TEXT INDEPENDENT SPEAKER VERIFICATION USING BINARY-PAIR PARTITIONED NEURAL NETWORKS

by

Claude A. Norton, III

A Thesis submitted to the faculty of Old Dominion
University in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE
ELECTRICAL ENGINEERING

OLD DOMINION UNIVERSITY
December 1995

Approved by:

____

Stephen A. Zahorian (Director)

____

____

____

ACKNOWLEDGMENTS

I would like to express my gratitude to my advisor, Dr. Stephen A. Zahorian, for his support and guidance. His expertise and guidance have made this research possible. I would also like to thank my family: Linda, Christopher and Scott, without their support and understanding this would not have been possible.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number
Page

CHAPTER ONE

INTRODUCTION

An acoustic speech signal contains a variety of information. It contains a textual message as well as information which distinguishes gender, identity, age and cultural origins. This information is based on acoustic differences which we, as humans, have relatively little difficulty in interpreting. We have no problem in discerning that a friend has a cold by hearing them speak. We can correctly recognize and identify the speaker, even though the interference caused by the cold has altered the acoustic speech signal. This is a truly remarkable feature of our auditory system. We, as engineers and researchers, would like to build our machine-based systems with these same remarkable qualities.

The research which lead up to this thesis focuses on the development of a neurally-based system for performing speaker verification. The system is based on a technique of partitioning a data space into a collection of smaller tasks called binary-pair partitioning. This approach partitions a large classification task into several smaller two-way tasks.

Following this introduction I will provide a brief background on automatic speaker recognition (both identification and verification). I will then explore recent research trends in the area of automatic speaker recognition. I will then provide methodology and motivation for the experiments which were conducted as a part of this research. The experiments and results are then presented followed by a conclusion. But, first let's introduce some basic background and terminology.

1.1 Speaker recognition

Automatic speaker recognition (ASR) can be considered a complement to speech recognition (O'Shaughnessy, 1986). For ASR we are interested in determining the identity of the person speaking based on the linguistic information contained in the speech signal. In comparison, speech recognition attempts to decode the linguistic information (i.e., the textual message) contained in an acoustic signal. For the task of speech recognition, variations dues to speaker identity are generally discarded as noise; because, the underlying textual message is all that is significant. However, in ASR, the textual part of the utterance is generally not as important as the speaker's identity. As a result, in text-independent recognition systems the textual information is considered as superfluous. How do we extract the necessary information from the signal to accomplish this task?

Generally speaking, the acoustic cues which distinguish voices are difficult to separate from those cues which reflect the identity of the sounds. The elements in the speech signal which carry this information are called phonemes. In ASR we are interested in exploiting the variability contained in the speech signal. It is this variability which contains speaker-specific information. The acoustic variation we seek to exploit is a result of physical differences in the vocal cords and vocal tract. However, it should be noted there are no acoustic cues which are exclusively the result of, or due to, speaker identity. This fact alone makes ASR a very challenging task. ASR is generally subdivided into two broad categories: *automatic speaker identification* (ASI) and *automatic speaker verification* (ASV).

Automatic speaker identification is a task whose primary objective is to identify the current talker out of a set of speakers. It asks the question, "Which of these speakers is talking?"; *automatic* s*peaker verification* or authentication asks, "Is this speaker who they claim to be?" Both tasks are similar in the fact that they make use of the information in the speech signal which emphasizes the speaker's identity and discard the textual information which is unused (except in the case of an authentication system which requires a specific utterance).

This information is carried in features of the speech signal; even though, as previously mentioned, no features exist which convey specifically speaker identity, there are those which convey physical characteristics of the vocal tract. Since variations occur from person to person in the physical composition of the vocal tract, these features are well suited for ASR. These variations are result of different resonance characteristics of the vocal cavity. These differences are usually attributed to variations in the physical layout (e.g., length and cross-sectional area of the vocal cavity) as well as characteristic of the vocal chords (e.g., thickness, resilience and deformation characteristics).

Both ASI and ASV use a stored database of speech reference patterns (templates) for N known speakers and both systems use similar analysis and decision techniques. Both ASV and ASI systems are based on traditional pattern classification techniques. First there is a training phase, followed by an evaluation phase. During the training phase samples of speech are sent through the ASR system, features are extracted from the acoustic signal, these feature sets are averaged over several training samples and a reference template is developed (i.e., a pattern of how a similar sample from the same speaker should look). During the evaluation phase an unknown speech segment is presented to the system, features are extracted from this segment and used to create a sample template. This sample template is then compared against all the reference templates constructed during the

Figure 1: A block diagram depicting a typical pattern classification algorithm.

Figure 2: Venn diagram showing universe N, M users and N-M impostor speakers.

training phase. Some type of similarity measure is made with respect to the sample and each reference pattern. Then a decision is made, usually based on the reference pattern with the greatest similarity. Figure 1 depicts a typical block diagram for a pattern recognition-based speaker recognition system.

While verification and identification share many commonalties they also differ in several areas which will be covered below. Most of these differences are due to the very nature of the tasks. For example, typically, law enforcement agencies are interested in identification systems (for obvious reasons); while verification systems are used to restrict access to a few select individuals.

An easy way to graphically represent a speaker space associated with both ASI and ASR is through the use of a Venn diagram (see figure 2). This Venn diagram shows an infinite universe of N speakers, M of which are considered users (or customers). Users are the speakers in our set of interest. For many ASI systems the concept of impostor speakers does not exist (Rudasi and Zahorian, 1992); thus, they deal with a closed set of speakers (e.g., M speakers). However, in ASV the set of impostor speakers is infinitely large and given by N-M, where N is infinitely large.

1.1.1 Speaker identification

A typical identification system starts with prototypes (templates) for all speakers, let’s assume there are N speakers. In this case, several samples of speech are recorded for each possible speaker. These samples are used to construct templates of each speaker; this is referred to as a training phase. An unknown sample of speech is presented to the system. From this unknown sample, a test template is constructed which must be compared against all N templates before a decision can be reached. Generally, the comparison takes the form of some type of distortion measure between the template and the unknown speech sample. After all comparisons are made the most likely match is chosen. In the case of a N-speaker system, N comparisons must be made for each unknown sample of speech.

Several sources of signal variability exist in a typical ASI system, the most important being: speaker vocal variations and channel variations. Because of the very nature of an identification system the subject may be reluctant to participate. This poses concerns about the subject attempting to disguise their voice (to prohibit correct identification). Unless speech samples are gathered in a clandestine manner a subject altering their voice can generally 'trick' the system. The communication channel presents another source of variability. In ASI, the channel is not controlled; signals are typically gathered after transmission across media such as: telephone lines, recording tape, or air waves. As a result of these transmission characteristics the signal-to-noise ratio is often low, leading to a poor quality signal. ASI systems are generally not real-time systems due to the number of comparisons necessary to make an accurate decision. Usually, there is no control over the content of the spoken text. As a result systems must be text-independent. These characteristics result in the fact the usually many sample speech patterns must be tested before identification is achieved.

1.1.2 Speaker verification

By contrast, in speaker verification, generally a speaker is cooperative. The system is used to verify that speaker's claim of identity. Although, disguising the voice is not a major concern, mimicry of someone else's voice is a concern. Most verification systems would operate in a real-time mode; thus, speed is an important factor. Usually, although there are exceptions, the communication channels characteristics can be controlled; eliminating background noise, crosstalk and others channel impairments. The vocabulary can be restricted (i.e. the verification of a digit string). Since most applications of a verification system would be in a controlled environment, signal-to-noise ratio can be somewhat predetermined.

A typical verification system starts with speaker templates (just as in identification). For the N speaker case, again we have N templates. Then a sample of unknown speech is presented to the system and a sample template is constructed; however, associated with this presentation is also a claim of identity. The system simply compares the sample template against the template belonging to the speaker whose identity is being claimed and renders a binary decision (yes or no). The fact that the system does not need to compare this sample of speech against all possible templates leads to a faster response time. Because of this fact real-time verification systems are not only possible but highly desirable.

1.2 Applications of a robust verification system

Speaker verification is the logical extension of the identification task. A verification system which is accurate, fast and reliable is an important objective with many possible applications. A good speaker verification system should not sacrifice reliability and accuracy for speed; but, should work toward the goal of consistently good verification with a minimum of input speech. To achieve such performance a speaker verification system must have a solid foundation built upon past work with speaker identification. The approach in this research is based on a robust technique for speaker identification which has been presented by Rudasi and Zahorian (Rudasi and Zahorian, 1992).

There are numerous applications for a speaker verification system which is fast and accurate. Such a system would find use in the commercial banking industry, the automotive industry and in government applications. A reliable verification system could be used for: voice controlled access, voice-activated control of machinery and equipment, as well as voice controlled computer systems. A simple but effective use for a hybrid system (a verification system that also serves another main objective) is in the voice control of a wheelchair. The physically challenged person with limited range-of-motion has significant difficulty in operating a motorized wheelchair. Some current techniques involve the use of a straw and puffs of air to operate a control system that in turn operates the device. A simple but reliable hybrid system that could reliably recognize operational commands from a specific individual would be a major improvement and would provide additional mobility to the individual.

1.3 Objectives of this research

The motivation for this thesis and its supporting research was to develop the methodology and algorithms necessary to apply the binary pair-partitioned neural network classifier to the speaker verification task. This thesis will present the methodologies used to perform this research, it will also convey the motivation behind why a particular algorithm was modified or enhanced.

As in all work based on pattern classification techniques, the obvious goal is to consistently achieve a high classification rate. However, in this research this goal was tempered by two other important factors: to achieve consistently high results regardless of the number of speakers and with a minimum amount of input speech.

CHAPTER TWO

BACKGROUND ON AUTOMATIC SPEAKER RECOGNITION

2.1 Current trends

2.2 Binary-pair partitioned neural networks for speaker identification

Binary-pair partitioned neural networks are based on a form of group partitioned called binary-pair partitioning. This scheme was first introduced by Rudasi and Zahorian (Rudasi and Zahorian, 1992) and applied to speaker identification. Group partitioning has been an accepted method of partitioning a large classification task into smaller, more manageable tasks. Binary-pair partitioning is a special case of group partitioning in which each of the elemental tasks are two-way classifications.

Figure 3: (Insert figure 3 near here) Examples of group partitioning and binary-pair partitioning

This approach partitions an N-way classification task into N-1 two-way tasks. A simple group partitioning scheme and how it compares to binary-pair partitioning is shown in Figure 3. The advantage of using BPP networks comes in the form of reduced training time for each pairwise network (sub-classifier). However, this advantage comes at a cost of increasing the number of elemental classifiers required for a task. For each N-way classification task N*(N-1)/2 elemental two-way classifiers are trained, each with a relatively short training time (as compared to a single large N-way network).

Figure 4 : Speaker identification performance of one large network versus a system of binary-pair networks (page 39 Lazlo's dissertation)

Figure 5: Training time comparison of one large network versus a system of binary-pair networks. (page 41 Lazlo's dissertation)

Rudasi and Zahorian applied this technique to closed-set speaker identification. A closed-set identification task is one in which the set of speakers is limited to only those in the group of interest. The experiments they conducted show that BPP networks are a reasonable alternative to a single large-scale network; provides excellent recognition rates with the advantage of reduced training time (Figures 4, 5).

With the motivation for using binary-pair networks firmly in view, let's look at the application of this technique to the ASI task. The basic approach follows a typical pattern classification model: feature extraction, training, evaluation. Each will be discussed as it applies to the identification task using BPP networks.

Proper feature selection is of great importance to the overall quality of the identification system. Features should accurately represent the acoustic signal in a compact manner. To this end the features of choice have been cepstral coefficients (Rudasi and Zahorian, 1992). The features selected and 'front-end' processing remained constant for all experiments reported in this work. In all cases raw acoustic data was sample at 16,000 Hz, the signal was partitioned into windowed frames which consisted of forty milliseconds of speech. These frames were extracted at twenty millisecond intervals. This led to a twenty millisecond overlap at the frame level. A Kaiser window was used for smoothing:

$$w_k(n) = \begin{cases} {I_0(b)}/{I_0(a)}, & |\mathrm{n}| \le Q \\ 0, & \text{otherwise} \end{cases} \tag{1.}$$

Where ***$w_k(n)$*** is the window function, ***a*** is empirically determined to be 5.33 (Zahorian, Nossair and Norton, 1993), b is given by:

$$b = a\left[1 - \left(\frac{n}{Q}\right)^2\right]^{\frac{1}{2}} \tag{2.}$$

and ***$I_0(x)$*** is the modified Bessel function of the first kind and given by the following series expansion:

$$I_0(x) = 1 + \sum_{n=1}^{\infty}\left[\frac{\left({x}/{2}\right)^n}{(n!)}\right]^2 \tag{3.}$$

Each frame was then represented by a twenty-nine term Discrete Cosine Transform coefficient vector (DCTC). The DCTC expansion is used because of its ability to compactly represent the magnitude frequency components of a signal as a series of coefficients. A twenty-nine term expansion was used based on previous work (Zahorian, et. al., 1993). The coefficients were then scaled, linearly, using:

$$X_i' = \frac{X_i - \overline{X}}{5 \cdot \sigma_X} \tag{4.}$$

Where $\boldsymbol{X_i}$ is the scaled value and equal to the original coefficient minus the original mean, divided by five standard deviations. This type of scaling normalizes the feature space and promotes better overall performance. It has been effectively used in previous speaker identification work (Rudasi and Zahorian, 1991; Rudasi and Zahorian, 1992).

Once the ‘front-end’ processing has been performed a system of binary-pair networks is trained to make pairwise discriminations based on the spectral features extracted. For the N-way task, the networks make discriminations such as: is the speech from speaker 1 or 2, speaker 1 or 3, ..., speaker 1 or N. Thus, there are N-1 networks trained to discriminate speaker 1 from all other N-1 possible speakers.

During the evaluation phase, different speech is used. A segment of speech is

2.3 Application of BPP to speaker verification

CHAPTER THREE

ALGORITHM DEVELOPMENT

In this chapter I will present the experiments which were performed. I will give a basic methodology which is followed for each experiments, then I will explain each of the experiments which were conducted, giving rationale for changes in the methodology where appropriate. Each of the experiments was used as a building block to improve either the reliability or the robustness of the verification system.

3.1 Basic methodology

After the acoustic data was extracted and scaled as mentioned above an initial binary-pair neural network was trained. The training parameters for this first network were determined empirically. This phase required several runs of the neural network, modifying only certain parameters, monitoring performance as a function of the varied parameter. For this initial training case a small number of speakers was used (ten). The small number facilitated faster network training, yet still was a representative sample of the actual training data. Several factors pertaining to the performance of the network were adjusted, namely: the number of nodes in the hidden layer, number of training iterations and the network learning rate. The neural network was trained in increments of 75,000 iterations, from 75,000 to 300,000. Overall recognition rate increased modestly from 75,000 to 300,000 iterations at the cost of training time. For the remainder of the experiments the number of iterations was fixed at 150,000 as this provided a tradeoff between reasonable recognition rate and reasonable training time. Several different hidden layer configurations were tested: 5, 10, 25, 50 and 100 nodes being the most significant. Again, the overall recognition rate improvement was modest at the cost of training time.

3.2 Algorithm development

3.2.1.1 Simple thresholding technique

The first experiment cen

3.2.1.2 Variations of simple thresholding

3.2.2.1 A template-based approach

3.2.3.1 Use of long term statistics to the basic template

3.2.4.1 Templates based on statistics of network activation levels

3.2.5.1 Addition of segmental information to the template

CHAPTER FOUR

EXPERIMENTS AND RESULTS

CHAPTER FIVE

CONCLUSIONS

BIBLIOGRAPHY

Norton, C. A., Zahorian, S. A., Nossair, Z. B., (1994) "The Application of Binary-pair Partitioned Neural Networks to the Speaker Verification Task," ***ANNIE '94,*** pp. 441-446

O'Shaughnessy, D. (1986). "Speaker Recognition," IEEE ASSP Magazine, October 1986, pp. 4-17.

Parsons, Thomas W., (1987), **Voice and Speech Processing**, McGraw-Hill.

Rabiner, L., Juang, B., (1994), **Fundamentals of Speech Recognition**, Prentice-Hall.

Rudasi, L., Zahorian, S. A. (1991). "Text-independent talker identification with neural networks," ***ICASSP-91***, pp. 389-392.

Rudasi, L., Zahorian, S. A. (1992). "Text-independent speaker identification using binary-pair neural networks," ***IJCNN-92***, pp. IV: 697-684.

Zahorian, S. A., Nossair, Z. B., & Norton, C. A. (1993). "A partitioned neural network approach for vowel classification using smoothed time/frequency features," ***Eurospeech-93,*** pp. II: 1225-1228.