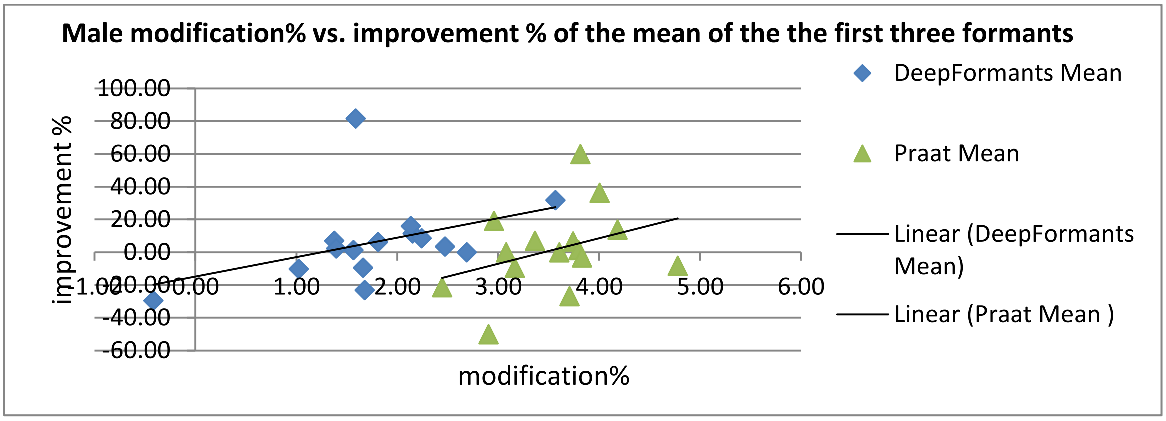


Figure 6.3.2-29 Male modification% vs. improvement% of the mean of the first three formants

Figure 6.3.2-29 plots the results, which are summarised below:

- For male speakers, increasing the mean of the first three formants by the estimated formant value from the mean of the formant values extracted by *DeepFormants*, by between 1.4% and 3.57%, improved the WER by between 1.35% and 81.63%.
- For speakers 24, 23 and 16, increasing the mean of the first three formants by 1.02%, 1.66% and 1.68% respectively resulted in the improvement% dropping by 10.26, 9.39 and 23.08, while for speaker 21 an increment of 2.69 gave no improvement. For speaker 20, however, the estimated modification values decreased the formants, resulting in decreasing the improvement by 29.55%.
- For male speakers, increasing the mean of the first three formants by the estimated formant value from the mean of the formant values extracted values by *Praat*, by between 2.9% and 4.18%, improved the WER by between 1.36% and 60%.
- For speakers 20, 24, 23, 16, 22 and 21, increasing the mean of the first three formants by 2.44%, 2.9%, 3.16%, 3.71%, 3.83% and 4.78% respectively (via *Praat*) resulted in the improvement% dropping by 21.46, 50, 9.39, 26.79, 3.09 and 8.21, while for speakers 14 and 12, increments of 3.08% and 3.6% gave no improvement.
- For speakers 20, 12, 24, 17, 13, 23, 21 and 16, increasing the mean of the first three formants by 3.6%, 3.97%, 3.99%, 4.10%, 4.16%, 4.19%, 4.86% and 5.02% respectively (via *Praat*) resulted in the improvement% dropping by 29.55, 2.86, 20.51, 4.69, 10.75, 3.29, 8.21 and 28.91.

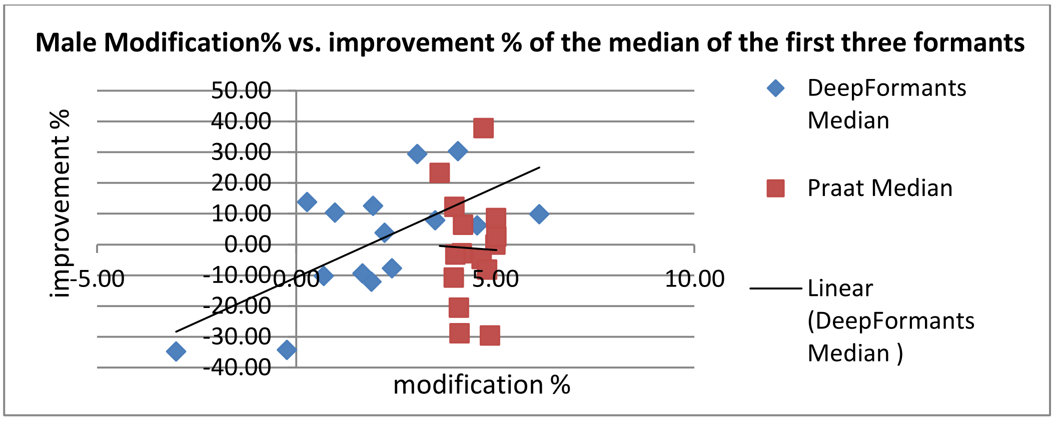


Figure 6.3.2-30 Male modification% vs. improvement% of the median of the first three formants

Figure 6.3.2-30 plots the results, which are summarised below:

- For male speakers, increasing the median of the first three formants by the estimated formant value from the median of the formant values extracted by *DeepFormants*, by between 0.97% and 6.10%, improved the WER by between 3.81% and 30.34%.
- For speakers 24, 23, 21 and 16, increasing the median of the first three formants by 0.69, 1.66, 1.89, and 2.41 respectively resulted in the improvement% dropping by 10.26, 9.39, 12.16 and 7.69. For speakers 12 and 20, however, the estimated modification values decreased the formants by 0.24% and 3.02% respectively, resulting in decreasing the improvement by 34.29% and 34.82% respectively.
- For male speakers, increasing the median of the first three formants by the estimated formant value from the median of the formant values extracted by *Praat*, by between 3.6% and 5%, improved the WER by between 9.4% and 58.8%.

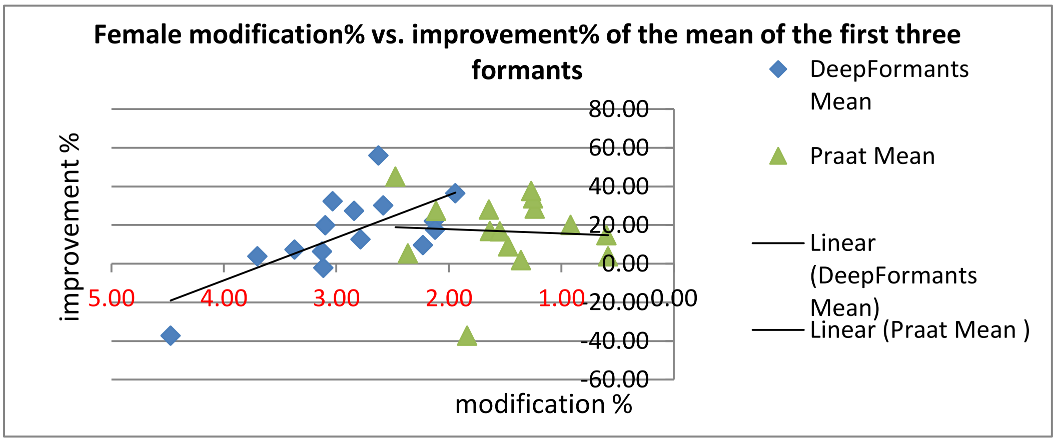


Figure 6.3.2-31 Female modification% vs. improvement% of the mean of the first three formants

Figure 6.3.2-31 plots the results which are summarised below

- For female speakers, decreasing the mean of the first three formants by the estimated formant value from the mean of the formant values extracted by *DeepFormants*, by between 1.94% and 3.7%, worsened the WER by between 6.38% and 56.05%.
- For speakers 28 and 35, decreasing the mean of the first three formants by 3.11% and 4.47% respectively resulted in the improvement dropping by 2.14% and 34.25%.
- For female speakers, decreasing the mean of the first three formants by the estimated formant value from the mean of the values extracted by *Praat*, by between 0.6% and 2.47%, worsened the WER by between 1.9% and 45%.
- For speaker 34, decreasing the mean of the first three formants by 1.84% (via *Praat*) resulted in the improvement dropping by 37.3%.

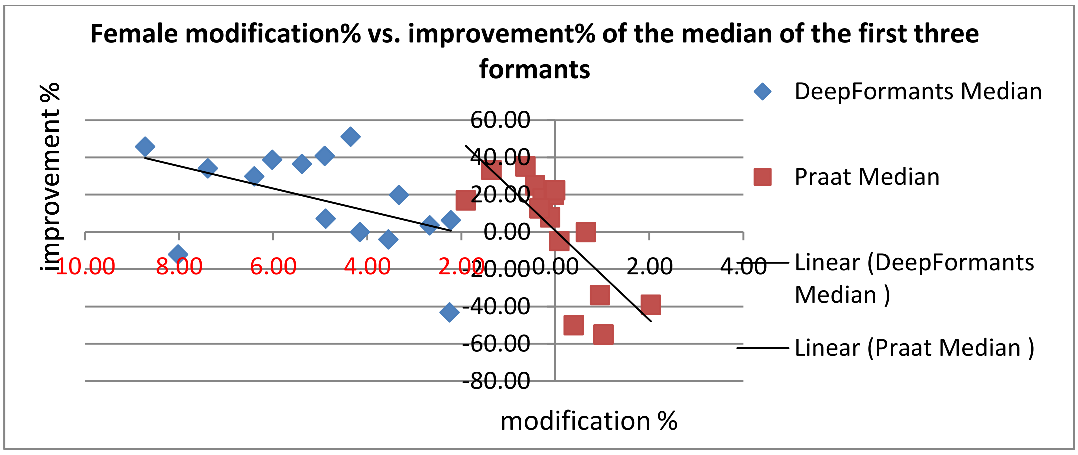


Figure 6.3.2-32 Female modification% vs. improvement% of the median of the first three formants

Figure 6.3.2-32 plots the results which are summarised below

- For female speakers, decreasing the median of the first three formants by the estimated formant value from the median of the formant values extracted by *DeepFormants*, by between 2.22% and 8.02%, worsened the WER by between 3.62% and 51.24%.
- For speakers 34, 28 and 29, decreasing the median of the first three formants by 2.25%, 3.55% and 8.02% respectively resulted in the improvement% dropping by 43.14, 4.04 and 12.02, while for speaker 35 a decrement of 4.15% gave no improvement.
- For female speakers, decreasing the median of the first three formants by the estimated formant value from the median of the formant values extracted by *Praat*, by between 0.1% and 1.9%, worsened the WER by between 7.8% and 25%.
- For female speakers, increasing the median formants by between 0.1% and 2%.

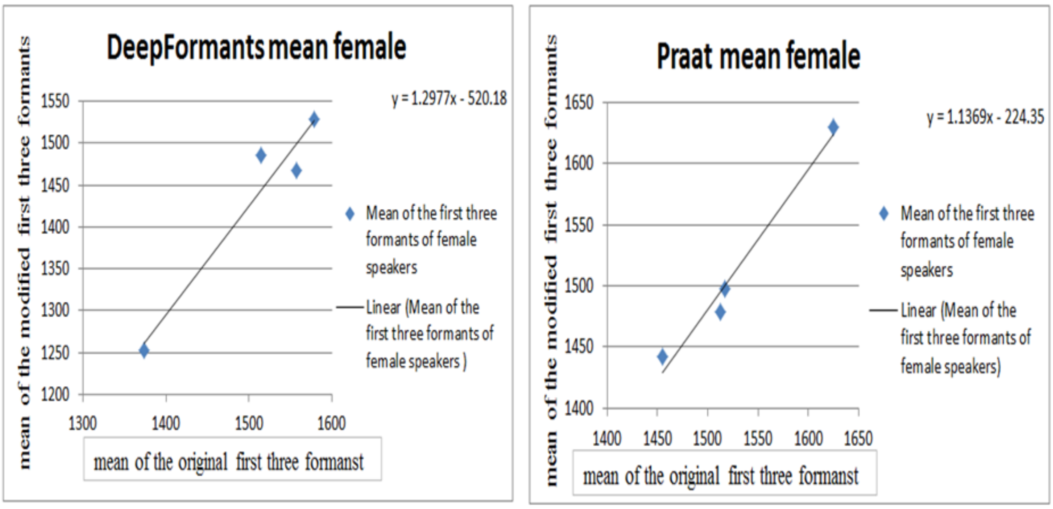


Figure 6.3.2-33 Linear regression approach used to estimate the mean of recording modification values for female speakers

Figure 6.3.2-33 shows the mean of the first three formants which have been extracted by *DeepFormants* and *Praat* from the original recordings, as x coordinates, and those extracted from modified recordings as y coordinates for 4 female speakers used as training data. Each straight line that fits these data sets was used separately to estimate the new female speakers' formant modification values.

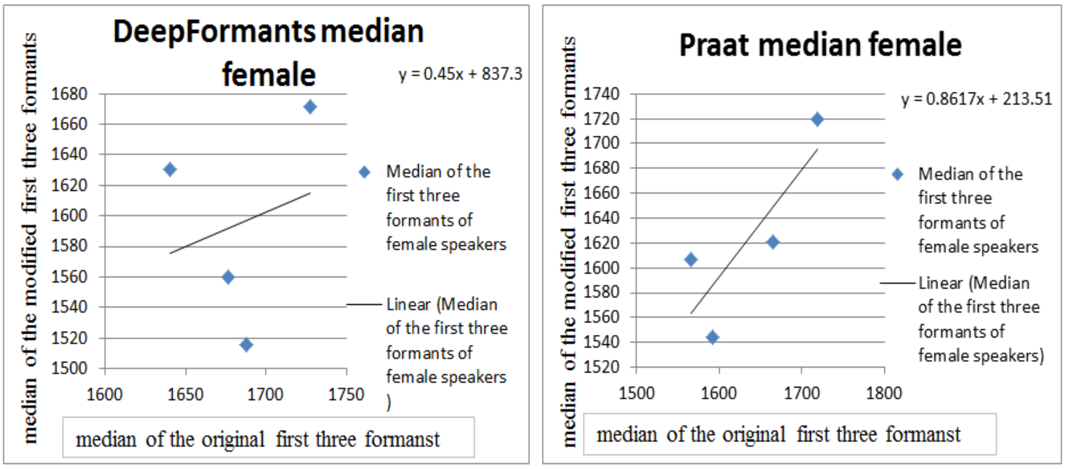


Figure 6.3.2-34 Linear regression approach used to estimate the median of recording modification values for female speakers

Figure 6.3.2-34 shows the median of the first three formants, which have been extracted by *DeepFormants* and *Praat* from the original recordings, as x coordinates, and those extracted from modified recordings as y coordinates, for 4 female speakers used as training data. Each straight line that fits these data sets was used separately to estimate the new female speakers' formant modification values.

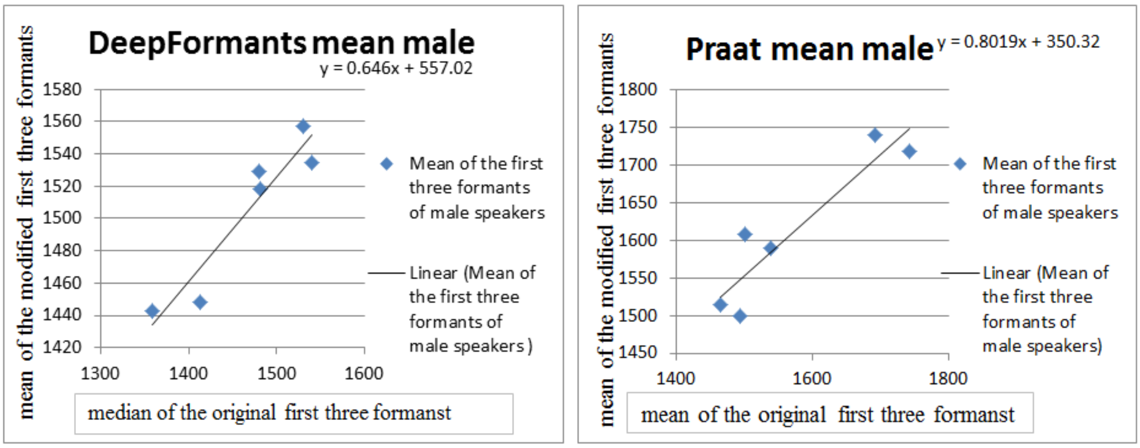


Figure 6.3.2-35 Linear regression approach used to estimate the mean of recording modification values for male speakers

Figure 6.3.2-35 shows the mean of the first three formants which have been extracted by *DeepFormants* and *Praat* from the original recordings, as x coordinates, and those extracted from modified recordings as y coordinates, for 6 male speakers used as training data. Each straight line that fits these data sets was used separately to estimate the new male speakers' formant modification values.

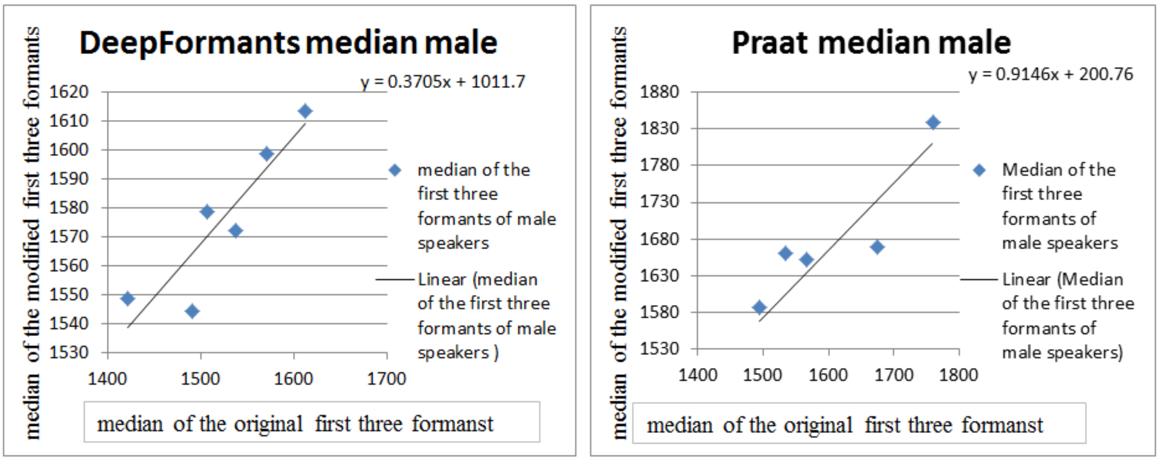


Figure 6.3.2-36 Linear regression approach used to estimate the median of recording modification values for male speakers

Figure 6.3.2-36 shows the median of the first three formants which have been extracted by *DeepFormants* and *Praat* of the original recordings, as x coordinates, and those extracted from modified recordings as y coordinates, for 6 male speakers used as training data. Each straight line that fits these data sets was used separately to estimate the new male speakers' formant modification values.

6.3.3: Formant frequency (F1, F2, F3) in isolation

In In order to understand the importance of each formant for normalizing speech, this chapter studies the individual benefits of using each formant separately.

The original formants of each speaker were extracted from their original recording, while modified formants were extracted from the best warping formant recording – that is, the one that gives the minimum WER for that speaker when transcribing this recording via ASR.

The formants were extracted by analysing each 100 ms interval for each original and modified recording in the training set, and computing the median and mean for F_2 and F_3 for the whole recording. Linear modifying functions were then defined from the median and mean for each formant from F_2 and F_3 separately. The median of the original and modified values for each formant were plotted on a graph, where the x coordinates were the median of the original values of the formant, and the y coordinates the median of the modified values. The straight line that best fitted the values defined the function for modifying the recording. A similar function for the mean was computed. Table 16 presents the results for each of these linear functions.

Speaker	Feature	Formant	Improvement of WER based on formant values extracted by		
			DeepFormants	***Praat***	***Manual***
Female	Mean	F_2	15.6%	18%	-
		F_3	16%	Worse by 26.7%	10%
	Median	F_2	17%	No improvement	-
		F_3	15%	Worse by 43%	17%
Male	Mean	F_2	4.7%	No improvement	-
		F_3	4.5%	3.3%	Worse by 0.8%
	Median	F_2	1%	Worse by 0.8%	-
		F_3	Worse by 6.4%	Worse by 1.2%	Worse by 11%

Table 16 Improvement% using linear regression functions

Table 16 shows large differences in improvement. The results in Table 16 can be sorted on the relative benefit to ASR performance WER, as shown in Table 17, where the values that resulted in similar improvement of WER are grouped together.

Order	**Linear**
Best Female	Mean F_2 *Praat*
	Median F_2 *DeepFormants* & Median F_3 *Manual*
	Mean F_3 *DeepFormants*
	Mean F_2 *DeepFormants*
	Median F_3 *DeepFormants*
	Mean F_3 *Manual*
Worst Female	Median and Mean F_3 *Praat*
	Median F_2 *Praat*
Best Male	Mean F_2 & F_3 *DeepFormants*
Worst Male	Median F_3 *Manual*

Table 17 Formants sorted by improvement% – cells present statistically significant differences

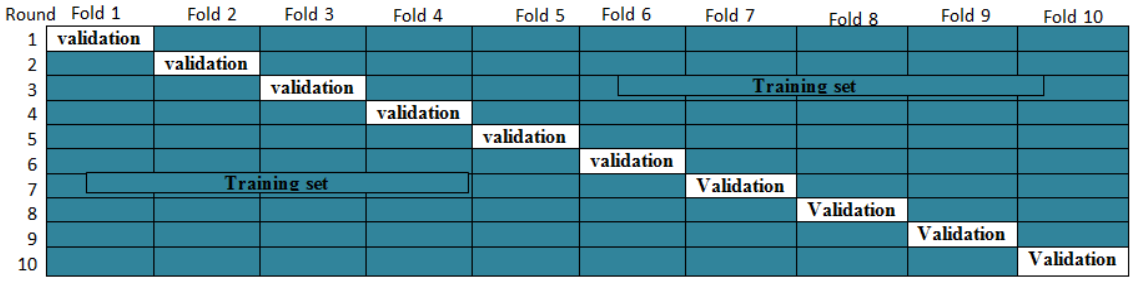
Table 17 indicates that means tend to be better than the medians of each formant, and also that the statistics of F_2 tend to be better than those of F_3.

However, the extraction of formants is not an exact process. This drawback was demonstrated when using different tools, where bad estimates for the formant values caused a greater effect on the modification function that was estimated using those values, and resulted in increasing the WER when using this function. The experiment in this section found that the mean and median of the second formant, extracted by different tools, gave better functions than the mean and median of the third formants. The improvement% achieved with the mean of the second formant extracted by *Praat* is the best for female speakers, while that from *DeepFormants* is the best for male speakers. The reason for this might be that the warping of the formants before ASR processing may interact with the automatic warping methods built into the ASR system, which mostly use F_3 for its warping function since this has the least variation across different phonemes uttered by a single speaker. Therefore, it is theoretically possible that implementing a re-warping function that uses F_2 as its guiding parameter would further improve WER.

6.4: Establishing a function for male recordings with 10-fold cross validation

The experiment in this section uses the same recordings that were used in the experiments of Chapter 5, but as a training set to help in estimating the modification factors for recordings of new male speakers. This experiment uses 10-fold cross validation (CV), a technique used for partitioning the sample into a training set for the predictive model and a test set to evaluate that model. The sample was divided into 10 equal-sized subsamples, and in each round, a single subsample was retained as the validation data to test the model while the other 9 subsamples were used as the training data (OpenML, 2016), as shown in Table 18.

The reasons for using CV are to decide which fraction of data is testing and which is training, while maximizing the number of items in both sets, which gets the best learning results in the training set and the best validation in the test set, and avoids a trade-off between the sets. This avoids the problem of finding a straight line function that fits the points of the training set where it starts to diverge when applied to the recordings in the test set.



Round	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
1	validation									
2		validation								
3			validation			Training set				
4				validation						
5					validation					
6						validation				
7	Training set						Validation			
8								Validation		
9									Validation	
10										Validation

Table 18 10-fold cross validation (CV)

The data from 240 males were segmented into 10 different folds that each contained 24 recordings, and 10 separate learning experiments run. In each experiment, one of the folds was picked as a test set while the other 9 folds were used as a training set. For each round, the formants extracted from the 216 recordings of the training set were plotted on a graph, whose x axis was the original values and y axis was the modified values. The straight line that fitted the values define the modifying function. The formants from the 24 recordings of the test set were modified by the function estimated by the training set of that round. This procedure was repeated 10 times using a different fold as the test set. Finally, the training set function that gave both the minimum root mean square error (RMSE), which measures the difference between the original and estimated values, and gave the highest WER improvement compared with the other functions, was used to estimate the modified formant values for 15 new male speakers. Table 19 shows the RMSE and improvement% results of each fold.

Training data			Validation set
Training set 90%	Test set 10%		15 new males
Folds	RMSE	Improvement%	Improvement%
1	18.62	13	5
2	20.37	2	0
3	34.13	8	5
4	13.17	15	7
5	6.48	9	6
6	10.95	5	0
7	20.30	13	2
8	38.89	27	3
9	21.79	12	0
10	13.05	19	1.11

Table 19 RSME and improvement% of each fold

Table 19 shows that although the estimation of folds 5 and 6 resulted in the minimum RMSE, which means that these folds' values were the closest to the manual estimation values for modifying recordings than the other folds, the estimated functions of these folds did not result in a higher improvement% than the other folds' functions. The Pearson correlation coefficient was $r = 0.4$ between RMSE values and the improvement% of the test set, which indicates a positive association between these variables, while $r = 0$, between RMSE values and the improvement% of the validation set, indicates that there is no association between these variables. This might be caused by high sensitivity in the improvement resulting from modifications, i.e. a small change in the best modification value resulted in a huge decrease in the improvement% and might lead to a worse performance of ASR than by transcribing the original version of that recording.

Using CV showed that the estimated function from training set of fold 4 led to a significant 7% improvement in the average WER of the 15 male speakers, with T-test $p < 0.05$.

6.5: Recording modification applied to different speech recognition systems

This section briefly explores the differences in ASR performance when modifying the recordings before transcribing them with different systems.

The same recordings that were used, either for male or female speakers, in the previous section have been used again to find the best modification of each recording based on different speech recognition systems YouTube, Siri, and Dragon Dictation.

The differences in performance of each speech recognition system are explored by applying a trial and error approach to modify each speaker's recording to find that which results in the best improvement in performance for each.

6.5.1: Impact of recording modification on YouTube speech recognition

To improve the performance of YouTube speech recognition, the trial and error approach was applied first to find the best modification for each speaker recording. Then the original and the best modified versions of each recording were used as a training set to find an appropriate function for estimating the modification values for new recordings, and also gave the minimum WER when processing them through YouTube's transcribing system.

Each of the 4 recordings of females, from 2 to 5 minutes long and contained 3,893 words in total, was processed through trial and error approach iteratively until the best modification was found. The recordings had their pitch changed with *SoundTouch* in increments of 0.1 in the range

[–3,3] and by changing the frequency in increments of 0.1 in the range [0.8,1.3] through *Rubber Band* . They were then transcribed by ASR and the one with the lowest WER taken. The original and modified versions of the recordings were converted to video with *Windows Movie Maker* and uploaded to YouTube for transcription. Their transcripts were then collected and compared with the correct transcripts by using NIST tool that gives the computed transcript WER. Figure 6.5.1-37 shows the average WER of the original and best modified versions of the 4 female speakers' recordings.

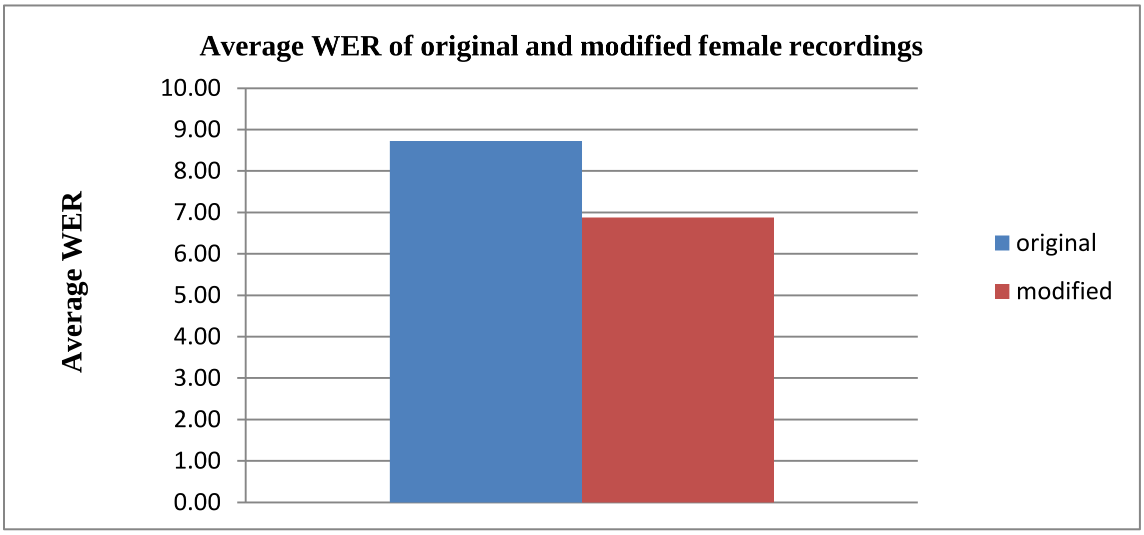


Figure 6.5.1-37 Average WER of original and modified females

Modifying the recordings of 4 females before converting these recordings to video and submitting them to YouTube's transcribing system resulted in an improvement in the system WER of 22%.

The same procedure was carried out for 4 male speakers' recordings which varied in length from 1 minute to 11h49m, and contained 4,445 words in total. Figure 6.5.1-38 shows the average WER of the original and best modified versions of the 4 male speakers' recordings.

Modifying the recordings of 4 males before converting these recordings to video and submitting them to YouTube's transcribing system resulted an improvement in the system WER of 17%.

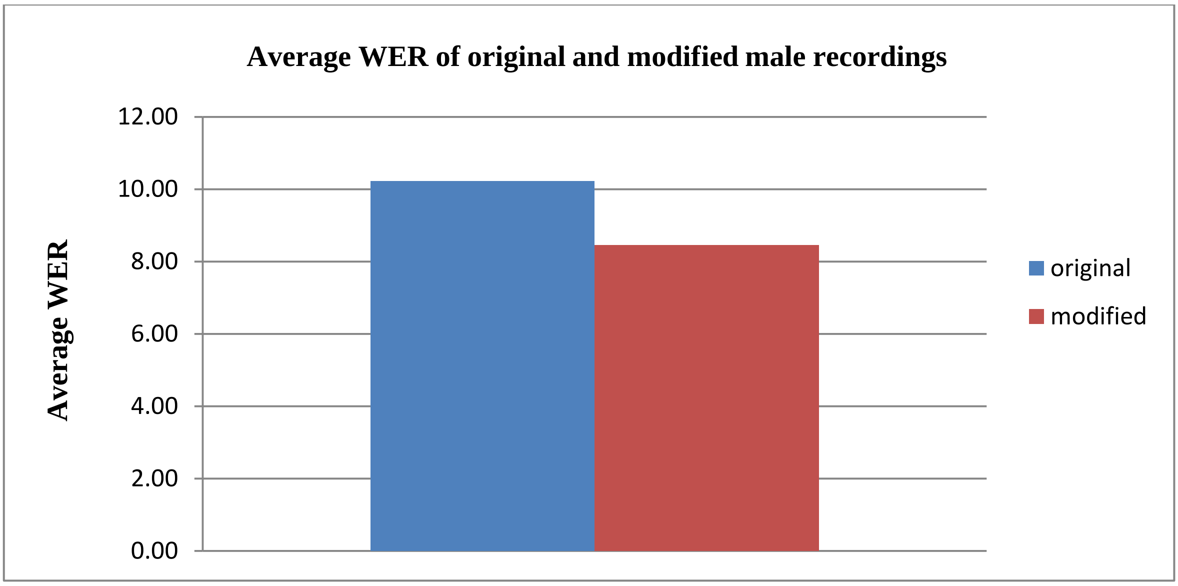


Figure 6.5.1-38 Average WER of original and modified male recordings

The original and best modified versions which resulted from this investigation were used in the following section as a training set to find an appropriate estimated function for the modification values for new speakers' recordings.

6.5.2: Linear regression and experimental results

An experiment was made using YouTube's ASR system with a database of 8 different speakers' recordings, consisting of 4 males and 4 females. The mean of the extracted formants of the original recordings and the mean of the formants extracted from the best modified versions were used as the training data for two separate simple linear regression approaches, one for each gender. In each approach, the mean of the original formant values is the explanatory variable denoted x and the mean of the modified formant values is the scalar dependent variable y. The resulting function from each simple linear regression was used to estimate the mean of the modified formant values for a set of 30 new speakers' recordings, consisting of 15 males and 15 females. The improvement in the average WER of the male recordings was 7%, and this is significant with $p < 0.03$ in the T-test. Modifying the 15 recordings of females led to an improvement in the average WER of 6%, with $p < 0.01$.

6.6: Summary and Discussion

In this chapter *DeepFormants* and *Praat* were used to extract the original and modified formants from the original and modified versions of 10 recordings. When comparing the results of extracting the formants through *DeepFormants* with *Praat*, the T-test showed no significant difference between the tools. The mean and median of the formants extracted by each tool for the

6 male speakers and for the 4 female speakers were each used as training data, and the functions of the straight lines which fitted these data sets were used to estimate 15 new male and 15 new female speaker modified formant values. Although there was no significant difference between the formants extracted by *DeepFormants* and *Praat* (p = 0.12 for both), improvement only resulted when they were modified based on the mean of the first three formants using *DeepFormants*.

Although the training set of the female speakers comprised 4 speakers and the training set of the male speakers comprised 6 speakers, and all were tested under the same conditions, the 15 female speakers in the test set improved significantly by 16% with $p < 0.01$, while the 15 male speakers in the test set improved by 6% which was not significant at $p = 0.05$. Therefore, a further experiment was implemented in section 6.4, in order to find a better training set for the male speakers.

The results of this experiment showed that one fold of the 10-fold cross validation was able to produce a function for the male recordings which gave a significant 7% improvement after transcribing these modified recordings through ASR, with $p < 0.05$.

Even though the improvement in YouTube’s ASR system with the functions is small compared to that in Dragon Server Edition 11, it is encouraging because these functions can be improved, for example by investigating the use of different training sets of recordings, or by improving the training data by enhancing the formant extraction by filtering in the spectral domain. While the formants here were calculated for each 0.1 interval that was chunked automatically using the *ffmpeg* tool, it may be possible to improve formant extraction with other methods for segmenting the sound files, for example the *wavelet* mathematical function, which is used to divide a continuous-time signal into different scale components by assigning a frequency range to each scale component.

Chapter 7: Conclusions and Future Work

The contribution of the research described in this thesis is to present a new approach to improving the accuracy of ASR transcription of lecture recordings through pre-processing the recordings by warping the formants and speech rate before transcribing these recordings using ASR. This modification of speakers' recordings is carried out through the use of sound processing libraries that adjust formants through adjusting pitch and adjusting speech rate. This research also contributes an efficient approach to selecting the adjustments.

7.1: Conclusions

This report has described the implementation of pre-processing speakers' recordings before transcribing those recordings by modifying the fundamental frequency and speech rate (SR) through the use of pitch shifting and time stretching libraries. Pre-processing speakers' recordings has been implemented via three approaches. Firstly, pre-processing recordings via an iterative 'trial and error' approach which iteratively modifies each speaker recording by different pitch and SR modification percentages and then corrects the transcript of each modified version of each recording in order to find the minimum WER. The pitch was modified by between –12% and +19% since modifying the pitch by less than –12% or more than 19% result in that the recordings were difficult or impossible to recognize and the WER increased by an average of 80%. Also, the SR of the recordings were modified iteratively by –10% and +50% in order to find the modification % that resulted in an improvement in the performance of ASR. The results of this approach have been analysed from pre-processing two male and two female recordings. Shifting the pitch of the speakers' recordings led to warping of the formants of the original speakers' recordings, resulting in an improvement in WER by an average of 20.4% and 25.8% for male and female speakers respectively. The modification of SR led to performing warping of the time series, resulting in an improvement in WER by an average of 10.1% and 17.1% for male and female respectively. Modifying both fundamental frequency and speech rate led to significant improvement in the ASR system WER. The accuracy of ASR improved by an average of 39.5% because increasing the pitch and decreasing the speech rate of the male speakers resulted in stretching the formants and compressing the time series. For female speakers the system improved by an average of 36.2% with decreasing both pitch and speech rate, because this resulted in compressing both the formants and time series.

To find the significant difference between the average WER of the original recordings and that of the modified recordings, 320 recordings and their modified versions were analysed statistically where the results showed with 95% confidence that the mean WER of all speakers improved by between 15% and 20% with p <0.01.

The second approach investigated the ability to eliminate the correction time taken in the ‘trial and error’ approach by picking the transcripts with the lowest WER, without correcting these transcripts. This investigation showed the ability to distinguish the transcripts that had the minimum word errors when their modified recordings gave an improvement of 11% or higher.

The third approach used, as a training set, both the original recordings and the best modified version of those recordings resulting from the ‘trial and error’ approach for only those speakers that were improved by an average of 15% in the ‘trial and error’ approach. The formants of the original and modified recordings for each speaker were extracted and then the mean and median of those formants were used as training sets, and the functions that were defined by these training sets were used to compute the new speaker modification mean and median formant values.

The results showed that the function estimated from the mean of the first three formants extracted by *DeepFormants* improved the WER significantly, by an average of 16% with p <0.01 for female speakers, and by an average of 7% for male speakers. The function estimated from the mean of the first three formants extracted by *Praat* improved the WER significantly, by an average of 17% with p <0.01 for female speakers, and by an average of 2% for male speakers.

Also, using the second and the third formants separately to estimate the modification functions resulted in an improvement in the WER. The functions estimated from the mean of the second formant extracted by *Praat* and *DeepFormants* significantly improved the WER by an average of 18% and 16% respectively with p <0.01 for female speakers. The functions estimated from the median of the second formant extracted by *DeepFormants* significantly improved the WER, by an average of 17% and 5% for female and male speakers respectively with p <0.01.

Moreover, the functions estimated from the mean of the third formant extracted by *DeepFormants* significantly improved the WER, by an average of 16% with p <0.01 for female speakers. Also, the functions estimated from the mean of the third formant extracted by *Praat* and *DeepFormants* improved the WER by an average of 5% and 3% respectively for male speakers.

Finally, it can be concluded that transcribing pre-processed recordings via ASR resulted in reducing the number of transcript errors.

7.2: Future work

All the experiments of this thesis have taken into account improving the performance of ASR. Optimal modification for each speaker's recording has been achieved via the iterative procedure for the trial and error approach, and this can be a computational cost or may be in terms of elapsed time. A sub optimal solution, using a simple linear regression approach, might be the best method for modifying some speaker recordings in terms of improving the performance of ASR for transcribing these recordings.

In this thesis, the emphasis throughout has been on limiting the time for finding the optimal modification for each speaker's recording via eliminating the iteration procedure. Nevertheless, it has been noted in Chapter 6 that an improvement can be obtained via using the knowledge and information that is provided by the iteration procedure for the other approaches that were used to modify the new speakers' recordings.

One of the recordings was of particularly bad quality, and neither Dragon Dictation system nor YouTube were able to transcribe this recording, though after modifying this recording the YouTube system gave a transcript with accuracy of 60%, which suggested investigating the improvements given by recording modification for such low quality recordings.

An experiment which studied the impact of modifying 4 male and 4 female recordings on different ASR systems, such as IBM, Siri and Dragon Dictation, showed an improvement on these systems' WER, which suggests future investigation for these systems in terms of finding a modification function for recordings for each system that results in better transcripts.

In the experiment that was implemented in Chapter 6 using *Praat* and *DeepFormants* tools, to extract the formants from original and modified recordings and use these formants as training data to find an optimal modification function, the function that was estimated by values extracted by *DeepFormants* gives a better improvement in WER compared to that estimated by *Praat*. Balusu & Adamantios (2010) pointed out that *Praat* formant tracker does make mistakes and it does not always successfully identify the formants. Because the formant values are extracted by *Praat* at each time step of a selection by first making a selection in the speech signal and then extracting the formants of that selection, a further investigation into more accurate methods for segmenting the sound files is suggested, for example the wavelet mathematical function which is used to divide a continuous-time signal into different scale components by assigning a frequency range to each scale component (Alías, et al., 2016), which might improve the modification function which is estimated by values extracted via *Praat*.

Acoustic features are statistical measurements of a speaker showcasing some specific formant distributions and this is the basis on which the recording modification function is determined. There are usually only a small number of dissimilar phonemes in a set of sentences delivered by a speaker. In such cases, where there is less variability in the speech, the distribution of formants will be skewed. But, if we take into consideration only one particular set of phonemes, there will be distortions even if the distributions are skewed. If we compute statistics of these distorted distributions, then the modification function will give a better estimation for the mean modified formants for the new speaker's compared to if we do not account for these distortions and other effects.

Moreover, the investigation of using a Bayesian approach to the linear regression estimates can be performed as an extension to this thesis. In Chapter 6, the results of the experiment that modified the new speaker recordings based on the gender specific modifying functions showed significant improvement and advancement as compared to transcribing their original recordings. Therefore these functions can be implemented if no other speaker specific information is available as it gives improvement. However, from the practical point of view the Bayesian approach may be worthy of investigation as it would take advantage of the modifying functions based on speaker gender.

Appendix A

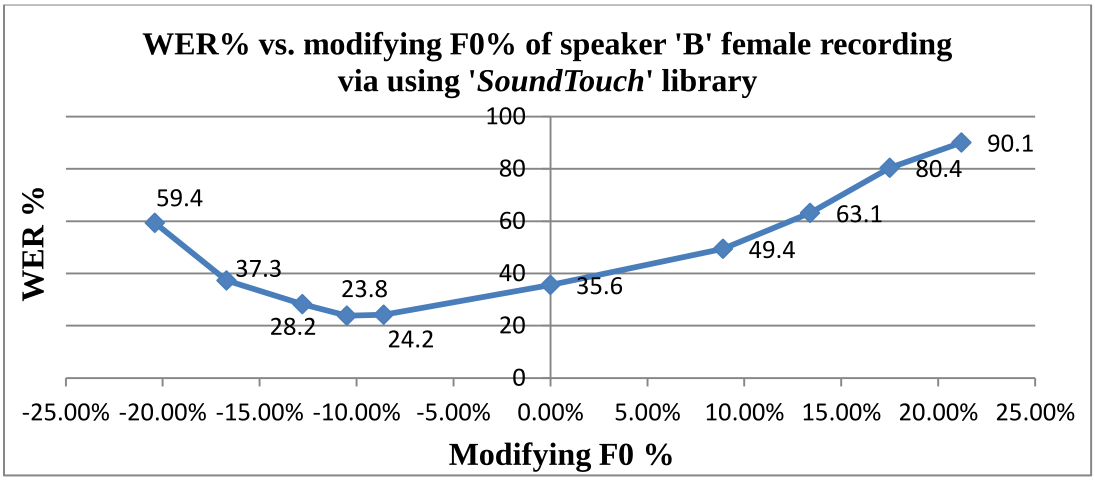
The further results of the experiment that was implemented and discussed in details in chapter 4, the other three different original recordings for three different speakers which were pre-processed by increasing or decreasing either pitch or speech rate, or both pitch and speech rate.

A.1 Modifying the fundamental frequency of a recording of female speaker 'B' by using '*SoundTouch*' library

Changing of F_0 percentage	Pitch (AV freq Hz)	WER	AR
–20.40%	182.4	59.4	40.6
–16.70%	190.9	37.3	62.7
–12.80%	199.9	28.2	71.8
–10.50%	205.3	23.8	76.2
–8.60%	209.6	24.2	75.5
0.00%	229.2	35.6	64.4
8.90%	249.7	49.4	50.6
13.40%	260	63.1	36.9
17.50%	269.3	80.4	19.6
21.20%	277.7	90.1	9.9

Table_Apx A-1 Shows the summary results for WER, resulting from modifying pitch of female 'B' speech by increasing and decreasing the fundamental frequencies of the speech, using '*SoundTouch*' library

Table_Apx A-1 shows WER and AR of generating transcripts via a Dragon ASR system for each pre-processed recording of speaker B. Each pre-processed recording of 'B' was altered by modifying the fundamental frequencies of his speech. The changes in the original pitch percentage were between –20.4% and 21.2%. The modification percentages result in an average fundamental frequency between 182.4.1Hz and 277.7 Hz. The results of WER of the pre-processed recording's transcripts were found to be between 90.1% (Worst value) and 23.8% (best values) and therefore the AR between 9.9%, and 76.2% respectively.



Figure_Apx A-1WER percentages vs. modifying the F0 of the original female speaker 'B' via 'SoundTouch' library

Figure_Apx A-1 plots the results which are summarised below:

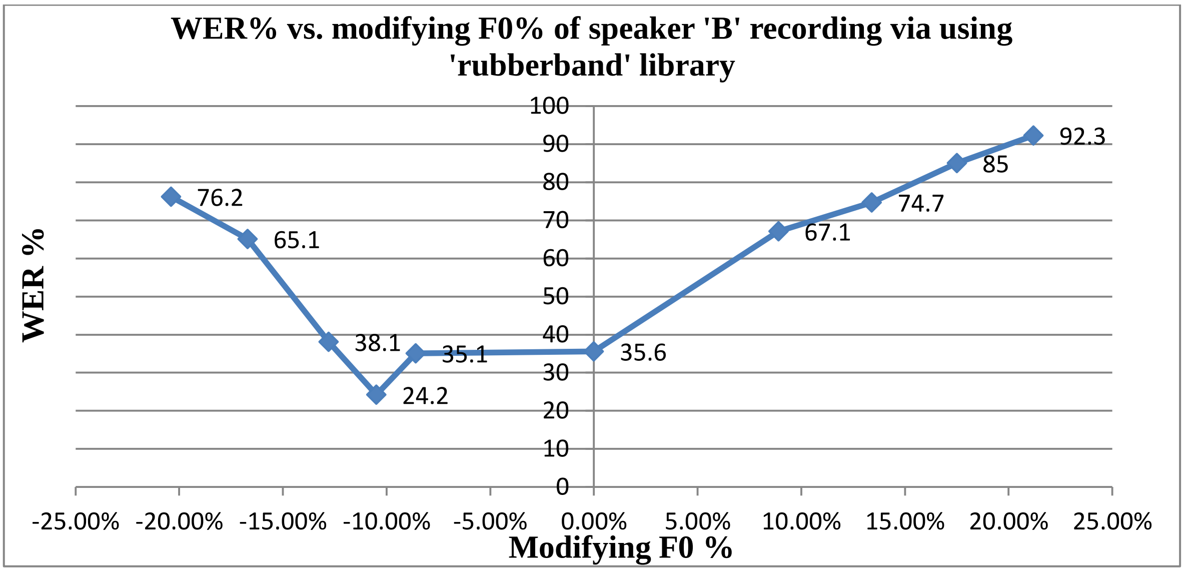
- Modifying F0% by –8.6% reduced WER to a minimum of 23.8%, which was a reduction of 33.5% from the original 35.6% WER.
- Modifying F0% further increased the WER.
- A positive change in F0 % increased WER.

A.2 Modifying fundamental frequency of female speaker 'B' recording by using 'Rubber Band' library

Changing F_0 percentage	Pitch (AV freq Hz)	WER	AR
–18.1%	187.8	76.2	23.8
–14.70%	195.4	65.1	34.9
–10.80%	204.5	38.1	61.9
–8.6%	209.4	24.2	75.5
–6.60%	214	35.1	64.9
0.00%	229.2	35.6	64.4
11.2%	254.8	67.1	32.9
16.40%	266.8	74.7	25.3
20.9%	277	85	15
25.1%	286.8	92.3	7.7

Table_Apx A-2 Shows the summary results for WER, resulting from modifying pitch of female 'B' speech by increasing and decreasing F0, via '*Rubber Band* ' library

Table_Apx A-2 shows that modifiying F_0 of the original speaker's recordings by between –18.1% and 25.1% resulted in different pre-processed recordings of different average frequency values between 187.8 Hz and 286.8 Hz. Transcribing the pre-processed recodings by the Dragon speech recognition system produced different WER values between 92.3 %(worst value) and 24.2% (best value) and therefore corresponding accuracy rate values between 7.7% and 75.5%.



Figure_Apx A-2 WER percentages vs. modifying F0 of the original speaker 'B' recording via '*Rubber Band* ' library

Figure_Apx A-2 plots the results which are summarised below:

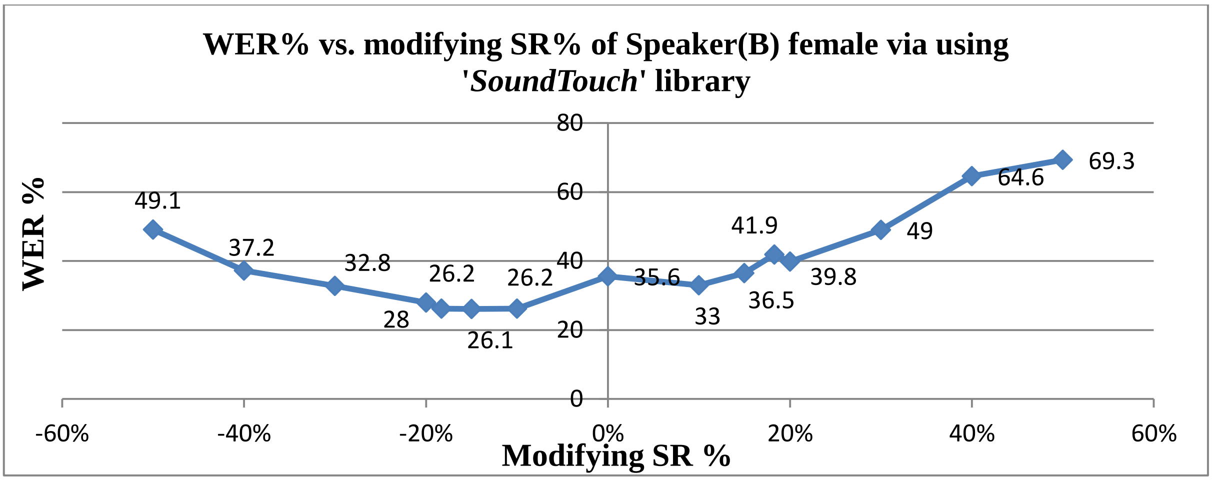
- decreasing the F_0 of B's speech by 10.5% leads to decreasing the WER to 24.2% (best value), an improvement of 32%
- The WER increased with further decreasing of F_0

A.3 Modifying speech rate of a recording of female speaker 'B' by using '*SoundTouch*'

Speed rate (percentage)	Pitch (AV freq Hz)	WER	AR	Length (min)
–50	229	49.1	50.9	09:59
–40	229.6	37.2	62.8	08:19
–30	228.5	32.8	67.2	07:08
–20	226.7	28	72	06:14
–18.3	225.9	26.2	73.8	06:07
–15	226.4	26.1	73.9	05:52
–10	227.4	26.2	73.8	05:33
0	229.2	35.6	64.4	04:59
10	229.3	33	67	04:32
15	229.8	36.5	63.5	04:20
18.3	230	41.9	58.1	04:13
20	229.9	39.8	60.2	04:09
30	230	49	51	03:50
40	229.9	64.6	35.4	03:34
50	230.6	69.3	30.7	03:19

Table_Apx A-3 Shows the summary results for WER, resulting from modifying SR of female 'B' speech by increasing and decreasing the original SR, via using '*SoundTouch*' library

Table_Apx A-3 shows the effects of modifying the speech rate of the audio file of speaker 'B' by between –50% and 50%. The changes on the audio file were made only on the speech rate while trying to keep the F0 stable. However, it was noticed that modifying SR was still causing between –4% and 1% changes on F0%. The WER was showing changes between the worst value 69.3%, with increasing the SR by 50%, and the best value 26.1%, with decreasing the SR by 15%. Also, decreasing the SR by 10% caused an improvement on WER by 26.2%.



Figure_Apx A-3 Shows WER % vs. modifying SR% of female speaker 'B' by using '*SoundTouch*' library

Figure_Apx A-3 plots the results which are summarised below:

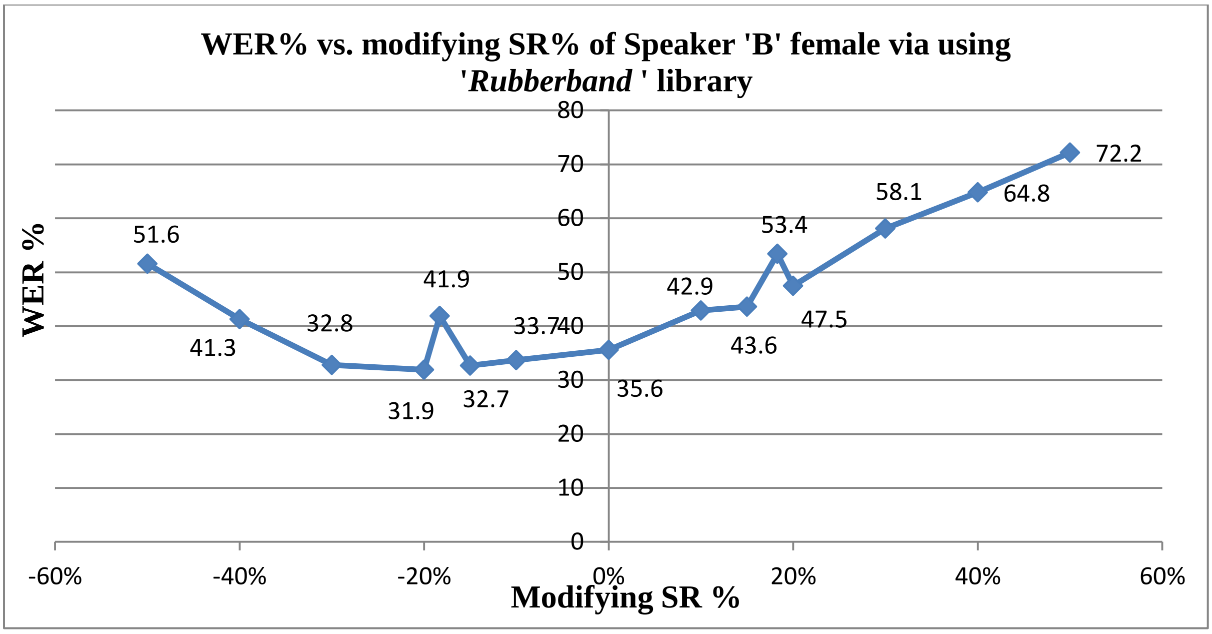
- Changing the SR by – 10% could improve the WER from 35.6% to 26.2% which is an 32% improvement
- On the other hand, the figure shows increasing the SR by 10% decreased the WER to 33%.

A.4 Modifying speech rate of a recording of female speaker 'B' by using 'Rubber Band'

Speech rate (percentage)	Pitch (AV freq Hz)	WER	AR	Length (min)
–50	231.2	51.6	48.4	09:59
–40	232.5	41.3	58.7	08:19
–30	232.1	32.8	67.2	07:08
–20	232.9	31.9	68.1	06:14
–18.3	233.1	41.9	58.1	06:07
–15	232	32.7	67.3	05:52
–10	233.1	33.7	66.3	05:33
0	229.2	35.6	64.4	04:59
10	233.1	42.9	57.1	04:32
15	232.9	43.6	56.4	04:20
18.3	233.4	53.4	46.6	04:13
20	234.2	47.5	52.5	04:09
30	234	58.1	41.9	03:50
40	233.7	64.8	35.2	03:34
50	234.3	72.2	27.8	03:19

Table_Apx A-4 Shows the summary results for WER, resulting from modifying SR of female 'B' speech by increasing and decreasing the original SR, via using '*Rubber Band* ' library

Table_Apx A-4 shows that the WER was showing changes between the worst value 72.2%, with increasing the SR by 50%, and the best value 31.9%, with decreasing the SR by 20%.



Figure_Apx A-4 Shows WER % vs. modifying SR% of female speaker ‘B’ by using ‘*Rubber Band* ’ library

Figure_Apx A-4 plots the results which are summarised below:

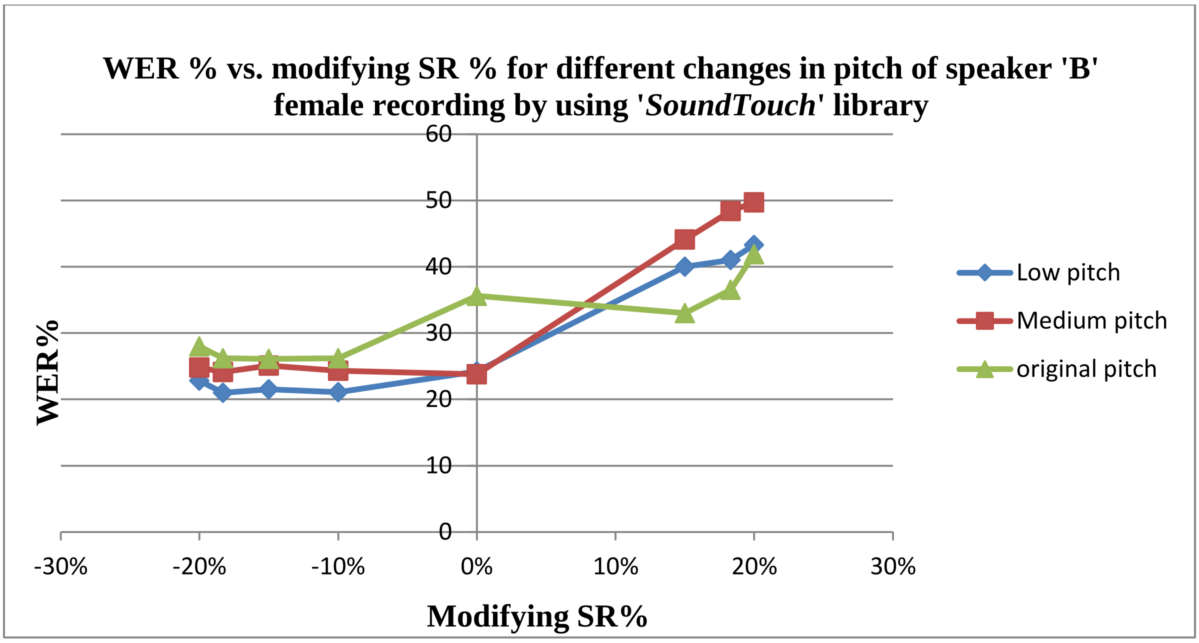
- Changing the SR by – 10% could improve the WER from 35.6% to 33.7% which is an 5% improvement even though that decreased SR % caused increasing on f0% by 1.7%.
- The results on Figure_Apx A-1 and Figure_Apx A-2 showed that any increasing of the female fundamental frequency leads to increase the WER %.
- However, the result in Figure_Apx A-4 shows decreasing SR% could decrease the WER even if F0% increased.
- On the other hand, the figure shows increasing the SR by 10% increased the WER to 42.9%.

A.5 Modifying fundamental frequency and speech rate of female speaker ‘B’ by using ‘*SoundTouch*’ library

Speech rate (percentage)	Changing F_0 percentage	Pitch (AV freq Hz)	WER	AR
–20	–8.6%	207.7	22.8	77.2
–20	–12.8%	199.1	24.8	75.2
–18.3	–8.6%	208.1	21	79
–18.3	–12.8%	199.1	24.1	75.9
–15	–8.6%	208.8	21.5	78.5
–15	–12.8%	199.2	25.1	74.9
–10	–8.6%	209.2	21.1	78.9
–10	–12.8%	199.7	24.3	75.7
0	0	229.2	35.6	64.4
15	–8.6%	210.1	40	60
15	–12.8%	200.9	44.1	55.9
18.3	–8.6%	209.6	41	59
18.3	–12.8%	200.4	48.4	51.6
20	–8.6%	210.4	43.3	56.7
20	–12.8%	200.3	49.7	50.3

Table_Apx A-5 Shows the summary results for WER, resulting modifying SR for different decreases in F0 of female ‘B’ speech by using ‘*SoundTouch*’ library

Table_Apx A-5 shows the WER and AR of speech recognition resulting from modifying both fundamental frequencies and speech rate of the recording of speaker ‘B’. The changes in the original pitch were between –8.6% and –12.8%. The modification percentages have resulted in an average F_0 of between 199.1 and 229.2 Hz. The results of WER of pre-processed audio files were found to be between 21% (best) and 48.4% (worst) and therefore the AR was between 79% and 51.6%, respectively.



Figure_Apx A-5 Plots WER percentages vs. the modifications of SR percentages for different changes in pitch percentages of speaker 'B' female via '*SoundTouch*' library

Figure_Apx A-5 plots the results which are summarised below:

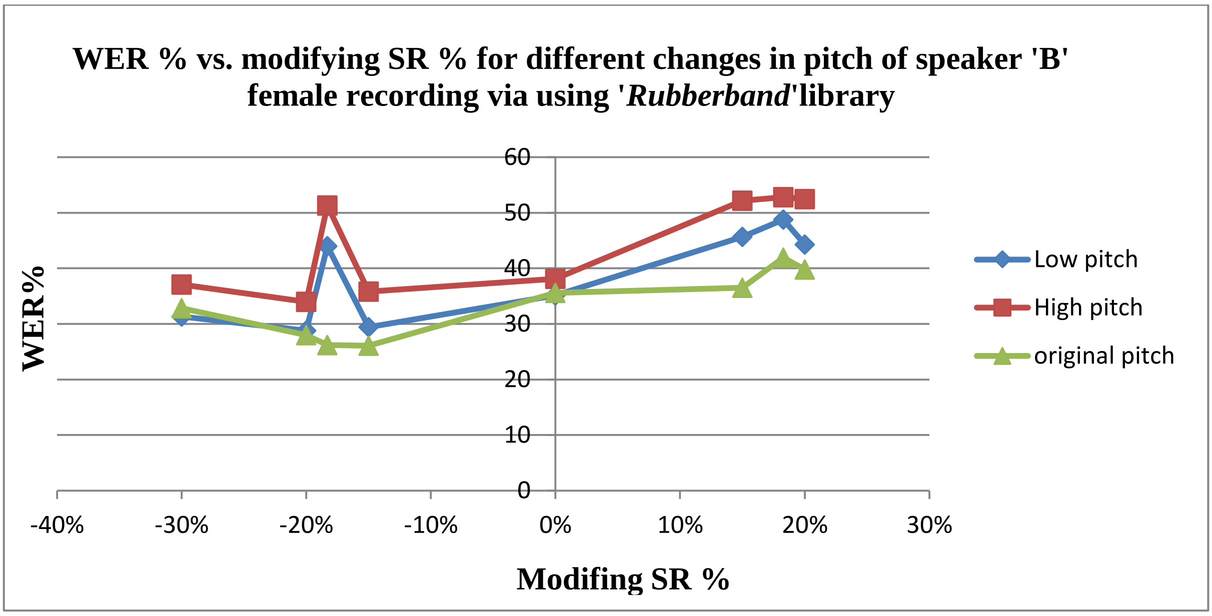
- Modifying SR by –18.3% and decreasing the fundamental frequency by 8.6% improved the WER to 21(best value), a change of 41%
- The WER was increasing and decreasing significantly by between 21% and 49.7% with changing speech rate by between –20% and 20% and decreasing the pitch by between 8.6% and 12.8 %.

A.6 Modifying fundamental frequency and speech rate of female speaker 'B' by using 'Rubber Band' library

Speed rate (percentage)	Changing F_0 percentage	Pitch (AV freq Hz)	WER	AR
–30	–6.6%	213.5	31.3	68.7
–30	–10.8%	204.6	37.1	62.9
–20	–6.6%	213.2	28.8	71.2
–20	–10.8%	203.1	34	66
–18.3	–6.6%	213.3	44	56
–18.3	–10.8%	202.4	51.3	48.7
–15	–6.6%	211.9	29.4	70.6
–15	–10.8%	204.1	35.8	64.2
0	0	229.2	35.6	64.4
15	–6.6%	215.6	45.7	54.3
15	–10.8%	205.2	52.2	47.8
18.3	–6.6%	214.6	48.8	51.2
18.3	–10.8%	205.5	52.8	47.2
20	–6.6%	214.7	44.3	55.7
20	–10.8%	206.2	52.5	47.5

Table_Apx A-6 Shows the summary results for WER, resulting modifying SR for different decreases in F0 of female 'B' speech by using '*Rubber Band* ' library

Table_Apx A-6 shows the WER and AR of speech recognition resulting from modifying both fundamental frequencies and speech rate of the recording of speaker ‘B’. The changes in the original pitch were between –6.6% and –10.8%. The modification percentages have resulted in an average F_0 of between 203.1 Hz and 215.6Hz. The results of WER of pre-processed audio files were found to be between 29.4% (best) and 44.3% (worst) and therefore the AR was between 70.6% and 55.7%, respectively.

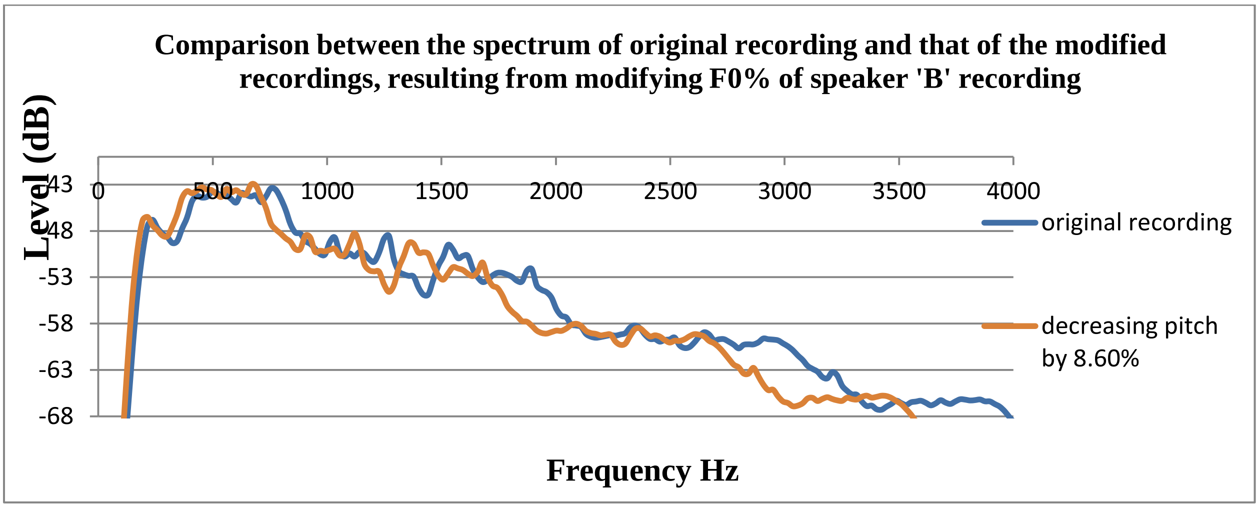


Figure_Apx A-6 Plots WER percentages vs. the modifications of SR percentages for different changes in pitch percentages of speaker ‘B’ female via ‘*Rubber Band* ’ library

Figure_Apx A-6 plots the results which are summarised below:

- Modifying the SR by –20% and decreased the fundament frequency by 6.6% improved the WER by 19.1% (best improvement).
- The resulting effects on WER were sensitive to modification of fundamental frequency and speech rate; the figure shows that modifying the SR by –15% and decreasing the fundamental frequency by between 6.6 and 10.8 caused an improvement on the WER of 26.6%, while keeping the same values of the fundamental frequency and decreasing the SR by 18% leads to increasing the WER again by between 23.6% and 44.1%.
- Both increasing and decreasing SR caused decreasing f0% of between 5% and 12%.

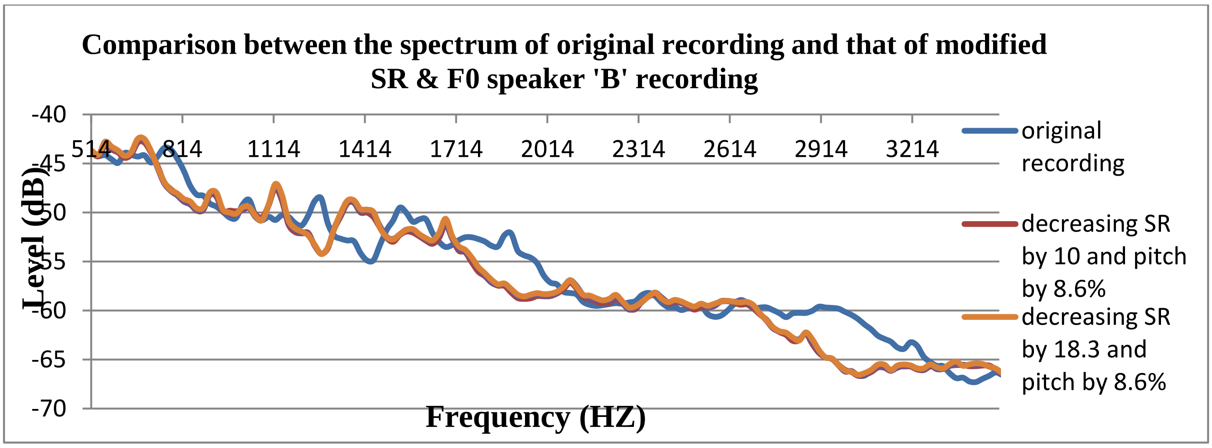
A.7 Spectral frequency warping, resulting from modifying fundamental frequency of speaker 'B' recording



Figure_Apx A-7 Comparison between the spectrums of speaker 'B' recording and that of the modified pitch recording. The formants and peak positions offset result from the modified F0%

Figure_Apx A-7 shows the spectral frequency warping caused by decreasing the fundamental frequency of speaker 'B' recording by 8.6 respectively, which gives the best improvement in the WER. It can be seen that this shifts the formant resonance peaks to the left (i.e. decreases their frequencies).

A.8 Spectral frequency warping resulting from modifying SR and fundamental frequency % of speaker 'B' recording



Figure_Apx A-8 Comparison between the spectrums of original 'B' recording and that of modified speech rate and fundamental frequency recording

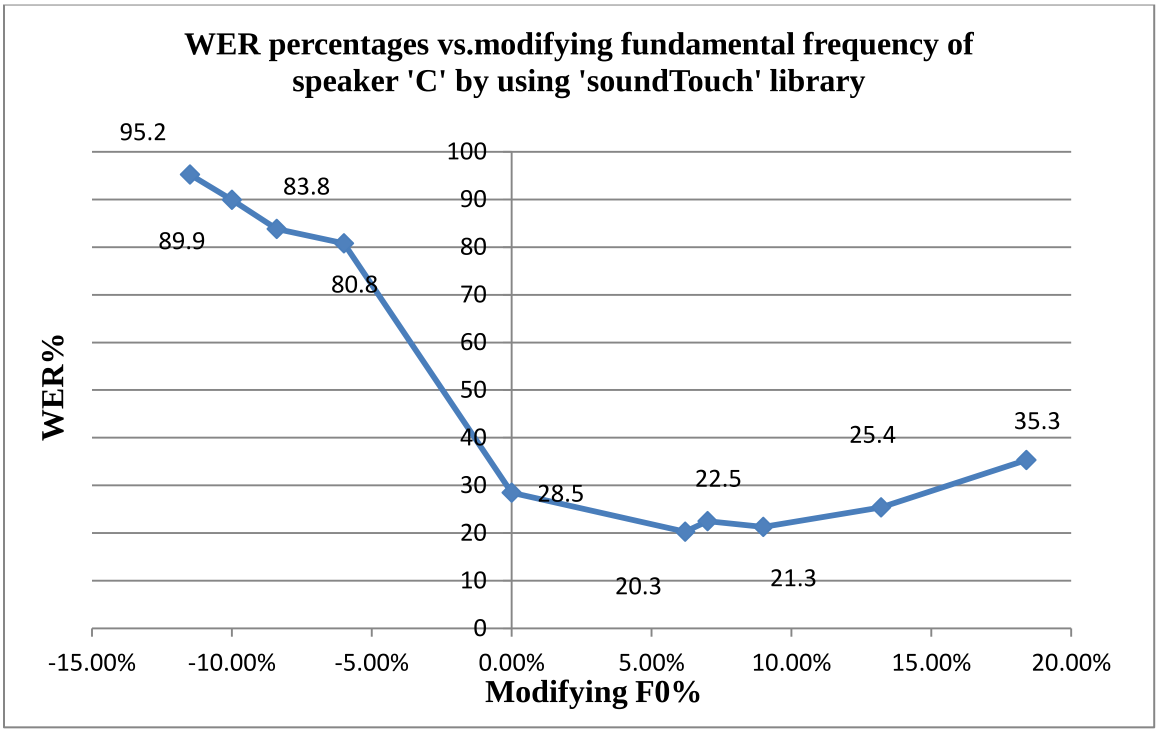
Figure_Apx A-8 shows how modifying the recording of female speaker 'B' by increasing the pitch by 8.6% and decreasing SR by either 10 % or 18.3% resulted in compressions of the 'frequency spectrum (i.e. shifts formant resonance peaks to the left) to give the best WER.

A.9 Modifying the fundamental frequency of a recording of male speaker ‘C’ by using ‘*SoundTouch*’ library

Changing F_0 percentage	Pitch (AV freq Hz)	WER	AR
-14.4%	131	95.2	4.8
–12.2%	134.4	89.9	10.1
–10.6%	137.1	83.8	16.2
–7.8%	141.2	80.8	19.2
0%	153.1	28.5	71.5
8.8%	166.5	20.3	79.7
9.1%	167	22.5	77.5
12.8%	172.7	21.3	78.7
17.1%	179.3	25.4	74.6
21.6%	186.1	35.3	64.7

Table_Apx A-7 Shows the summary results for WER, resulting from modifying pitch of male ‘C’ speech by increasing and decreasing the fundamental frequencies of the speech, using ‘*SoundTouch*' library

Table_Apx A-7 shows WER and AR of generating transcripts via a Dragon ASR system for each pre-processed recording of speaker C. Each pre-processed recording of ‘B’ was altered by modifying the fundamental frequencies of his speech. The changes in the original pitch percentage were between –14.4% and 21.6%. The modification percentages result in an average fundamental frequency between 131 Hz and 186.1 Hz. The results of WER of the pre-processed recording’s transcripts were found to be between 35.3% (Worst value) and 20.3% (best values) and therefore the AR between 64.7%, and 79.7% respectively.



Figure_Apx A-9 WER percentages vs. modifying the F0 of the original male speaker 'C' via '*SoundTouch*' library

Figure_Apx A-9 plots the results which are summarised below:

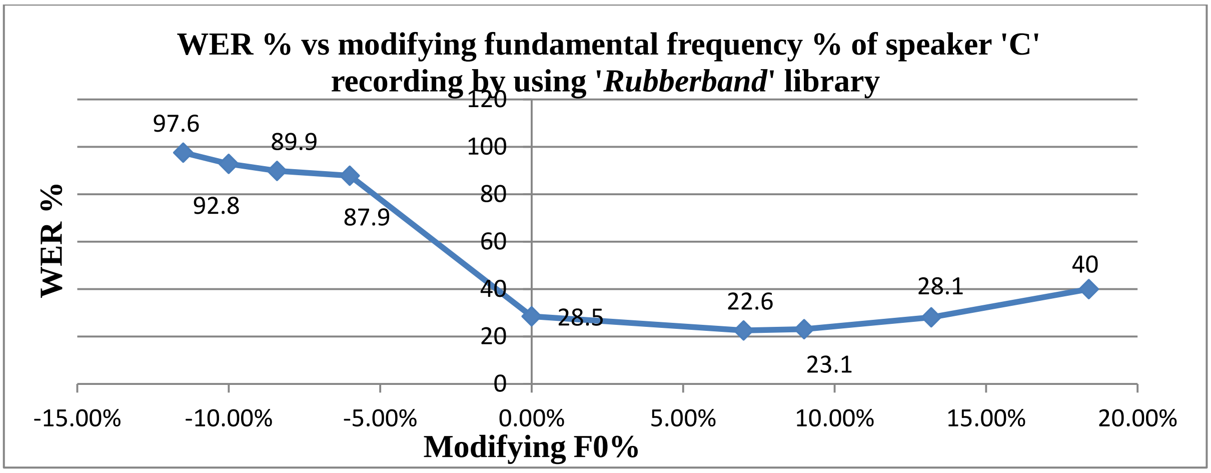
- Modifying F0% by 8.8% reduced WER to a minimum of 20.3%, which was a reduction of 28.8% from the original 28.5% WER.
- A negative change in F0 % increased WER.

A.10 Modifying fundamental frequency of male speaker 'C' recording by using 'Rubber Band' library

Changing F_0 percentage	Pitch (AV freq Hz)	WER	AR
–13.7%	132.1	97.6	2.4
–11.8%	135.1	92.8	7.2
–9%	139.3	89.9	10.1
–6.1%	143.7	87.9	12.1
0%	153.1	28.5	71.5
12.8%	172.7	22.6	77.4
18.9%	182.1	23.1	76.9
25.3%	191.9	28.1	71.9
31.9%	201.9	40	60

Table_Apx A-8 Shows the summary results for WER, resulting from modifying pitch of male 'C' speech by increasing and decreasing F0, via 'Rubber Band ' library

Table_Apx A-8 shows that modifiying F_0 of the original speaker's recordings by between 13.7% and 31.9% resulted in different pre-processed recordings of different average frequency values between 132.1Hz and 201.9Hz. Transcribing the pre-processed recodings by the Dragon speech recognition system produced different WER values between 97.6 %(worst value) and 22.6% (best value) and therefore corresponding accuracy rate values between 2.4% and 77.4%.



Figure_Apx A-10 WER percentages vs. modifying F0 of the original speaker ‘C’ recording via ‘*Rubber Band* ’ library

Figure_Apx A-10 plots the results which are summarised below:

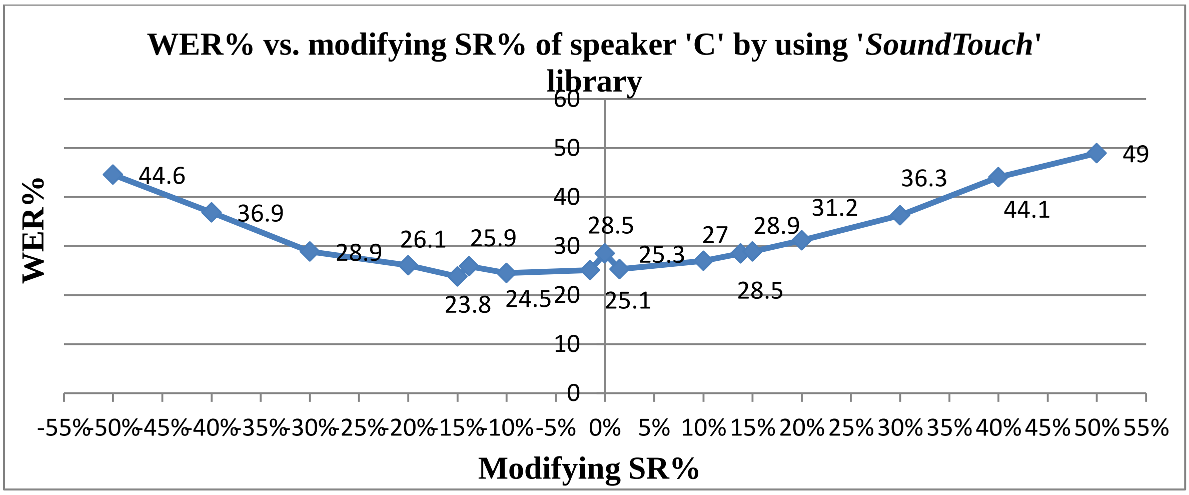
- Increasing the F_0 of C’s speech by 7% leads to decreasing the WER to 22.6% (best value), an improvement of 20.7%
- The WER increased with further increasing of F_0

A.11 Modifying speech rate of a recording of male speaker ‘C’ by using ‘*SoundTouch*’

Speed rate (percentage)	Pitch (AV freq Hz)	WER	AR	Length (min)
–50	154.1	44.6	55.4	23:39
–40	153.7	36.9	63.1	19:42
–30	151.5	28.9	71.1	16:53
–20	148	26.1	73.9	14:46
–15	148.3	23.8	76.2	13:54
–13.8	148.2	25.9	74.1	13:43
–10	150	24.5	75.5	13:08
–1.5	152.6	25.1	74.9	12:00
0	153.1	28.5	71.5	11:49
1.5	153.1	25.3	74.7	11:39
10	154.2	27	73	10:45
13.8	154.2	28.5	71.5	10:23
15	154.2	28.9	71.1	10:16
20	154.7	31.2	68.8	09:51
30	155.4	36.3	63.7	09:05
40	155.8	44.1	55.9	08:26
50	156.3	49	51	07:53

Table_Apx A-9 Shows the summary results for WER, resulting from modifying SR of male ‘C’ speech by increasing and decreasing the original SR, via using ‘*SoundTouch*’ library

Table_Apx A-9 shows the effects of modifying the speech rate of the audio file of speaker ‘B’ by between –50% and 50%. The changes on the audio file were made only on the speech rate while trying to keep the F0 stable. However, it was noticed that modifying SR was still causing between –3.3% and 2% changes on F0%. The WER was showing changes between the worst value 49%, with increasing the SR by 50%, and the best value 23.8%, with decreasing the SR by 15%. Decreasing the SR by 15% caused an improvement on WER by 16.5%.



Figure_Apx A-11 Shows WER % vs. modifying SR% of male speaker ‘C’ by using ‘*SoundTouch*’ library

Figure_Apx A-11 plots the results which are summarised below:

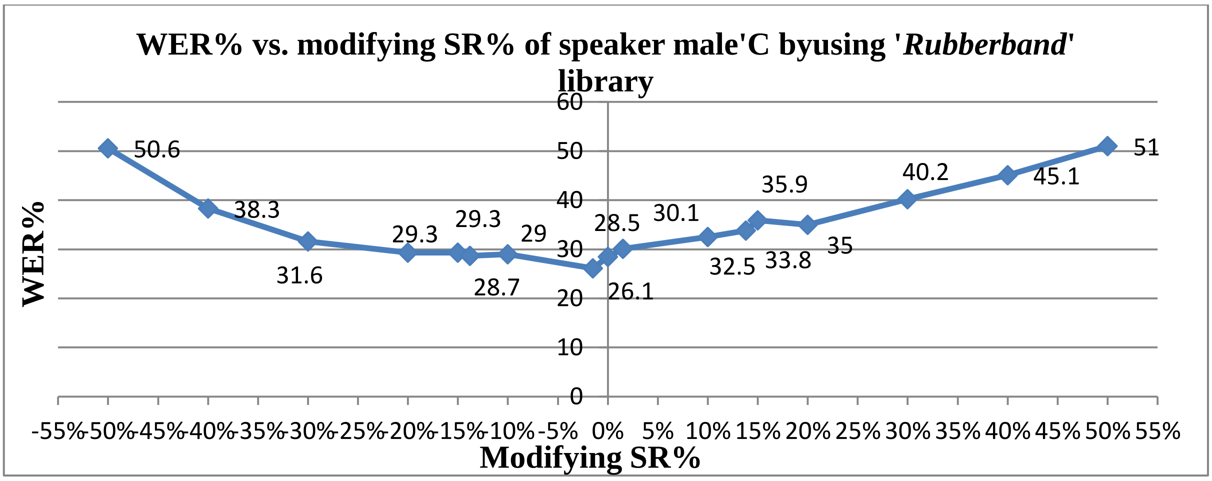
- Changing the SR by – 1.5% could improve the WER from 28.5% to 25% which is an 12% improvement. Also, increased the SR by 1.5 led to improve the WER by 11.9%.
- On the other hand, the figure shows decreasing the SR by 15% decreased the WER to 23.8%.

A.12 Modifying speech rate of a recording of male speaker ‘C’ by using ‘Rubber Band ’

Speed rate (percentage)	Pitch (AV freq Hz)	WER	AR	Length (min)
–50	152.1	50.6	49.4	23:39
–40	156.3	38.3	61.7	19:42
–30	156.2	31.6	68.4	16:53
–20	156	29.3	70.7	14:46
–15	155.3	29.3	70.7	13:54
–13.8	155.4	28.7	71.3	13:43
–10	155.5	29	71	13:08
–1.5	153.5	26.1	73.9	12:00
0	153.1	28.5	71.5	11:49
1.5	152.9	30.1	69.9	11:39
10	157	32.5	67.5	10:45
13.8	157.8	33.8	66.2	10:23
15	158	35.9	64.1	10:16
20	158.6	35	65	09:51
30	160.1	40.2	59.8	09:05
40	161.6	45.1	54.9	08:26
50	162.9	51	49	07:53

Table_Apx A-10 Shows the summary results for WER, resulting from modifying SR of male ‘C’ speech by increasing and decreasing the original SR, via using ‘*Rubber Band* ’ library

Table_Apx A-10 shows that the WER was changing between the worst value 51%, with increasing the SR by 50%, and the best value 26%, with decreasing the SR by 1.5%.



Figure_Apx A-12 Shows WER % vs. modifying SR% of male speaker ‘C’ by using ‘*Rubber Band* ’ library

Figure_Apx A-12 plots the results which are summarised below:

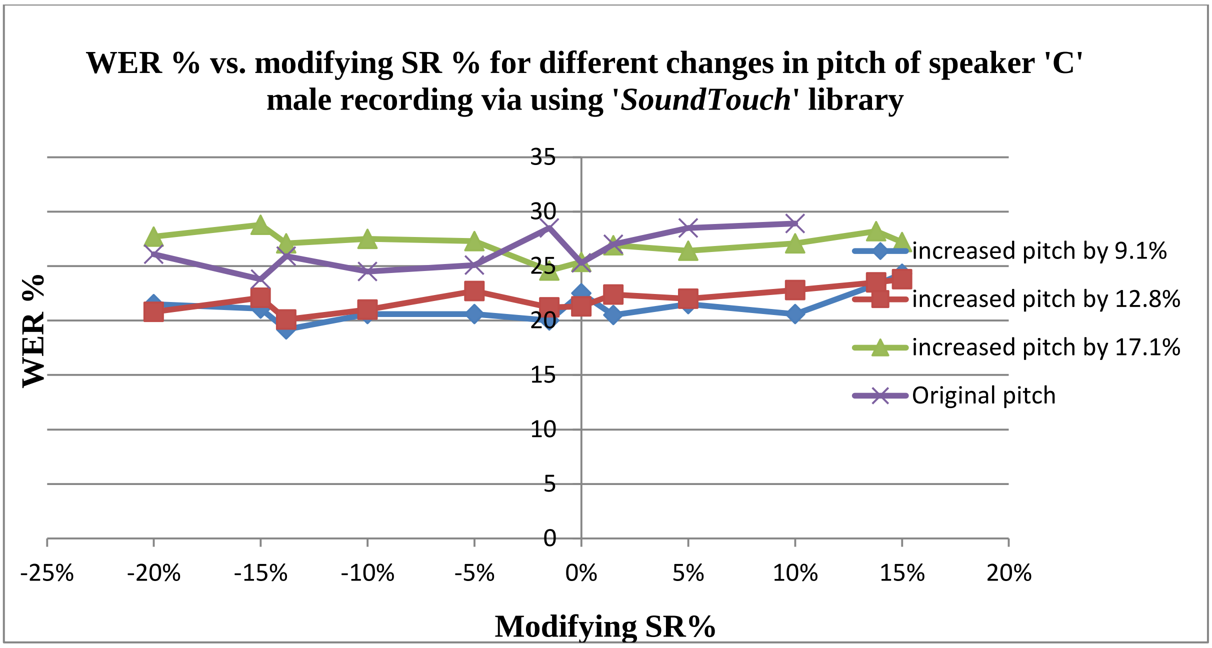
- decreasing the SR of C’s speech by 1.5% leads to decreasing the WER to 26% (best value).
- The WER increased with further increasing and decreasing of SR

A.13 Modifying fundamental frequency and speech rate of male speaker ‘C’ by using ‘*SoundTouch*’ library

Speech rate (percentage)	Changing F_0 percentage	Pitch (AV freq Hz)	WER	AR
–20	9.1%	165.8	21.5	78.5
–20	12.8%	175.6	20.8	79.2
–20	17.1%	186.3	27.7	72.3
–15	9.1%	162.9	21.1	78.9
–15	12.8%	172.6	22.1	77.9
–15	17.1%	183.4	28.8	71.2
–13.8	9.1%	163.1	19.2	80.8
–13.8	12.8%	172.1	20.1	79.9
–13.8	17.1%	183	27.1	72.9
–10	9.1%	162.7	20.6	79.4
–10	12.8%	170.4	21	79
–10	17.1%	180.1	27.5	72.5
–5	9.1%	164.3	20.6	79.4
–5	12.8%	171	22.7	77.3
–5	17.1%	178.1	27.3	72.7
–1.5	9.1%	165.9	20	80
–1.5	12.8%	172.1	21.2	78.8
–1.5	17.1%	178.5	24.6	75.4
0	0	153.1	28.5	71.5
1.5	9.1%	166.6	20.5	79.5
1.5	12.8%	173.6	22.4	77.6
1.5	17.1%	179.8	26.9	73.1
5	9.1%	168.1	21.5	78.5
5	12.8%	175	22	78
5	17.1%	181.1	26.4	73.6
10	9.1%	169.3	20.6	79.4
10	12.8%	176.2	22.8	77.2
10	17.1%	183.5	27.1	72.9
13.8	9.1%	169.8	23.3	76.7
13.8	12.8%	177.1	23.5	76.5
13.8	17.1%	184.5	28.2	71.8
15	9.1%	170.2	24.3	75.7
15	12.8%	177.9	23.8	76.2
15	17.1%	185.3	27.2	72.8

Table_Apx A-11 Shows the summary results for WER, resulting modifying SR for different increases in F0 of male ‘C’ speech

Table_Apx A-11 shows the WER and AR of speech recognition resulting from modifying both fundamental frequencies and speech rate of the recording of speaker ‘C’. The changes in the original pitch were between 6% and 9%. The modification percentages have resulted in an average F_0 of between 123.6 and 139.1 Hz. The results of WER of pre-processed audio files were found to be between 20% (best) and 28.2% (worst) and therefore the AR was between 80% and 71.8%, respectively.



Figure_Apx A-13 Plots WER percentages vs. the modifications of SR percentages for different changes in pitch percentages of speaker 'C' male via '*SoundToutch*' library

Figure_Apx A-13 plots the results which are summarised below:

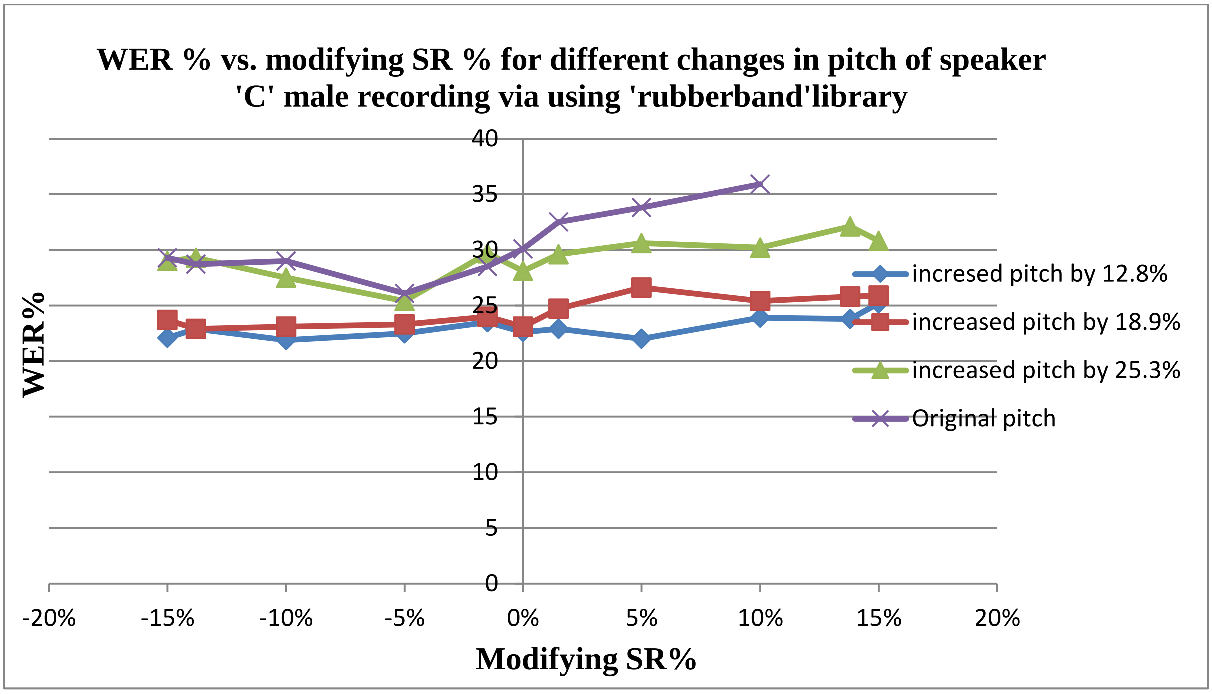
- Modifying SR by –1.5% and increasing the fundamental frequency by 9.1% improved the WER to 20(best value), a change of 29.8%
- The WER was increasing and decreasing significantly by between 24.6% and 4.6% with changing speech rate by between –20% and 15% and increasing the pitch by between 9.1% and 17.1 %.

A.14 Modifying fundamental frequency and speech rate of male speaker ‘C’ by using ‘Rubber Band’ library

Speech rate (percentage)	Changing F_0 percentage	Pitch (AV freq Hz)	WER	AR
-15	12.8%	172.6	22.1	77.9
–15	18.9%	182.5	23.7	76.3
–15	25.3%	192.4	29	71
–13.8	12.8%	172.7	22.9	77.1
–13.8	18.9%	182.4	22.9	77.1
–13.8	25.3%	192.3	29.3	70.7
–10	12.8%	172.8	21.9	78.1
–10	18.9%	181.9	23.1	76.9
–10	25.3%	192.4	27.5	72.5
–5	12.8%	173	22.5	77.5
–5	18.9%	182.6	23.3	76.7
–5	25.3%	191.3	25.4	74.6
–1.5	12.8%	172.8	23.5	76.5
–1.5	18.9%	182.7	24	76
–1.5	25.3%	192.4	29.7	70.3
0	0	153.1	28.5	71.5
1.5	12.8%	172.2	22.9	77.1
1.5	18.9%	182.5	24.7	75.3
1.5	25.3%	191.9	29.6	70.4
5	12.8%	172.1	22	78
5	18.9%	182.1	26.6	73.4
5	25.3%	191.6	30.6	69.4
10	12.8%	171	23.9	76.1
10	18.9%	181.5	25.4	74.6
10	25.3%	191.5	30.2	69.8
13.8	12.8%	167.9	23.8	76.2
13.8	18.9%	180.3	25.8	74.2
13.8	25.3%	191.2	32.1	67.9
15	12.8%	171.1	25.2	74.8
15	18.9%	179.8	25.9	74.1
15	25.3%	190.5	30.8	69.2

Table_Apx A-12 Shows the summary results for WER, resulting modifying SR for different increases in F0 of male ‘C’ speech by using ‘*Rubber Band* ’ library

Table_Apx A-12 shows the WER and AR of speech recognition resulting from modifying both fundamental frequencies and speech rate of the recording of speaker ‘C’. The changes in the original pitch were between 12.8% and 18.9%. The modification percentages have resulted in an average F_0 of between 172 Hz and 192Hz. The results of WER of pre-processed audio files were found to be between 21.9% (best) and 32.1% (worst) and therefore the AR was between 78.1% and 67.9%, respectively.

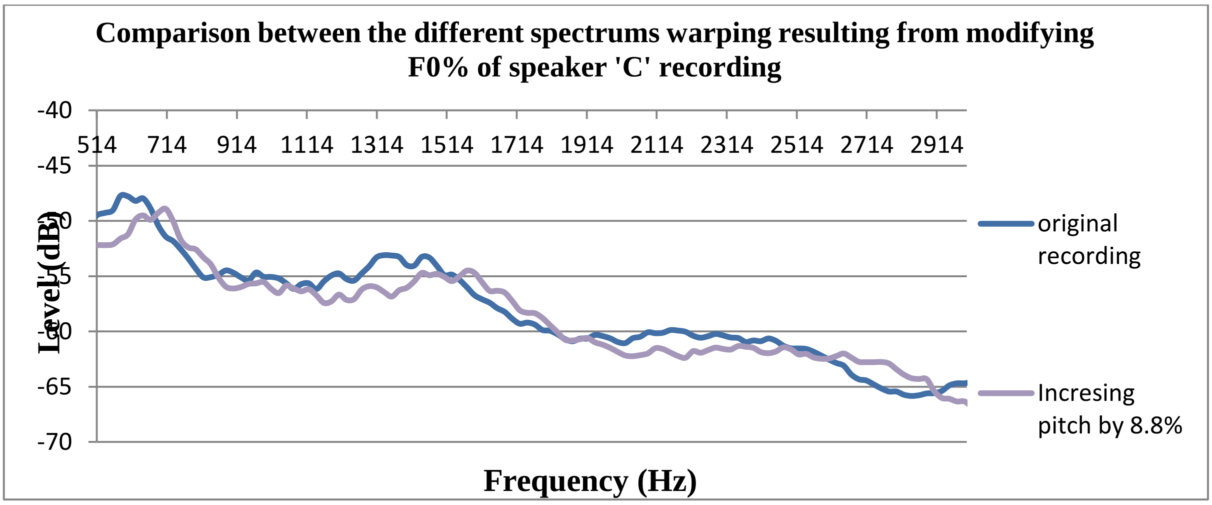


Figure_Apx A-14 Plots WER percentages vs. the modifications of SR percentages for different changes in pitch percentages of speaker 'C' male via '*Rubber Band* ' library

Figure_Apx A-14 plots the results which are summarised below:

- Modifying the SR by –10% and decreased the fundament frequency by 12.8% improved the WER to 21.9% (best value), a change of 23.2%.
- Increasing the SR caused a slightly more increasing on the f0% of between 9.7% and 25.7%.

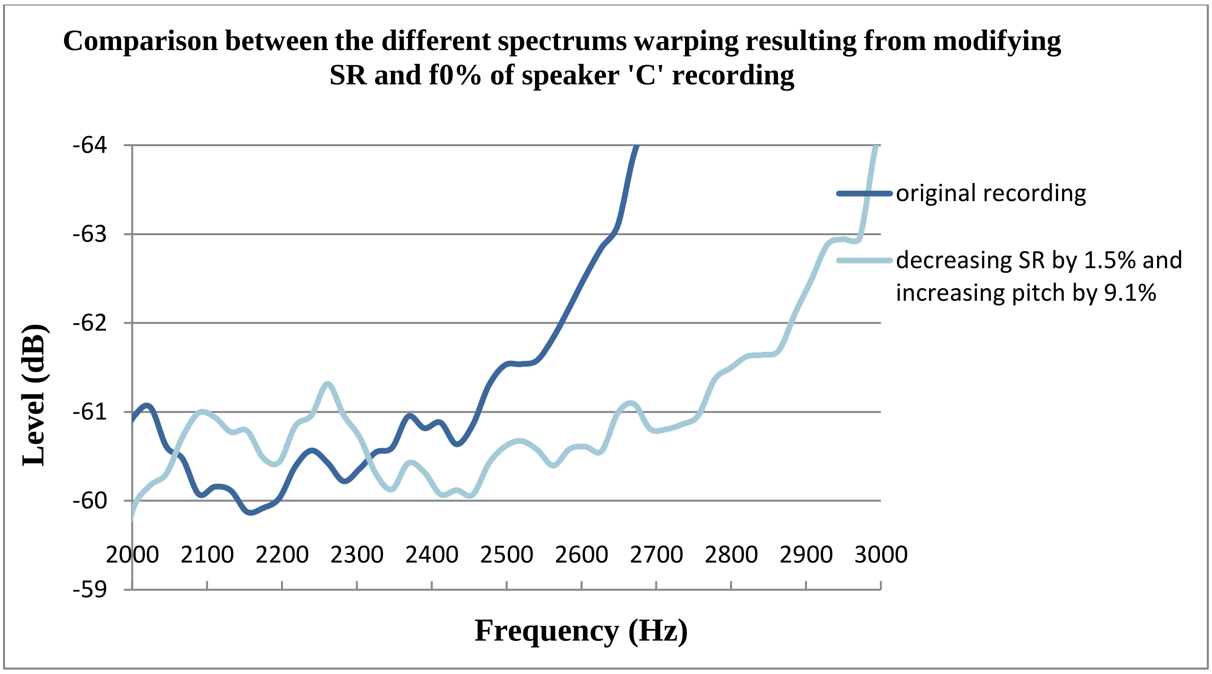
A.15 Spectral frequency warping, resulting from modifying fundamental frequency of speaker 'C' recording



Figure_Apx A-15 Comparison between the spectrums of speaker 'C' recording and that of the modified pitch recording. The formants and peak positions offset result from the modified F0%

Figure_Apx A-15 shows the spectral frequency warping caused by increasing the fundamental frequency of speaker ‘C’ recording by 8.8%, which gives the best improvement in the WER. It can be seen that this shifts the formant resonance peaks to the right (i.e. raises their frequencies).

A.16 Spectral frequency warping resulting from modifying SR and fundamental frequency % of speaker ‘C’ recording



Figure_Apx A-16 Comparison between the spectrums of original ‘C’ recording and that of modified speech rate and fundamental frequency recording

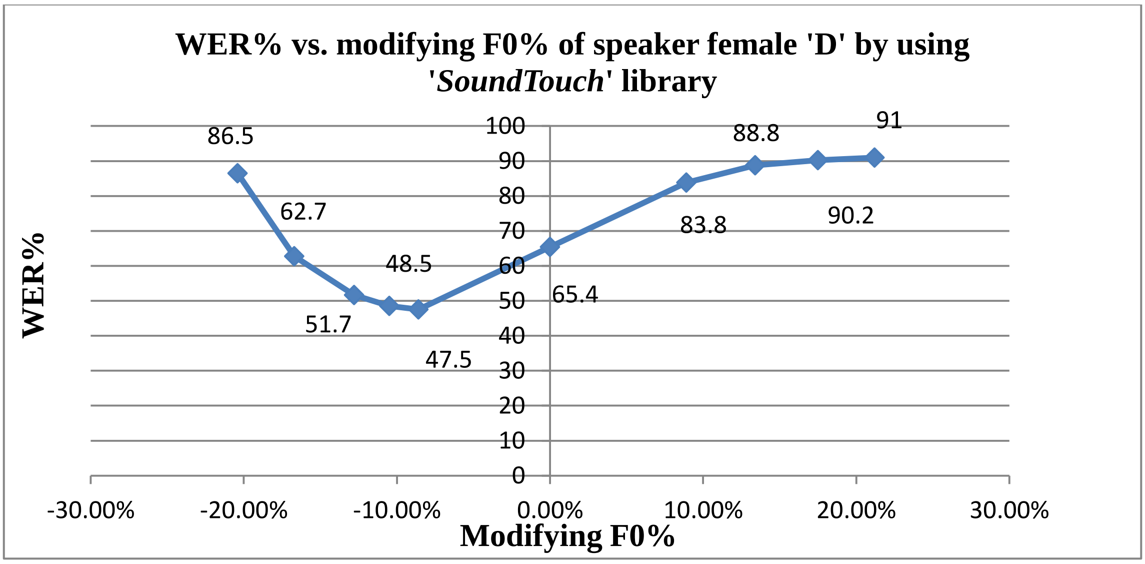
Figure_Apx A-16 shows how modifying the recording of male speaker ‘C’ by increasing the pitch by 9.1% and decreasing SR by 1.5 % resulted in expansions of the ‘frequency spectrum (i.e. shifts formant resonance peaks to the right) to give the best WER.

A.17 Modifying the fundamental frequency of a recording of female speaker ‘D’ by using ‘*SoundTouch*’ library

Changing F_0 percentage	Pitch (AV freq Hz)	WER	AR
–22.3%	158.5	86.5	13.5
–18.5%	166.1	62.7	37.3
–14.5%	174.4	51.7	48.3
–12.4%	178.8	48.5	51.5
–10.2%	183.2	47.5	52.5
0.00%	204	65.4	34.6
9.4%	223.2	83.8	16.2
14.6%	233.8	88.8	11.2
19.2%	243.2	90.2	9.8
25%	254.9	91	9

Table_Apx A-13 Shows the summary results for WER, resulting from modifying pitch of female ‘D’ speech by increasing and decreasing the fundamental frequencies of the speech, using ‘*SoundTouch*' library

Table_Apx A-13 shows WER and AR of generating transcripts via a Dragon ASR system for each pre-processed recording of speaker D. Each pre-processed recording of ‘D’ was altered by modifying the fundamental frequencies of his speech. The changes in the original pitch percentage were between –22.3% and 25%. The modification percentages result in an average fundamental frequency between 158.5Hz and 254.9Hz. The results of WER of the pre-processed recording’s transcripts were found to be between 91% (Worst value) and 47.5% (best values) and therefore the AR between 9%, and 52.5% respectively.



Figure_Apx A-17 WER percentages vs. modifying the F0 of the original female speaker ‘D’ via '*SoundTouch*' library

Figure_Apx A-17 plots the results which are summarised below:

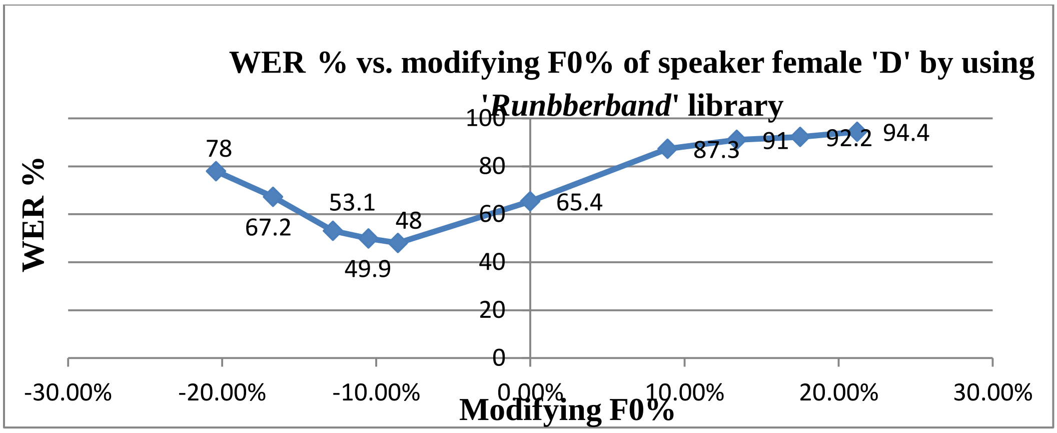
- Modifying F0% by –8.6% reduced WER to a minimum of 47.5%, which was a reduction of 27% from the original 65.4% WER.
- Modifying F0% further increased the WER.
- A positive change in F0 % increased WER.

A.18 Modifying fundamental frequency of female speaker 'D' recording by using 'Rubber Band' library

Changing F_0 percentage	Pitch (AV freq Hz)	WER	AR
–21%	161.2	78	22
–17.5%	168.4	67.2	32.8
–13.2%	177.1	53.1	46.9
–11.2%	181.1	49.9	50.1
–8.7%	186.2	48	52
0.00%	204	65.4	34.6
13%	230.5	87.3	12.7
19.8%	244.4	91	9
26.1%	257.3	92.2	7.8
32.6%	270.6	94.4	5.5

Table_Apx A-14 Shows the summary results for WER, resulting from modifying pitch of female 'D' speech by increasing and decreasing F0, via '*Rubber Band* ' library

Table_Apx A-14 shows that modifiying F_0 of the original speaker's recordings by between –21% ad 32.6% resulted in different pre-processed recordings of different average frequency values between 161.2 Hz and 270.6 Hz. Transcribing the pre-processed recodings by the Dragon speech recognition system produced different WER values between 94.4 %(worst value) and 48% (best value) and therefore corresponding accuracy rate values between 5.5% and 52%.



Figure_Apx A-18 WER percentages vs. modifying F0 of the original speaker 'D' recording via '*Rubber Band* ' library

Figure_Apx A-18 plots the results which are summarised below:

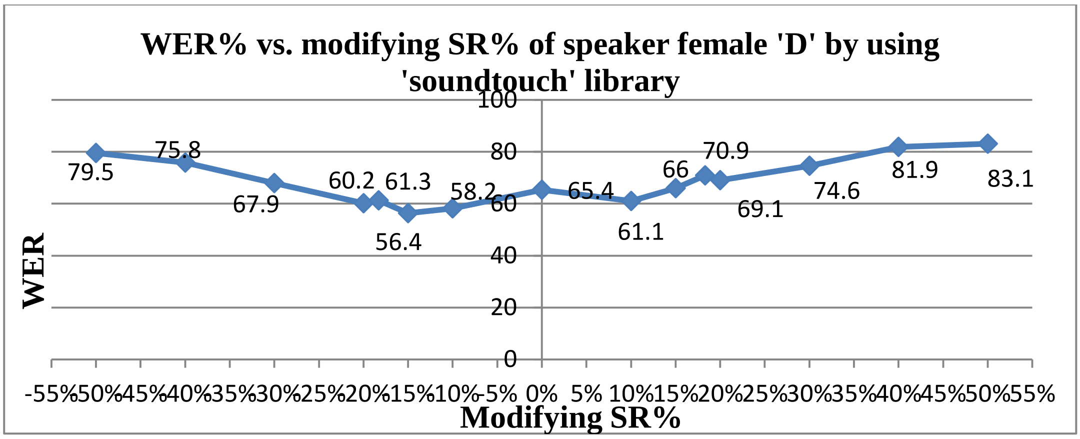
- decreasing the F_0 of D's speech by 8.6% leads to decreasing the WER to 48% (best value), an improvement of 26.6%
- The WER increased with further decreasing of F_0

A.19 Modifying speech rate of a recording of female speaker 'D' by using '*SoundTouch*'

Speed rate (percentage)	Pitch (AV freq Hz)	WER	AR	Length (min)
–50	204.4	79.5	20.5	09:59
–40	203.7	75.8	24.2	08:19
–30	202.2	67.9	32.1	07:08
–20	198.1	60.2	39.8	06:14
–18.3	197.9	61.3	38.7	06:07
–15	198.9	56.4	43.6	05:52
–10	200.7	58.2	41.8	05:33
0	204	65.4	34.6	4:95
10	204.2	61.1	38.8	04:32
15	204.5	66	34	04:20
18.3	204.5	70.9	29.1	04:13
20	205	69.1	30.9	04:09
30	204.8	74.6	25.4	03:50
40	205.4	81.9	18.1	03:34
50	206.2	83.1	16.9	03:19

Table_Apx A-15 Shows the summary results for WER, resulting from modifying SR of female 'D' speech by increasing and decreasing the original SR, via using '*SoundTouch*' library

Table_Apx A-15 shows the effects of modifying the speech rate of the audio file of speaker 'D' by between –50% and 50%. The changes on the audio file were made only on the speech rate while trying to keep the F0 stable. However, it was noticed that modifying SR was still causing between –2.9% and 1.1% changes on F0%. The WER was showing changes between the worst value 83.1%, with increasing the SR by 50%, and the best value 56.4%, with decreasing the SR by 15%.



Figure_Apx A-19 Shows WER % vs. modifying SR% of female speaker 'D' by using '*SoundTouch*' library

Figure_Apx A-19 plots the results which are summarised below:

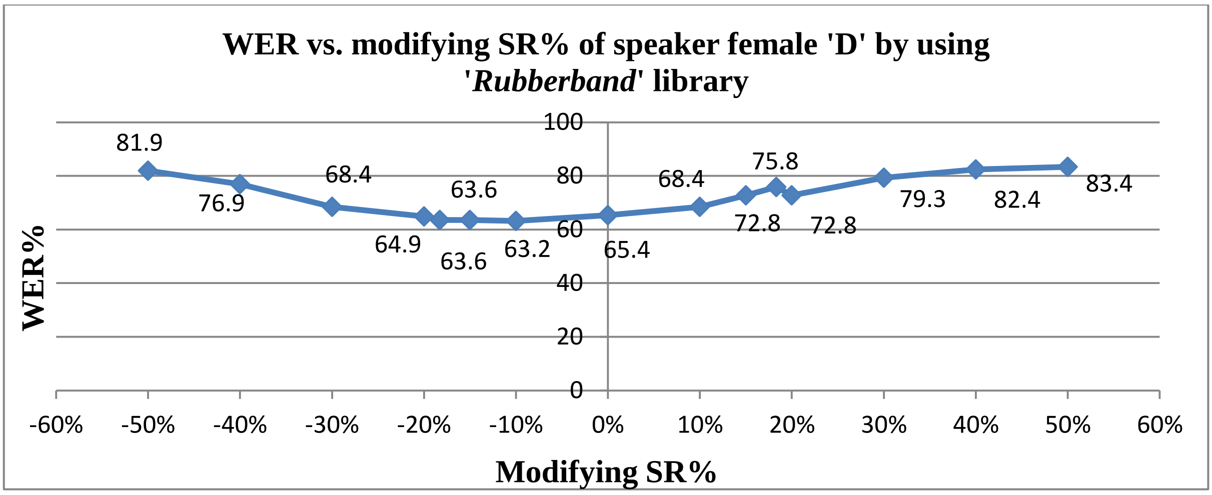
- Changing the SR by – 15% could improve the WER from 65.4% to 56.4% which is an 13.8% improvement
- On the other hand, the figure shows increasing the SR by 10% decreased the WER to 61.1% from 65.4%.

A.20 Modifying speech rate of a recording of female speaker 'D' by using 'Rubber Band'

Speech rate (percentage)	Pitch (AV freq Hz)	WER	AR	Length (min)
–50	206.5	81.9	18.1	09:59
–40	206.9	76.9	23.1	08:19
–30	207.3	68.4	31.6	07:08
–20	207.4	64.9	35.1	06:14
–18.3	207.3	63.6	36.4	06:07
–15	206.6	63.2	36.8	05:52
–10	207.2	63.6	36.4	05:33
0	204	65.4	34.6	4:59
10	207.5	68.4	31.6	04:32
15	207.9	72.8	27.2	04:20
18.3	208	75.8	24.2	04:13
20	207.8	72.8	27.2	04:09
30	208.6	79.3	20.7	03:50
40	209	82.4	17.6	03:34
50	209.4	83.4	16.6	03:19

Table_Apx A-16 Shows the summary results for WER, resulting from modifying SR of female 'D' speech by increasing and decreasing the original SR, via using '*Rubber Band* ' library

Table_Apx A-16 shows that the WER was changing between the worst value 83.4%, with increasing the SR by 50%, and the best value 63.2%, with decreasing the SR by 15%.



Figure_Apx A-20 Shows WER % vs. modifying SR% of female speaker 'D' by using '*Rubber Band* ' library

Figure_Apx A-20 plots the results which are summarised below:

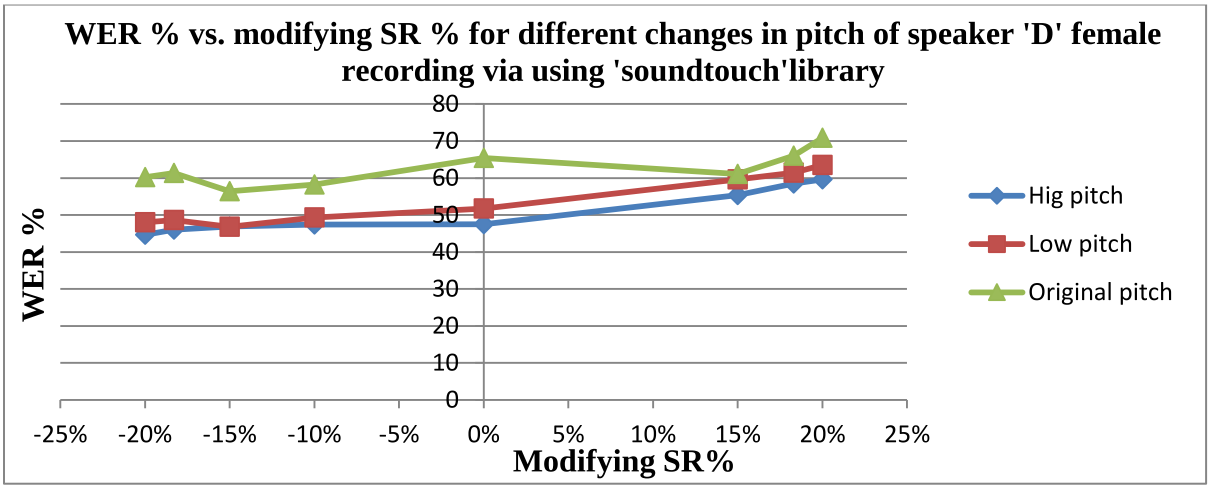
- Changing the SR by – 10% and –15 could improve the WER from 65% to 63% which is an 2% improvement even though that decreased SR % caused increasing on f0% by 1.6%.
- The results on Figure_Apx A-17 and Figure_Apx A-18 showed that any increasing of the female fundamental frequency leads to increase the WER %.
- However, the result in Figure_Apx A-20 shows decreasing SR% could decrease the WER even if F0% increased.
- On the other hand, the figure shows increasing the SR by 10% increased the WER to 68.4%.

A.21 Modifying fundamental frequency and speech rate of female speaker 'D' by using '*SoundTouch*' library

Speech rate (percentage)	Changing F_0 percentage	Pitch (AV freq Hz)	WER	AR
–20	–10.2%	180.4	44.6	55.4
–20	–14.5%	172.6	48	52
–18.3	–10.2%	181.1	46	54
–18.3	–14.5%	173.2	48.6	51.4
–15	–10.2%	182.2	46.9	53.1
–15	–14.5%	173.5	46.8	53.2
–10	–10.2%	183.2	47.4	52.6
–10	–14.5%	173.8	49.3	50.6
0	0	204	65.4	34.6
15	–10.2%	183.9	55.4	44.6
15	–14.5%	175.2	59.6	40.4
18.3	–10.2%	184	58.5	41.5
18.3	–14.5%	175.1	61.4	38.6
20	–10.2%	184.1	59.6	40.4
20	–14.5%	175	63.5	36.5

Table_Apx A-17 Shows the summary results for WER, resulting modifying SR for different decreases in F0 of female 'D' speech by using '*SoundTouch*' library

Table_Apx A-17 shows the WER and AR of speech recognition resulting from modifying both fundamental frequencies and speech rate of the recording of speaker ‘D’. The changes in the original pitch were between –10.2% and –14.5%. The modification percentages have resulted in an average F_0 of between 173.2 and 184.1 Hz. The results of WER of pre-processed audio files were found to be between 44.6% (best) and 63.5% (worst) and therefore the AR was between 55.4% and 36.5%, respectively.



Figure_Apx A-21 Plots WER percentages vs. the modifications of SR percentages for different changes in pitch percentages of speaker ‘D’ female via ‘*SoundTouch*’ library

Figure_Apx A-21 plots the results which are summarised below:

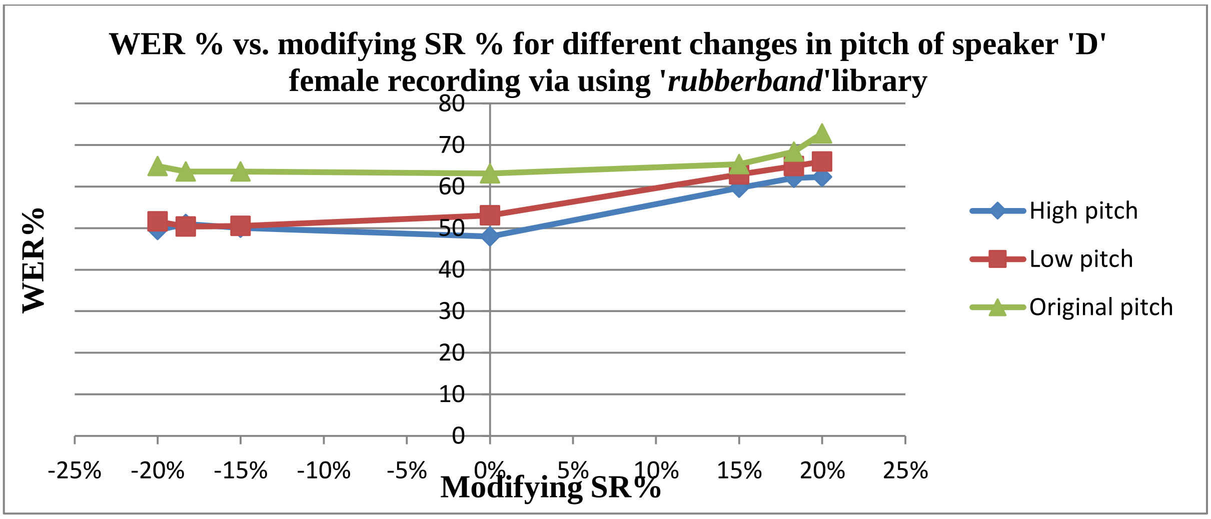
- Modifying SR by –20% and decreasing the fundamental frequency by 10.2% improved the WER to 44.6(best value), a change of 31.8%
- The WER was increasing and decreasing significantly by between 49.3% and 44.6% with changing speech rate by between –20% and 20% and decreasing the pitch by between 10.4% and 14.5 %.

A.22 Modifying fundamental frequency and speech rate of female speaker ‘D’ by using ‘Rubber Band’ library

Speed rate (percentage)	Changing F_0 percentage	Pitch (AV freq Hz)	WER	AR
–20	–8.7%	185.1	49.5	50.5
–20	–13.2%	175.2	51.6	48.4
–18.3	–8.7%	185.7	51	49
–18.3	–13.2%	175	50.4	49.6
–15	–8.7%	184.5	50.1	49.9
–15	–13.2%	174.9	50.5	49.5
0	0	204	65.4	34.6
15	–8.7%	187.2	59.7	40.3
15	–13.2%	178	62.9	37.1
18.3	–8.7%	187.2	62.1	37.9
18.3	–13.2%	178.4	64.9	35.1
20	–8.7%	187.2	62.3	37.7
20	–13.2%	178.2	66	34

Table_Apx A-18 Shows the summary results for WER, resulting modifying SR for different decreases in F0 of female ‘D’ speech by using ‘*Rubber Band* ’ library

Table_Apx A-18 shows the WER and AR of speech recognition resulting from modifying both fundamental frequencies and speech rate of the recording of speaker ‘D’. The changes in the original pitch were between –8.7% and –13.2%. The modification percentages have resulted in an average F_0 of between 174.9Hz and 187.2Hz. The results of WER of pre-processed audio files were found to be between 50.1% (best) and 66% (worst) and therefore the AR was between 49.9% and 34%, respectively.

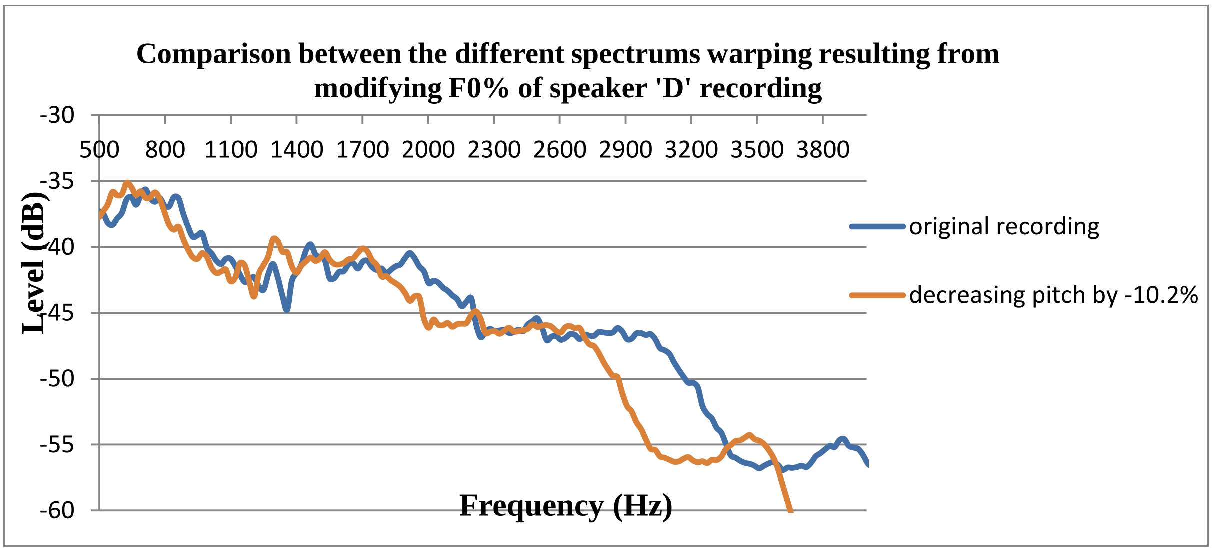


Figure_Apx A-22 Plots WER percentages vs. the modifications of SR percentages for different changes in pitch percentages of speaker ‘D’ female via ‘Rubber Band ’ library

Figure_Apx A-22 plots the results which are summarised below:

- Modifying SR by –20% and decreasing the fundamental frequency by –8.7% improved the WER to 49.5(best value), a change of 24.3%

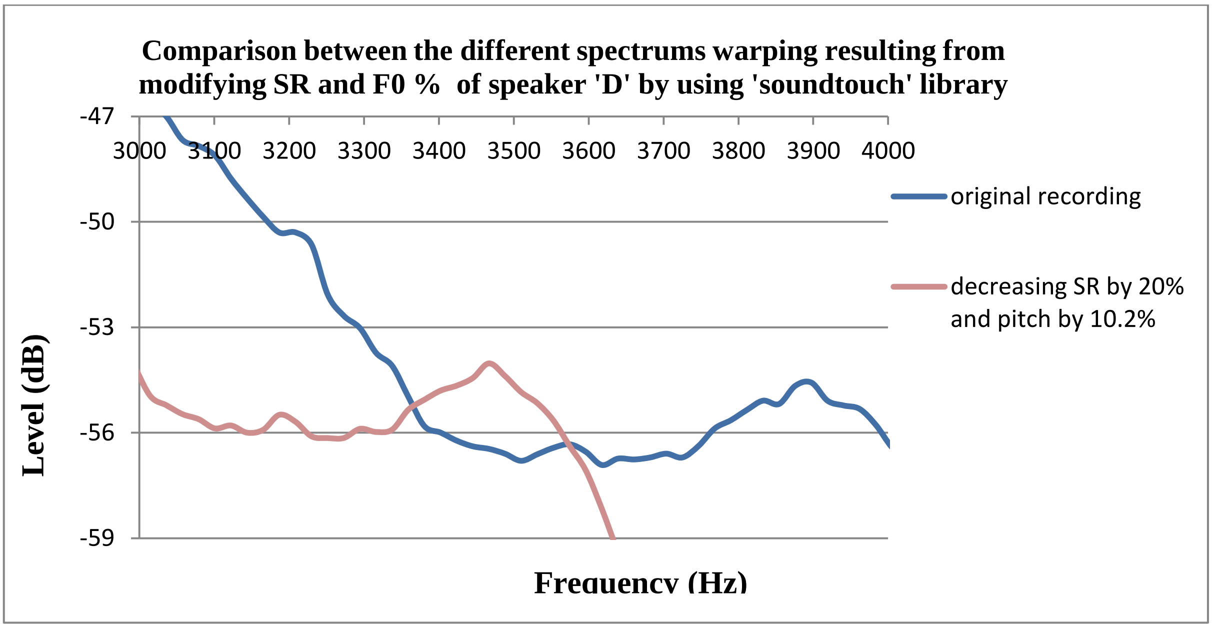
A.23 Spectral frequency warping, resulting from modifying fundamental frequency of speaker ‘D’ recording



Figure_Apx A-23 Comparison between the spectrums of speaker ‘D’ recording and that of the modified pitch recording. The formants and peak positions offset result from the modified F0%

Figure_Apx A-23 shows the spectral frequency warping caused by decreasing the fundamental frequency of speaker ‘d’ recording by 10.2 respectively, which gives the best improvement in the WER. It can be seen that this shifts the formant resonance peaks to the left (i.e. decreases their frequencies).

A.24 Spectral frequency warping resulting from modifying SR and fundamental frequency % of speaker ‘D’ recording

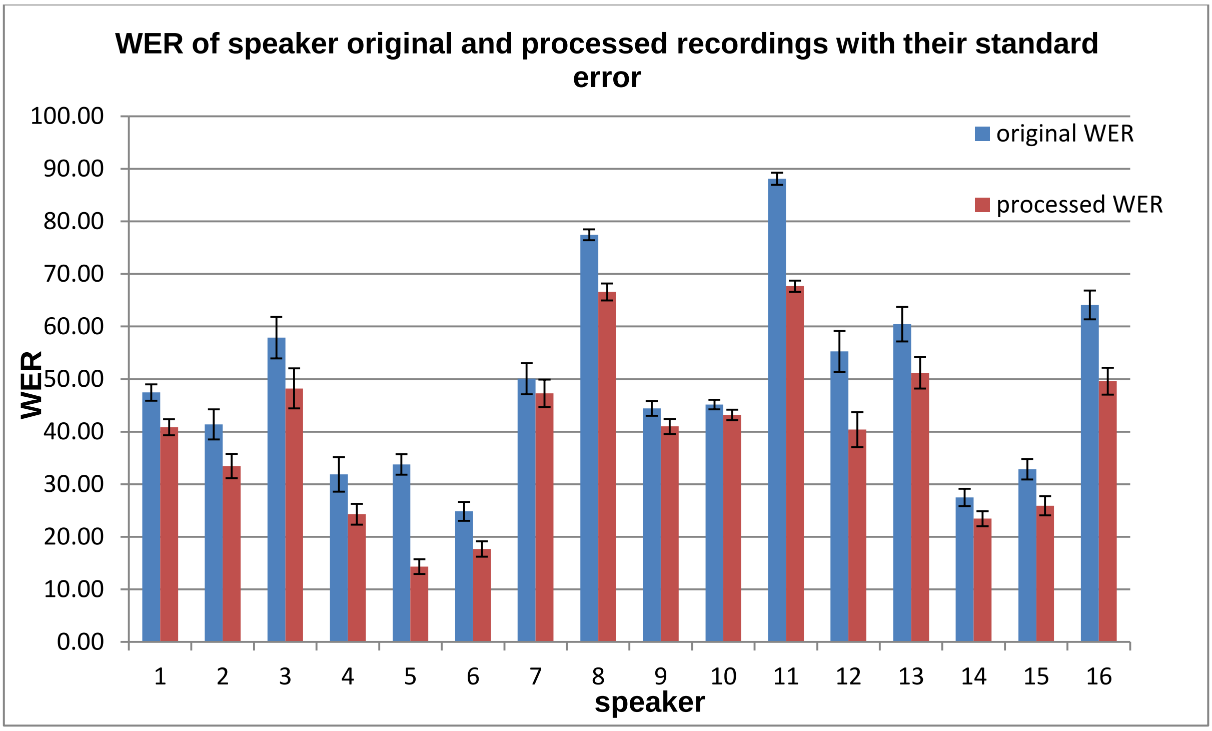


Figure_Apx A-24 Comparison between the spectrums of original ‘D’ recording and that of modified speech rate and fundamental frequency recording

Figure_Apx A-24 shows how modifying the recording of female speaker ‘D’ by increasing the pitch by 10.2% and decreasing SR by either 20 % resulted in compressions of the ‘frequency spectrum (i.e. shifts formant resonance peaks to the left) to give the best WER.

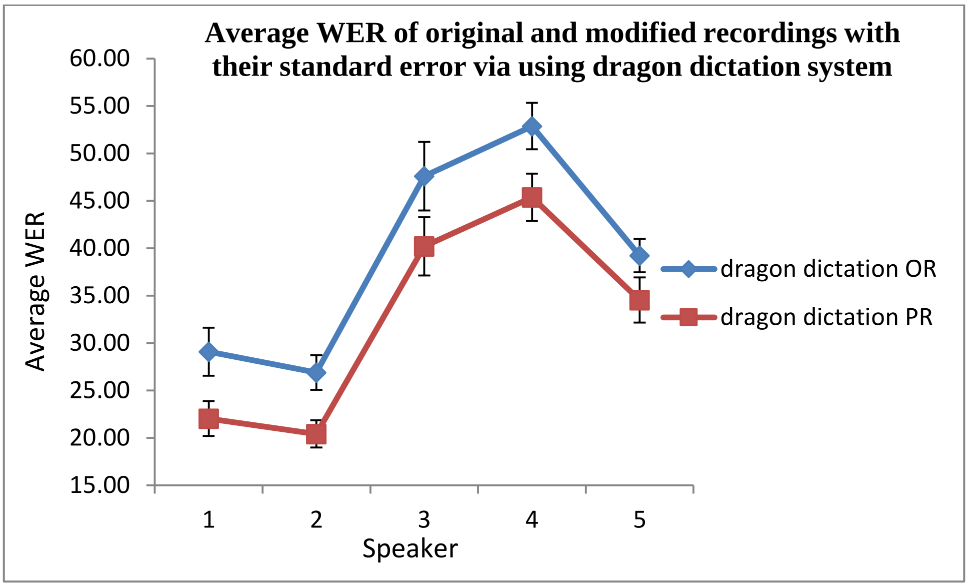
Appendix B

Figure_Apx B-1 shows the average WER of the original and best modified versions of the 16 speakers' recordings. The original and best modified versions which resulted from this section's investigation were used in the Chapter 6 as a training set to find an appropriate estimated function for the modification values for the new speakers' recordings.



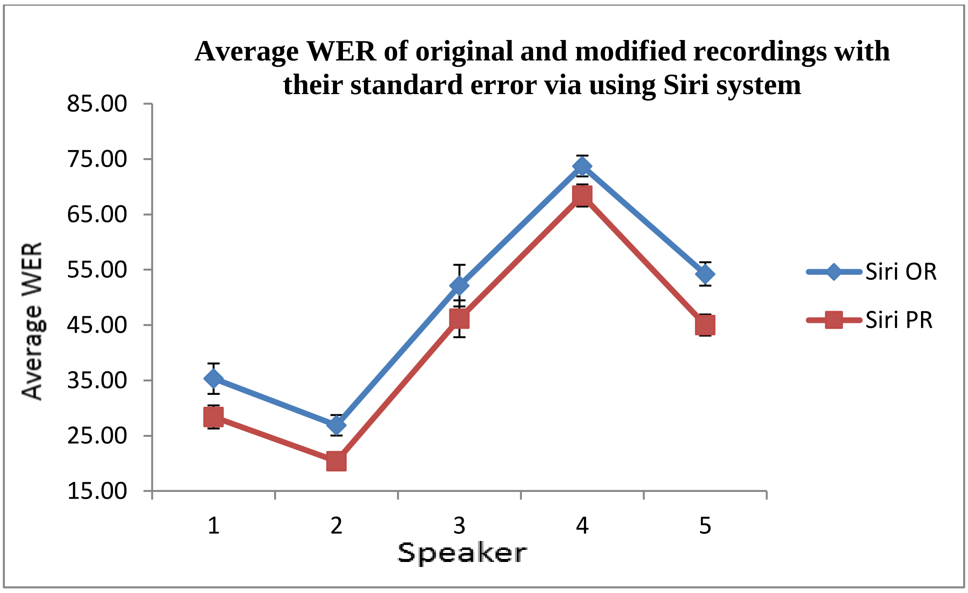
Figure_Apx B-1 Average WER of original and modified recordings with their error bars

Appendix C



Figure_Apx C-1 Average WER of original and modified recordings with their standard error via using dragon dictation system

Figure_Apx C-1 shows the average WER of the original and best modified versions that resulted from trial and error approach for 5 speakers' recordings, consisting of 3 males and 2 females, with their error bars via using dragon dictation system.



Figure_Apx C-2 Average WER of original and modified recordings with their standard error via using Siri system

Figure_Apx C-2 shows the average WER of the original and best modified versions that resulted from trial and error approach for 5 speakers' recordings, consisting of 3 males and 2 females, with their error bars via using Siri system.

Bibliography

Acero, A. & Stern, R., 1991. *Robust speech recognition by normalization of the acoustic space.* California, In Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on (pp. 893-896). IEEE., pp. 893-896.

Alías, F., Socoró, J. C. & Sevillano, X., 2016. A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. *Applied Sciences*, p. p.143.

Alku, P., Pohjalainen, J. & Vainio, M., 2013. Formant frequency estimation of high-pitched vowels using weighted linear prediction. *The Journal of the Acoustical society of America*, pp. 1295-1313.

Amento, B., Whittaker, S. & Hirschberg, J., 2002. *Scanmail: a voicemail interface that makes speech browsable, readable and searchable.* Minneapolis, Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 275-282). ACM..

American Standards Association, and Acoustical Society of America, 1960. American standard acoustical terminology:(including mechanical shock and vibration). *American Standards Association.*

Anguera, X., Luque, J. & Gracia, C., 2014. *Audio-to-text alignment for speech recognition with very limited resources.* Spain, In Fifteenth Annual Conference of the International Speech Communication Association.

Arsikere, H., Lulich, S. M. & Alwan, A., 2013. Non-linear frequency warping for VTLN using subglottal resonances and the third formant frequency. *ICASSP*, pp. 7922-7926.

Audacity Team, 2012. *Audacity®. Version 2.0.2. Audio editor and recorder.* [Online]
Available at: http://audacityteam.org/
[Accessed 20 10 2013].

Baecker, R., Munteanu, C. & Penn, G., 2006. *Measuring the acceptable word error rate of machine-generated webcast transcripts.* s.l., s.n.

Ban, S., Choi, B., Choi, Y. & Kim, H., 2014. *VTLN Based Approaches for Speech Recognition with Very Limited Training Speakers,* Bormio: In 2014 5th International Conference on Intelligent Systems, Modelling and Simulation (pp. 285-288). IEEE..

Beigi, H., 2011. *Fundamentals of speaker recognition.* 1st Edition ed. New York: Springer.

Bei, X., Chen, N. & Zhang, S., 2013. *On the complexity of trial and error.* Palo Alto, California, In Proceedings of the forty-fifth annual ACM symposium on Theory of computing (pp. 31-40). ACM.

Bell, J.-M., 2007. Enhancing accessibility through correction of speech recognition errors. *ACM SIGACCESS Accessibility and Computing.*

Benzeghiba, M., Mori, R. D. & Deroo, O., 2007. Automatic speech recognition and speech variability: A review. *Speech Communication,* 49(10), pp. 763-786.

Berke, J., 2009. Lipreading (or speechreading).

Boersm, P. & Weenink, D., 2002. Praat,asystemfordoingphonetics by computer. *Glot international,* 5(9/10), p. 341–345.

Boulain, P., Wald, M. & Bell, J.-M., 2007. *Correcting automatic speech recognition captioning errors in real time.* s.l., s.n.

Bradlow, A. R., Torretta, G. M. & Pisoni, D. B., 1996. Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication,* 20(3), pp. 255-272.

Chang, B., Young, C., Tsai, H. & Fang, R., 2011. Timed PR-SCTP for fast voice/video over IP in wired/wireless environments.. *Journal of Information Hiding and Multimedia Signal Processing,* 2(4), pp. 320-331.

Chauhan, Y., 2013. *Vowel.* [Online]
Available at: https://www.britannica.com/topic/vowel
[Accessed 2017].

Cohen, J., kamm, T. & Andreou, A., 1995. Vocal tract normalization in speech recognition: Compensating forVocal tract normalization in speech recognition: Compensating for systematic speaker variability. *The Journal of the Acoustical Society of America,* 97(5), pp. 176-178.

Connelly, L. M., 2008. Pilot studies. *Medsurg Nursing,* 17(2).

Daher, W., Lieberman, H. & Faaborg, A., 2005. *How to wreck a nice beach you sing calm incense.* s.l., s.n.

Dias, B., Matos, D., Davies, M. & Pinto, H., 2016. *Time Stretching & Pitch Shifting with the Web Audio API: Where are we at?,* Atlanta: Georgia Tech Conferences.

Dissen, Y. & Keshet, J., 2016 . Formant Estimation and Tracking using Deep Learning.. *Interspeech 2016,* pp. 958-962.

Dissen, Y. & Keshet, J., 2016. Formant Estimation and Tracking using Deep Learning.. *Interspeech 2016,* pp. 958-962.

Emori, T. & Shinoda, K., 2001. Rapid vocal tract length normalization using maximum likelihood estimation. *Proc. EuroSpeech,* pp. 1649-1652.

Faria, A. & Gelbart, D., 2005. *Efficient Pitch-based Estimation of VTLNWarp Factors,* Berkeley: International Computer Science Institute.

Fong, S., Lan, K. & Wong, R., 2013. Classifying human voices by using hybrid SFX time-series preprocessing and ensemble feature selection. *BioMed research international.*

Gaur, Y., Lasecki, W., Metze, F. & Bigham, J., 2016. *The effects of automatic speech recognition quality on human transcription latency.* s.l., ACM.

Given, L., 2008. *The Sage encyclopedia of qualitative research methods.* 2nd ed. CA: Sage Publications.

Gouvêa, E. B., 1998. *Acoustic-feature-based frequency warping for speaker normalization,* Mellon: (Doctoral dissertation, Carnegie Mellon University)..

Guoping, L., Lutman, M. E., Wang, S. & Bleeck, S., 2012. Relationship between speech recognition in noise and sparseness. *International journal of audiology,* 51(2), pp. 75-82.

Hertzog, M., 2008. Considerations in determining sample size for pilot studies. *Research in Nursing & Health,* Volume 31, pp. 180-191.

Hill, R., 1998. What sample size is “enough” in internet survey research?. *Interpersonal Computing and Technology: An Electronic Journal for the 21st Century,* Volume 6, pp. 3-4.

Huang, X. & Deng, L., 2004. *Challenges in adopting speech recognition,* s.l.: Communications of the ACM.

IBM Crop, Released 2013. *IBM SPSS Statistics for Windows, Version 22.0,* NY: Armonk, NY: IBM Corp.

Isaac, S. & Michael, W. B., 1995. *Handbook in research and evaluation..* San Diego: CA: Educational and Industrial Testing Services.

Jakovljević, N., Janev, M., Pekar, D. & Mišković, D., 2008. *Energy normalization in automatic speech recognition,* s.l.: Text, Speech and Dialogue. Springer Berlin Heidelberg.

Julious, S. A., 2005. Sample size of 12 per group rule of thumb for a pilot study. *Pharmaceutical Statistics,* Volume 4, pp. 287-291.

Kersting, S. A., 1997. Balancing between deaf and hearing worlds:Reflections of mainstreamed college students on relationships and social interaction.. *Journal of Deaf Studies and Deaf Education,* pp. 252-263.

LabMaster, 2013. *Speaker normalization in ASR: Vocal Tract Length Normalization (VTLN).* [Online]

Available at: http://dynadmic-lab.com/2013/10/21/speaker-normalization-in-asr-vocal-tract-length-normalization-vtln-5/

Lawrence, C., 2009. *Causal Theory and Research Design.* [Online]

Available at: http://www.lordsutch.com/pol251/shively-6.pdf.

[Accessed 01 Jan 2016].

Lee, C.-y. & Glass, J., 2011. *A Transcription Task for Crowedsourcing with Automatic Quality Control.* Cambridge , MIT Computer Science and Artificial Intelligence Laboratory .

Machlica, L., Zbyn, ˇ. Z. & Ale, ˇ. P., 2009. Methods of Unsupervised Adaptation in Online Speech Recognition. *SPECOM'2009,* pp. 1-6.

Madhavi, M., Sharma, S. & Patil, H., 2016. *VTLN Using Different Warping Functions for Template Matching.*. London: In Machine Intelligence and Big Data in Industry (pp. 111-121). Springer International Publishing.

MATLAB and Statistics Toolbox , Release 2012b. *MATLAB 8.0 and Statistics Toolbox 8.1,* United States: The MathWorks, Inc., Natick, Massachusetts.

McDonough, J., Metze, F., Soltau, H. & Waibel, A., 2001. *Speaker compensation with sine-log all-pass transforms.* Florida, In Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on (Vol. 1, pp. 369-372). IEEE., pp. pp. 369-372.

Melanie, T. & Cialdini, R., 1998. Social influence: Social norms, conformity and compliance.

Mimura, M. & Kawahara, T., 2011. *Fast speaker normalization and adaptation based on BIC for meeting speech recognition.* Xi'an, Proc. APSIPA.

Morandi, D., 2012. *Effect of pitch modification on the voice identification of the speakers,* Argentina: Acoustic laboratory, Untref university.

NIST Scoring Toolkit Version 0.1, 1996. *sclite - score speech recognition system output.* [Online]
Available at: http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm
[Accessed 04 February 2015].

OpenML, 2016. *10-fold Crossvalidation.* [Online]
Available at: https://www.openml.org/a/estimation-procedures/1
[Accessed 16 4 2017].

Osterrath, F., Boulianne, G. & Boisvert, M., 2008. Real-time speech recognition captioning of events and meetings. *Heritage Canada New Media Funds.*

Padrnanabhan, M. & Mangu, L., 2001. *Error corrective mechanisms for speech recognition.* s.l., s.n.

Papadopoulos, M. & Pearson, E., 2009. An Analysing Tool to Facilitate the Evaluation Process of Automatic Lecture Transcriptions. *Proceedings of the World Conference on e-Learning in Corporate, Government, Healthcare, and Higher Education (ELEARN 2009),* pp. 2189-2198.

Papadopoulos, M. & Pearson, E., 2011. *A System to Support Accurate Transcription of Information Systems Lectures for Disabled Students.* Sydney, ACIS 2011 Proceedings. 35. http://aisel.aisnet.org/acis2011/35.

Parviainen, O., 2001. *SoundTouch Audio Processing Library.* [Online]
Available at: http://www.surina.net/soundtouch/index.html
[Accessed 28 01 2014].

Pearson, E. & Papadopoulos, M., 2011. *A system to support accurate transcription of information systems lectures for disabled students,* Sydney : Accessibility Research Center.

Quey, B. f., 2012. *Rubber band library.* [Online]
Available at: http://breakfastquay.com/rubberband/index.html
[Accessed 28 01 2014].

Radnitzky, G., Bartley, W. & Popper, K., 1987. *Evolutionary epistemology, rationality, and the sociology of knowledge.* Chicago: Open Court Publishing.

Romenesko, E., Vorperian, H., Kent, R. & Austin, D., 2016. Aging Effects on Acoustic Characteristics of Adult Speech.

Sahoo, D., Mishra, S., Panda, G. & Dash, P., 2013. Estimation Of Formant Frequency Of Speech Signal By Linear Prediction Method And Wavelet Transform. *In International Journal of Engineering Research and Technology (Vol. 2, No. 3 (March-2013)). ESRSA Publications..*

Shields, P. M. & Rangarajan, N., 2013. *A playbook for research methods: Integrating conceptual frameworks and project management.* 1st ed. Oklahoma: New Forums Press.

Six, J., Cornelis, O. & Leman, M., 2014. *TarsosDSP, a real-time audio processing framework in Java.* s.l., In Audio Engineering Society Conference: 53rd International Conference: Semantic Audio. Audio Engineering Society., p. Belgium.

Sommers, M. S., Nygaard, L. C. & Pisoni, D. B., 1994. Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude. *The Journal of the Acoustical Society of America,* 96(3), pp. 1314-1324.

Soy, S. K., 1997. *The case study as a research method.,* Austin: Unpublished paper, University of Texas at Austin.

Stathopoulos, E., Huber, J. & Richardson, K., 2014. Increased vocal intensity due to the Lombard effect in speakers with Parkinson's disease: Simultaneous laryngeal and respiratory strategies.. *Journal of communication disorders,* pp. 1-17.

Statistics, L., 2013. *Pearson Product-Moment Correlation.* [Online]
Available at: https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php
[Accessed 25 12 2016].

Szczypiorski, K. & Zydecki, W., 2017. *StegIbiza: Steganography in Club Music Implemented in Python..* NY, arXiv preprint arXiv:1705.07788.

Taylor, P., Caley, R. & Black, A. W., 1999. *Edinburgh Speech Tools Library.* [Online]
Available at: http://festvox.org/docs/speech_tools-1.2.0/x2152.htm
[Accessed 14 04 2014].

Treece, E. W. & Treece, J. W., 1982. Elements of research in nursing. *Nursing2016,* 7(6), pp. 12-13.

Trochim, W., 2010. *Research method knowledge base.* [Online]
Available at: http://www.socialresearchmethods.net/kb/statdesc.php
[Accessed 1 Oct 2014].

UCL, 2016. *Speakers and Accents.* [Online]
Available at: http://www.phon.ucl.ac.uk/courses/spsci/iss/week10.php
[Accessed 22 Dec 2016].

Uebel, L. & Woodland, P., 1999. *An investigation into vocal tract length normalisation.* Hungary, In Eurospeech.

Umesh, S., 2011. Studies on inter-speaker variability in speech and its application in automatic speech recognition. *Sadhana,* 36(5), pp. 853-883.

van Belle, G., 2002. *Statistical rules of thumb.* New York: John Wiley.

Wald, M., 2007. *correcting automatic speech recognition errors in real time.,* Southampton: University of Southampton.

Wald, M., 2007. *correcting automatic speech recognition errors in real time.,* s.l.: University of Southampton.

Wald, M., 2010. *Synote: Designed for all Advanced Learning Technology for Disabled and Non-Disabled People.* Sousse, The 10th IEEE International Conference on Advanced Learning Technologies, 10, 716-717.

Wang, S., Alwan, A. & Lulich, S. M., 2008. Speaker normalization based on subglottal resonances. *Acoustics, Speech and Signal Processing. ICASSP. IEEE International Conference on . IEEE.,* pp. 4277-4280.

Weenink, D., 2015. Improved formant frequency measurements of short segments..

Welling, L., Ney, H. & Kanthak, S., 2002. *Speaker adaptive modeling by vocal tract normalization.* New York, IEEE Transactions on Speech and Audio Processing, 10(6), pp.415-426., pp. 415-426.

Xuemin Chi, M. S., 2007. Subglottal coupling and its influence on vowel formants. *The Journal of the Acoustical Society of America* , 122(3), pp. 1735-1745.

Zhan, P. & Westphal, M., 1997. Speaker normalization based on frequency warping. *IEEE International Conference on,* Volume 2, pp. 1039-1042.