



2011

EVALUATION OF INTELLIGIBILITY AND SPEAKER SIMILARITY OF VOICE TRANSFORMATION

Anusha Raghunathan

University of Kentucky, anusha.raghunathan@uky.edu

[Right click to open a feedback form in a new tab to let us know how this document benefits you.](#)

Recommended Citation

Raghunathan, Anusha, "EVALUATION OF INTELLIGIBILITY AND SPEAKER SIMILARITY OF VOICE TRANSFORMATION" (2011). *University of Kentucky Master's Theses*. 101.
https://uknowledge.uky.edu/gradschool_theses/101

This Thesis is brought to you for free and open access by the Graduate School at UKnowledge. It has been accepted for inclusion in University of Kentucky Master's Theses by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.

ABSTRACT OF THESIS

EVALUATION OF INTELLIGIBILITY AND SPEAKER SIMILARITY OF VOICE TRANSFORMATION

Voice transformation refers to a class of techniques that modify the voice characteristics either to conceal the identity or to mimic the voice characteristics of another speaker. Its applications include automatic dialogue replacement and voice generation for people with voice disorders. The diversity in applications makes evaluation of voice transformation a challenging task. The objective of this research is to propose a framework to evaluate intentional voice transformation techniques. Our proposed framework is based on two fundamental qualities: intelligibility and speaker similarity. Intelligibility refers to the clarity of the speech content after voice transformation and speaker similarity measures how well the modified output disguises the source speaker. We measure intelligibility with word error rates and speaker similarity with likelihood of identifying the correct speaker. The novelty of our approach is, we consider whether similarly transformed training data are available to the recognizer. We have demonstrated that this factor plays a significant role in intelligibility and speaker similarity for both human testers and automated recognizers. We thoroughly test two classes of voice transformation techniques: pitch distortion and voice conversion, using our proposed framework. We apply our results for patients with voice hypertension using video self-modeling and preliminary results are presented.

KEYWORDS: Audio Privacy Protection, Voice Transformation, Speech Recognition, Speaker Identification, Speech Therapy.

(Anusha Raghunathan)

(April 2011)

EVALUATION OF INTELLIGIBILITY AND SPEAKER SIMILARITY OF
VOICE TRANSFORMATION

By

Anusha Raghunathan

Dr.Sen-ching Samson Cheung

(Director of Thesis)

Dr.Stephen Gedney

(Director of Graduate Studies)

April 2011

(Date)

RULES FOR THE USE OF THESIS

Unpublished thesis submitted for the Master's degree and deposited in the University of Kentucky Library are as a rule open for inspection, but are to be used only with due regard to the rights of the authors. Bibliographical references may be noted, but quotations or summaries of parts may be published only with the permission of the author, and with the usual scholarly acknowledgements.

Extensive copying or publication of the thesis in whole or in part also requires the consent of the Dean of the Graduate School of the University of Kentucky.

A library that borrows this thesis for use by its patrons is expected to secure the signature of each user.

NameDateThis image shows a single sheet of white paper with horizontal ruling lines. The lines are evenly spaced and run across the width of the page. There are no margins, text, or other markings on the paper.

THESIS

Anusha Raghunathan

The Graduate School

University of Kentucky

2011

EVALUATION OF INTELLIGIBILITY AND SPEAKER SIMILARITY OF
VOICE TRANSFORMATION

THESIS

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical Engineering
in the College of Engineering
at the University of Kentucky

By
Anusha Raghunathan
Lexington, Kentucky

Director: Dr.Sen-ching Samson Cheung , Department of Electrical and Computer
Engineering
Lexington, Kentucky

2011

Copyright © Anusha Raghunathan 2011

This work is dedicated to my mother.

ACKNOWLEDGEMENTS

First I would like to express my sincere gratitude to my advisor, Dr. Sen-Ching Cheung for his excellent guidance, support and encouragement throughout the thesis work. It has been a wonderful learning experience having worked with him both in terms of research and for personal growth. Next I would like to thank my thesis advisory committee members, Dr. Kevin Donohue and Dr. Rita Patel for their valuable comments and for their time. I would also like to thank the DGS, Dr. Stephen Gedney for his time and support.

I would like to thank all the members of the MIALAB group and the Vis Center who have helped me and supported me throughout my work. I would like to thank all my friends and relatives who have helped me through my journey here. I am extremely grateful to my family particularly my mother without whose love, support and blessings this work would not have been completed.

Table of Contents

Acknowledgements	iii
List of Tables	vii
List of Figures	ix
Chapter 1 Introduction	1
1.1 Voice Transformation Techniques and Applications	1
1.2 Evaluation Framework of Voice Transformation	3
1.3 Contributions of thesis	5
1.4 Organization of the Thesis	7
Chapter 2 Related Work	8
2.1 Evaluation of speech transformation techniques	8
2.2 Speaker privacy protection	11
2.3 Video self-modeling	13
Chapter 3 Speech Transformation	15
3.1 Why pitch distortion and voice conversion.	15
3.2 Voice Transformation	16
3.3 Pitch distortion Technique	20
3.4 Voice conversion algorithm	25
Chapter 4 Evaluation Framework	28
4.1 System Overview	28
4.2 Novelty of the Evaluation framework	29
4.2.1 Automatic recognizers trained for Same Model and Cross Model	30
4.2.2 Subjective testing	30
4.2.3 Subjective experiments with Same Model and Cross Model . .	31
4.3 Intelligibility evaluation with Automatic Speech Recognition	31
4.3.1 Computing Intelligibility	31
4.3.2 Speech Recognition	32

4.3.3	Voice Transformation and Intelligibility	38
4.4	Automatic Speech Recognizer - Julius	38
4.5	Speaker Identification with GMM	39
4.5.1	Speaker similarity vs. Speaker Identification	39
4.5.2	Speaker Identification	40
4.6	Speaker Identification ALIZE	44
Chapter 5 Experimental results for Privacy Protection		46
5.1	Experimental data and parameters	46
5.1.1	TIMIT Dataset	46
5.1.2	Pitch Distortion parameters	47
5.1.3	Voice Conversion parameters	47
5.2	Intelligibility evaluation results	49
5.2.1	Pitch Distortion Algorithm results	49
5.2.2	Voice Conversion results	49
5.2.3	Statistical significance test for Intelligibility	51
5.3	Speaker similarity using Speaker Identifier	54
5.3.1	Pitch Distortion Algorithm results	55
5.3.2	Voice Conversion results	55
5.3.3	Statistical significance test for Speaker similarity	56
5.4	Intelligibility evaluation with Subjective Experiments	59
5.4.1	Subjective experiment results	60
5.4.2	Statistical significance test	61
5.5	Speaker similarity evaluation with Subjective Experiments	64
5.5.1	Subjective experiment results	65
5.5.2	Statistical significance test	66
5.6	Observations from the Privacy Protection experiments	69
Chapter 6 Experimental Results for Voice Therapy		71
6.1	VSM for voice therapy	71
6.2	Experimental Setup	72
6.3	Audio Segmentation Algorithm	73
6.4	Speech synthesis	74
6.5	VSM Experimental Results	75
6.5.1	Speaker similarity results	76
6.5.2	Subjective evaluation	78

Chapter 7 Conclusions and Future Work	80
Bibliography	83
Vita	90

List of Tables

3.1	Average formant frequencies for some basic vowels	20
5.1	Pitch distortion parameters	47
5.2	Female to Female Voice distortion parameters	48
5.3	Female to Male Voice distortion parameters	48
5.4	Male to Female Voice distortion parameters	48
5.5	Male to Male Voice distortion parameters	49
5.6	Average WER% with Pitch Distortion for Cross Model	50
5.7	Average WER% with Pitch Distortion for Same Model	50
5.8	Average WER% with Voice Conversion for Cross Model	50
5.9	Average WER% with Voice Conversion for Same Model	51
5.10	z-test parameter values for WER using ASR	52
5.11	z-test results for WER with Pitch Distortion for Cross Model	53
5.12	z-test results for WER with Pitch Distortion for Same Model	53
5.13	z-test results for WER with Voice Conversion for Cross Model	54
5.14	z-test results for WER with Voice Conversion for Same Model	54
5.15	Average Speaker rank with Pitch Distortion	55
5.16	Average Speaker rank with Voice Conversion	56
5.17	z-test parameter values for Speaker rank using Speaker Identifier	57
5.18	z-test results for Speaker rank with Pitch Distortion for Cross Model	58
5.19	z-test results for Speaker rank with Pitch Distortion for Same Model	58
5.20	z-test results for Speaker rank with Voice Conversion for Cross Model	59
5.21	z-test results for Speaker rank with Voice Conversion for Same Model	59
5.22	Average recognition accuracy% for Subjective test with Pitch Distortion	60
5.23	Average recognition accuracy% for Subjective test with Voice Conversion	61
5.24	t-test parameter values for Recognition accuracy using Subjective tests	62
5.25	t-test results for intelligibility with Pitch Distortion for Cross Model	63
5.26	t-test results for intelligibility with Pitch Distortion for Same Model	63
5.27	t-test results for recognition accuracy with Voice Conversion for Cross Model	64
5.28	t-test results for recognition accuracy with Voice Conversion for Same Model	64

5.29	Average Speaker accuracy% for Subjective test with Pitch Distortion	65
5.30	Average Speaker accuracy% for Subjective test with Voice Conversion	66
5.31	t-test parameter values for Speaker Accuracy using Subjective test . .	67
5.32	t-test results for Speaker Accuracy with Pitch Distortion for Cross Model	68
5.33	t-test results for Speaker Accuracy with Pitch Distortion for Same Model	68
5.34	t-test results for Speaker Accuracy with Voice Conversion for Cross Model	69
5.35	t-test results for Speaker Accuracy with Voice Conversion for Same Model	69
6.1	Likelihood of synthesized voice compared to healthy voice	77
6.2	Results of forced choice test	79
6.3	Results of rank test	79

List of Figures

3.1	Speech signal of <i>“By the look of him he wasnt that far gone”</i>	17
3.2	Speech signal of <i>“He”</i>	18
3.3	Spectrum of the word <i>“He”</i>	19
3.4	Male and female formant spectrum [32].	19
3.5	Pitch distortion algorithm	21
3.6	Step 1 of Time stretching	22
3.7	Step 2 of Time Stretching	22
3.8	Step 3 of Time Stretching	23
3.9	Step 4 of Time Stretching	23
3.10	Step 5 of Time Stretching	24
3.11	Step 11 of Time Stretching	24
3.12	Warping function from two spectra	27
4.1	Training system of Evaluation Framework	29
4.2	Testing System of Evaluation Framework	29
4.3	Mel frequency cepstrum computation	34
4.4	HMM Example	36
4.5	HMMs concatenation example	36
4.6	Male and female cepstrum for vowels [38]	39
4.7	Linear Cepstral Coefficients computation	42
4.8	Example Gaussian mixture model with individual Gaussian densities [43]	42
5.1	Bar chart showing Intelligibility accuracy% with Pitch Distortion using ASR.	51
5.2	Bar chart showing Intelligibility accuracy% with Voice Conversion using ASR.	51
5.3	z-test results for accuracy of Intelligibility using ASR with Pitch Distortion and Voice Conversion	53
5.4	Bar chart showing speaker rank% with Pitch Distortion using Speaker Identifier	56
5.5	Bar chart showing speaker rank% with Voice Conversion using Speaker Identifier	56

5.6	z-test results for speaker rank using Speaker Identifier with Pitch Distortion and Voice Conversion.	58
5.7	Bar chart showing Recognition accuracy% with Pitch Distortion using Subjective tests.	61
5.8	Bar chart showing Recognition accuracy% with Voice Conversion using Subjective tests.	61
5.9	t-test results for Recognition accuracy using Subjective tests with Pitch Distortion and Voice Conversion.	63
5.10	Bar chart showing speaker accuracy% with Pitch Distortion using Subjective tests.	66
5.11	Bar chart showing speaker accuracy% with Voice Conversion using Subjective tests.	66
5.12	t-test results for Speaker accuracy using Subjective tests with Pitch Distortion and Voice Conversion.	68
6.1	Video capture for VSM	72
6.2	VSM content generation process	73
6.3	Audio Segmentation Algorithm.	74
6.4	Bar chart showing Likelihood of synthesized voices against the normal healthy voice.	77

Chapter 1

Introduction

In this chapter we discuss the speech transformation techniques, its applications and the practical challenges in the evaluation of the speech transformation techniques. We then present the motivation behind this research and the main objective of the thesis. We present an overview of our evaluation framework, the novelty of our approach to evaluate the intelligibility and speaker similarity and an application based on our evaluation results. We present the outline of the thesis at the end of the chapter.

1.1 Voice Transformation Techniques and Applications

Voice transformation refers to a class of techniques that modify the characteristics of the speech such that the identity of the actual speaker is concealed or the actual speakers voice is modified to resemble another speakers voice. Years of research in the area of voice transformation has resulted in a number of techniques to achieve high quality voice transformation. These include pitch synchronous overlap and add, vector quantization by code book mapping and Gaussian mixture models [1] [2] [4] [6] [8] [9]. The applications of voice transformation in various fields include

Automatic dialogue replacement in movies and TV shows: Movies and TV shows are frequently dubbed into other languages. Voice transformation can replace the voices of the actors in any target language.

Cross language voice conversion: Similar to dialogue replacement this is an ap-

plication where the same persons voice is generated in a language that the person cannot speak.

Text-to-speech (TTS): TTS itself being a stand- alone technique, natural human like voices or voices resembling a specific person can be generated by using voice conversion to the persons voice. This type of application requires very few training samples from the particular speaker.

Concealing the speaker identity: To protect the audio identity of the speaker and to protect his/her privacy voice transformation techniques are applied to obfuscate the characteristics of the speakers identity.

Healthcare applications: Voice transformation is also applied to generate voice samples for training and rehabilitation of people suffering from various voice disorders.

When voice transformation techniques are applied on a speaker's voice and the voice characteristics are transformed to another target speaker's voice, it is called voice conversion. Mathematical or statistical models of the source speaker's voice and target speaker's voice are constructed. Then a mapping function from the source to the target model is built. This mapping function is used as the voice conversion function. Different techniques of voice conversion work with one or more characteristics of speech and aim at building a mapping function that can produce a converted voice which resembles the target speaker most accurately.

When voice transformation is performed on the source speaker's voice with no specific target speaker, then there is a transformation function applied on the source speaker's voice model such that the speaker's identity is modified. As specified earlier voice transformation techniques aim at modifying the voice characteristics or speaker

identity of a speaker. These voice transformation techniques need to be evaluated so as to determine the effectiveness of the applied technique.

The voice transformation techniques need to be evaluated so as to determine the effectiveness of the applied transformation. But there are some conditions which pose a technical challenge to building a standard methodology for evaluation of voice transformation techniques.

The effectiveness of noise removal in speech or speech compression is frequently evaluated with techniques like “Mean square error” (MSE) and “Signal to Noise Ratio” (SNR). The original and the processed speech are compared in this case. But similar approaches cannot be used to evaluate voice transformation techniques to compare the original and the transformed speech.

Concealing the speaker’s identity is based on a number of speech characteristics like tone, timbre, prosody and rhythm of speech. When an attacker tries to access a voice transformed speech a minor speech sample from the family members might be sufficient to identify the person.

Speaker similarity evaluation is also a function of the human auditory system. Similarity seen in the waveform domain of speech may not necessarily mean a similarity in the quality of speech.

1.2 Evaluation Framework of Voice Transformation

Given the technical challenges in evaluating the voice transformed speech, the motivation behind this research is to build an evaluation framework for evaluating intentional voice transformation techniques. The evaluation framework is based on two

fundamental qualities of the transformed speech: intelligibility and speaker similarity.

Intelligibility refers to the clarity of the content of the modified speech. As many of the voice transformation techniques alter the speech characteristics, intelligibility is a quality of primal importance. This evaluates whether the modified speech is still useful with regard to the content of the speech.

In this research we study the intelligibility of “intentional” modifications. Speech could be less intelligible due to the speakers physical characteristics like voice disorders or voice imperfections. Intelligibility studies include factors like noise, attenuation occurring during transmission, telephonic or mobile transmission of speech. But in our research we focus on voice transformation techniques that are specifically applied to conceal the identity of the speaker. In those cases, when the motive is to alter the speaker identity, the modified speech should still be useful and be understood.

Intelligibility is measured in terms of accuracy or word error rate of the speech that is recognized or understood correctly. The accuracy of speech can be evaluated objectively or subjectively. Automatic speech recognizers are widely used to automatically evaluate the intelligibility of speech. CMU Sphinx [11], HTK [12] are some widely used speech recognizers.

Speaker similarity refers to the effectiveness of the transformation technique to disguise the actual speakers voice identity. This is also of a great importance when we use the voice transformation techniques to conceal the identity of the speaker.

Speaker similarity, in the context of concealing the identity of the speaker, is a measure of how similar the modified speech is to the actual speaker. In applications where the objective is to convert the speech to resemble a target, it is a measure of

how close the modified speech is to the target speaker. Effective voice conversion techniques which very closely resemble the target speaker are best suited for the applications discussed previously but on the other hand these techniques pose a threat to speaker authentication verification applications.

Voice transformation techniques with the objective of concealing the speaker identity or resembling a specific target speaker are typically evaluated with subjective tests with human listeners. ABX tests and opinion tests are most widely used evaluation techniques. In ABX test, a test speech sample X is rated by choosing whether it resembles the source speaker A or the target speaker B. In opinion test, the listeners usually rate the quality of the modified speech based on specific questions on a given scale.

1.3 Contributions of thesis

In our research we propose an evaluation framework that evaluates two voice transformation techniques: pitch distortion and voice conversion to demonstrate the effectiveness of these techniques in maintaining the intelligibility of the modified speech and concealing the identity of the speaker. Protecting the audio privacy of the speaker is an important concern in many audio applications. Medical therapy sessions might be recorded for medical records but the privacy of the patient is an important concern in this case. Surveillance cameras, tapping of phone lines for security are very common these days but the audio privacy of the speakers need to be protected from the attackers.

A complete evaluation framework with both subjective experiments and auto-

mated tools for objective evaluation for privacy and intelligibility is yet to be undertaken. In our research we modify the speech data with pitch distortion [30] and voice conversion [31] techniques. We have performed extensive experiments for both intelligibility and speaker similarity. Intelligibility is evaluated with automatic speech recognizers and listening experiments by human testers. Speaker similarity is evaluated with automated speaker identifier and listening experiments by human testers. The detail discussion of the techniques and the automatic tools is done in Chapters 3 and 4.

The novelty of our approach is to study the performance of the techniques in situations where the modified speech data is available to the automatic tools and human testers. This assumption and the evaluation provide a comprehensive and objective approach in measuring any voice transformation techniques.

In addition, as a novel application of voice transformation, we propose a system for voice therapy of people suffering from voice hypertension using video self-modeling technique. Video self-modeling is a technique based on behavioral therapy where the learner watches a video that models an unseen behavior and this provides a feedback of the improved behavior. We use this modeling technique to build a video sequence for people with voice hypertension. This video depicts the speaker with an improved speech track. The voice transformation techniques discussed previously are used to generate this new speech track. The effectiveness of this approach and the results of the study are discussed in Chapter 6.

1.4 Organization of the Thesis

After providing the introduction and background of the research in this chapter, the rest of the thesis is organized as follows. The related research in the areas of voice transformation, their evaluation methodologies, speaker identification and video self-modeling studies are discussed in Chapter 2. Chapter 3 describes the voice transformation techniques and the how they modify the voice characteristics of the speaker. Chapter 4 presents our evaluation framework and the automatic tools that constitute our framework. Experiments for privacy protection and its results are discussed in Chapter 5. Chapter 6 discusses the voice therapy study using video self-modeling, the experiments conducted and its results. conclude the research in Chapter 7 with a summary of the results and the observations from them and present the suggestions for future work.

Chapter 2

Related Work

In this chapter we discuss the research works related to speech transformation, audio privacy protection, methods of evaluation and studies for video self- modeling. The first section discusses the various voice transformation techniques and their evaluation methods. The second section discusses the techniques applied to protect the audio privacy of the speaker and their evaluation methods. In last section we discuss the works related to video-self modeling and its applications for behavioral therapy.

2.1 Evaluation of speech transformation techniques

Research in the area of voice transformation over the past years has given rise to various approaches which modify different characteristics of human speech [1]- [9]. Pitch detection of the source and target speakers and voice conversion by PSOLA method is discussed in [1]. The voice conversion algorithm is applied to English and Arabic speech and the results are compared. The voice conversion is demonstrated with time and frequency plots of actual and the voice converted waveforms at frame level and complete speech level. Pitch conversion method for voice conversion is discussed in [2]. Pitch extraction, pitch mark mapping, pitch scaling and “Pitch Synchronous Overlap and Add” (PSOLA) are performed on male and female speech data. The results of pitch conversion are demonstrated with pitch contour plots.

Voice transformation based on residual prediction methods for voice conversion is presented in [3] and [4]. In [3], the authors discuss and compare the different residual

predictions and evaluate the results subjectively using ABX test and “Mean Opinion Score” (MOS) tests. In [4], the authors discuss a GMM based residual prediction method using residual codebook mapping. The authors evaluate the performance subjectively using ABX tests.

In ABX tests, an unknown sample, a test speech sample in this case, is compared with two samples A and B and is rated as being similar to A or B. In listening experiments, A is normally original speech or a speech sample from the source speaker and B is the speech sample of the target speaker. The experiments are conducted with test takers who are not aware of the processing techniques and they rate the speech samples by choosing whether they sound similar to the source or the target speaker.

MOS or Mean Opinion Score is another subjective evaluation technique using listening experiments in which the quality of the transformed speech is evaluated. The quality is rated on a numerical scale typically from 1 to 5 with 1 as the worst and 5 as the best quality. The mean score is used to decide the quality of the transformation technique.

In [5], the authors propose a voice conversion technique based on spectral mapping and residual prediction. The results of voice conversion are evaluated through listening experiments. In [6], the authors propose Gaussian Mixture Model (GMM) based dynamic frequency warping of the spectrum to avoid the over-smoothing in baseline GMM methods. Additionally they propose weighted residual spectrum to avoid the conversion accuracy deterioration on speaker individuality. The conversion accuracy is objectively evaluated using cepstral distortion (CD) between the converted and

target speech. Subjective evaluation is performed with ABX tests.

In [7], the authors propose a vocal tract parameter estimation technique for high quality voice conversion. Spectral comparison, time domain waveform comparison and SNR are used to evaluate the performance objectively. ABX tests through listening experiments are used for subjective evaluation. Eigen voice conversion based on GMM is proposed in [8]. Cepstral distortion is used to measure the performance of the voice conversion algorithm. Voice conversion through nonparallel training based on GMM parameter adaptation is proposed in [9]. The authors compare the voice conversion with parallel training and non-parallel training. The results are objectively evaluated using mean square error measure and by using a GMM based speaker identifier. Subjective evaluation is also performed using ABX tests.

As we can see, in most of the voice transformation techniques discussed above, the quality of the transformed voice or the effectiveness of the technique is evaluated using either subjective listening experiments or some objective measures like distortion. Only in [9], objective measurement is performed using automated speaker identifier. But it has been demonstrated in [10] that both subjective methods and automatic tools are necessary to show the effectiveness of intentional voice modification schemes.

We can also note that these techniques aiming at voice conversion evaluate only the quality of the converted speech in terms of its naturalness and its similarity to the target voice. There is no evaluation for the quality of the speech content or whether it remains unmodified due to the transformation.

The objective measures used in the works cited above are based on the characteristics of the waveform and measure the difference between the source and the

transformed speech or transformed and the target speech. The objective measures do not evaluate the quality of the resulting transformed speech or how natural they sound to the ear. The naturalness and the quality are evaluated only with subjective experiments.

Human listeners have their own limitations in terms of gender, race and age. Also large data sets cannot be evaluated efficiently with human listening experiments. This stresses the need for an objective or automatic evaluation techniques which are computationally more efficient than human listeners. Additionally the evaluations by automated techniques should be supported by subjective experiments to take advantage of the fact that though automated techniques are efficient computationally, the natural auditory capability, understanding and perception are best in the case of human and can only modeled towards perfection in the automatic techniques.

2.2 Speaker privacy protection

In this section we discuss the voice transformation techniques that aim in concealing the voice identity of the speaker. In [13], the authors propose a voice morphing technique that can hide the speaker’s gender, age and identity. This technique is based on spectral mapping and residual prediction of syllables. In this case also the performance of the algorithm is evaluated with listening experiments using ABX and MOS tests. The evaluation for speaker identity being morphed is not performed in this case. In [14], the authors propose a complete voice morphing system to overcome the voice artifacts generated by spectral mapping systems and thus improving the quality of the morphed speech. The evaluation is performed subjectively to compare

the baseline systems and the enhanced system. ABX tests and preference tests are performed to judge the quality of the modified speech.

Privacy protection linked to speech recognition corpora content is studied in [15] where generalization and iterative distortion are performed to protect sensitive data. The results are evaluated with automatic speech recognition tools but no speaker identification testing is conducted. Privacy protection of speech content and anonymity of the speaker/patient using keyword spotting and replacement is proposed in [16]. This preliminary work is specifically designed for clinical data and the performance evaluations are still underway.

In [17], the authors analyze a pitch shifting algorithm that distorts the speech and the speaker’s identity but maintains the speech content. The evaluations for the clarity of the speech and the privacy of the speaker are evaluated with only subjective listening experiments. Recently, voice transformation algorithms and their effectiveness in protecting the privacy of the speaker are studied in [18]. The performance on protecting speakers’ identity is measured using both subjective testing and a GMM and phonetic based speaker identification system. The clarity of the speech, however, is measured only by human testers.

A complete evaluation study using both subjective testing and automatic tools for a wide range of voice transformation tools for privacy protection has yet to be conducted.

2.3 Video self-modeling

Watching video to learn or model a target positive behavior is in fact a proven technique in behavior therapy. This is called the Video Modeling (VM) intervention, widely used for rehabilitation and education of patients recovered from surgery [19] and cancer [20] as well as job and safety training for hospital staffs [21] and office workers [22]. VM is also effective in a school setting to teach children and young adolescents various skills including social interactions, communication, self-monitoring and emotional regulation [23].

Besides watching others, we can also watch and learn from our own positive behaviors. Such form of self-modeling is classically done with a mirror and one of the most prominent examples is the use of the “mirror box” in treating phantom limb pain among amputees [24]. Seeing or visualizing oneself accomplishing the target behavior provides the most ideal form of behavior modeling. This is particularly true to one of the major groups of VM users, namely children with autism, as it has been shown that only face images of themselves rather than others can trigger the appropriate neural response in their brain [25]. With the use of a camcorder and video editing software, clips of positive behaviors can be spliced together to create video material for behavioral modeling.

Compared with traditional VM, researchers have argued that such form Video Self Modeling (VSM) is more effective at boosting confidence, capturing and maintaining attention as well as shaping positive memories of the learner [26]. Though still in its early development, effectiveness of VSM has been studied for many different types of disabilities and behavioral problems ranging from stuttering, inappropriate social

behaviors, autism, selective mutism to sports training. A summary of these researches can be found in [26].

There are two forms of VSM: positive self-review and feed forward [27]. In positive self-review, the portions of the recorded video showing bad or poorly executed routines are removed leaving only the positive target behaviors. The resulting video will be reviewed to enhance fluency of the skills that have already been acquired by the learner but not yet perfected. On the other hand, the feed forward VSM shows novel skills that have never been observed but still within the reach of the learner. The goal is to teach new skills to a learner. An example of feed forward VSM can be found in [26] where he splices short clips to form a long video of a full sentence from a child who can only speak one or two-word utterances. It is the feed forward approach that shows more dramatic learning effect than the positive self-review approach as there is more room for improvements during the initial stages of learning.

With the given researches using VSM for various types of learning therapy, applying this VSM technique for voice therapy applications for people with voice hypertension is a novel idea. The effectiveness of the VSM approach and its potential to reduce the length of the treatment and the number of therapy sessions is a yet to be studied.

Chapter 3

Speech Transformation

In this chapter we discuss the voice transformation techniques used in our evaluation framework. The main objective of our research is to compare and evaluate the performance of two voice transformation techniques in maintaining the intelligibility of the modified speech but conceal the identity of the source speaker. In this research we have demonstrated the performance of the two voice transformation techniques using the proposed framework. But the framework is not limited by the voice transformation technique used and can be used to evaluate the performance of any transformation techniques. We initially present the theory behind the selection of pitch distortion and voice conversion as the voice transformation techniques for our evaluation. We then describe the mechanism through which the pitch distortion and voice conversion algorithms alter the speech characteristics. At the end of the chapter we illustrate the modifications in the voice characteristics with samples of actual and transformed speech.

3.1 Why pitch distortion and voice conversion.

Source - filter model of speech [28] is an approximation model for the speech production process. According to this theory, airstream from the lungs is modulated by the vocal folds vibration and this acts as the source of sound. The vocal tract acts as a filter for the sound source and by taking different shapes produces different

sounds. So the source filter theory models the source as a series of impulses which pass through a linear filter to produce speech.

In [29] the author provides a survey of the various voice transformation techniques. According to this source-filter model, the author classifies the voice transformation techniques as source modification techniques, filter modification techniques and combined source and filter modification techniques.

1. Source modification techniques are prosodic modifications and include time scale modifications, pitch modifications and energy modifications.
2. The magnitude spectrum of the frequency response of speech carries the speaker identity. Filter modification techniques alter the magnitude spectrum and these modifications can be general or towards a specific target speaker.
3. Combined source and filter modification techniques are used in voice conversion to a particular target speaker and they combine both the prosody and vocal tract modifications.

For our evaluation framework we have chosen one source modification technique and one filter modification techniques - pitch distortion [30] and voice conversion based on vocal tract mapping [31] respectively. We compare the performance of these techniques in maintaining the clarity of speech while concealing the identity of the speaker.

3.2 Voice Transformation

Before we discuss the voice transformation techniques, here we try to illustrate how speech is modified when voice transformation techniques are applied and its impact on

intelligibility and speaker identity. The illustration provides a better understanding in terms of the specific characteristics of speech that determine the speech content, speaker identity and also how the transformation affects them. This also helps in understanding the theory behind the automatic techniques of recognition. Figure 3.1 shows a sample speech waveform in time domain corresponding to “*By the look of him he wasn’t that far gone*”. It shows the original speech on top, pitch distorted speech for $\alpha = 0.75$ in the middle and female to female voice converted speech with $\alpha=0.9$ and $\rho=0.98$ at the bottom. The segment highlighted corresponds to the word “*He*” and is extracted and shown in Figure 3.2.

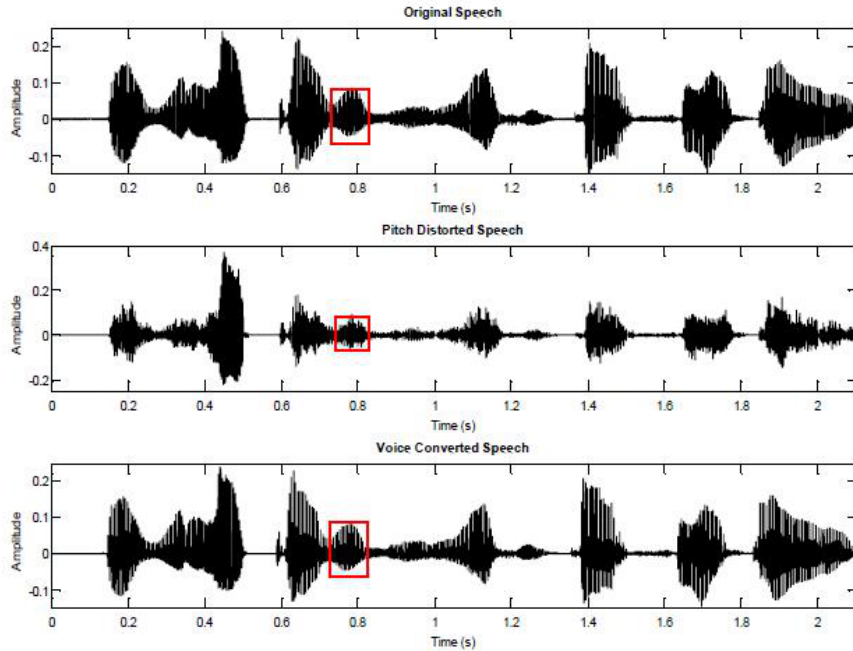


Figure 3.1: Speech signal of “*By the look of him he wasnt that far gone*”.

Figure 3.1 and 3.2 highlight the modification of the speech signal in the time domain. Both pitch distortion and voice conversion waveforms differ from the original speech. In addition we see that pitch distorted speech shows higher degree of modification compared to voice converted speech. Figure 3.3 shows the spectrum of the

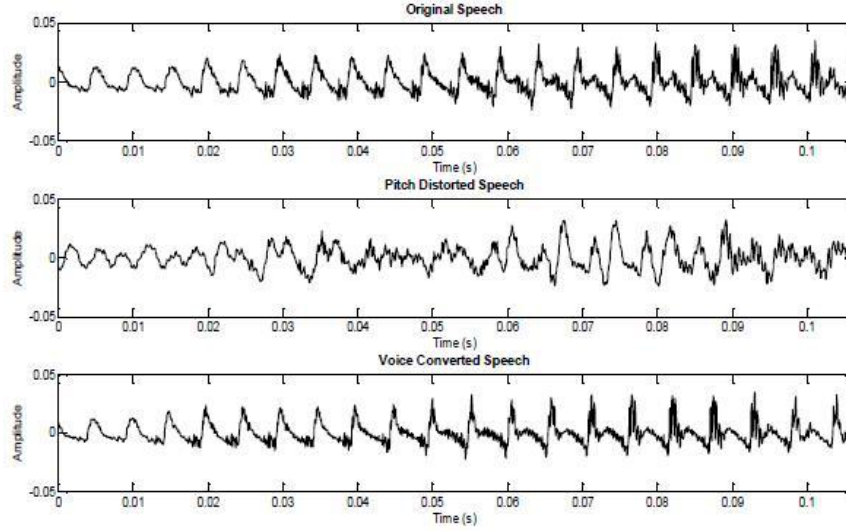


Figure 3.2: Speech signal of “He”.

word “He”. The frequency spectrum gives the frequency content of the speech signal and the spectral envelope of the speech. Comparing the shape of the spectrum and the spectral envelope, again we can see that pitch distorted speech shows a higher modification than voice converted speech. The speech waveforms are modified due to pitch distortion and voice conversion. Though the speaker identity is modified, they still represent the word “He”. The specific characteristics or features of speech which determine the speech content and the speaker identity are discussed below.

The spectral envelope and its formants or resonant frequencies are the key for recognizing the phonemes and the speaker characteristics. Vowels and consonants are associated with a particular frequency pattern and frequency range. This pattern and frequency range is used in identifying the corresponding phonemes. Similarly the formant frequencies itself are unique for a speaker and are used in identifying a speaker. The figure below shows how the formant frequency is associated with both the phonemes and the speaker characteristics.

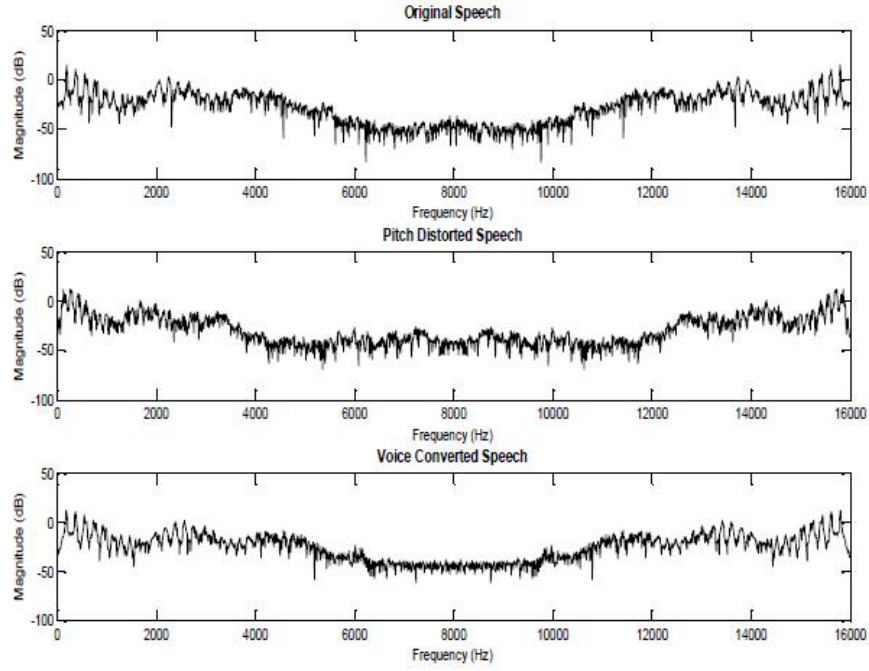


Figure 3.3: Spectrum of the word “He”.

Figure 3.4 illustrate the formant peaks for male and female speakers for the vowel /ae/ and /i/ respectively. Table 3.1 shows a list of average formant frequencies for

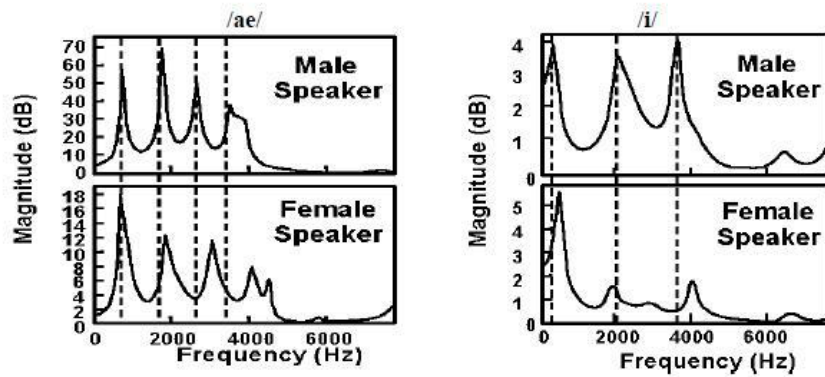


Figure 3.4: Male and female formant spectrum [32].

some basic vowels for male and female [32]. In the case of voice conversion, the locations of the formant frequencies get modified and so the speaker’s identity is modified. But the spectral shape is preserved thus preserving the content of the

Table 3.1: Average formant frequencies for some basic vowels

Vowel	Gender	F1	F2	F3
ee	Male	270	2290	3010
	Female	310	2790	3310
ae	Male	660	1270	2410
	Female	850	2050	2850
oo	Male	300	870	2240
	Female	370	950	2670

speech. On the other hand, in pitch distortion, the degree of modification in the speech wave form is more compared to the original signal. Since the pitch of the signal is modified, the identity of the speaker gets modified. The speech content depends on how well the spectral envelope is preserved. The experimental results of the techniques are discussed in Chapter 5.

3.3 Pitch distortion Technique

The pitch distortion technique comprises of two steps time stretching and resampling by “Pitch Synchronous Overlap and Add” (PSOLA) method.

Time Stretching: Expands or contracts the speech signal without perceptually modifying the signal. Time stretching operates on the number of samples of the speech waveform and it replicates or discards samples to expand or contract the speech signal. Time stretching is represented as

$$M' = \alpha \cdot M \quad (3.1)$$

In the above equation, M is the number of samples in the actual signal, M' is the number of samples of the time stretched signal and α is the time stretching factor.

Resampling: Resampling or time scaling or varying the speed of the signal

increases or decreases the actual time duration of the speech i.e. the original speech is played within a shorter or longer duration of time. This modifies the frequency and thus the perceived pitch of the signal. Time scaling is implemented by re-sampling the signal at a higher or lower rate and is represented as

$$n \cdot T_{out} = n \cdot T_{in}/v \quad (3.2)$$

$$f_{out} = f_{in} \cdot v \quad (3.3)$$

In the above equation, n is the number of samples, T_{in} and T_{out} are the input and output sampling time periods, f_{in} and f_{out} are the input and output sampling frequencies and v is the speed varying factor

The pitch distortion algorithm can be depicted as in the Figure 3.5 below.

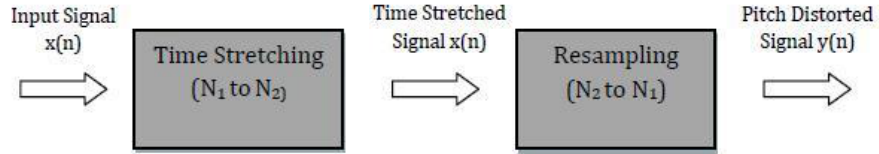


Figure 3.5: Pitch distortion algorithm

Time Stretching Module - Step by Step: Time stretching expands or compresses the input signal from length N_1 to N_2 .

1. The speech signal is divided into blocks of size N with a shift of S_a as shown in Figure 3.6 below.
2. The blocks are reshifted to a length $S_s = \alpha \cdot S_a$, where $\alpha = N_1/N_2$, the stretching factor as shown in Figure 3.7.

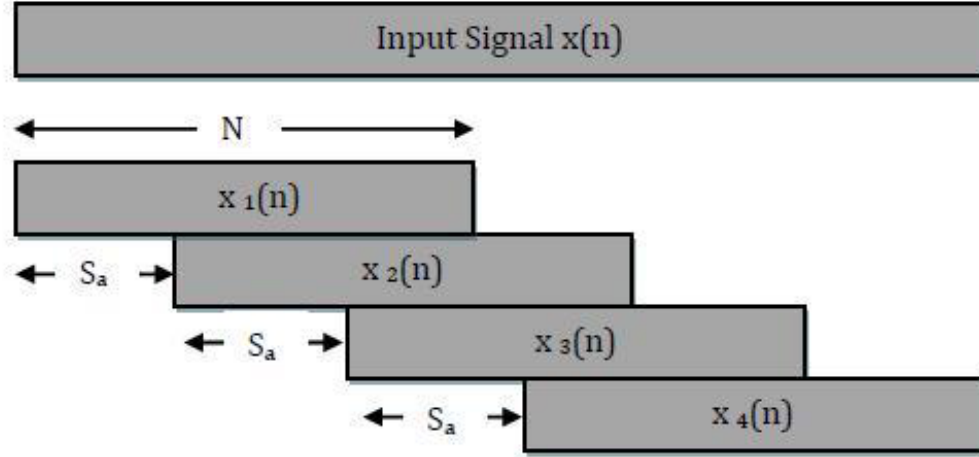


Figure 3.6: Step 1 of Time stretching

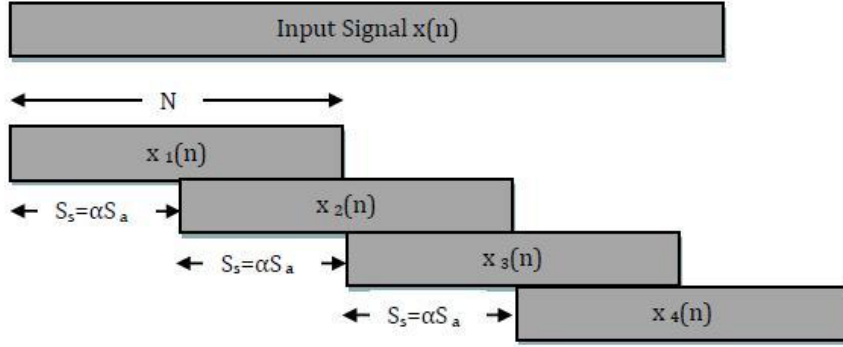


Figure 3.7: Step 2 of Time Stretching

3. Cross correlation between the two blocks is computed to find the point of maximum similarity in the overlapping region as shown in Figure 3.8. This is given by

$$r_{x_{L1}x_{L2}}(m) = \frac{1}{L} \sum_{n=0}^{L-m-1} x_{L1(n)}x_{L2(n+m)} \quad , 0 \leq m \leq L \quad (3.4)$$

In equation 3.4 and are the speech segments $x_{L1(n)}$ and $x_{L2(n)}$ and L is the overlap interval between the segments.

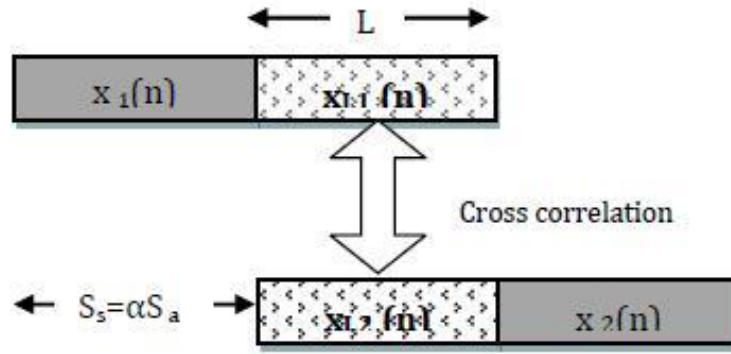


Figure 3.8: Step 3 of Time Stretching

4. The point of maximum cross correlation is calculated as k . The segments are again shifted to this point k as shown in Figure 3.9.

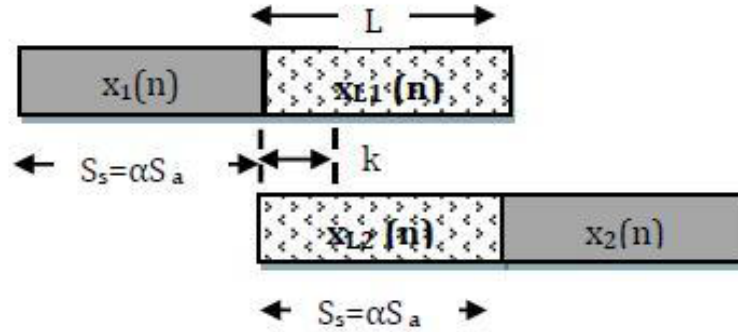


Figure 3.9: Step 4 of Time Stretching

5. The overlapping regions are processed as follows - the first segment is weighted by a fade out function from the point k to the end of the segment and the second segment is weighted by a fade in function from the beginning of the segment to the end of overlapping region. The fade-in, fade-out functions are used to avoid transients in the signal. This process is shown in Figure 3.10.

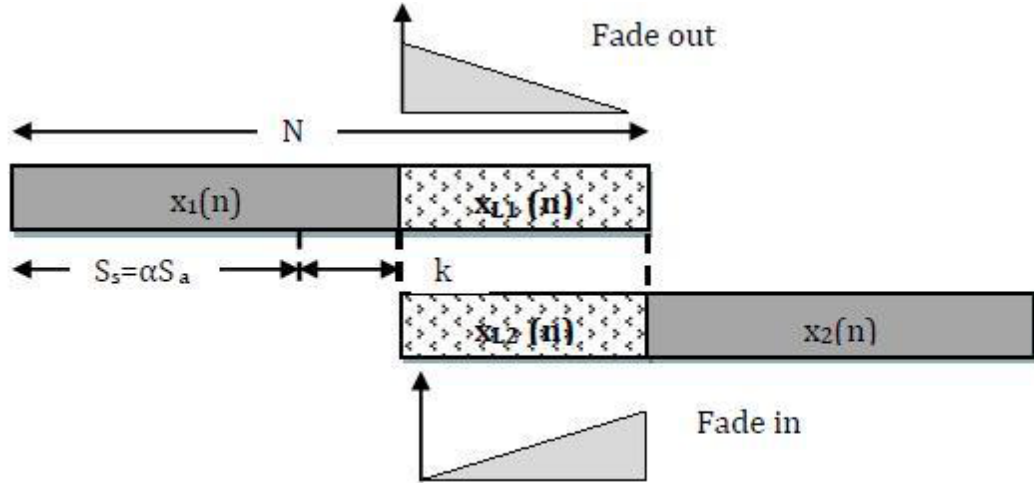


Figure 3.10: Step 5 of Time Stretching

6. The steps 3 to 6 are repeated for every pair of segments.
7. Finally the faded in faded out regions are added together ,sample by sample, for each pair of segments as shown in Figure 3.11.

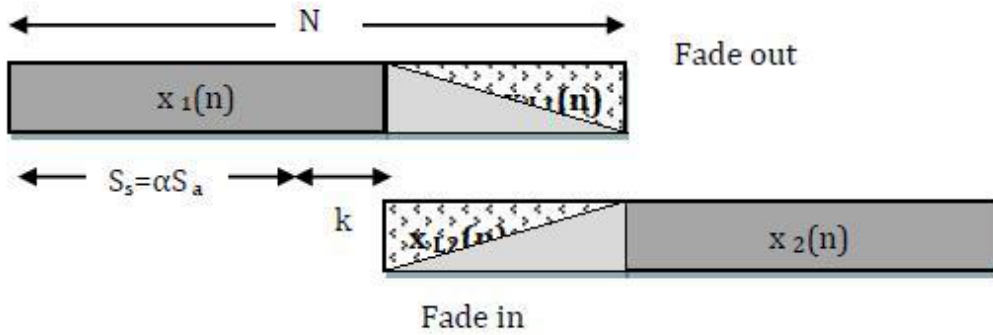


Figure 3.11: Step 11 of Time Stretching

Resampling Module - Step by Step: Resampling distorts the pitch of the speech signal and makes the signal back to the original length N_1 .

1. The time stretched signal from the previous step is resampled back to the orig-

inal length of the signal $x(n)$. Hanning window is used to segment the time stretched speech signal.

2. The segments are resampled to reduce the length i.e., samples are dropped at every pitch period so the pitch of the speech signal get distorted. The segments are then added to produce pitch distorted speech signal.

3.4 Voice conversion algorithm

Voice conversion transforms a source speaker's voice to a target speaker's voice using "Vocal Tract Length Normalization" (VTLN) technique. Vocal tract length is unique for a speaker and forms a major characteristic of the voice identity. The vocal tract length and shape impacts the spectral envelope and the formant structure which in turn influence the phonetic content of the speech signal. The voice conversion technique normalizes the vocal tract length by warping the frequency axis of the source speaker towards the target speaker thus modifying the speaker's identity.

The algorithm has training and testing phase. In the training phase the source and the target speech samples are analyzed and the warping factor, α and fundamental frequency ratio, ρ are estimated and warping function computed. In the testing phase the warping function is applied to the source speech samples and converted to target speaker's voice. The working of the algorithm for a linear warping function is given below.

In the training phase,

1. Pitch synchronous frames are extracted from the source and the target speech

using short duration windows and frequency spectrum of the signals are estimated.

2. The source and the target spectra are assumed to be piece-wise linear and the warping parameters are estimated in the frequency domain as follows.

If ω is the normalized frequency of the source spectrum and $\hat{\omega}$ is the normalized frequency of the target spectrum, then the warped frequency is given by [31]

$$\hat{\omega}(\omega) = \alpha_i \omega + \beta_i \quad , \omega_i \leq \omega \leq \omega_{i+1}; i = 0, \dots, I$$

$$\alpha_i = \frac{\hat{\omega}_{i+1} - \omega_i}{\omega_{i+1} - \omega_i}$$

$$\beta_i = \hat{\omega}_{i+1} - \alpha_i \omega_{i+1}$$

The equation can be rewritten as

$$\hat{\omega}^{-1}(\omega) = \sum_{i=0}^I \frac{\omega - \beta_i}{\alpha_i} R(\alpha_i \omega + \beta_i \mid \omega_i, \omega_{i+1})$$

where R is a rectangular function defined in terms of $\omega, \omega', \omega''$.

The warped spectrum is given by,

$$\tilde{X}(\omega) = \sum_{i=0}^I X\left(\frac{\omega - \beta_i}{\alpha_i}\right) R(\alpha_i \omega + \beta_i \mid \omega_i, \omega_{i+1}) \quad (3.5)$$

The warped spectrum in the time domain is the inverse Fourier transform of equation of equation (3.5).

$$\tilde{x}(t) = u(t) * r(t) = F^{-1}\left\{X\left(\frac{\omega - \beta_i}{\alpha_i}\right)\right\}(t) * F^{-1}\{rect\omega\}(t) = \alpha e^{i\beta t} x(\alpha t) * r(t) \quad (3.6)$$

In the testing phase,

1. Pitch synchronous frames are extracted from the source speech samples, to be converted, using short duration windows.

2. The segments are warped to the target speaker's voice using the warping parameters estimated in the training phase and are then combined back into transformed speech signal.

The voice conversion technique and the warping of spectrum is illustrated in Figure 3.12

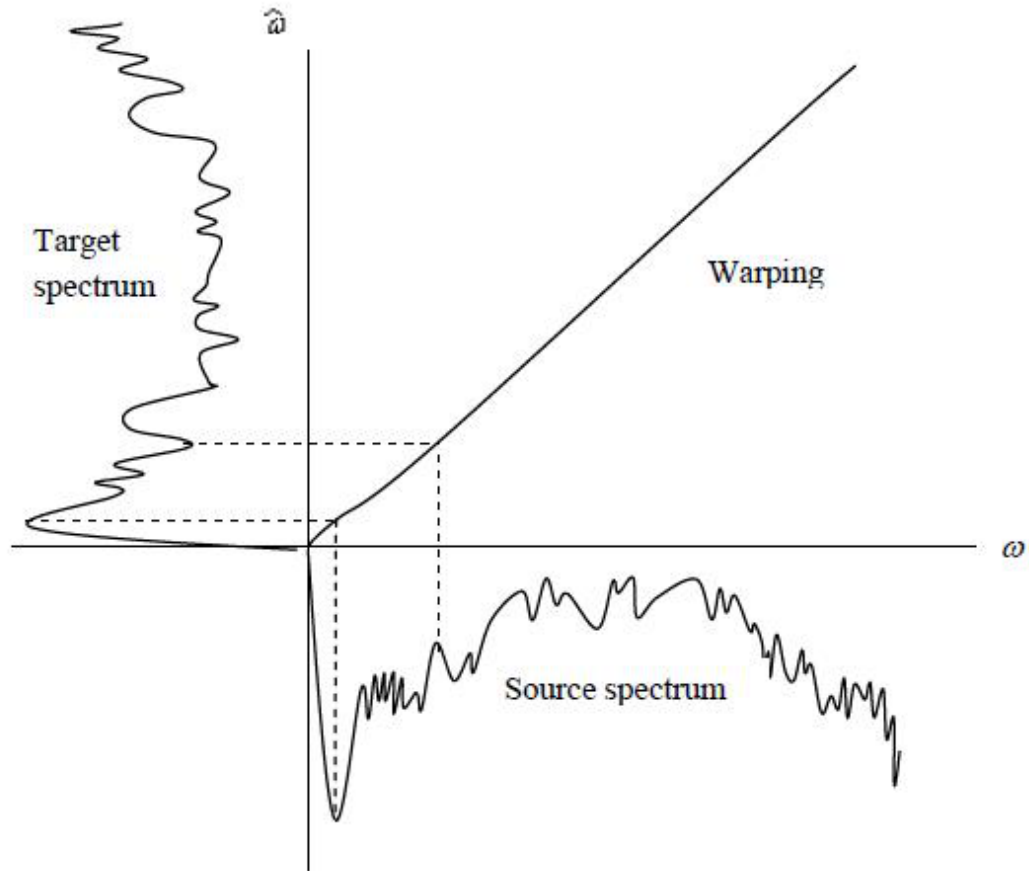


Figure 3.12: Warping function from two spectra

Chapter 4

Evaluation Framework

In this chapter we present the overview of our evaluation system and discuss each of its components in detail. We begin with the automatic speech recognition system, the mathematical model behind it, the mechanism of speech modeling using “Hidden Markov Models” (HMMs) and decoding of the speech samples into text. Similarly we discuss the speaker identification system, the methodology involved in training the system using mixture models and speaker identification by computing the likelihood ratio. We also illustrate how the voice transformed data is correctly decoded by speech recognizer but conceals the speaker identity and evades the speaker identifier.

4.1 System Overview

Figures 4.1 and 4.2 shows the block diagram of our evaluation framework. The system used in training is shown in Figure 4.1 and testing is shown in Figure 4.2. The voice transformation has been discussed in detail in Chapter 3. Intelligibility or clarity of transformed speech is evaluated with automatic speech recognition engines and its performance is measured in terms of “Word Error Rate” (WER) of the decoded speech. Speaker similarity is evaluated with automatic speaker identification system and the similarity to the source speaker is measured in terms of likelihood to the speaker models available previously. These automatic systems are discussed in detail in the following sections.

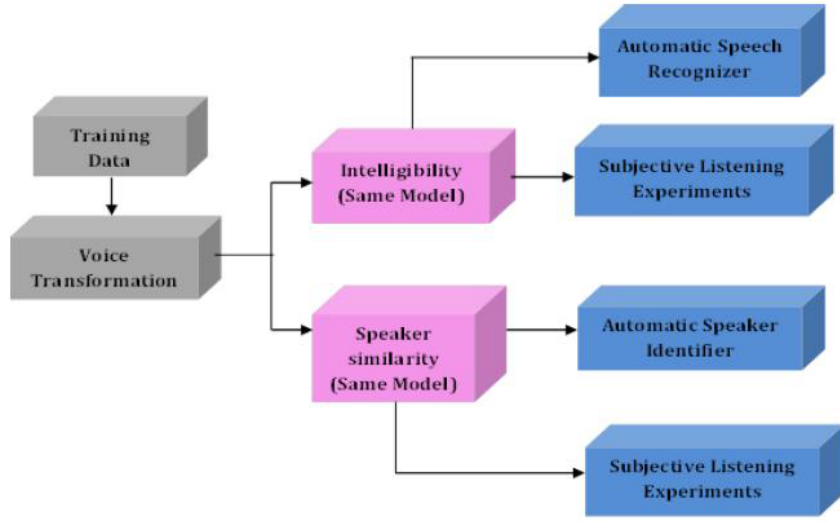


Figure 4.1: Training system of Evaluation Framework

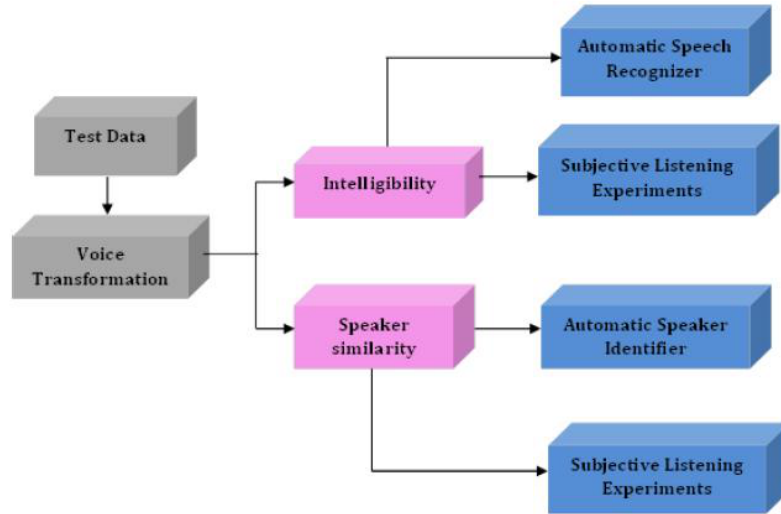


Figure 4.2: Testing System of Evaluation Framework

4.2 Novelty of the Evaluation framework

The novelty of this research work in evaluating the voice transformation techniques are

4.2.1 Automatic recognizers trained for Same Model and Cross Model

While evaluating the modified speech data using automatic speech recognizer and speaker identifier, we consider two cases.

1. The recognizer is trained with the original speech data and then tested with pitch distorted and voice converted data. With this set up, intelligibility and speaker similarity are measured.
2. The recognizer is trained with modified data for each type of transformed speech. In the testing phase again pitch distorted and voice converted data are used (with corresponding training data) and results are evaluated. This new approach where the recognizers have knowledge about the transformation parameters, provided us with some interesting results for both intelligibility and speaker similarity.

4.2.2 Subjective testing

The automatic recognizers are initially trained with a set of speech data. So the performance of the recognizers depends on the training data. In addition to the input speech the results are based on the probability determined by the acoustic and language models. On the other hand, human beings naturally have knowledge of the grammar, vocabulary and dialects of any particular language. To study this difference, we have performed subjective evaluations with listening experiments for both intelligibility and speaker similarity. Comparing the subjective test results with those of automatic techniques shows interesting characteristics of human listeners.

4.2.3 Subjective experiments with Same Model and Cross Model

Automatic recognizers have an advantage of taking modified data for training and have been tested for performance with specific training data. Similar to this we have performed subjective experiments with human testers by making them listen to modified data and then evaluate the performance of the transformation techniques. The results of experiments listening to actual data and modified data are discussed in the experiments section.

4.3 Intelligibility evaluation with Automatic Speech Recognition

4.3.1 Computing Intelligibility

Intelligibility refers to the clarity of the speech content and we use “Automatic Speech Recognition” (ASR) systems to evaluate clarity of the speech that has been modified by pitch distortion and voice conversion. Speech recognition is a technique that decodes speech into text/words uttered. Performance of a speech recognizer is specified in terms of accuracy which is measured in terms of “Word Error Rate” (WER) of the decoded content [34]. When comparing with the original text, lower the WER of the decoded speech, better the accuracy of the ASR. WER is computed in terms of number of insertions, substitutions and deletions in the decoded text compared to the original text.

$$WER = \frac{S + D + I}{N} \quad (4.1)$$

In the above equation S is the number of substitutions, D is the number of deletions, I is the number of insertions and N is the number of words.

Let us consider an example.

Original utterance: *“Scholastic aptitude is judged by standardized tests”.*

Recognized text: *“Scholastic aptitude is that by standardized tests”.*

The values of the parameters are $S = 1$, $D = 0$, $I = 1$ and $N = 7$ so the $WER\%$ is 28.57. The accuracy W_{Acc} is computed from the WER as

$$W_{Acc} = 1 - WER = \frac{N - D - S - I}{N} \quad (4.2)$$

The accuracy for the above example is 71.43

4.3.2 Speech Recognition

For recognizing the speech, the ASR system builds acoustic and language models with the training speech data during the training phase which is a learning phase for the recognizer [35] - [37]. These statistical models could be word level or phoneme level depending on the complexity of speech and are based on the grammar, the possible words and the probability of their occurrence in a sentence. So during the training phase, the recognizer builds acoustic models act as reference templates, and in the testing phase, the ASR acts as a classifier and decodes the test speech samples by computing the likelihood to the reference templates.

The training phase

1. Generates feature vectors from the input speech samples.
2. Builds acoustic models with these observed feature vectors.

and the testing phase

1. Generates feature vectors for the test speech samples.

2. Finds the most likely sequence of words (or phonemes) for this sequence of feature vectors to the acoustic models.

4.3.2.1 Generating feature vectors:

First we present some basic concepts involved in speech processing. Speech is a highly dynamic waveform with variations in pitch, rate of speech and speaker characteristics. Mathematical representation of speech with its variations is a challenging task. Statistical methods used currently to model speech, provide a close enough approximation of speech. To reduce the dynamics of the speech and to reduce the effects of noise, speech is analyzed in short time segments across which it is assumed to be stationary. Speech is segmented with short duration windows and transformed into a different domain in which the segments are represented as vectors named *features*.

Phonemes are basic unit of sound and smallest unit of speech. Each phoneme is associated with a particular vocal tract shape and each shape produces a set of resonant frequencies which are unique for a speaker.

The resonant frequencies are represented by feature vectors and can be displayed by computing the power spectrum of the speech signal. Power spectrum when computed in the logarithmic domain is called as *cepstrum*.

When processing speech, cepstral features are widely used and they model the spectral envelope of speech. The advantages of cepstral features include smaller number of features sufficient to represent the speech signal, features being uncorrelated with each other and ease of computation.

Human ear has different responses for different frequencies. Mel scale and Bark scale represent the perception of human ear at different frequencies. The feature

vectors computed using these scales are called “Mel Frequency Cepstral Coefficients” (MFCC) and “Perceptual Linear Prediction coefficients” (PLP) respectively. We also have “Linear Prediction Coefficients” (LPC) which is based on autocorrelation of speech.

Speech recognizers are built based on the above mentioned characteristics of speech. The ASR used in our experiments uses MFCC coefficients. The cepstrum has components from both the source and the filter of speech (based on the source filter model of speech production) i.e., spectral envelope and the excitation. The lower coefficients of the cepstrum model the vocal tract’s spectral envelope and so only the first few coefficients of the cepstrum are sufficient to form a template of phonemes for the recognizer. The Mel cepstrum is computed by applying logarithm to the magnitude spectrum, converting to Mel scale, followed by discrete cosine transform. The resulting coefficients are in a time-like domain called the *quefrequency* domain. The MFCC computation is shown in Figure 4.3

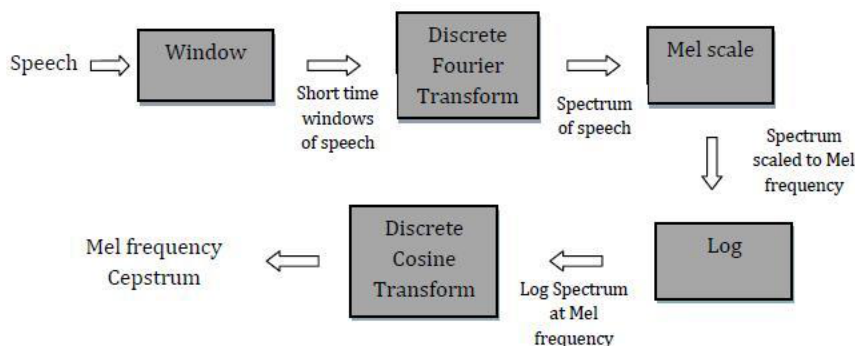


Figure 4.3: Mel frequency cepstrum computation

4.3.2.2 Building acoustic models from the feature vectors:

From the feature vectors extracted from the training speech data, the recognizer builds the acoustic model. As specified earlier, the feature vectors represent the resonant frequencies which in turn constitute different phonemes. So the acoustic model represents the phonemes and the probability of occurrence of the phonemes in the training speech data.

Hidden Markov Model is used to model this acoustic data which is doubly stochastic i.e., phoneme corresponding to a set of feature vector is stochastic and given a phoneme x , the phoneme that could possibly follow the phoneme x is also stochastic. The HMMs model the phonemes as states and the observed sequence of feature vectors as the outcome of the states. The outcome of the states and the transition between the states both are stochastic for the speech recognition problem.

At the end of the training phase, HMMs are generated and form the acoustic model for the recognition. The training phase, while building the acoustic model/HMMs, computes the probability of occurrence of every phoneme (or sequence of feature vectors) and the probability for transition between the phonemes. Figure 4.4 shows an example of HMM.

In Figure 4.4, $S_1, S_2, S_3, \dots, S_6$ are the states each existing at a time instant t , O_1, O_2, \dots, O_6 are the observed sequence of feature vectors, a_{ij} is the transition probability from state i to state j and $b_j(o_t)$ is the output probability distribution at time instant t and when state j is entered.

As specified before, the training phase builds HMMs and computes the probabilities. The training phase also computes the most likely state sequence for a given set

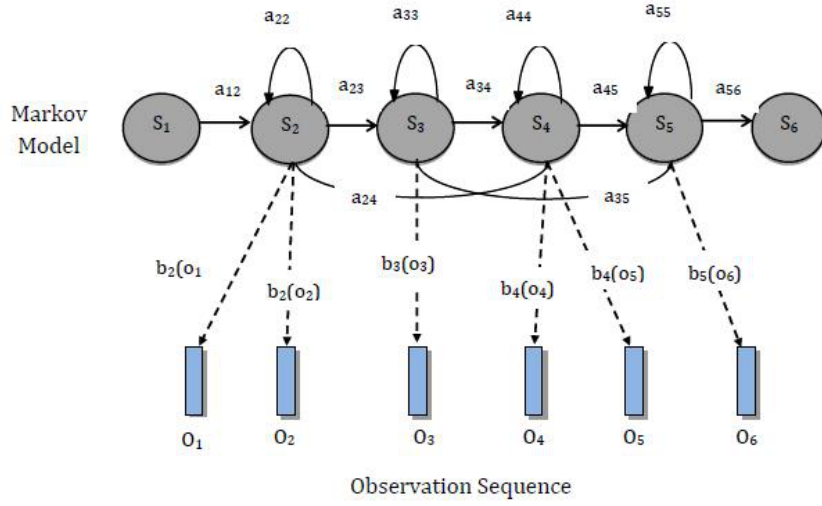


Figure 4.4: HMM Example

of observation vectors. If X is the unknown state sequence, then the likelihood for the observation O given the model M is $P(O | M)$ and is expressed as the product of transition probabilities and output probabilities.

$$P(O | M) = \max_x \{a_{x(0)x(1)} \prod_{t=1}^T b_{x(t)}(o_t) a_{x(t)x(t+1)}\} \quad (4.3)$$

Typically recognizers use 5 state HMMs per phoneme and these HMMs are concatenated to form a complete acoustic model for the entire training data. A concatenated two word HMM (at phoneme level) example for “*She carry*” is shown in Figure 4.5.

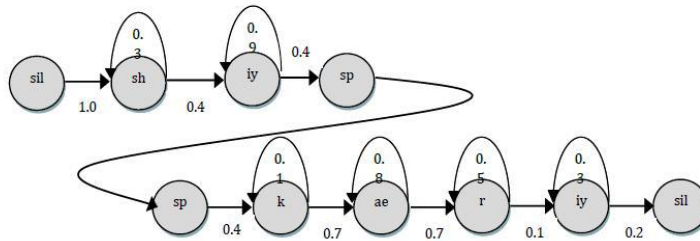


Figure 4.5: HMMs concatenation example

In addition to the acoustic model, a speech recognizer also uses a language model for its decoding. A language model is also a stochastic model but is related to the grammar of the language. It represents the probability of the words and the sequence in which they can occur in any sentence. The probability of any group of words is defined as,

$$p(W) = p(w_1, w_2, w_3, \dots, w_n) = \prod_{i=1}^n p(w_i \mid w_0, w_1, \dots, w_{i-1}) \quad (4.4)$$

In equation (4.4) W represents any group of words and w_1, w_2, \dots, w_n represent a sequence of n words that constitute W .

4.3.2.3 Decoding the test speech samples:

In the recognition phase, feature vectors are extracted from the speech samples to be recognized. With these feature vectors, decoding involves searching the HMM to find the state sequence from that has the maximum probability of occurrence. Using this state sequence and the language model, phonemes and words are decoded. If $O_T = o_1, o_2, \dots, o_n$ is a sequence of observed feature vectors, then the recognition phase involves finding the probability of a word W given the observation O is

$$\hat{W} = \arg \max_W p(W \mid O) = \arg \max_W \frac{p(O \mid W)p(W)}{p(O)} \quad (4.5)$$

In the above equation, $p(O \mid W)$ is the acoustic model and $p(W)$ is the language model.

4.3.3 Voice Transformation and Intelligibility

In Chapter 3 , we discussed the speech characteristics and how the voice transformation affects the characteristics. The spectral envelope, formant frequencies and their structure for different phonemes and for male and female were shown using the Figure 3.4

Here we relate the intelligibility of speech and the cepstral features. The phonemes are determined by the spectral envelope, its formants/resonant frequencies and their pattern. The feature vectors used by the recognizers model this spectral envelope of speech. Thus the initial coefficients of the cepstrum are used in building the mathematical or probabilistic model which is in turn used in the recognition by the automatic techniques.

Figure 4.6 shows the vowel cepstrum for male and female. The initial huge peak in the cepstrum corresponds to the spectral envelope of the speech and the periodic peaks correspond to the excitation of the speech. From the number of peaks in the female cepstrum it is evident that the female pitch is higher than the male.

4.4 Automatic Speech Recognizer - Julius

For our experiments we use Julius [39] large vocabulary continuous speech recognition engine. Julius is a two pass speech recognition engine and can work with acoustic and language models generated using “HMM Tool Kit” (HTK).

In our research we use HTK to generate the HMM definition files for acoustic and language models. With training speech samples, a dictionary and transcription file, HTK generates MFCC feature vectors and builds acoustic model. HTK also generates

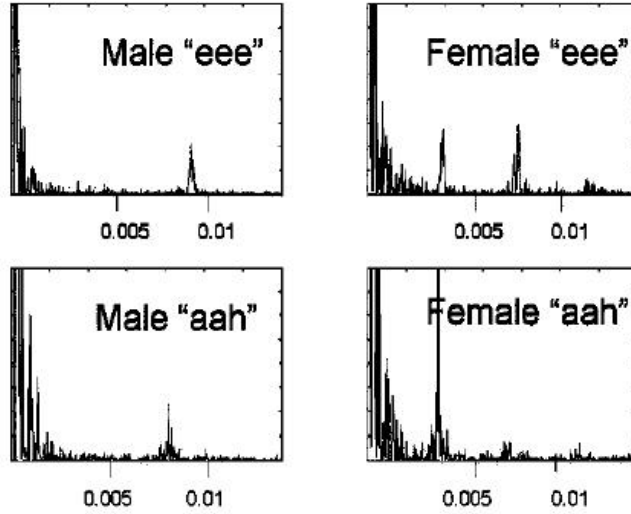


Figure 4.6: Male and female cepstrum for vowels [38]

the MFCC feature vectors for the test speech samples to be recognized. Julius uses the acoustic model and decodes the test feature vectors.

4.5 Speaker Identification with GMM

Speaker similarity refers to the similarity between the speaker characteristics of actual speech and the voice transformed speech. With speaker similarity we evaluate the effectiveness of the voice transformation techniques in protecting the privacy of the speaker by concealing the identity of the speaker. Speaker identification technique using “Gaussian Mixture Model” (GMM) is used to measure the speaker similarity.

4.5.1 Speaker similarity vs. Speaker Identification

Speaker identification is a technique which examines a speech sample and identifies the source speaker from the set of speakers available in the system [40] [41]. Source speaker is identified using some features extracted from the speech sample and com-

paring against the same features extracted from each speaker in the database. Likelihood is computed for each feature pair (speech sample and speaker from database) and the mostly likely speaker is selected based on a threshold value.

In our experiment we use the speaker similarity to measure the similarity between the source speaker and the database of speakers. The features are extracted, compared against the features of each speaker in the database and likelihood computed similar to speaker identification. But here based on the likelihood we rank the speakers in the database and compare the rank of the source speaker.

4.5.2 Speaker Identification

Speaker identification as mentioned earlier compares a speech sample against a database of speaker. For this the system requires knowledge of voice characteristics of each speaker. The speaker identification system is also a statistically modeled system and has a training phase and testing phase. During the training phase, with the training speech data, the system learns the speaker characteristics and builds mixture models which model the voice characteristics of each speaker. In the testing phase the system computes likelihood for the test speech sample against the speaker classes built during the training phase.

The speaker identification system extracts cepstral features from the speech samples and builds the mixture models. As discussed in section 4.3.2.1, the lower part of the cepstral coefficients represent the spectral envelope carry the identity of the speaker and first few coefficients are used to build a template for each speaker in the training process.

The training phase,

1. Extracts the feature vectors from the training speech.
2. Builds world model for all the speakers using Gaussian mixture model.
3. Adapts the world model to build speaker models.

and the testing phase,

1. Extracts feature vectors from the test speech.
2. Computes the likelihood of the features and identifies the speaker.

4.5.2.1 Feature Extraction:

The characteristics of speech, features, cepstral feature extraction and their advantages have been discussed in detail in section 4.3.2.1. For our evaluation we use “Linear Frequency Cepstral Coefficients” (LFCC) [42] as the feature vectors. Similar to MFCC, LFCC are also computed by applying logarithm to the power spectrum of short duration speech segments followed by discrete cosine transform. The LFCC computation is shown in Figure 4.7.

4.5.2.2 Building World Model:

The speaker identification system builds statistical models to represent the speaker characteristics from the extracted features. Gaussian mixture models are widely used for text-independent speaker identification. The choice of GMMs is mainly due to the fact that no prior knowledge about the distribution is available. In addition they are suited for cases where there are a set of data points which appear in groups.

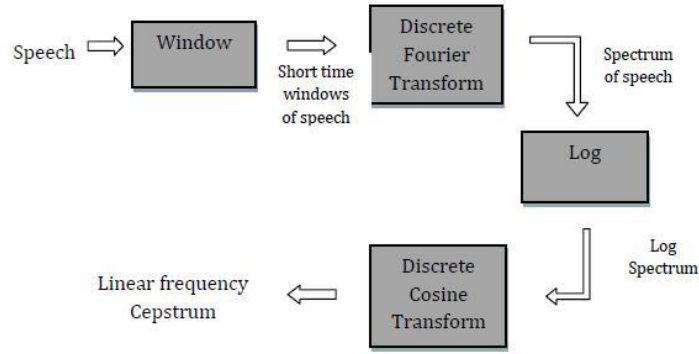


Figure 4.7: Linear Cepstral Coefficients computation

GMMs are well suited to model these kinds of data where each group of data can be represented as a Gaussian distribution and the total distribution is given by the mixture of these Gaussians and called a GMM. GMMs are represented in terms of number of components, weight of each component and the density of each Gaussian component. Figure 4.8 shows an example of GMM.

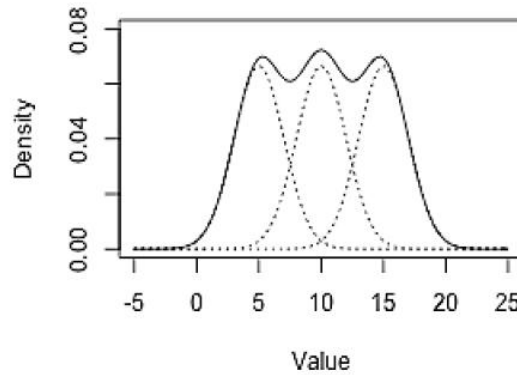


Figure 4.8: Example Gaussian mixture model with individual Gaussian densities [43]

In case of speaker identification, each component Gaussian can be assumed to model a single speaker and the GMM modeling the complete set of feature vectors.

A GMM distribution is defined as

$$p(x | \theta) = \sum_{i=1}^M w_i p_i(x) \quad (4.6)$$

In equation (4.6), θ is the model parameter, M is the number of components in the mixture model, w is the weight of the component, x is the feature vector, $p_i(x)$ is the Gaussian distribution density.

The GMM has no prior knowledge of the distribution and so starts with approximate model parameters and refines the model parameters iteratively using “Expectation Maximization” (EM) algorithm [44]. The algorithm computes the parameters of the model iteratively so as to maximize the likelihood of the model to represent the given feature vectors. The likelihood ratio is given by

$$M = p(X | \theta) \quad (4.7)$$

In the above equation M is the estimated model, X is the feature extracted from the test speech sample and θ is the model parameter. Initially the feature vectors from all the speakers are used to build the model and this is called the world model.

4.5.2.3 Adapting World Models to build Speaker Models:

“Maximum A Posteriori” (MAP) estimate is used to build distributions that more closely model the voice characteristics of individual speaker. This is done by adapting the world model built previously. MAP differs from the “Maximum Likelihood Estimation” (MLE) in that it has a prior knowledge of the model parameters to begin with. So with the world model, the EM algorithm iterates and refines the model parameters now using the feature vectors of the individual speaker to build speaker

models. These speaker models are more close approximations of each speaker’s voice characteristics. MAP estimation is given by

$$\tilde{\theta} = \arg \max_{\theta} p(\theta | X) = \arg \max_{\theta} p(X | \theta)p(\theta) \quad (4.8)$$

In equation (4.8), $p(\theta)$ is the posterior probabilityworld model.

4.5.2.4 Classifying the features of test Speakers:

In the testing phase, the LFCC features are extracted from the test speech samples. These features now have to be classified by checking for similarity with the speaker classes built in the training. The decision for the speaker identity is accomplished through hypothesis testing. Consider a speech sample s , which is hypothesized to be from a speaker T . The decision about the speaker of s is based on the following hypothesis.

Nullhypothesis, H_0 : s is from the speaker T .

Alternat ehypothesis, H_1 : s is not from speaker T .

If θ is the decision threshold, the likelihood is computed as [40].

$$\frac{p(s|H_0)}{p(s|H_1)} = \begin{cases} \geq \theta & \text{accept } H_0 \\ < \theta & \text{reject } H_0 \end{cases}$$

4.6 Speaker Identification ALIZE

For our experiments we use a text- independent speaker identification tool ALIZE [45]. ALIZE is a GMM based speaker identifier using LFCC features. 2048 components for the mixture model are used to compute the world and speaker models.

We generate feature vectors for all the speakers from the training speech samples.

Then from these features we first generate the world model. Speaker models are generated using the features specific to the each speaker. In the testing phase, features are extracted from the test speech samples. Then we compute the likelihood ratio for each test speech sample against each speaker in the database. We then compute the rank of the speakers in the database in the order the likelihood ratio beginning from highest. So higher the rank, the speech sample is most likely to be from that speaker. Based on these ranks we identify the rank of the actual speaker of the speech sample and compare this rank for the pitch distortion and voice conversion algorithms.

Chapter 5

Experimental results for Privacy Protection

In this chapter we discuss the experimental setup of our evaluation framework. First we describe the transformation parameters used for the pitch distortion and voice conversion techniques. The intelligibility evaluation with the ASR, training and testing procedure and the results of the speech recognition are presented for pitch distortion and voice conversion. Similarly the speaker similarity evaluation with the speaker identifier, its training and testing are also discussed.

5.1 Experimental data and parameters

5.1.1 TIMIT Dataset

For our evaluation experiments we use “Texas Instruments and Massachusetts Institute of Technology” (TIMIT) [46] speech corpora. TIMIT speech data is widely used for acoustic-phonetic knowledge and for evaluation of automatic speech recognition systems. TIMIT corpus contains speech data from 630 speakers, both male and female, from 8 major dialect regions of North America, thus making a total of 6300 sentences. We use this data set for both intelligibility/speech recognition and speaker similarity/speaker identification experiments. Our evaluation experiments with the automatic recognizers contain a set of training data and a set of test data. The training data consists of utterances from 461 speakers: 136 female and 325 male.

The test data has 5 male speakers and 5 female speakers, 10 utterances from each speaker making a total of 100 utterances.

5.1.2 Pitch Distortion parameters

The parameter involved in pitch distortion techniques is α and for our experiments we consider 5 different values of α (based on allowable α range [17]). The complete training data is distorted with these 5 α values. Similarly all the 100 utterances of test data are distorted by the 5 α values and these test data sets are named *SetA*, *SetB*, *SetC*, *SetD* and *SetE* corresponding to $\alpha = 1.0$, $\alpha = 0.5$, $\alpha = 0.75$, $\alpha = 1.25$ and $\alpha = 1.40$. *SetA* with $\alpha = 1.0$ corresponds to the test data set with no distortion. These 5 sets of data are then evaluated for intelligibility and speaker similarity. The data sets and α values are tabulated in Table 5.1.

Table 5.1: Pitch distortion parameters

Dataset	α
Set A	1.0
Set B	0.5
Set C	0.75
Set D	1.25
Set E	1.4

5.1.3 Voice Conversion parameters

The 5 male and 5 female speaker utterances are converted towards a single target male speaker and a single target female speaker. These target speakers are only reference speakers and randomly chosen. The choice of a particular speaker as the target does not have a significant influence on the evaluation because all the speakers are uniformly converted towards the target speakers.

In the case of voice conversion, we have 4 sets of data named FF , FM , MF and MM corresponding to “Source Female to Target Female”, “Source Female to Target Male”, “Source Male to Target Female” and “Source Male to Target Male”. The warping parameters for each of the data sets are listed in Tables 5.2, 5.3, 5.4 and 5.5. SF and SM are source female and source male respectively.

Table 5.2: Female to Female Voice distortion parameters

FF		
Speaker	α	ρ
SF1	0.92	0.93
SF2	0.90	0.98
SF3	0.81	0.89
SF4	0.81	0.87
SF5	0.98	0.95

Table 5.3: Female to Male Voice distortion parameters

FM		
Speaker	α	ρ
SF1	1.27	0.89
SF2	1.22	0.94
SF3	0.98	0.86
SF4	1.20	0.83
SF5	1.34	0.91

Table 5.4: Male to Female Voice distortion parameters

MF		
Speaker	α	ρ
SM1	0.84	1.16
SM2	0.82	1.18
SM3	0.81	1.38
SM4	0.76	1.84
SM5	0.85	1.62

Table 5.5: Male to Male Voice distortion parameters

MM		
Speaker	α	ρ
SM1	1.09	1.11
SM2	1.14	1.13
SM3	1.03	1.32
SM4	1.10	1.76
SM5	1.21	1.56

5.2 Intelligibility evaluation results

The intelligibility of the pitch distorted and voice converted speech is measured in terms of *Word Error Rate (WER)* and *Phoneme Error Rate*. The ASR is trained initially with the training data and then tested with different test data sets. For both pitch distortion and voice conversion techniques we consider two cases: ASR trained with the original data and ASR trained with modified data.

5.2.1 Pitch Distortion Algorithm results

1. **Cross Model:** The ASR is trained with original undistorted training data and tested with the distorted *SetA*, *B*, *C*, *D* and *E* data sets.
2. **Same Model:** The ASR is trained using the same five α values as the test data and tested with the distorted *SetA*, *B*, *C*, *D* and *E* data sets.

The average WER and phoneme error rate for both the cases are listed in Tables 5.6 and 5.7

5.2.2 Voice Conversion results

1. **Cross Model:** The ASR is trained with the original training data and tested with the voice converted *FF*, *FM*, *MF* and *MM* data sets.

Table 5.6: Average WER% with Pitch Distortion for Cross Model

Cross Model		
Dataset	<i>WER%</i>	<i>PhonemeErrorRate%</i>
A	29.07	33.01
B	92.37	79.72
C	65.52	58.14
D	61.52	53.13
E	78.42	66.38

Table 5.7: Average WER% with Pitch Distortion for Same Model

Same Model		
Dataset	<i>WER%</i>	<i>PhonemeErrorRate%</i>
A	29.07	33.01
B	31.76	34.8
C	29.81	33.01
D	30.25	33.66
E	32.46	34.03

2. **Same Model:** The ASR is trained with the same α and ρ values as the test data and tested with the voice converted *FF*, *FM*, *MF* and *MM* data sets.

The average WER and phoneme error rate for both the cases are listed in Tables 5.8 and 5.9

Table 5.8: Average WER% with Voice Conversion for Cross Model

Cross Model		
Dataset	<i>WER%</i>	<i>PhonemeErrorRate%</i>
FF	44.94	43.61
FM	39.69	38.77
MF	44.99	41.53
MM	39.47	39.18

Bar charts showing the accuracy% of Intelligibility evaluated using Automatic Speech Recognizer with Pitch Distortion and Voice conversion are shown in Figures 5.1 and 5.2

Table 5.9: Average WER% with Voice Conversion for Same Model

Same Model		
Dataset	WER%	PhonemeErrorRate%
FF	28.94	32.35
FM	30.43	32.87
MF	29.64	33.21
MM	31.34	33.95

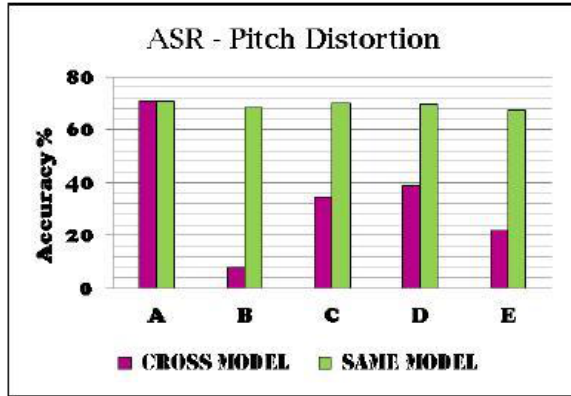


Figure 5.1: Bar chart showing Intelligibility accuracy% with Pitch Distortion using ASR.

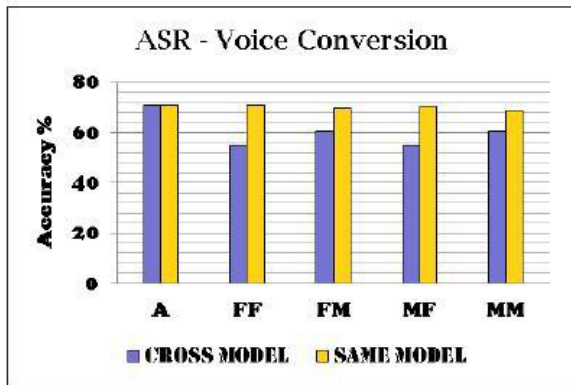


Figure 5.2: Bar chart showing Intelligibility accuracy% with Voice Conversion using ASR.

5.2.3 Statistical significance test for Intelligibility

To verify the statistical significance of the intelligibility results from the ASR, we performed the statistical z- tests. We consider the original speech data and compare

its WER with that of the voice transformed data.

Let population proportion p_1 be the average WER of original clips and population proportion p_2 the average WER of the voice transformed clips. Our null and alternate hypotheses are:

NullHypothesis, H_0 : $p_1 = p_2 \implies$ Average WER does not change due to voice transformation.

AlternateHypothesis, H_1 : $p_1 \neq p_2 \implies$ Average WER changes due to voice transformation.

We use standard two tail z-test to compare the populations since we consider the WER difference in either direction. We use 95% confidence level to compare the populations. If the calculated z value is less than the critical value (+1.96 or greater than -1.96) then the null hypothesis is true. Otherwise we reject the null hypothesis. Considering the average WER of the compared populations to be x_1 and x_2 , the z-test parameters are given in Table 5.10

Table 5.10: z-test parameter values for WER using ASR

Parameter	Value
Z_{obs}	$\frac{\bar{x}_1 - \bar{x}_2}{SE(x_1 - x_2)}$
Population size	100 (Pitch distortion), 50 (Voice Conversion)
Confidence Level	95%
Critical z-value ($z_{\frac{\alpha}{2}}$)	± 1.96
Rejection Criteria for H_0	$z_{obs} \leq -z_{\frac{\alpha}{2}}$ or $z_{obs} \geq z_{\frac{\alpha}{2}}$

Tables 5.11, 5.12, 5.13 and 5.14 show the z-test results on the intelligibility evaluation results of pitch distortion and voice conversion techniques.

Figure 5.3 shows the z-test results for accuracy of Intelligibility evaluated using Automatic Speech Recognizer with both Pitch Distortion and Voice conversion.

From the intelligibility evaluation results we can observe the following:

Table 5.11: z-test results for WER with Pitch Distortion for Cross Model


Cross Model	
Population	z-test values
A vs. B	-12.03 $\Rightarrow H_a$
A vs. C	-5.54 $\Rightarrow H_a$
A vs. D	-4.87 $\Rightarrow H_a$
A vs. E	-8.05 $\Rightarrow H_a$

Table 5.12: z-test results for WER with Pitch Distortion for Same Model


Same Model	
Population	z-test values
A vs. B	-0.41 $\Rightarrow H_0$
A vs. C	-0.11 $\Rightarrow H_0$
A vs. D	-0.18 $\Rightarrow H_0$
A vs. E	-0.51 $\Rightarrow H_0$

Automatic Recognizer

Pitch Distortion			Voice Conversion		
	Cross Model	Same Model		Cross Model	Same Model
A vs. B			A vs. FF		
A vs. C			A vs. FM		
A vs. D			A vs. MF		
A vs. E			A vs. MM		



**Alternate Hypothesis True –
Intelligibility changes**



**Null Hypothesis True –
Intelligibility does not
change**

Figure 5.3: z-test results for accuracy of Intelligibility using ASR with Pitch Distortion and Voice Conversion

- In the cross model, for the pitch distortion algorithm, all α values give a poor recognition with *SetB* having the least recognition.
- When the ASR is trained with similar α values, all the sets of data perform very well with *SetC* and *SetD* very close to the original.
- On the other hand, voice conversion gives an overall good performance in terms of intelligibility.

Table 5.13: z-test results for WER with Voice Conversion for Cross Model

Cross Model	
Population	z-test values
A vs. FF	-1.89 $\implies H_0$
A vs. FM	-1.28 $\implies H_0$
A vs. MF	-1.90 $\implies H_0$
A vs. MM	-1.25 $\implies H_0$

Table 5.14: z-test results for WER with Voice Conversion for Same Model

Same Model	
Population	z-test values
A vs. FF	0.02 $\implies H_0$
A vs. FM	-0.17 $\implies H_0$
A vs. MF	-0.07 $\implies H_0$
A vs. MM	-0.28 $\implies H_0$

- The cross model results are only slightly lower than the same model indicating that voice converted speech can be understood even without prior training.

5.3 Speaker similarity using Speaker Identifier

The speaker similarity of the pitch distorted and voice converted speech is measured in terms of *average rank* of the actual speaker. For every test speech data, we find the rank of the actual speaker and then compute the average rank among all the speakers in the database (136 female speakers and 325 male speakers). We compare this average rank for pitch distortion and voice conversion. The speaker identifier is trained initially with the training data and then tested with different test data sets. For both pitch distortion and voice conversion techniques we consider two cases: speaker identifier trained with the original data and the speaker identifier trained with the modified data.

5.3.1 Pitch Distortion Algorithm results

1. **Cross Model:** The speaker identifier is trained with original undistorted training data and tested with the distorted *SetA*, *B*, *C*, *D* and *E* data sets.
2. **Same Model:** The speaker identifier is trained using the same five α values as the test data and tested with the distorted *SetA*, *B*, *C*, *D* and *E* data sets.

The average rank of the speakers for both the cases is given in Table 5.15.

Table 5.15: Average Speaker rank with Pitch Distortion

Dataset	Cross Model		Same Model	
	Male	Female	Male	Female
A	33.2	7.40	33.2	7.40
B	59.0	61.8	40.6	11.7
C	56.7	58.9	36.4	11.1
D	51.2	60.5	39.6	16.7
E	54.2	47.2	40.2	20.9

5.3.2 Voice Conversion results

1. **Cross Model:** The speaker identifier is trained with the original training data and tested with the voice converted *FF*, *FM*, *MF* and *MM* data sets.
2. **Same Model:** The speaker identifier is trained with the same α and ρ values as the test data and tested with the voice converted *FF*, *FM*, *MF* and *MM* data sets.

The average rank of the speakers for both cases is listed in Table 5.16.

Bar charts showing the speaker rank% of evaluated using Speaker Identifier with Pitch Distortion and Voice conversion are shown in Figures 5.4 and 5.5.

Table 5.16: Average Speaker rank with Voice Conversion

Dataset	Cross Model	Same Model
FF	36.9	10.5
FM	48.2	8.0
MF	55.4	41.7
MM	40.9	31.9

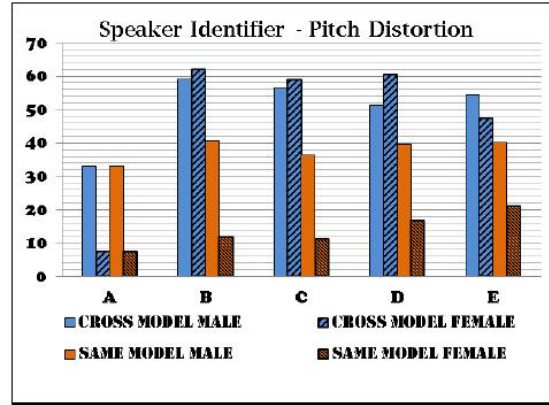


Figure 5.4: Bar chart showing speaker rank% with Pitch Distortion using Speaker Identifier

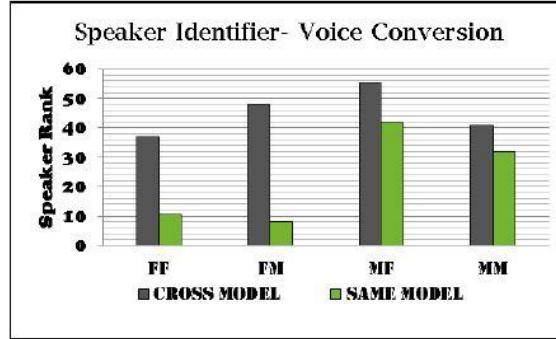


Figure 5.5: Bar chart showing speaker rank% with Voice Conversion using Speaker Identifier

5.3.3 Statistical significance test for Speaker similarity

Similar to the intelligibility, to verify the statistical significance of the speaker similarity results from the speaker identifier, we performed the statistical z-tests. We consider the original speech data and compare its speaker rank with that of the voice

transformed data.

Let population proportion p_1 be the average speaker rank of original clips and population proportion p_2 the average speaker rank of the voice transformed clips.

Our null and alternate hypotheses are

NullHypothesis, $H_0: p_1 = p_2 \implies$ Average speaker rank does not change due to voice transformation.

AlternateHypothesis, $H_1: p_1 \neq p_2 \implies$ Average speaker rank changes due to voice transformation.

We use standard two tail z-test to compare the populations since we consider the WER difference in either direction. We use 95% confidence level to compare the populations. If the calculated z value is greater than the critical value (-1.96) then the null hypothesis is true. Otherwise we reject the null hypothesis. Considering the average WER of the compared populations to be x_1 and x_2 , the z-test parameters are given in Table 5.17

Table 5.17: z-test parameter values for Speaker rank using Speaker Identifier

Parameter	Value
Z_{obs}	$\frac{\bar{x}_1 - \bar{x}_2}{SE(x_1 - x_2)}$
Population size	50 (Pitch distortion), 50 (Voice Conversion)
Confidence Level	95%
Critical z-value ($z_{\frac{\alpha}{2}}$)	-1.96
Rejection Criteria for H_0	$z_{obs} \leq -z_{\frac{\alpha}{2}}$

Tables 5.18, 5.19, 5.20 and 5.21 show the z-test results on speaker similarity evaluation results of pitch distortion and voice conversion techniques.

Figure 5.6 shows the z-test results for speaker rank evaluated using Speaker Identifier with both Pitch Distortion and Voice conversion.

From the speaker similarity evaluation we can observe the following


Table 5.18: z-test results for Speaker rank with Pitch Distortion for Cross Model

Cross Model		
Population	z-test values	
	Male	Female
A vs. B	$-2.79 \Rightarrow H_a$	$-6.96 \Rightarrow H_a$
A vs. C	$-2.43 \Rightarrow H_a$	$-6.53 \Rightarrow H_a$
A vs. D	$-1.85 \Rightarrow H_0$	$-6.77 \Rightarrow H_a$
A vs. E	$-2.17 \Rightarrow H_a$	$-5.92 \Rightarrow H_a$


Table 5.19: z-test results for Speaker rank with Pitch Distortion for Same Model

Same Model		
Population	z-test values	
	Male	Female
A vs. B	$-0.77 \Rightarrow H_0$	$-0.73 \Rightarrow H_0$
A vs. C	$-0.34 \Rightarrow H_0$	$-0.64 \Rightarrow H_0$
A vs. D	$-0.67 \Rightarrow H_0$	$-1.44 \Rightarrow H_0$
A vs. E	$-0.74 \Rightarrow H_0$	$-1.99 \Rightarrow H_a$

Speaker Identifier						
Pitch Distortion					Voice Conversion	
	Cross Model		Same Model			
	Male	Female	Male	Female		
A vs. B					A vs. FF	
A vs. C					A vs. FM	
A vs. D					A vs. MF	
A vs. E					A vs. MM	



Null Hypothesis True –
Speaker Identifiability does not
change



Alternate Hypothesis True –
Speaker Identifiability
changes

Figure 5.6: z-test results for speaker rank using Speaker Identifier with Pitch Distortion and Voice Conversion.

- In general speaker's identification is better in the case of female speakers most likely due to smaller number of female speakers.
- The cross model performs worse in identifying the speaker as expected in the case of both pitch distortion and voice conversion.

Table 5.20: z-test results for Speaker rank with Voice Conversion for Cross Model

Cross Model	
Population	z-test values
A vs. FF	-3.80 $\implies H_a$
A vs. FM	-5.12 $\implies H_a$
A vs. MF	-2.30 $\implies H_a$
A vs. MM	-0.79 $\implies H_0$

Table 5.21: z-test results for Speaker rank with Voice Conversion for Same Model

Same Model	
Population	z-test values
A vs. FF	-0.53 $\implies H_0$
A vs. FM	-0.12 $\implies H_0$
A vs. MF	-0.89 $\implies H_0$
A vs. MM	0.14 $\implies H_a$

- In the cross model itself there are two cases-*SetD* in pitch distortion and *SetMM* in voice conversion where the identity of the speaker is not protected.
- In the case of same model, it is unsettling to note that except for two cases-*SetE* in pitch distortion and *SetMM* in voice conversion, the speaker’s identity is not protected.
- This means that if an attacker is able to acquire some prior training data, then the voice transformation mat not be able to conceal the speaker’s identity.

5.4 Intelligibility evaluation with Subjective Experiments

The performance of the voice transformation techniques in maintaining the intelligibility is evaluated with subjective experiments also. For the subjective tests, we perform listening experiments and evaluate the recognition accuracy with human testers. In the case of intelligibility evaluation we have 8 testers who are randomly selected and are a balanced set of normal people, voice experts, male and female. The

listening experiments consist of two parts similar to the cross model and same model of the ASR.

5.4.1 Subjective experiment results

The experiment consists of 10 tests. In each test the test takers were asked to listen to 4 reference speech clips from 4 different speakers and then identify the words uttered in by the test data clips. The test data consists of 2 clips and one of the test clips is from one of the speakers of the reference clips. In addition the test takers were provided with a set of words for each test clip and asked to provide the transcription of the utterance.

The first test is similar to the cross model test of the ASR. The reference clips are original speech data and the test data consists of 16 clips, 2 clips each from the sets - *SetsB, C, D, E, FF, FM, MF* and *MM*. The other 9 tests are similar to the same model test of the ASR and each test corresponds to the *SetsA, B, C, D, E, FF, FM, MF* and *MM*. So in this case the reference and the test speech clips are from same set data [47]. The average recognition accuracy of the listening experiments is given in Tables 5.22 and 5.23.

Table 5.22: Average recognition accuracy% for Subjective test with Pitch Distortion

Dataset	Cross Model	Same Model
A	96.87	96.87
B	38.75	43.08
C	76.25	88.28
D	76.63	70.99
E	57.82	70.42

Bar charts showing the Recognition accuracy% evaluated using Subjective tests with Pitch Distortion and Voice conversion are shown in Figures 5.7 and 5.8.

Table 5.23: Average recognition accuracy% for Subjective test with Voice Conversion

Dataset	Cross Model	Same Model
FF	80.68	78.57
FM	77.64	90.74
MF	80.21	89.32
MM	85.40	86.53

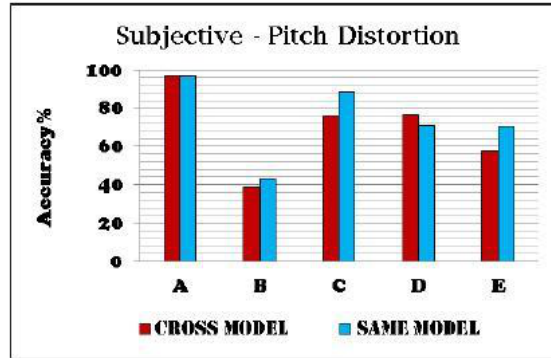


Figure 5.7: Bar chart showing Recognition accuracy% with Pitch Distortion using Subjective tests.

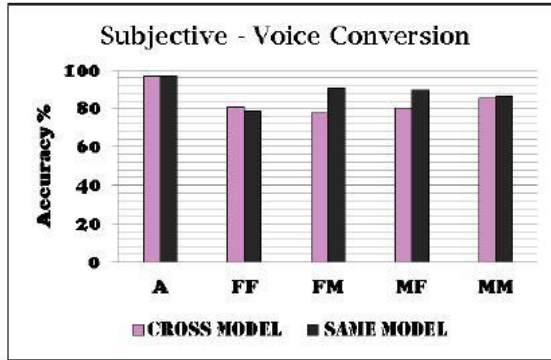


Figure 5.8: Bar chart showing Recognition accuracy% with Voice Conversion using Subjective tests.

5.4.2 Statistical significance test

We verify the statistical significance of the subjective tests for intelligibility through the statistical t-tests. We consider the accuracy results for the original speech data

and compare it with that of the voice transformed data.

Let population proportion p_1 be the average recognition accuracy of original clips and population proportion p_2 the average recognition accuracy of the voice transformed clips. Our null and alternate hypotheses are

NullHypothesis, H_0 : $p_1 = p_2 \implies$ Average accuracy does not change due to voice transformation.

AlternateHypothesis, H_1 : $p_1 \neq p_2 \implies$ Average accuracy changes due to voice transformation.

We use paired, two sample, two tailed t-test to compare the populations since we consider the accuracy difference in either direction. We use 95% confidence level to compare the populations. If the calculated t value is less than the critical value (+2.36 or greater than -2.36) then the null hypothesis is true. Otherwise we reject the null hypothesis. Considering the average recognition accuracy of the compared populations to be x_1 and x_2 , the z-test parameters are given in Table 5.24

Table 5.24: t-test parameter values for Recognition accuracy using Subjective tests

Parameter	Value
t_{obs}	$\frac{\bar{x}_1 - \bar{x}_2}{SE(x_1 - x_2)}$
Population size	8
Degrees of freedom	7
Confidence Level	95%
Critical t-value ($t_{\frac{\alpha}{2}}$)	± 2.36
Rejection Criteria for H_0	$t_{obs} \leq -t_{\frac{\alpha}{2}}$ or $t_{obs} \geq t_{\frac{\alpha}{2}}$

Tables 5.25, 5.26, 5.27 and 5.28 show the t-test results on recognition accuracy of subjective experiments for pitch distortion and voice conversion techniques.

Figure 5.9 shows the t-test results for Recognition accuracy evaluated using Subjective tests with both Pitch Distortion and Voice conversion.

Table 5.25: t-test results for intelligibility with Pitch Distortion for Cross Model

Cross Model	
Population	t-test values
A vs. B	3.17 $\Rightarrow H_a$
A vs. C	1.29 $\Rightarrow H_0$
A vs. D	1.25 $\Rightarrow H_0$
A vs. E	2.11 $\Rightarrow H_0$

Table 5.26: t-test results for intelligibility with Pitch Distortion for Same Model

Same Model	
Population	t-test values
A vs. B	2.92 $\Rightarrow H_a$
A vs. C	0.69 $\Rightarrow H_0$
A vs. D	1.52 $\Rightarrow H_0$
A vs. E	1.56 $\Rightarrow H_0$

Subjective Experiments

Pitch Distortion			Voice Conversion		
	Cross Model	Same Model		Cross Model	Same Model
A vs. B			A vs. FF		
A vs. C			A vs. FM		
A vs. D			A vs. MF		
A vs. E			A vs. MM		



 Alternate Hypothesis True – Intelligibility changes
  Null Hypothesis True – Intelligibility does not change

Figure 5.9: t-test results for Recognition accuracy using Subjective tests with Pitch Distortion and Voice Conversion.

From the subjective experiment results for intelligibility we can observe that,

- For both the same model and the cross model case the pitch distortion Set B performs poor in terms of recognition accuracy.
- All the other sets have good performance accuracy for both same and cross model.
- We provided the set of possible words for the listeners to choose from while transcribing the utterances. This was done to build a platform similar to the

Table 5.27: t-test results for recognition accuracy with Voice Conversion for Cross Model

Cross Model	
Population	t-test values
A vs. FF	1.06 $\Rightarrow H_0$
A vs. FM	1.20 $\Rightarrow H_0$
A vs. MF	1.11 $\Rightarrow H_0$
A vs. MM	0.86 $\Rightarrow H_0$

Table 5.28: t-test results for recognition accuracy with Voice Conversion for Same Model

Same Model	
Population	t-test values
A vs. B	1.15 $\Rightarrow H_0$
A vs. C	0.51 $\Rightarrow H_0$
A vs. D	0.63 $\Rightarrow H_0$
A vs. E	0.75 $\Rightarrow H_0$

automatic speech recognizers operation, where the training data is available as template and the recognition is based on those templates.

- Though the speech clips were modified, the human listeners were able to identify most of the words correctly which could be attributed to the fact that they were able to identify the sounds and then choose the words from the word pool. In addition human beings possess a natural knowledge of the vocabulary of the language, correctness and meaning of the sentence construction - which is a technical challenge for the large vocabulary automatic speech recognizers.

5.5 Speaker similarity evaluation with Subjective Experiments

The performance of the voice transformation techniques in concealing the identity of the speaker is evaluated with subjective experiments also. For the subjective tests, we perform listening experiments and evaluate the performance with human testers.

In the case of speaker similarity evaluation we have 12 testers who are randomly selected and are a balanced set of normal people, voice experts, male and female. The listening experiments consist of two parts similar to the cross model and same model of the speaker identifier.

5.5.1 Subjective experiment results

The experiment consists of 11 tests. In each test the test takers were asked to listen to 10 reference speech clips corresponding to 10 speakers from the test data and then identify the speaker of the test data clips. The test data consists of 5 clips chosen randomly from the TIMIT test data.

The first 9 tests are similar to the same model test of the speaker identifier and each test corresponds to the *SetsA, B, C, D, E, FF, FM, MF* and *MM*. So in this case the reference and the test speech clips are from same dataset [48]. The test 10 and 11 are similar to the cross model test of the speaker identifier. The reference clips are original speech data and the test data corresponds to the sets - *SetsB, C, D, E* for test 10 and *SetsFF, FM, MF, MM* for test 11 [49]. The average speaker accuracy of the listening experiments are given in Tables 5.29 and 5.30.

Table 5.29: Average Speaker accuracy% for Subjective test with Pitch Distortion

Dataset	Cross Model	Same Model
A	80.00	80.00
B	0	28.33
C	33.33	63.33
D	0	65.00
E	75.00	73.33

Bar charts showing the Speaker accuracy% of evaluated using Subjective tests with Pitch Distortion and Voice conversion are shown in Figures 5.10 and 5.11.

Table 5.30: Average Speaker accuracy% for Subjective test with Voice Conversion

Dataset	Cross Model	Same Model
FF	58.33	88.33
FM	16.67	75.00
MF	8.33	85.00
MM	83.33	86.67

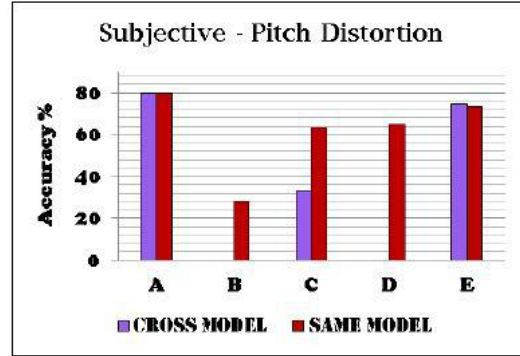


Figure 5.10: Bar chart showing speaker accuracy% with Pitch Distortion using Subjective tests.

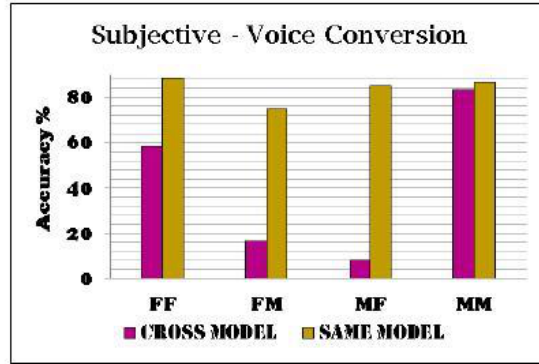


Figure 5.11: Bar chart showing speaker accuracy% with Voice Conversion using Subjective tests.

5.5.2 Statistical significance test

We verify the statistical significance of the subjective tests for intelligibility through the statistical t-tests. We consider the speaker recognition accuracy results for the original speech data and compare it with that of the voice transformed data.

Let population proportion p_1 be the average speaker recognition accuracy of original clips and population proportion p_2 the average speaker recognition accuracy of the voice transformed clips. Our null and alternate hypotheses are

NullHypothesis, H_0 : $p_1 = p_2 \implies$ Average accuracy does not change due to voice transformation.

AlternateHypothesis, H_1 : $p_1 \neq p_2 \implies$ Average accuracy changes due to voice transformation.

We use paired, two sample, two tailed t-test to compare the populations since we consider the accuracy difference in either direction. We use 95% confidence level to compare the populations. If the calculated t value is less than the critical value ($+2.20$ or greater than -2.20) then the null hypothesis is true. Otherwise we reject the null hypothesis. Considering the average recognition accuracy of the compared populations to be x_1 and x_2 , the t-test parameters are given in Table 5.31

Table 5.31: t-test parameter values for Speaker Accuracy using Subjective test

Parameter	Value
t_{obs}	$\frac{\bar{x}_1 - \bar{x}_2}{SE(x_1 - x_2)}$
Population size	12
Degrees of freedom	11
Confidence Level	95%
Critical t-value ($t_{\frac{\alpha}{2}}$)	± 2.20
Rejection Criteria for H_0	$t_{obs} \leq -t_{\frac{\alpha}{2}}$ or $t_{obs} \geq t_{\frac{\alpha}{2}}$

Tables 5.32, 5.33, 5.34 and 5.35 show the t-test results on speaker accuracy of subjective experiments for pitch distortion and voice conversion techniques.

Figure 5.12 shows the t-test results for Speaker accuracy evaluated using Subjective tests with both Pitch Distortion and Voice conversion.

Table 5.32: t-test results for Speaker Accuracy with Pitch Distortion for Cross Model


Cross Model	
Population	t-test values
A vs. B	6.93 $\Rightarrow H_a$
A vs. C	2.64 $\Rightarrow H_a$
A vs. D	6.93 $\Rightarrow H_a$
A vs. E	0.29 $\Rightarrow H_0$

Table 5.33: t-test results for Speaker Accuracy with Pitch Distortion for Same Model


Same Model	
Population	t-test values
A vs. B	2.99 $\Rightarrow H_a$
A vs. C	0.94 $\Rightarrow H_0$
A vs. D	0.83 $\Rightarrow H_0$
A vs. E	0.41 $\Rightarrow H_0$

Subjective Experiments

Pitch Distortion			Voice Conversion		
	Cross Model	Same Model		Cross Model	Same Model
A vs. B			A vs. FF		
A vs. C			A vs. FM		
A vs. D			A vs. MF		
A vs. E			A vs. MM		



Null Hypothesis True –
Speaker Identifiability
does not change



Alternate Hypothesis True –
Speaker Identifiability
changes

Figure 5.12: t-test results for Speaker accuracy using Subjective tests with Pitch Distortion and Voice Conversion.

From the subjective experiment results for speaker similarity it can be observed that,

- In the case of cross model, the speaker identification is poor for the pitch distortion case except for *SetE*.
- But for the voice conversion case the speaker's identity is not protected for two cases *FF* and *MM*.

Table 5.34: t-test results for Speaker Accuracy with Voice Conversion for Cross Model

Cross Model	
Population	t-test values
A vs. FF	1.20 $\implies H_0$
A vs. FM	3.98 $\implies H_a$
A vs. MF	5.16 $\implies H_a$
A vs. MM	-0.19 $\implies H_0$

Table 5.35: t-test results for Speaker Accuracy with Voice Conversion for Same Model

Same Model	
Population	t-test values
A vs. B	-0.54 $\implies H_0$
A vs. C	0.29 $\implies H_0$
A vs. D	-0.32 $\implies H_0$
A vs. E	-0.46 $\implies H_0$

- The same model case is similar to the results of the speaker identifier and shows that when training samples are available, the listeners can identify the speaker.

5.6 Observations from the Privacy Protection experiments

Based on the experimental results from the automatic techniques and the subjective experiments, we can observe that pitch distortion conceals the identity of the speaker comparatively more than voice conversion. On the other hand though voice conversion provides less protection to the privacy of the speaker comparatively, it is better in terms of the intelligibility of the modified speech. So among the two techniques we have trade-off between the degree of privacy protection and intelligibility. For applications where privacy is the priority than intelligibility pitch distortion particularly the *SetC* provides a good performance for intelligibility and speaker's privacy. For applications where human listeners are the end users and intelligibility is the main requirement, the voice conversion technique can be used.

We apply these observations for a video self-modeling therapy for people with voice hypertension. In this application, the people undergoing the voice therapy are the end users and so we use the voice conversion technique for generating a clean voice.

Chapter 6

Experimental Results for Voice Therapy

In this chapter we discuss the proposed video self-modeling (VSM) technique for voice therapy of people with voice hypertension. We propose a novel system that re-renders new talking-head sequence that is suitable for the voice therapy. We describe the experimental set up for this VSM: the audio-video capture of the patient undergoing voice therapy, the audio tracks segmentation phase for extracting the time markers and the voice transformation techniques for generating the new speech track. For our preliminary experiments, we also perform subjective and objective evaluation on the synthesized speech. We present the evaluation results which demonstrate the effectiveness of this approach.

6.1 VSM for voice therapy

Video self-modeling (VSM) is a behavioral intervention technique in which a learner models a target behavior by watching a video of him or herself. Vocal hyper function is one type of voice disorders that is defined as the use of excessive muscle force and physical effort in the production of voice [50]. Generally, vocal hyper function can be effectively treated with behavioral voice treatment or voice therapy. Participation in voice therapy requires regular (at least once a week) voice therapy sessions with the speech-language pathologist over a period of at least two months [51]. Access to voice therapy services in rural areas and developing countries is particularly lacking, due to difficulties in recruiting and retaining speech-language

pathologists and the expenses of time and travel for the required voice therapy program. The use of VSM for voice therapy is a novel application where the pathologist can use the proposed system to create videos of a patient speaking with an improved voice. This new form of treatment has the potential of reducing the length of the treatment program and the number of therapy sessions, thereby reducing health disparities in rural populations. A clinical study is currently underway at the University Of Kentucky Clinical Voice Center to test the effectiveness of the proposed VSM therapy.

6.2 Experimental Setup

The first phase of the voice therapy experiment is audio-video capture where a raw video of the patients head and his voice are recorded. Figure 6.1 shows the raw video capture setup. The red box highlights that only the head of the patient is being focused and captured.

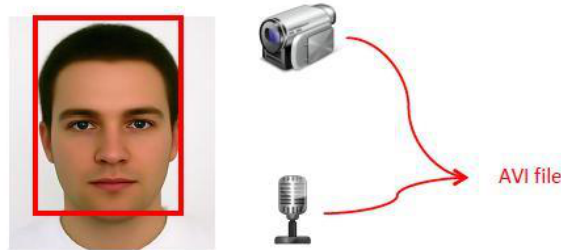


Figure 6.1: Video capture for VSM

After the raw footage is captured, the audio track is extracted. The audio is segmented to extract the time markers corresponding to word boundaries. A new speech track is synthesized using either text-to-speech or selecting a similar voice from a database of clean speech. Voice conversion is then applied to match the new

speech to the patient’s voice. Time markers extracted from the original and new speech track are passed to the video module which produces new video sequences for lip synchronization. Details of the algorithms used for video and audio processing are provided in the next two sections. Figure 6.2 shows the VSM’s audio content generation process. In the figure, *Track 1* is the speech track generated using text-to-speech voice, *Track 2* is the text to speech voice with voice conversion applied, *Track 3* is the similar voice from the clean speech database and *Track 4* is the similar voice with voice conversion applied.

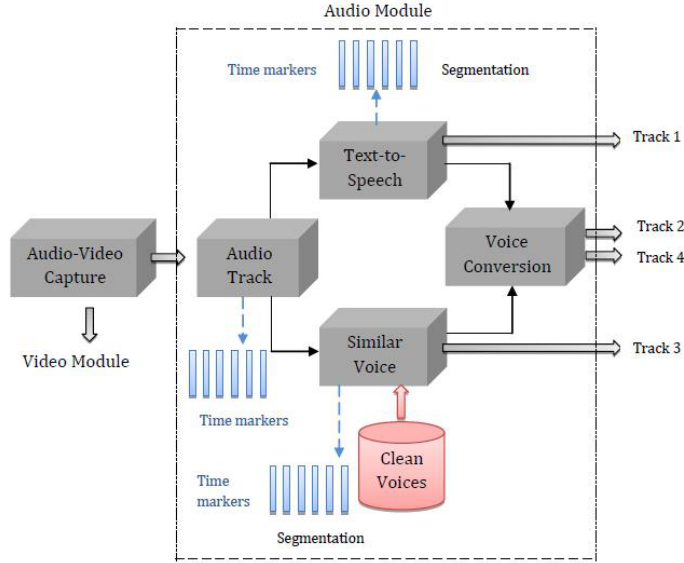


Figure 6.2: VSM content generation process

6.3 Audio Segmentation Algorithm

The audio segmentation module is used to extract the time markers in the audio track. Currently we detect the word boundaries as the time markers and use them for the synchronization in the video module.

The audio segmentation module has an initial training period for a few seconds

before the patient starts speaking. During this training period, the module determines the ambient noise amplitude level which acts as the threshold level for the segmentation. The remaining audio is segmented using this threshold value. The module extracts the time markers at instances when the amplitude increases above the threshold as the beginning of the word, and at instances when it decreases below the threshold for a sufficiently long period of time as the end of word. For our experiment we have 30s as the initial training period. The segmentation algorithm is shown in Figure 6.3.

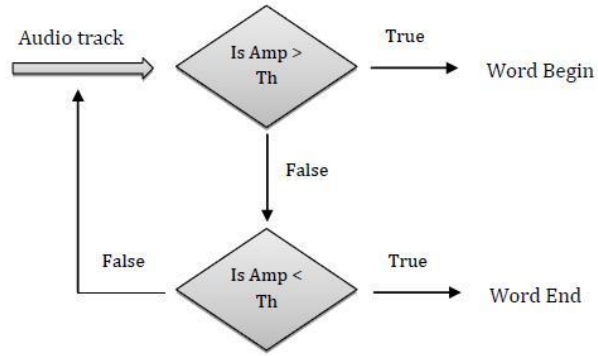


Figure 6.3: Audio Segmentation Algorithm.

6.4 Speech synthesis

To generate the replacement speech track, we have tested two different approaches. The scripts used in a typical therapy session are usually fixed and are available to us before the actual experiments with the VSM. In our first approach, we use a commercially available text-to-speech synthesizer from Cereproc [52]. While the text-to-speech engine offers great flexibility in generating arbitrary scripts, the quality of synthetic speech is still not as good as human voices. The other approach is to first collect speech clips from a diverse set of individuals with healthy voice reciting the

same script used in the therapy session and then identifying the clip that is most similar to the patient’s voice. These two approaches generate the speech tracks 1 and 3 shown in Figure 6.2.

For the second approach, to compute the similarity of the voices in database with the patients voice, we use the text independent speaker identification system, ALIZE [45] discussed in Chapter 4. The speech data collected from the speakers are used to construct a 2048 component world GMM model, which is then adapted to individual speaker models. In the testing phase, we use the patient’s voice as input and find the speaker with the maximum likelihood among all speakers in the database.

To make the selected similar voice or text-to-speech generated speech signal sound even closer to the patients voice, the two speech tracks are further processed using the time domain voice conversion technique [31] discussed in Chapter 3. In this case text-to-speech generated speech and the selected similar voice are the source voices and they are voice converted towards target voice: patient’s voice to generate speech tracks 2 and 4 as shown in Figure 6.2.

6.5 VSM Experimental Results

In our preliminary experiment, we capture two video sequences of a voice expert who is familiar with the voice characteristics of vocal hyper function. He prepared two sequences one using his natural voice and the other mimicking that of a patient with vocal hyper function. The two videos are on average 43 seconds long and are captured at a video frame rate of 30 fps and audio sampling rate at 16 kHz. The script recited consists of isolated words and a couple of sentences with pauses between

each.

Despite the fact that we have the normal voice of our voice expert, this normal-voice clip is never used in any of the training of our speaker identification system for identifying similar voices or the vocal tract modeling used in the voice conversion module. In all cases, we use the mimicked voice as the target as it would have been the only data available if he was a true patient. The normal-voice is only used for comparison.

For the CereProc text-to-speech synthesizer, we manually identify one character with southern English accent named “William” to be the closest match to our voice expert. For the clean speech database, we have identified three individuals in the same gender, age and race group as our voice expert and have recorded their voices reciting the same script. None of these three individuals claim to have any voice disorder.

6.5.1 Speaker similarity results

Apart from using ALIZE for selecting the most similar voice from the clean speech database, we also use it to evaluate the quality of the generated speech tracks by comparing with the normal healthy voice. We use *log likelihood ratio* measured by ALIZE in ranking the similarity between the normal healthy voice, s and the hypothetical target speaker model, L , which in our case is the speaker model generated for different speech tracks.

$$LLR(S | L) = \log\left(\frac{l(s | L)}{l(s | L)}\right) \quad (6.1)$$

In the above equation, $l(s | L)$ is the likelihood of s against the target speaker model L and $l(s | W)$ is the likelihood of s against the world model W .

For our experiment ALIZE was trained with the text-to-speech voice, text-to-speech voice with voice conversion, best human voice from database, best human voice with voice conversion and the mimicked voice. The testing was performed with the normal voice of the speaker. The synthesized voice closest to the normal healthy voice is evaluated based on the computed likelihood. Table 6.1 shows the likelihood values for the different voices synthesized.

Table 6.1: Likelihood of synthesized voice compared to healthy voice

Synthesized speech	LLR
Mimicked Voice	9.03e-2
Best Human	2.26e-2
Text-to-speech	1.39e-2
Best Human with Voice conversion	0.47e-2
Text-to-speech with Voice conversion	-0.63e-2

A bar chart showing the Likelihood of the synthesized voices against normal healthy voice is shown in Figure 6.4.

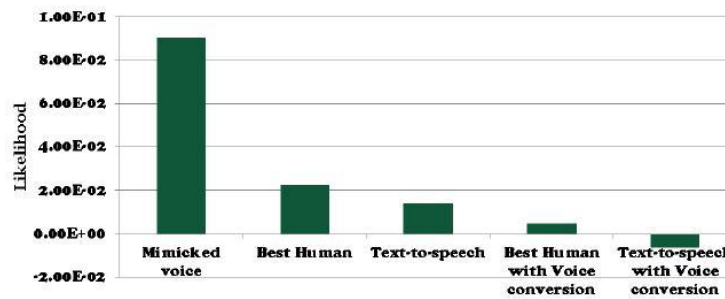


Figure 6.4: Bar chart showing Likelihood of synthesized voices against the normal healthy voice.

From Table 6.1, it is not surprising to see that the mimicked voice is the one closest to the healthy voice. Among the synthetic voices, the best human voice is ranked

top followed by the text-to-speech version. On the other hand, the application of voice conversion seems to have a detrimental effect. One possible explanation is the warping parameter α and the fundamental frequency ratio ρ . The parameters computed directly by the software produce voices that are non-human like, mostly likely due to the unnatural hoarseness in the mimicked voice. We have tuned the parameters in such a way that the voice is more human but it adversely affects the overall similarity to the target voice.

6.5.2 Subjective evaluation

We have also performed a series of subjective evaluation tests to determine the overall quality of the synthesized speech [53]. Two sets of questions are administered to five test takers who are unaware of the details of our proposed system. The test takers were asked to watch and compare the different videos to determine the voice that is more natural and of high quality.

The first test aims in comparing the different speech synthesis techniques and select the better one in terms of quality and naturalness. The test takers were made to watch pairs of videos and then select one of them as a better technique. The average results of the preferred techniques of the test takers are tabulated in Table 6.2.

The second test is to rate the different synthesized speech tracks, including the normal healthy voice, in the order of most likely normal voice of the speaker. The test takers were first made to view the mimicked voice video. Then, they were asked to view the five different videos and rank them based on their likelihood of being the

Table 6.2: Results of forced choice test

Common parameter	Test	% favoring 1st technique
Text-to-speech	without Voice Conversion vs. with Voice Conversion	100
Best Human	without Voice Conversion vs. with Voice Conversion	100
Voice Conversion	Best human voice vs. Text-to-speech voice	100
With Voice Conversion	Best human voice vs. Text-to-speech voice	100

voice without disorder. The results of the subjective evaluation are given in Table 6.3.

Table 6.3: Results of rank test

Test	Average Rank
Healthy Voice	1.8 ± 1.8
Best Human	2.0 ± 0.7
Best Human with Voice Conversion	3.4 ± 1.3
Text-to-speech	3.2 ± 0.4
Text-to-speech with Voice Conversion	4.6 ± 0.5

The results of the tests are as expected. In the first test, the testers prefer best-human voice over text-to-speech voice and the absence of voice conversion. In the second test, while most users choose the “correct” answer, i.e. the video with healthy voice, the synthetic video with best human voice comes quite close. This result is promising as it demonstrates the possibility of using synthetic video in depicting unseen behavior of an individual, which is precisely the goal of the VSM therapy. We are currently conducting more experiments with a larger user group.

Chapter 7

Conclusions and Future Work

In this research, we have presented an evaluation framework for the voice transformation techniques. We evaluated two techniques which operate on different characteristics of the voicepitch and the spectral envelope and both these characteristics control the identity of the speaker. We have evaluated the performance of the two algorithms: pitch distortion and voice conversion, based on two key qualities of the modified speech: intelligibility or clarity of the modified speech and privacy of the speaker or degree to which the speaker's identity is concealed.

We have performed extensive experiments for the evaluation of the transformation techniques with both subjective and objective experiments. The intelligibility evaluation was done with automatic speech recognizers and with human testers. The speech recognizers outperform human listeners when trained with modified data. But on the other hand human listeners have an advantage of the better grammatical knowledge and sentence construction compared to the speech recognizers which rely on the training data. In the case of speaker identification also, the automated speaker identifier can find a similar match to the speakers for some cases where human testers could not identify.

Among the two techniques, the voice converted speech has an overall better performance in terms of intelligibility. But when specific training models are used there is no difference in intelligibility between pitch distortion and voice conversion tech-

niques. The privacy of the speaker is better protected when pitch distortion is used as the transformation technique. But in the case of speaker identity too when specific training speech samples are used, neither technique effectively conceals the identity of the speaker. In real world application, we cannot assume a scenario where the attackers may not acquire the training samples. So we see that better voice transformation techniques are needed to protect the privacy of the speaker such that both human users and automatic techniques fail to identify the actual speaker.

Based on our evaluation results we propose a novel voice therapy technique using video self-modeling for people with voice hypertension. We have conducted preliminary experiments for this behavioral intervention technique which uses video feedback of the patient with an improved good voice. The experiments performed with a mimicked voice and alternate speech tracks generated using text-to-speech and similar voice from a clean speech database are evaluated subjectively and objectively. The evaluation results show that these synthesized voice tracks provide a close approximation to the actual voice of the speaker.

In our future work, we plan to build an interface for the audio-video capture which is automated and operating in real time. This interface captures the video with head position alignment and the script to be spoken as a scrolling text. The other phases of the audio processing, audio segmentation and voice transformation will generate the new speech tracks in real time. Better voice conversion techniques to closely resemble the patients voice need to be applied. The advantage of using multimedia techniques in the voice therapy and its effectiveness needs to be studied with a clinical test. While our proposed system is domain specific, we believe that the concept of using

multimedia techniques for video self -modeling has far-reaching importance in many different areas of health care that can benefit from behavioral modification.

Bibliography

- [1] Allam Mousa. Voice Conversion using Pitch Shifting Algorithm by Time Stretching with PSOLA and Re-sampling. *Journal of Electrical Engineering, Vol 61, pages 57-61, 2010.*
- [2] Srikanth Mangayyagari and Ravi Sankar. Pitch Conversion Based on Pitch Mark Mapping. *Proceeding of the IEEE SoutheastCon 2007, pages 8-12, Apr 2007.*
- [3] David Sundermann, Antonio Bonafonte, Hermann Ney and Harald Hoge. A study on residual prediction techniques for voice conversion. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pages 13-16, May 2005.*
- [4] Jing Xia and Junxun Yin. A GMM based residual prediction method for Voice Conversion. *Proceedings of International Symposium on Intelligent Signal Processing and Communication Systems, Dec 2005.*
- [5] Alexander Kuin and Michael W Macon. Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2001, pages 813-816, May 2001.*
- [6] Kun Liu, Jianping Zhang and Yonghong Yan. High Quality Voice Conversion through Combining Modified GMM and Formant Mapping for Mandarin. *Second International Conference on Digital Communications, July 2007.*

- [7] Ning Xu and Zhen Yang. A Precise Estimation of Vocal Tract Parameters for High Quality Voice Morphing. *9th International Conference on Signal Processing 2008, pages 684-687, Oct 2008.*
- [8] Tomoki Toda, Yamato Ohtani and Kiyohiro Shikano. Eigenvoice Conversion Based on Gaussian Mixture Model. *Proceedings of the International Conference on Spoken Language Processing, pages 2446-2449, Oct 2006.*
- [9] Athanasios Mouchtaris, Jan Van der Spiegel, and Paul Mueller. Nonparallel Training for Voice Conversion Based on a Parameter Adaptation Approach. *IEEE Transactions on Audio, Speech and Language Processing, Vol 14, May 2006.*
- [10] Sachin S. Kajarekar, Harry Bratt, Elizabeth Shriberg and Rafael de Leon. A Study of Intentional Voice Modifications for Evading Automatic Speaker Recognition. *The Speaker and Language Recognition Workshop, IEEE Odyssey 2006, pages 1-6, Jun 2006.*
- [11] CMU Sphinx. Open Source Toolkit for Speech Recognition. <http://cmusphinx.sourceforge.net/>
- [12] The Hidden Markov Model Toolkit. <http://htk.eng.cam.ac.uk/>
- [13] Yinqiu Gao and Zhen Yang. Pitch modification based on syllable units for voice morphing system. *The IFIP International Conference on Network and Parallel Computing Workshops, 2007, pages 135-139, Sep 2007.*

- [14] Hui Ye and Steve Young. High quality voice morphing. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2004, pages 9-12, May 2004.*
- [15] Scot Cunningham and Traian Marius Truta. Protecting Privacy in Recorded Conversations. *Proceedings of ACM Conference on Privacy and anonymity in information society, 2008*
- [16] Jianfeng Chen, Dat Tran Huy, Koksoon Phua, Jit Biswas and Maniyeri Jayachandran. Using keyword spotting and replacement for speech anonymization. *IEEE Conference on Multimedia and Expo 2007, pages 548, Jul 2007.*
- [17] M. Vijay Venkatesh, Jian Zhao, Larry Profitt and Sen-ching S. Cheung. Audio-visual Privacy Protection for Video Conference. *IEEE International Conference on Multimedia and Expo 2009, pages 1574-1575, Jul 2009.*
- [18] Qin Jin, Arthur R. Toth, Tanja Schultz and Alan W Black. Voice convergin: speaker de-identification by voice transformation. *IEEE International Conference on Acoustics, Speech and Signal Processing 2009, pages 3909, Apr 2009.*
- [19] H. J. Krouse. Video modeling to educate patients. *Journal of Advanced Nursing, Vol. 33, pages. 748-757,2001.*
- [20] R. W. McDaniel and v. A. Rhodes. Development of a preparatory sensor information videotape for women receiving chemotherapy for breast cancer. *Cancer Nursing, Vol. 21, pages. 143-148, 1998.*

- [21] D. Nielsen, S. O. Sigurdsson, and J. Austin. Preventing back injuries in hospital settings: the effects of video modeling on safe patient lifting by nurses. *Journal of Applied Behavioral Analysis*, Vol. 42, pages. 551-561, 2009.
- [22] A. M. Alvero and J. Austin. The effects of conducting behavioral observations on the behavior of the observer. *Journal of Applied Behavior Analysis*, Vol. 37, pages. 457-468, 2004.
- [23] C. H. Hitchcock, P. W. Dowrick, and M. A. Prater. Video self-modeling intervention in school-based settings: A review. *Remedial and Special Education*, Vol. 24, pages. 36-45, January/February 2003.
- [24] V. S. Ramachandran, D. C. Rogers-Ramachandra, and S. Cobb. Touching the phantom. *Nature*, Vol. 377, pages 489-490, 1995.
- [25] Lucina Q. Uddin, Mari S. Davies, Ashley A. Scott, Eran Zaidel, Susan Y. Bookheimer, Marco Iacoboni, and Mirella Dapretto. Neural basis of self and other representation in autism: An fmri study of self-face recognition. *PLoS ONE*, Vol. 3, 2008.
- [26] T. Buggy. Seeing Is Believing: Video Self Modeling for people with Autism and other developmental disabilities *Wodbine House*, 2009.
- [27] P. W. Dowrick. Self-modeling.Using Video: Psychological and Social Applications. New York: Wiley, 1983.
- [28] Steve Cassidy. Speech Recognition. <http://web.science.mq.edu.au/~cassidy/comp449/html/index.html>

- [29] Yannis Stylianou. Voice transformation: a survey. IEEE International Conference on Acoustics, Speech and Signal Processing, Apr 2009.
- [30] Udo Zoelzer Dafx: Digital Audio Effects, John Wiley and Sons, Inc., New York, NY, USA, 2002.
- [31] David Sundermann, Antonio Bonafonte, Hermann Ney and Harald Hoge. Time Domain Vocal Tract Length Normalization. *Proceedings of the Fourth IEEE International Symposium on Signal Processing and Information Technology, 2004, pages 191-194, Dec 2004.*
- [32] Male and Female Formant Cepstrum. home.iitk.ac.in/~rhegde/ee627_2010/lec_3.4.pdf
- [33] Average formant frequencies for some basic vowels for male and female. <https://ccrma.stanford.edu/~kglee/m220c/formant.html>
- [34] Speech Recognition. http://en.wikipedia.org/wiki/Speech_recognition
- [35] Tony Robinson. Speech Analysis. <http://mi.eng.cam.ac.uk/~ajr/SA95/SpeechAnalysis.html>
- [36] Steve Young, Julian Odell, Dave Ollason, Valtcho Valtchev and Phil Woodland. The HTK Book <http://www.ee.columbia.edu/ln/rosa/doc/HTKBook21/HTKBook.html>
- [37] John- Paul Hosom. Automatic Speech Recognition with Hidden Markov Models. <http://www.cslu.ogi.edu/people/hosom/cs552/>

- [38] Male and Female Cepstrum for Vowels <http://cnx.org/content/m12469/latest/>
- [39] A. Lee and T. Kawahara. Recent Development of Open Source Speech Recognition Engine Julius. *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, 2009*
- [40] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. *Journal on Digital Signal Processing, Vol 10, pages 19-41, January 2000.*
- [41] Jean-Luc Gauvain and Chin-Hui Lee. Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transactions on Speech and Audio Processing, Vol 2, pages 291, Apr 1994.*
- [42] Howard Lei and Eduardo Lopez-Gonzalo. Mel, Linear, and Antimel Frequency Cepstral Coefficients in Broad Phonetic Regions for Telephone Speaker Recognition. www.icsi.berkeley.edu/pubs/speech/lei_lopez_melantimel.pdf
- [43] Gaussian Mixture Model <http://upload.wikimedia.org/wikipedia/commons/d/dd/Gaussian-mixture-example.png>
- [44] Moon T.K. The expectation maximization algorithm. *IEEE Signal Processing Magazine, Vol 13, pages 47-60, Nov 1996.*
- [45] Jean-Francois Bonastre, Nicolas Scheffer, Corinne Fredouille and Driss Matrouf. NIST04 speaker recognition evaluation campaign: new LIA speaker detection

- platform based on ALIZE toolkit. *Proceedings of NIST Speaker Recognition Evaluation, 2004*.
- [46] TIMIT acoustic-phonetic continuous speech corpus. <http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1>
- [47] Speech Recognition Subjective Experiment web page <http://tinyurl.com/4agdt5b>
- [48] Speaker Identification Same Model Subjective Experiment web page. <http://tinyurl.com/2g7xqy5>.
- [49] Speaker Identification Cross Model Subjective Experiment web page. <http://tinyurl.com/2bjtazx>.
- [50] D. R. Boone and S. C. McFarlane. The Voice and Voice Therapy, Prentice Hall, 2006.
- [51] L. O. Ramig and K. Verdolini. Treatment efficacy: Voice disorders. *Journal of Speech, Language and Hearing Research, Vol. 41, pages 101-116, 1998*.
- [52] CereProc, Text to Speech Technology. <http://www.cereproc.com>.
- [53] VSM Subjective Experiment web page. <http://tinyurl.com/63esvzk>.

VITA

Name: Anusha Raghunathan

Bachelors in Electronics and Communication Engineering

Anna University, Chennai, India

Date of birth: 23rd April 1984

Position held:

1. Associate, Cognizant, India