UNIVERSITI PUTRA MALAYSIA

PHONEME BASED SPEAKER VERIFICATION SYSTEM BASED ON TWO STAGE SELF-ORGANIZING MAP DESIGN

ANG CHEE HUEI

PHONEME BASED SPEAKER VERIFICATION SYSTEM BASED ON TWO STAGE SELF-ORGANIZING MAP DESIGN

By

ANG CHEE HUEI

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirement for Degree of Master of Science

December 2001

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

PHONEME BASED SPEAKER VERIFICATION SYSTEM BASED ON TWO STAGE SOM DESIGN

By

ANG CHEE HUEI

December 2001

Chairman : Abdul Rahman bin Ramli, Ph.D.

Faculty : Engineering

Speaker verification is one of the pattern recognition task that authenticate a person by his or her voice. This thesis deals with a relatively new technique of classification that is the self-organizing map (SOM). Self-organizing map, as an unsupervised learning artificial neural network, rarely used as final classification step in pattern recognition task due to its relatively low accuracy. A two-stage self-organizing map design has been implemented in this thesis and showed improved results over conventional single stage design.

For speech features extraction, this thesis does not introduce any new technique. A well study method that is the linear prediction analysis (LPA) has been used. Linear predictive analysis derived coefficients are extracted from segmented raw speech signal to train and test the front stage self-organizing map. Unlike other multistage or hierarchical self-organizing map designs, this thesis utilized residual vectors generated from front stage self-organizing map to train and test the second stage self-organizing map.

The results showed that by breaking the classification tasks into two level or more detail resolution, an improvement of more than 5% can be obtained. Moreover, the computation time is also reduced greatly.

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

PENGENALPASTIAN PENGUCAP MELALUI 'PHONEME' DENGAN PENGKELAS DUA ARAS 'SELF-ORGANIZING MAP'

Oleh

ANG CHEE HUEI

Disember 2001

Pengerusi : Abdul Rahman bin Ramli, Ph.D.

Fakulti : Kejuruteraan

Pengenalpastian pengucap merupakan salah satu cabang dalam bidang pengecaman corak yang cuba mengenalpasti identiti seseorang melalui suaranya. Tesis ini mengenegahkan teknik pengkelasan yang agak baru iaitu 'self-organizing map' (SOM). Sebagai rangkaian neural tiruan yang berasaskan pembelajaran tanpa pengawasian, SOM jarang digunakan sebagai langkah akhir pengkelasan disebabkan kejituan yang rendah. Dalam tesis ini, SOM dua aras telah diguna dan menunjukkan keputusan yang lebih baik daripada SOM searas yang biasa.

Untuk ekstraksi fitur pertuturan, tesis ini tidak memperkenalkan teknik baru. Satu Teknik yang dikenali iaitu analisis penganggaran linear telah digunakan. Koefisi yang diekstrak daripada isyarat pertuturan mentah dengan analisis penganggaran linear digunakan untuk melatih dan menguji SOM aras pertama. Berlainan dengan SOM muitiaras atau hirarki yang lain, tesis ini menggunakan vektor ralat yang dihasilkan daripada SOM pertama untuk melatih dan menguji SOM kedua.

Keputusan tesis ini menunjukkan dengan memecahkan tugas pengkelasan kepada dua aras atau resolusi, peningkatan keputusan lebih daripada 5% boleh diperolehi. Tambahan pula, masa pengkomputasian dapat dikurangkan dengan banyak.

ACKNOWLEDGEMENTS

I wish to thank my supervisor, Dr. Abdul Rahman Ramli, for the guidance that he managed to extend to me during my Master study, and for his efforts to provide a good research environment, which made this thesis possible.

I am grateful to the researchers at Multimedia and Imaging System Research Group , Department of Computer and Communication, for their helpful suggestions and discussions. I also wish to acknowledge valuable interactions I have had with Mr. Andrew Teoh, who was extremely generous in sharing with me his knowledge in speaker recognition. I would also like to thank those at Otago University, New Zealand, who had involved in preparing the Otago Speech Corpus and made it available online. Without these data, this work may not have been possible.

Last but certainly not least, I wish to thank my parents and other family's members for having raised me in such a stable and loving environment, which has enabled me to come so far.

This thesis submitted to the Senate of Universiti Putra Malaysia has been accepted as fulfilment on the requirement for the degree of Master of Science.

AINI IRERIS, Ph.D.
Professor/Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

DECLARATION

I hereby declare that the thesis is based on my original work except for quotations and citations, which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other at UPM or other institutions.

_____

Ang Chee Huei

Date:

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANNs	-	Artificial Neural Networks
ART	-	Adaptive Resonant Theory
ASV	-	Automatic Speaker Verification
BP	-	Back-propagation
CAVE	-	The European Caller Verification Project
DTW	-	Dynamic Time Warping
GMM	-	Gaussian Mixture Model
HMM	-	Hidden Markov Model
LPA	-	Linear Predictive Analysis
LVQ	-	Learning Vector Quantization
MLP	-	Multi-layer Perceptron
NTN	-	Neural Tree Network
PICASSO	-	Pioneering Caller Authentication for Secure Service Operation
SOM	-	Self-Organizing Map
SV	-	Speaker Verification
TDNN	-	Time Delay Neural Network
VQ	-	Vector Quantization

CHAPTER 1

INTRODUCTION

1.1 An Overview

Speech conveys information on several levels. It contains a message generically expressed as a sequence of words, information specific to the speaker that produced the speech, and information about the environment in which the speech was produced and transmitted. Speaker specific information includes the identity of the speaker, the gender of the speaker, the language or dialect of the speaker and possibly the physical and emotional condition of the speaker. With this richness of information it comes as no surprise that, with the advent of computers, speech has found wide spread application in human-computer communication. In particular, automatic speech recognition is the process of extracting the underlying message and automatic speaker recognition is the process of verifying the identity of the speaker. Applications range from using voice commands over the telephone to control financial transactions and verifying the identity of the speaker, to continuous dictation and speaker detection in multi-party dialogues. The application generally dictates the types of information in the speech signal that are useful. For example, for the purpose of extracting the underlying message in automatic speech recognition, the presence of speaker and environmental information may actually lead to confusions and degrade system accuracy. Similarly, message and environmental information may degrade speaker recognition accuracy. For an application to be successful, an accurate modeling of the desired type of information is therefore important.

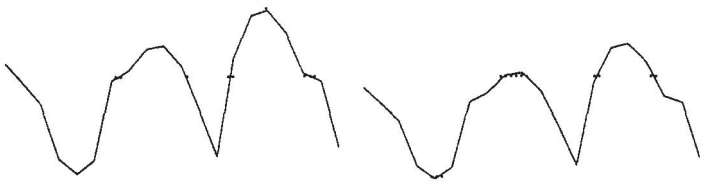
1.2 Basic Problem in Speaker Verification

1.2.1 The Variability of Speech

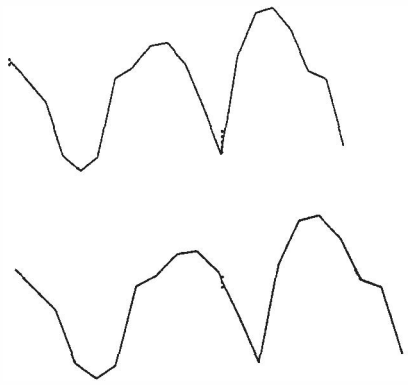
The main problem in speaker verification (SV) is the variability of speech. Different phonemes have different acoustic characteristic. Although in SV, we only consider inter-speaker variation, intra-speaker variation in most of the time, pose a big problem. This is because nobody can utter an utterance exactly the same every time. The utterance will usually have variation in speaking rate and amplitude if we consider it in time domain; and will have variation in formant frequencies and formant bandwidth in frequency domain. Furthermore, in digital signal processing, it is common to process speech signal in fixed size frame which will cause time displacement problem. Figure 1.1 illustrate these issues in graphics.

1.2.2 Feature Selection

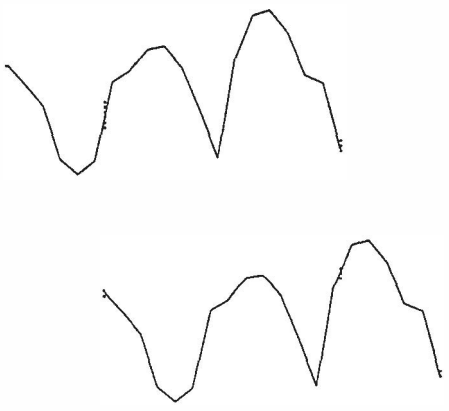
To reduce the amount of data, it is necessary to extract features instead of using raw speech waveforms. However, it is not currently known what features carry the most speaker specific information. Different people have different vocal cords and vocal tracts hence produced different formant frequencies. This serves the basic for finding suitable candidates in SV. Commonly used features such as linear predictive coefficients (Chen et al. 1993), cepstrum coefficients (Rosenberg 1994) and mel-scaled filter banks output (Che et al. 1996) produce almost satisfactory results. But the search for an ideal feature which has the best discriminating ability is going on.



(a)



(b)



(c)

Figure 1.1 Illustrate Variation of an Utterance by A Same Speaker Due To (a) Amplitude, (b) Speaking Rate and (c) Time Displacement The x-axis is time and y axis is amplitude

1.2.3 Channel / Background Noise

Channel or background noise can distort the speech and reduce the verification rate. A system which is trained in a specific environment (room, indoor and quiet) will perform poorly in another environment (outdoor, noisy background). It is necessary to train the system to adapt to different environments. However, noise reduction or environment adaptation is not an easy task as well.

1.2.4 Classification Model

After the features are extracted, a classification method has to be selected. The effectiveness of a feature depends on the classification model used and vice versa. Techniques that have been used regularly include template matching (Euclidean Distance), Hidden Markov Model (HMM) (Markov et al. 1998, Bimbot et al. 1997, Chen et al. 1996) and Artificial Neural Networks (ANNs) (Bennani et al. 1994, Farrell et al. 1994, Fakotakis et al. 1996, Wouhaybi et al. 1999). Occasionally, time alignment needs to be carried out before classification can be made. This is true particularly if the system is word or sentence based. As for feature selection, different types of classification model is available and there is no ideal or standard model. A very careful selection of feature used and classification method has to be made.

1.3 Research Objective

The ultimate goal of speaker verification research is to develop a user-friendly, high performance system, that is computationally efficient and robust in all environments. The objective of the present thesis is to study a phoneme based speaker verification system. Phoneme based systems have advantages in simplicity and text independent. A two stage self-organizing map (SOM) design is proposed in this thesis which can solve the problems of time-alignment and signal normalization without the needs of more computational heavy algorithm.

1.4 Outline of the Thesis

The rest of the thesis is organized as follows :

Chapter 2 makes literature review on SV which include pre-processing of speech, feature extraction methods and classification techniques.

Detail discussion is carried out on linear predictive analysis (LPA) in chapter 3. LPA coefficients are used as speaker specific speech features in this thesis. It is followed by discussion on classification technique used, which is the self-organizing map (SOM) in chapter 4.

Chapter 4 presents methodology of this research work. The system architecture as well as each processing module will be discussed in details. This includes the concept and implementation of multi stage SOM.

Chapter 5 describes the experimental part of the thesis as well as discussion of its results. The results were obtained from the system described in Chapter 4.

Chapter 6 concludes and gives suggestions for extensions and future research.

CHAPTER 2

LITERATURE REVIEW

In this chapter, literature review of automatic speaker verification (ASV) task and its current research achievement is explained. In the first part of this chapter, the task is defined accordingly. The following parts discussed the general model of a ASV system. A brief review on various processing stages or steps is given. The last part of this chapter discusses a few speaker verification research projects in Europe and some commercially available ASV systems.

2.1 Automatic Speaker Verification (ASV)

Speaker verification can be considered within the wider context of speaker recognition. Speaker recognition collectively describes the tasks of extracting or verifying the identity of the speaker (Atal 1974, Doddington 1985). In speaker identification, the task is to use a speech sample to select the identity of the person that produced the speech from among a set of candidate identities, or population of speakers. This task involves classification from N-possibilities, where $N > 1$ is the population of speakers. In speaker verification, the task is to use a speech sample to test whether a person who claims to have produced the speech did in fact do so. For example in tele-banking, it is vital to verify that the caller is the exact account holder. This task involves a two-way classification which is a test of whether the claim is correct or not. In speaker identification the number of possible choices are the number of speakers in the population, whereas in speaker verification the outcome is limited to one of two choices. Closed-set speaker identification is the task

where every speaker in the population is known to the system at the time of use. Open-set identification is the task where some speakers in the population are unknown to the system at the time of use and hence must be rejected on the basis of being unknown. Open-set identification is therefore a combination of closed-set identification and speaker verification. An example where speaker identification has found use is audio indexing, which involves the automatic detection and tagging of speakers in a small multi-party dialogue. In this thesis the focus will be on the task of speaker verification, but it should be understood that the techniques investigated here can be readily applied to speaker identification.

Taking a broader view, speaker identification and verification themselves can be placed in the field of biometric identification and verification (Campbell 1997), where the goal is to use any of a number of person-specific cues to classify that person. Examples of commonly used cues are as diverse as a facial image, iris pattern, finger print, genetic material or even keyboard typing pattern. The advantage of using a biometric cue for access control is that it is always accessible, unlike a key or password that can be misplaced, forgotten or stolen.

Using a speaker recognition system is usually a two-step process (Furui 1996). The user first enrolls by providing the system (computer) with one or more representative samples of his or her speech. These training samples are then used by the system to train (construct) a model for the user. In the second step the user provides a test sample that is used by the system to test the similarity of the speech to the model(s) of the user(s) and provide the required service. In this second step the

speaker associated with the model that is being tested is termed the target speaker or claimant (Martin et al. 1998).

In speaker verification, when the person is constrained to speak the same text during both training and testing the task is text-dependent (Furui 1996). For example, the verification phrase may be a unique password or a fixed string of digits. Applications requiring access control, such as voice-mail, telephone banking and credit card transactions have successfully used this type of verification (Campbell 1997, Boves 1998). A similar system using fixed phrases is currently being tested at a US border crossing at Otay Mesa, in San Diego, California, that would allow frequent travelers to gain clearance by speaking into a hand-held computer inside the car. While text-dependent verification potentially requires only a small amount of speech it requires the user to faithfully produce the required text. As such it requires a cooperative user and a structured interaction between the user and system (Campbell 1997). When the person is not constrained to speak the same text during training and testing the task is text-independent (Furui 1996). This is required in many applications where the user may be uncooperative or applications where speaker recognition occurs as a secondary process unknown to the speaker as in audio indexing. For example, a forensic application may require verifying the identity of a speaker based on speech from a recorded telephone conversation and the speaker may not actually be aware of this process. In both text-dependent and text-independent modes of operation the verification decision can be sequentially refined as more speech is input until a desired significance level is reached (Furui 1996, Lund et al. 1996, Fukunaga 1990). The word "authentication" has sometimes been used for "verification" and "talker" or "voice" for "speaker". Similarly, "text-free" has been used for "text-independent" and "fixed-text" for "text-dependent".

2.2 General Model of ASV System

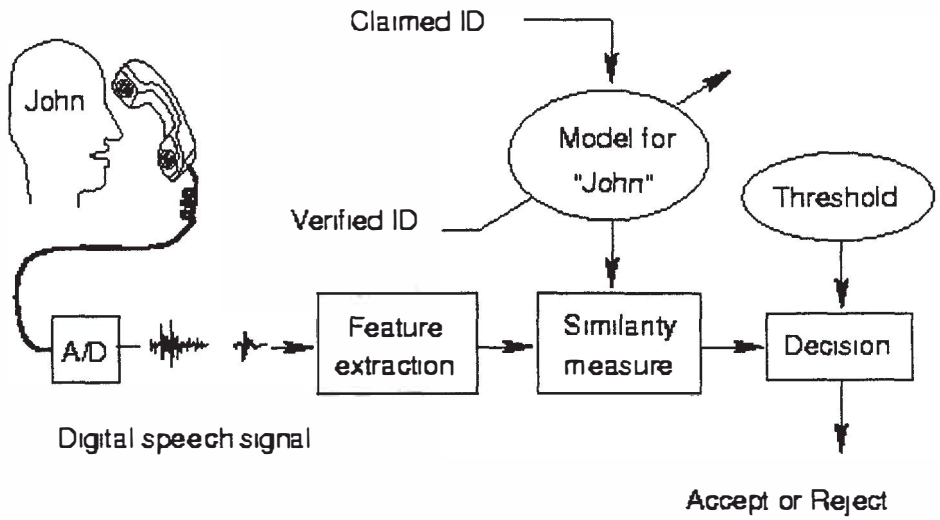


Figure 2.1 : Block Diagram of The Major Processes in A Speaker Verification System

A block diagram of the major stages in a speaker verification system is shown in Figure 2.1. First is the acquisition stage, where the speech produced by the speaker is converted from a sound pressure waveform into an electrical signal using a transducer. This acoustic signal is digitized and sampled at a suitable rate. Second is the signal processing and feature extraction stage, where salient parameters conveying speaker identity are extracted from the acoustic speech signal. Design of the feature extraction stage is based on the existing body of knowledge of the speech process -- such as models of the articulatory and auditory systems (O'Shaughnessy 1987, Hermansky et al. 1994), theory of linguistics and phonetics (Ladefoged 1993), perceptual cues used by listeners (Voiers 1964, Drullman et al. 1994), transmission process (Rabiner et al. 1978), and application specific requirements. Currently, linear predictive analysis (LPA) derived cepstral coefficients is the most popular choice for feature extraction (Furui 1996). The third stage involves computing a similarity

measure (Fukunaga 1990) between the information retrieved from the speech of the current speaker and a previously constructed model representing the person the speaker claims to be. The model training (construction) forms a major component of the speaker verification system. It determines storage cost and computation and dictates accuracy of the similarity measure. The fourth and final stage is to compare the similarity measure to a predetermined value or threshold and decide whether to accept or reject the claimed identity of the speaker. In this last stage for example, if the model of the claimed speaker is deemed to represent the information retrieved from the acoustic signal accurately, i.e. the two are similar, then the decision is to accept the claim made by the speaker. The third and final stage usually are combined together to form classification stage. From 1980's until mid of 1990's, Hidden Markov Model (HMM) was the most popular technique (Furui 1996). Since late of 1990's, researchers started to replace HMM with ANN since ANN produced better results in most of the cases compared with HMM. Among various ANN models, multi-layer perceptron (MLP) with back-propagation (BP) training seems to be the most popular choice. Other choices include adaptive resonance theory (ART) network, time delay neural networks (TDNN) and radial basic function networks (RBF). Self-organizing map (SOM), on the other hand, although proved to be applicable to speech recognition (Kohonen 1988a), failed to attract researchers from speaker recognition field.

There has been, and continues to be, a great deal of interest in speaker verification with a vast number of speaker specific cues, feature extraction techniques, modeling techniques, and evaluation measures proposed. These are

covered in a number of tutorial papers such as Atal (1974), Doddington (1985), Gish et al. (1994), Furui (1996), Campbell (1997) and Lee (1997).

2.3 Research Project

2.3.1 The European Caller Verification Project (CAVE)

CAVE was a 2 year long project, funded by the European Union, to develop and test speaker verification systems for use in telephone applications like calling-card or financial services. It was terminated on November 1997. A follow up project called PICASSO is now continuing research on speaker verification.

The CAVE project has investigated many differing aspects of speaker verification, not only core algorithmic research, but also how to incorporate the results of this research into real telephone-based services which anyone can use. The project partners therefore include user organizations such as banks and telecommunications companies, a number of highly-regarded European research institutes, and a major provider of speech-enabled automatic telephone services.

This project was tested with two different databases of speech recordings for SV, one of which is the widely used YOHO database (Campbell 1995). The feature vectors used are LPA coefficients derived cepstrum coefficients. Different strategies were tested which include text dependent, text prompted and text independent with different approaches in classification model such as Ergodic HMMs, Vector Quantization (VQ), Gaussian Mixture Modeling (GMM) and Dynamic Time