UNIVERSITI PUTRA MALAYSIA

VOICE CONVERSION APPROACH THROUGH FEATURE STATISTICAL MAPPING

ABDULBASET M. NASR

VOICE CONVERSION APPROACH THROUGH FEATURE STATISTICAL MAPPING

By

ABDULBASET M. NASR

Thesis Submitted in Fulfilment of the Requirement for the Degree of Master of Science in the Faculty of Engineering
Universiti Putra Malaysia

January 2001

To the loving memory of my father, for instilling within me the thirst for knowledge and the quest for excellence.

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Master of Science.

VOICE CONVERSION APPROACH THROUGH FEATURE STATISTICAL MAPPING

By

ABDULBASET M. NASR

January 2001

Chairman: Md. Mahmud Hassan, Ph.D.

Faculty: Engineering

Over the past few decades the field of speech processing has undergone tremendous changes and grown to be important both theoretically and technologically. Great advances have already been made in a broad range of applications such as speech analysis and synthesis techniques, voice recognition, text to speech conversion and speech coding techniques to name a few.

On the process of development of these applications, voice conversion (VC) technique has recently emerged as a new branch of speech synthesis dealing with the speaker identity. The basic idea behind VC is to modify one person's speech so that it is recognized as being uttered by another person.

There are numerous applications of voice conversion technique. Examples include the personalization of text to speech (TTS) systems to reduce the need for a large speech database. It could also be used in the entertainment industry. VC

technology could be used to dub movies more effectively by allowing the dubbing actor to speak with the voice of the original actor but in a different language. Voice conversion can also be used in the language translation applications to create the identity of a foreign speaker.

This project proposes a simple parametric approach to VC through the use of the well-known speech analysis technique namely Linear Prediction (LP). LP is used as analysis tool to extract the most important acoustic parameters of a person's speech signal. These parameters are the pitch period, LP coefficients, the voicing decision and the speech signal energy. Then, the features of the source speaker are mapped to match those of the target speaker through the use of statistical mapping technique.

To illustrate the feasibility of the proposed approach, a simple to use voice conversion software was developed. The program code was written in C++ and implemented using Microsoft Foundation Class (MFC).

The proposed scheme to the problem has shown satisfactory results, where the synthesized speech signal has come as close as possible to match that of a target speaker.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Master Sains.

KAEDAH PENGALIHAN SUARA MENERUSI PEMETAAN STATISTIK

Oleh

ABDULBASET M. NASR

Januari 2001

Pengerusi: Md. Mahmud Hassan, Ph.D.

Fakulti : Kejuruteraan

Sejak beberapa dekad yang lepas, bidang pemprosesan pertuturan telah melalui perubahan yang besar dan menjadi amat penting secara teori dan juga teknologi. Perkembangan besar telah dijalankan di dalam pelbagai rangkaian besar applikasi, sebagai contohnya, analisis pertuturan and teknik sintesis, suaikenal suara, penukaran teks kepada pertuturan dan teknik koding pertuturan.

Di dalam proses pembangunan applikasi sedemikian, teknik pengalihan suara telah muncul sebagai satu cabang baru dalam sintesis pertuturan yang melibatkan identiti penutur sejak kebelakangan ini. Idea asas sebalik pengalihan suara adalah untuk mengubah petuturan seseorang supaya ia dapat dikenali sebagai pertuturan seseorang yang lain pula.

Terdapat pelbagai applikasi yang menggunakan teknik pengalihan suara. Contoh-contoh merangkumi sistem personalisasi teks kepada pertuturan untuk mengurangkan keperluan terhadap pengkalan data pertuturan yang besar. Ia juga boleh digunakan didalam industri hiburan. Teknologi pengalihan suara juga boleh

digunakan untuk mengalih bahasa sesuatu tayangan gambar lebih berkesan dimana pelakon yang mengalih suara boleh bercakap melalui suara pelakon yang asal tetapi didalam bahasa yang lain. Pengalihan suara juga boleh digunakan didalam applikasi penterjemahan bahasa untuk menghasilkan identiti pembual asing.

Projek ini mencadangkan suatu pendekatan parametrik yang mudah kepada pengalihan suara menerusi pengunaan teknik analisis pertuturan yang terkenal iaitu Jangkaan Linear. Jangkaan Linear digunakan sebagai alat analisis untuk mendapatkan parameter akoustik terpenting daripada signal tuturan seseorang. Parameter ini adalah tempoh nada, perangkaan jangkaan linear, keputusan suara dan tenaga isyarat pertuturan. Justeru itu, ciri-ciri daripada penutur asal akan dipetakan supaya menyerupai penutur sasaran menerusi penggunaan teknik pemetaan statistik.

Untuk mengambarkan kemampuan pendekatan yang dicadangkan, suatu software pengalihan suara yang mudah telah dibina. Kod program ini telah ditulis didalam C++ dan diimplementasikan dengan menggunakan Kelas Asas Microsoft. "Microsoft Foundation Class (MFC)"

Cadangan skema kepada masalah ini telah manghasilkan keputusan yang memuaskan di mana isyarat pertuturan yang disintesiskan berjaya menyerupai penutur sasaran.

ACKNOWLEDGEMENTS

I owe a debt of gratitude to my supervisor Dr. Md. Mahmud Hassan, for his guidance, advice and unlimited support throughout the course of this project.

My thanks also go to my supervisory committee members, for their advice and technical support that make this work come to reality.

Finally, my deepest and most sincere thanks to my family members, for their endless love, encouragement and moral support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

a_k	:	LP coefficients.
A	:	Mapping variable.
ACF	:	Auto-correlation function.
AMDF	:	Average magnitude difference function.
B	:	Mapping variable.
DFT	:	Discrete fourier transform.
DTW	:	Dynamic time warping.
$e(n)$	:	Prediction error.
EWSM	:	Elementary waveform speech model.
EGG	:	Electroglottograph.
F_0	:	Fundamental frequency.
FFT	:	Fast fourier transform.
FIR	:	Finite impulse response.
FT	:	Fourier transform.
G	:	Filter gain.
G_{mod}	:	Modified gain.
GCI	:	Instants of glottal closure.
GELP	:	Glottal excited linear predictor.
GUI	:	Graphical user interface.
HMM	:	Hidden Markov model.
IFT	:	Inverse fourier transform.
K	:	Pitch lag.
K_{mod}	:	Modified LP coefficient.
LAP	:	Line spectrum pair.
LAR	:	Log area ratios.
LP	:	Linear prediction.
MFC	:	Microsoft foundation class.
MSE	:	Mean squared error.
N	:	Frame size.
O	:	Frame overlap.
P	:	Prediction order.
p_{mod}	:	Modified pitch period.
PCM	:	Pulse code modulation.
PDA	:	Pitch determination algorithm.
PPE	:	Pitch period estimator.
PSOLA	:	Pitch synchronious overlap add.
RC	:	Reflection coefficients.
$s(n)$	:	Speech signal.
$\hat{s}(n)$	:	Predicted speech signal.
STFT	:	Short time fourier transform.
TTS	:	Text to speech.
$u(n)$	:	Glottal excitation signal.
var(x)	:	Variance.
VC	:	Voice conversion.
VQ	:	Vector quantization.
$w(n)$	:	Window function.

CHAPTER 1

INTRODUCTION

1.1 Preamble

Over the past decades the field of speech processing has undergone tremendous changes and grown to be important both theoretically and technologically. Revolutionary advances have already been made in a broad range of applications such as speech analysis and synthesis techniques, voice recognition, text to speech (TTS) conversion and speech coding techniques to name a few.

On the process of development of these applications, voice conversion (VC) technique has recently emerged as a new branch of speech synthesis dealing with speaker identity. The basic idea behind VC is to modify one person's speech so that it is recognized as being uttered by another person.

1.2 What is Voice Conversion?

VC, also known as voice transformation, is a new branch of speech synthesis dealing with the modification of certain speech parameters related to the speaker identity. This task can be accomplished by converting the extracted speech

parameters of one speaker (source speaker) to those parameters of another speaker (target speaker), as shown in Figure 1.1.

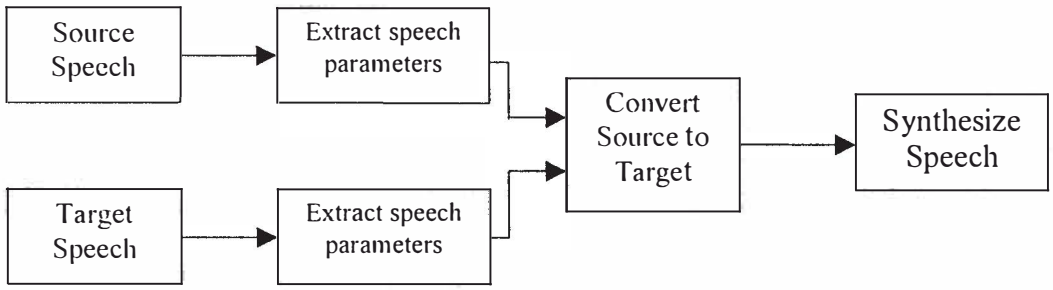

Figure 1.1: Basic scheme of voice conversion.

A voice conversion system, such as the one shown in Figure 1.1, is concerned with the speaker individuality in the sense that it transforms the acoustic speech parameters relevant to certain speaker while leaving the speech message content intact. This means that an utterance by the source speaker is modified to sound as if it had been uttered by the target speaker provided that the same information as being said by the source is left unchanged.

Among all the acoustic parameters related to the speaker individuality, pitch and formant frequencies are the two most important. Consequently, any attempt to VC is usually accomplished through the modification of these unique properties of the speech signal.

VC technique has numerous applications that show the importance of studying this field. These applications include personalization of text-to-speech synthesis system, aids to the handicap, improving the effectiveness of foreign

language training, and many others. A section in chapter 2 is devoted to give an idea of some possible applications of VC.

1.3 Thesis Objectives and Importance

1.3.1 Importance of Studying VC

In the recent literature of the speech signal processing field voice recognition, text to speech conversion and speaker identification techniques have been extensively studied. High quality voice recognition and speaker adaptation software is now commercially available. However, there are few real implementations of voice conversion systems. This has left a big room for researchers to study and develop techniques to challenge this problem.

The vast number of possible VC applications has also urged the researchers to study this immature field. The fact that any possible advance in this field will consequently enhance many other speech processing applications, has also been a strong reason for studying this interesting field.

1.3.2 Objectives of the Author's Work

In trying to tackle the problem of transforming one voice into another, various VC techniques were studied. The most basic and straightforward approach is basically to detect what word the source speaker has said and replace it with that same word being said by the target speaker. This method has certain obvious

limitations, including the need for large databases of speech and the fact that the reconstructed speech would sound broken up, because the natural flow between words that exists in natural speech would not be present.

In this project, a simple voice conversion scheme based on linear prediction (LP) analysis was developed. The LP analysis is performed on the speech signal to obtain the acoustical parameters related to the speaker identity. These parameters are the speech fundamental frequency, or pitch, voicing decision, signal energy, and vocal tract parameters.

Once these parameters are obtained for two different speakers designated as source and target speakers, statistical mapping functions are then applied to modify the extracted parameters. The mapping functions are derived from these parameters in such a way that the source parameters resemble those of the target. Finally the modified parameters are used to produce the new speech signal.

In summary, the main objectives of the author's work are:

- To study the field of voice conversion and its applications. This study involves a review of some of the existing speech analysis and voice conversion approaches.
- To apply the LP analysis algorithms to the speech signal in order to extract certain acoustic features.
- To develop a modification technique to be used in the conversion process.
- To design software that allows the implementation of the proposed system.

1.4 Structure of the Thesis

The rest of the thesis is organized as follows. Chapter 2 provides a general background to the topic. The chapter starts with a discussion of the speech production process since it is essential to understand the acoustic features of speech signal. These features are then presented in the following section. A literature review carried out by the author is inherent throughout the context of this chapter. A review of some of the speech analysis and feature extraction techniques is also included. Some of the existing approaches to VC including both, parametric and non-parametric methods are also discussed. The chapter ends with a brief discussion of some possible applications of VC systems.

Chapter 3 starts off with an overview of the framework followed by author to tackle the problem. A comprehensive description of the speech analysis algorithms is also provided. The chapter also gives a general description of the developed voice conversion software.

Chapter 4 contains the experiments and results of the proposed system. It also discussed the results obtained from the simulations carried out using MATLAB.

In Chapter 5, a general conclusion about the research work has been deduced, some recommendations for future work and the areas of possible improvements has also been discussed.

CHAPTER 2

BACKGROUND AND LITERATURE REVIEW

2.1 Introduction

The principal means of human communication is speech. It reflects the moods, the ideas, and identity of the speaker. Voice conversion techniques are concerned with the modification of speech signals in order to alter the perceived identity of the speaker. Therefore, the speech signal features related to the identity of the speaker are the main targets of any voice conversion system.

Many speech parameters had been proven to be related to the speaker identity. Examples include the speech fundamental frequency, or pitch, formant frequencies and bandwidths, prosody and many more. However, the most important feature of speech signal is the pitch. Consequently, any attempt to VC is usually accomplished through the modification of this unique property of the speech signal (Kuwabara and Takagi, 1991). Thus, a brief review of some of the most popular pitch determination algorithms (PDA) is discussed.

When developing speech analysis and synthesis systems for their many possible applications, it is essential to fully understand the fundamentals of the speech production process. The task is made much easier if one has a good

understanding of how humans generate speech, and how this human process can be modelled mathematically. The following section discusses the process of speech production and modeling.

2.2 Speech Production Process

The main organs of the human body responsible for producing speech sounds are the lungs, larynx, pharynx, oral cavity, nasal cavity and tongue, which are illustrated by the cross-section shown in Figure 2.1. The combination of these organs is known as the vocal tract.

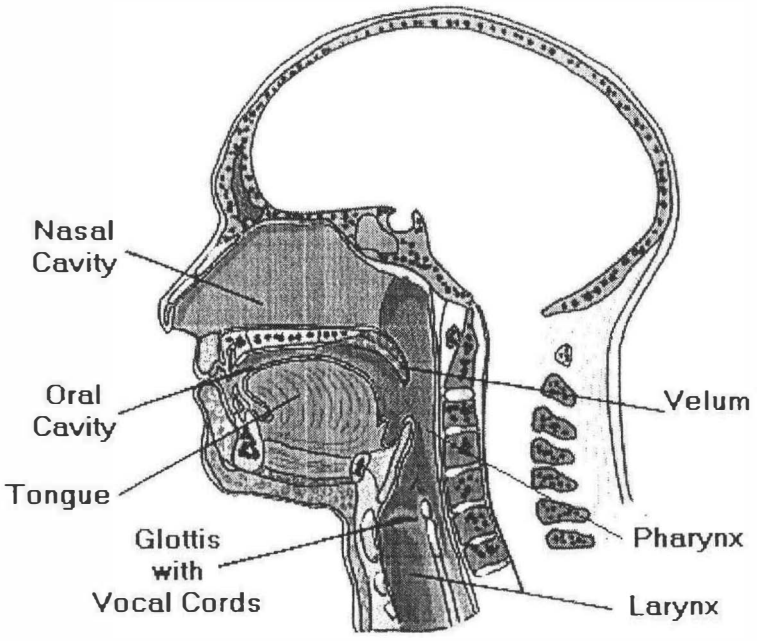

Figure 2.1: Speech production organs.

The larynx is the part of the respiratory tube containing the vocal cords, which are also known as vocal folds. The pharynx is the part between the larynx and the mouth that connects the larynx to the oral cavity. It has almost fixed dimensions,

but its length may be changed slightly by raising or lowering the larynx at one end and the velum at the other end.

The oral cavity is an irregular tube terminated at the front by the lips and at the back by the larynx. It is one of the most important parts of the vocal tract. Its size, shape and acoustics can be varied by the movements of the tongue, the lips, the cheeks and the teeth.

The nasal cavity is a non-uniform acoustic tube of fixed volume and length terminated at the front by nostrils and the rear by the velum. The velum controls acoustic coupling between the oral and nasal cavities.

The speech production process can be divided into three components. These components are: the generation of the excitation signal, the modulation of this signal by the vocal tract and the radiation of the final speech signal, as shown in Figure 2.2.

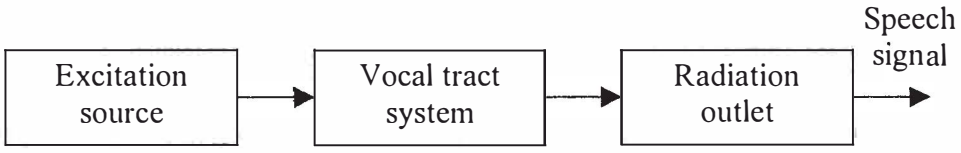

Figure 2.2: Main components of speech production.

The excitation signal is generated when the airflow from the lungs, which are the main energy source, is forced through the larynx to the main cavities of the vocal tract. As the excitation signal propagates through the vocal tract, its spectrum is shaped by the resonance and anti-resonance imposed by the physical shape of the tract. The produced signal is then radiated from the oral and nasal cavities through