An Efficient First Pass of a Two-Stage Approach for Automatic Language Identification of Telephone Speech

By

Jonathan K. Law

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Professional Studies
in Computing

at

School of Computer Science and Information Systems

Pace University

May 2002

Dissertation Signature (Approval) Page

We hereby certify that this dissertation, submitted by Jonathan K. Law, satisfies the dissertation requirements for the degree of Doctor of Professional Studies and has been approved.

_____-_____

Dr. Charles Tappert 20 May 2002
Chairperson of Dissertation Committee

_____-_____

Dr. Fred Grossman 20 May 2002
Dissertation Committee Member

_____-_____

Dr. Zhong-hua Wang 20 May 2002
Dissertation Committee Member

School of Computer Science and Information Systems
Pace University 2002

Abstract

An Efficient First Pass of a Two-Stage Approach for Automatic Language Identification of Telephone Speech

By
Jonathan K. Law

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Professional Studies
in Computing

June 2002

Automatic language identification, recognizing a speaker's language from a speech signal, is gaining increased importance in the context of economic globalization. The most accurate systems for such identification use multiple, large vocabulary continuous speech recognizers. But their scalability is limited, as each new language added requires complete recognizers and an enormous amount of training. For increased efficiency, we propose a Two-Stage approach: an efficient clustering algorithm (least cost) first selects the top candidates at an accuracy level of over 80%. Then, speech recognizers (high cost) would be used to narrow the field from approximately three candidates to correctly identify the language.

In this research, we describe the first phase of this Two-Stage process in which we are able to narrow the list of possible languages through an extremely scalable and flexible system. We cover how test and reference patterns (acoustic feature vectors) are extracted from speech utterances, how cepstral coefficients are used, and how reference models are generated from the reference patterns using Vector Quantization (VQ) clustering algorithm. Various distance measures are also examined in the selection phase to find the best method. We show using the top-N strategy in this first stage leads to substantial improvements over existing systems in discriminating between different languages, and our experiments showed the top three choices yield 87.2% in a 5-language task, 88.6% in a 7-language task, and the top five yield 87.5% in a 10-language task. The second iteration in the 10-language task easily narrows it down to three choices with 80% probability. Further research could improve the entire process, but combining this particular methodology with current best practices has proven an extremely efficient and cost effective way to address the challenges of automatic language identification.

Acknowledgments

It really did take a village to help me get through this program. There are many people to thank for their support and encouragement, without whose help this thesis would not have been possible.

First, I would like to thank my advisors. Dr. Tappert has diligently invited scholars and industry experts in the latest technological fields, which has further sparked my interest in speech recognition. Dr. Wang has tirelessly shown me his unquenchable curiosity and love for the subject of speech recognition technology. I must also thank Dr. Grossman for the reality check he provided, as his constant and justifiable rejections of my initially unachievable topics helped shape my work into something that could be accomplished. I must also thank the constant support and encouragement of the Computer Science department faculty, which kept me going over the last three or so years.

Next are all the friends I've made, a list is too long to mention, but whose friendship, Indian village jokes, and "Just-in-Time" motto seemed to work every time on our last minute assignments. My dear friend Patrick Wong, who introduced me to this program and also my "room mate" during these college years. My friends made those long, crazy monthly trips across the country not only enjoyable, but also memorable.

Finally, I thank my family, which has been wonderfully supportive and thoroughly believed my "midlife crisis" excuse for going back to school 20 years after my last college exam. I first thought about applying for my doctorate degree when my sister Joyce received her A.I.A. certification. Architecture was my own major in college, but with her degree she has done everything I always wanted to do. I thank her for our sibling rivalry and taking full advantage of the opportunities that our parent never had; this degree will hopefully once again put me one step ahead of my sister. My brother-in-law Oliver, who is not only a seasoned proofreader, but also works well under imposed deadlines. All errors in this paper are mine.

And the most important person behind this accomplishment is my loving wife Joan, who has put up with me all of these years and has taken great care in the raising of our five lovely college-bound children. I am indebted to her for her support and love.

Table of Contents

List of Tables

Figures List

Chapter 1.

Introduction

Due to the rapid growth in processing power of the computer chip in recent years, hardware and software manufacturers have begun incorporating speech recognition into their products. This trend is creating awareness of speech recognition as a viable tool. Corporate America, which is constantly looking to increase its bottom line by finding ways to increase productivity, is trying to leverage speech recognition. Most major corporations in the United States have business relations with other countries around the world, and the expansion of speech recognition in understanding human speech demands that communication cross the boundaries of languages.

An automatic language identification system can play an important role in a speech recognition system in which it acts as the front-end processor, such as routing foreign telephone calls to appropriate operators with language skill. It could even save lives when split second decisions alerting the right medical personnel during an emergency are the difference between life and death. In addition, all foreign government defense departments have a keen interest in using this capability for monitoring purposes. But duration of unknown utterances and high number of languages pose challenging tasks for researchers in multilinguality and language identification. This research tries to solve the scalability and performance problems by introducing a Two-Stage approach that

would efficiently discriminate between languages and easily expands to add many languages without those complex labeling or transcription work

1.1 Overview

The human brain is still the most advanced and accurate language identification system today, far exceeding the data storage and processing power of a supercomputer. Humans can easily determine those languages they already know or have heard before, usually with only a short duration of utterances. But using machines to automatically and accurately identify a sample of speech by an unknown speaker is still a huge challenge in speech recognition research. The biggest problem in designing a language identification system is the scalability issue. The current best systems use multiple, large vocabulary continuous speech recognizers. As these systems include a complete word and sentence level recognizer for each language, adding a new language would involve the enormous task of labeling speech to train phone recognizers. Another popular approach exploits the phonotactic properties of the languages with recognition and language modeling occurring at the phone level. Again, the requirement of labeled speech for a large subset of the language makes it very hard to extend beyond the most common languages.

We propose using a Two-Stage approach with an efficient front-end process that would narrow the unknown utterance down to the 2 or 3 top language choices, and then use other complex systems such as phone or word recognizers for an exact match. Most top performers involve heavy use of linguistic data. Our approach instead uses the acoustic models of phones in each language and finds the highest likelihood resulting from an efficient way of clustering algorithms. This dissertation reports how we adopted

speaker identification methodology via statistical pattern recognition and applies it to language identification using Vector Quantization approach for the First Pass. We show that this method would lead to substantial improvements over existing, fast systems in discriminating between different languages, and return those top candidates with a combined accuracy of over 80 percent. A next stage would then further narrow down the identification process from here.

To fully understand this approach, we first look at the differences among languages, the driving forces and the challenges.

1.1.1 Differences Among Languages

When we are listening to foreign languages, most of the time we can guess the language by the phonemes. A phoneme is the term defined by linguists to classify speech into a number of abstract categories for grouping together subsets of speech sound. For example, American English has about 40 phonemes. Even though no two speech sounds, or phones, are identical, all of the phones classified into one phoneme category are similar enough so that they convey the same meaning. But there is much overlap of the phoneme sets, and there can be differences in the way the same phoneme is interpreted in two different languages. In English, letters /l/ and /r/ are two different phonemes. Then the frequency of occurrence of phones and the phonotactic rules in languages can also differ significantly. For example, phoneme clusters /sr/ and /sp/ are quite common in Tamil and German, but not in English.[5] Prosodic features, rhythm, and intonation, also vary among languages. English is known to be stress-timed, and French is known to be

syllable-timed. All these unique features are the subject of research for automatic identification of languages by machine.

1.1.2 The Challenge

To formally distinguish one language from another, these are the main categories:

Phonetics –The study of speech sounds (phones) as physical entities on a sub-language basis. There are in any language a limited number of recurrent, fairly distinctive speech units. The number of phonemes in a language ranges about 15 to 50.

Phonotactic – The study of frequency distributions and combinations of phonemes. Some combinations that occur frequently in one language are illegal in another.

Prosodic – The study of sound patterns, which can be analyzed in terms of duration, pitch and stress. For example, one syllable of most words has a heavy stress or accent that sets if off from the other syllables in stress language.

Morphology – The study of ways in which words are built up from the smallest meaningful parts.

Syntax – The study of how words can be legally strung together.

The last forty years have produced technology capable of accurately processing spoken language, which has been marketed commercially in many industries. However, only a limited vocabulary and controlled environment are being utilized. Automatic language detection encounters most of the challenges of speech recognition. They are:

- Continuous flow of speech
- Volume of speech sound wave
- Paucity of information in speech sound wave

- Variability of voice and speech patterns
- Varying noise environment

In addition, languages have characteristic sound patterns that differ in the inventory of phonological units (speech sound categories) used to produce words. The frequency of occurrence of these units differs as well, and the order in which they occur in words[4] language identification, and the ability of systems to understand multiple languages are still in their infancy.

1.2 The Driving Force

Only in recent years has speech recognition achieved the reliability and flexibility to attract the interest of businesses willing to invest in the infrastructure that is needed for ongoing research in this field. External forces such as increasingly powerful microprocessors complement the significant technological advances that put speech recognition system into many small devices. The seamless integration of computer-telephony technology, and successful user acceptance of machine voice, has encouraged most retail businesses to use machine routing of calls today.

Moore's law has held true in the last twenty years of the microprocessor industry, which is the primary factor in the migration of advanced speech recognition from research lab into the business world. All computational experiments for this thesis were done on a Dell laptop computer with a 500 Mhz PIII Intel chip. Not only has the cost of equipment been decreasing, but also the size of the hardware is decreasing to the point where a digital signal-processing chip can be found in practically every consumer electronic device. In the near future, one should expect all paper user manuals or

operational instructions to be replaced by machine response in audio of a user's native language.

Global business also plays an important role in creating huge demand for telephone-based, multi-lingual speech recognition systems. It would provide a cost-effective way for businesses needing 24-hour telephone message and communication support as their business is actually non-stop amidst the world's time zones.

1.3 Related Researches

Most of the approaches to language identification have adopted the successfully proven techniques used in speaker-independent speech recognition systems. This paper will briefly discuss the major breakthroughs in speech recognition in groups of twenty years to give some context to the research conducted. The first attempts to devise systems for automatic speech recognition by machine were in 1950 at various research labs in the U.S. and Great Britain[6, 11, 30]. Research in language identification was actually started twenty years after the research began in the field of speech recognition.

1.3.1 The first twenty years

In the 1970's, a highly successful group effort in large vocabulary speech recognition at IBM called the New Raleigh language[40] for simple database queries yielded the laser patent text language[17] and the Tangora[16] for dictation of simple memos. The truly speaker independent research[34] was developed at AT&T labs in 1979. Here a wide range of sophisticated clustering algorithms was used to determine the number of distinct patterns required to represent all variations of different words across a wide samplings. This research has a significant influence on the later research where

refinement was achieved over the next decades, including the framework on which this thesis is based.

A significant technology shift from template-based approaches to statistical modeling methods led to the highly successful model of hidden Markov modeling (HMM)[9]. This widespread publication and theory has been applied in virtually every speech-recognition research laboratory in the world.

1.3.2 The last twenty years

In the 1990's, the major milestone for research was the availability of a public-domain multilingual speech corpus from Oregon Graduate Institute (OGI)[27]. Prior to this, the field of spoken language ID research has suffered from the lack of a common speech corpus that could be used to evaluate different approaches to the problem. A corpus is a collection of linguistic data, either compiled as written texts or as a transcription of recorded speech. The initial corpus contained 11 languages, which was used by the National Institute of Standards and Technology (NIST)[29] to conduct an annual common evaluation of spoken language ID algorithms.

Many papers have been published during the last 20 years on this subject, and model training has involved neural networks or HMMs. Zissman[49, 50] exploited the fact that a stochastic grammar for one language can be developed based on the acoustic models of a different language. Li [21] has applied speaker recognition techniques for language identification by classifying incoming utterances based on the similarity of the speaker of that utterance with the most similar speakers of the target languages. He based this similarity measure on spectral features extracted from experimentally determined

syllabic nuclei within the utterances. Muthusamy examined pitch variation within and across broad phonetic segments in his doctoral thesis[24] in which he found other prosodic information such as duration and syllabic rate to be more useful.

1.4 Overview of Proposed Two-Stages Approach

It is clear that there is really neither a defacto methodology to spoken language identification, nor a standard set number of languages to be tested. Moreover, the level of performance is not ready for commercial usage. As of today, there is no application in the market that can claim to identify even 2 languages. Since human identification performance can be very accurate with a very short duration of speech[25], it is clear that better-automated systems can be developed. We propose a simple 2-stage approach; first using an efficient clustering algorithm to select the top candidates, and then following up with speech recognizers to correctly identify the language

1.4.1 First Pass - Efficient Algorithm in Selecting Top Candidates

In order to find the most efficient algorithm and cost effective way in selecting the top candidates among many unknown languages, we first need to understand how human beings do it. When learning to speak and understand different languages, most human beings are exposed to one particular language constantly during their "learning phase". Over time, our data store (brain) records patterns and when the same pattern is detected, our nerve system naturally communicates with the speaker and acoustic background. Using this same basic pattern-recognition approach with machines, speech parameters are obtained using a type of spectral analysis that concentrates on the "short duration" characteristics of the speech signal. The short-time spectral measurements are performed

sequentially over time, producing a sequence of spectral feature vectors, called a speech pattern. This speech pattern is then compared with each class reference pattern and a measure of similarity between the unknown pattern and each reference pattern is calculated.

Vector Quantization (VQ) is a very efficient source-coding technique and has been widely used and successful in speech recognition research. It is a procedure that encodes a vector of input, a segment of waveform or a parameter vector that represents the segment spectrum, into an integer that is associated with an entry of a collection (codebook) of reproduction vector. The main advantage is that it requires less storage for signal analysis and reduces computation for determining similarity of speech analysis vectors. This would work well if also proven successful in language identification; as it can be easily expanded to add more languages without lots of front-end work, or transcriptions.

The following figure illustrates the concept with 2 languages and 3 dimensions of the acoustic space. The circles refer to the acoustic vectors from the Language ID #1 and the triangles are from #2. In the training phrase, a language-specific VQ codebook is generated for each known language by clustering the training acoustic vectors. Solid circles and solid triangles for language ID #1, and #2 show the result codeword (centroids), respectively. The distance from a vector to the closest codeword of a codebook is called a VQ-distortion. During the recognition phase, an input utterance of an unknown language is "vector quantized" using each trained codebook and the total VQ distortion is computed. The language input corresponding to the VQ codebook with the smallest total distortion is identified.

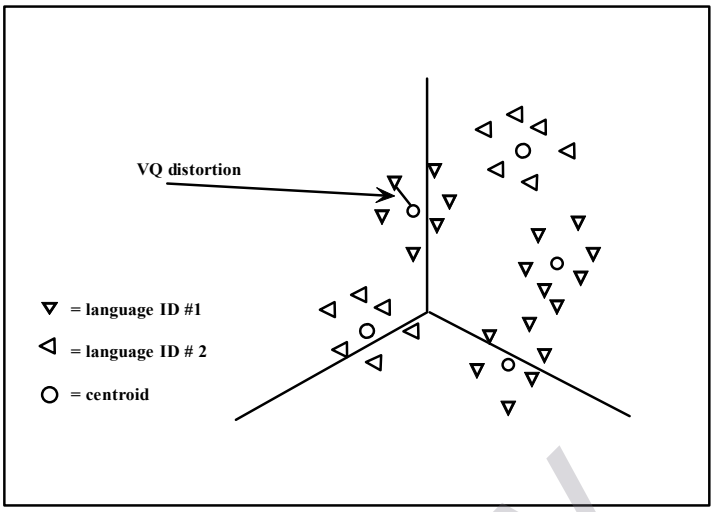

Figure 1-1 Conceptual diagram of Vector Quantization Distortion

The VQ distortion is the distance between the vector and the centroids of the same region, in this case, the same language model.

Given an utterance of unknown language, we will use the following algorithm to calculate the score against the model for each language. The one with the smallest score is regarded as the language of this utterance. The following figure depicts the feature matching process. The unknown test utterance is matched against all language models, and the language with minimum distortion is selected as the winning language. Using the nearest-neighbor approach, we can select those top candidates with combined probability of over 80% for the next stage.

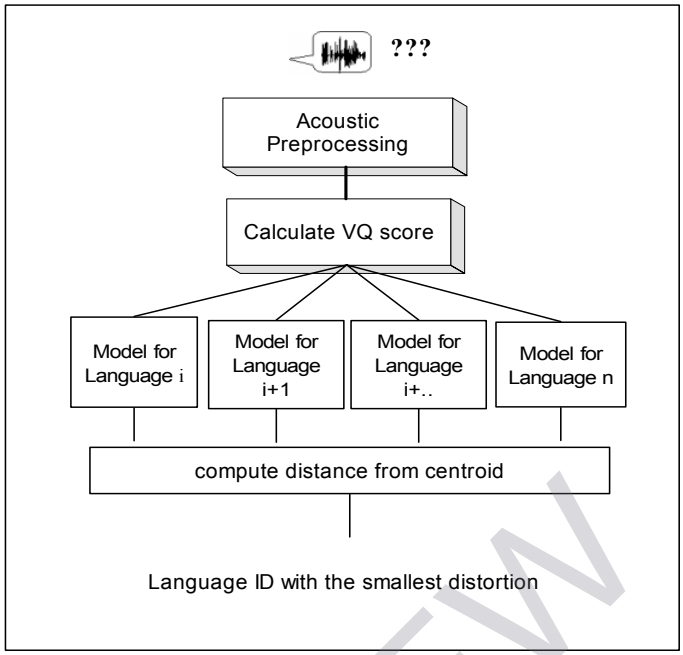

Figure 1-2 Conceptual diagram of nearest neighbor selection

The unknown test utterance is matched against all language models, and the language with minimum distortion is selected as the winning language

1.4.2 Second Stage – Accurately Identify unknown signal

The second stage is the process of trying to accurately identify the unknown speech signal using various high performance approaches. Since we have narrowed down to a handle full of top candidates, there are a few options available to correctly identifying them. The most common and easiest way is to allow users to select from the option prompt generated by VoiceXML, and responded either by DTMF or speech. The multi-lingual framework by Law [20] has provided a template-based engine that would create a user-friendly native language prompt using the languages of the unknown top candidates.

Another proven successful approach, which fits nicely within our framework, in terms of feature extraction, is the parallel language dependent phone recognition by Yan [47]and Zissman et al[50]. This type of approach exploits the phonotactic properties of the languages, and also does not need to recognize words. The recognition and language modeling are done at the phone level. The following figure depicts the basic concept.

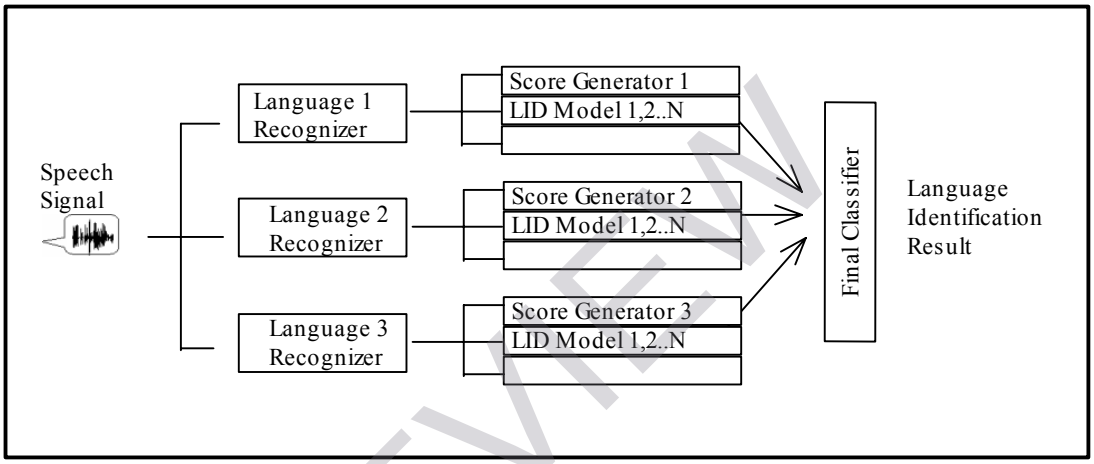

Figure 1-3 Overview of a Language-Dependent Phone Recognition system

The approach that exploits the phonotactic properties of the languages and not words

This approach is able to achieve identification accuracy in excess of 80% using 10-second utterances in 6 languages[47], and over 90% on 3 or less languages. The biggest drawback is the requirement of labeled speech for a large subset of languages used. Since it is based on multiple language-specific phone recognizers, it requires labeled speech to train those recognizer.

1.5 Outline of the Dissertation

This chapter briefly describes an idea of how and why we use this simple Two-Stage approach for language identification. Chapter 2 will describe speech recognition

and the most common recognition applications. We will then present comparative studies, experiments, and results of those approaches we have discussed earlier. Chapter 3 provides details on the Vector Quantization approach, and the data set from the Oregon Graduate Institute and corpus structure. The technical specifications and implementation process is outlined in Chapter 4. Details of the improvement in both configuration and algorithms, and how results were generated for a 5-language set (NIST 1996), 7-language set and 10-language set as the baselines. We then performed initial improvements on various configurations, which are described in Chapter 5. The use of various distance measures and data fusion to the output scoring is detailed in Chapter 6 and revised baselines are generated. Conclusions and future work, based on our findings from this research, are discussed in chapter 7. The last section is the appendix and references.