SPEAKER RECOGNITION OF DISGUISED VOICES: A PROGRAM FOR RESEARCH

Robert D. Rodman
(rodman@csc.ncsu.edu)
Department of Computer Science
North Carolina State University
Raleigh, North Carolina, 27695-8206
U.S.A.

ABSTRACT

A program for carrying out research on the speaker recognition of disguised voices is proposed. Such a program would consist of the following:
1) Definition and classification of disguises
2) Creation of databases of disguised voices
3) Testing of conventional speaker recognition systems on the disguised-voice database
4) The investigation of which basic methods of modeling the vocal tract -- HMMs, vector quantization, neural nets, etc. -- are most effective on the various types of disguised voices
5) The feasibility of automatically detecting the use of disguised voice
6) Methods for computationally "undisguising," or compensating for, a disguised voice
7) The use of several disguise methods at the same time

1. INTRODUCTION

Speech processing by computer is a major field of endeavor. It is multidisciplinary, encompassing electrical engineering, computer science, linguistics, speech communication, telecommunications, among others. There are three distinct sub-fields of speech processing: *speech synthesis*, *speech recognition*, and *speaker classification*.

This paper focuses within the area of speaker classification. Speaker classification is concerned with extracting information about individuals from their speech. It is a deep and growing area of research. From speech alone good guesses can be made as to whether the speaker is male or female, adult or child [15]. A person's mood, emotional state and attitude may be indicated in their speech. Anger, fear, belligerence, sadness, indignation, reluctance, elation may all be detectable in the speech signal [6] [16] [18] [19] [28] [30] [65] [82].

What language a person is speaking, whether that person is speaking formally or informally, to intimates or to strangers, to persons of higher social rank or lower social rank, to children or to adults, to foreigners or to nationals, may be determined from the speech signal [7] [33]. Evidence for a person's nationality, region of upbringing, social standing, and education level may be found in the speech signal [3] [53].

The identity of a person may be determined or validated by that person's speech. *Speaker recognition* (or *voice recognition*) is the most heavily investigated sub-area of speaker classification, and is the general topic of this paper [1] [5] [9] [13] [21] [34] [36] [43] [70] [73].

This paper is specifically concerned with the identification of speakers whose voices are "disguised," either deliberately or non-deliberately, and either electronically or non-electronically. These notions will be defined precisely in a later section.

Speaker recognition is complementary to speech recognition. Both techniques use similar methods of speech signal processing up to a point, but speech recognition, if it is to be speaker independent, must purposefully ignore any idiosyncratic speech characteristics of the speaker, and focus on those aspects of the speech signal richest in linguistic information. Conversely, speaker recognition must amplify those idiosyncratic speech characteristics that individuate a person.

Humans are adept at speaker recognition, even when voices are disguised [57]. A human can identify familiar speakers on the telephone after listening to a very short segment of speech. Humans are less effective at recognizing the voices of less familiar speakers, but can be "trained" to improve through additional exposure to individuals and their speech.

Like humans, computer speaker recognition systems need to be trained to learn how each person "sounds." The more training data, the better, just as the better one knows a person, the more likely one is to recognize them by their speech.

The range of sounds that can be produced by a human being is related to the physical size and shape of the speaker's vocal tract [29] [61] [80]. The vocal tract consists of the oral and nasal cavities, the glottis, tongue, velum or soft palate, hard palate, teeth, and lips. The elasticity of the tissue in the vocal tract also affects the sounds that are produced by an individual. With so many physical parameters contributing to the range of sounds that each individual can make, there is reason to believe that a person can be uniquely identified by voice alone [80]. Moreover, most of these physical features remain essentially unchanged when the voice is disguised, so that in many cases identification is still possible [53]. This is the underlying thesis of our approach to the speaker recognition of disguised voices.

2. DESIGN TRADEOFFS IN SPEAKER RECOGNITION

References to the discussion in this section are [21] [36] [43].

The design of a speaker recognition system can be greatly affected by the targeted application area of the system. Many tradeoffs are necessary to build a system that meets the constraints of the application. Some of the most common tradeoffs are listed below:

2.1. Speaker verification versus speaker identification

Speaker verification is determining whether a speaker is who they claim to be, for example, to gain entry to a secure area. Verification systems must deal with two kinds of errors: false rejection and false approval. False rejection occurs when a legitimate person is denied access. False approval, a more serious error, occurs when an impostor is granted access. The designers of speaker verification systems must adjust the decision criteria so that false approval is as low as possible without causing the false rejection rate to be unacceptably high.

Speaker identification is the process of determining which speaker, if any, in a group of known speakers, closely matches an unknown speaker. The identification may be *closed set*, where it is assumed that the unknown is in the set of known speakers; or *open set*, where the unknown speaker may or may not be in the set of known speakers. For closed set identification, the speaker recognition system can simply choose the known speaker that most closely matches the unknown, providing there are no close runners-up. Open set identification is more difficult. It is equivalent to performing a closed set identification followed by verification. Verification is needed to ensure that the match between the unknown and the "winner" of the identification task is close enough to be the same speaker.

2.2. Text-Dependent vs. Text-Independent:

Text-dependent speaker recognition systems are trained by having each speaker read a short, prescribed text of no more than several words. The text may be repeated once or twice, but the overall training period is brief. During the recognition (testing) phase, the unknown speakers must speak the same prescribed text that was used for training. These systems are suitable for security applications where the valid speakers are cooperative, and the security requirements are non-critical.

Text-independent systems allow the user to read any text during both training and testing. Typical text-independent systems require more training data than text-dependent systems. This is necessary to ensure that the full range of vocal sounds of a speaker can be captured during training. Text-independent systems are suitable for applications where the speakers are not cooperative, such as ones occurring in law enforcement. Often, the "testing" phase is a recorded message having to do with illegal activity, such as a threat. The "training" phase is drawn from interviews with suspects. This shows that the testing phase may precede the training phase under certain circumstances.

2.3. Ideal recording environment versus noisy environment

Ideal recording environments consist of high quality microphones used in rooms with little or no background noise or reverberation. The same microphone and room is used for both training and testing sessions. Using the same equipment for both training and testing eliminates any channel variations that might be falsely used as characteristics for identification.

Unfortunately many practical uses of speaker recognition occur in noisy environments, and in situations where channel variation is unavoidable. A bomb threat recorded by a 911 logging device, a surveillance tape of a drug deal, a wire tap and a personal threat on a home answering machine all engender noise and channel variation. Much of today's research in speaker recognition addresses the issues raised by noisy environments and idiosyncratic channels.

2.4. Real-time operation versus off-line operation

The nature of security applications requires that the speaker recognition system respond within a short period of time. Other applications, such as those occurring in law enforcement, may not have this constraint [36].

3. BACKGROUND OF SPEAKER RECOGNITION

Research on speaker recognition began in the 1960's when scientists attempted to use the speech spectrogram as a tool for speaker recognition. [5] [34] [73 [76]. Even with human experts interpreting the spectrograms, the results were limited. At the time computer technology was not sufficiently advanced to aid the process.

Advances in computer technology in the post-1960s triggered a series of research projects on speaker recognition. Although progress was made in the area of text-dependent speaker recognition, text-independent systems that could deal with channel and speaker variability were not as successful.

One of the first techniques proposed for speaker recognition was the long-term averaging of features extracted from the speech signal, both in the time and frequency domains. In this technique, a large number of feature vectors is obtained from each known speaker. The average and variance of each component of the feature vector are computed for all of the examples from an individual. The similarity of speakers is determined by computing a weighted distance measure between the average feature vectors of two speakers [50].

The accuracy of speaker recognition systems using long-term averaging is highly dependent on the duration of the training and test utterances. With shorter utterances, the intra-speaker variance increases due to differences in the content of the utterances. Using the long-term averaging technique, one investigator reported an error rate of 80 percent

for 0.06 seconds of test data, 34 percent for 2.5 seconds of test data, and 6 percent for 40 seconds of test data [81]. Researchers have used, and still use, the long-term averaging approach with several different kinds of features, such as inverse filter spectral coefficients, line spectrum pair (LSP) frequency features, pitch, and cepstral coefficients. [9] [13] [67].

Vector Quantization (VQ) is a more effective method than long-time averaging. Rather than a single cluster of data for each speaker's model, VQ segregates data into multiple clusters and determines their centroids. When VQ is used for speaker recognition, a codebook is created for each known speaker by applying the VQ algorithm to a set of feature vectors derived from training utterances. For testing purposes, a comparison of the unknown speaker's vectors is made with the codebooks of each known speaker. The accumulated distortion between the unknown and the codebook determines identification or probability of verification. [40] [69].

Hidden Markov Models (HMMs) have also been used for speaker recognition. Since HMMs can model the stationary and transient properties of a signal, they are a good choice for modeling speech signals. Vowels are relatively stationary whereas consonants are relatively transient. The probabilistic nature of HMMs enables them to represent speech that contains variability with accuracy [52] [74].

Artificial Neural Networks have also been used for performing speaker recognition. In one such system a feed-forward network was created for each known speaker. Each network contained one output that was trained to be active for its speaker only. For speaker identification each input vector was fed forward through each network. The network with the highest accumulated output values determined the identification. For speaker verification the input vectors for the unknown were fed forward through the network belonging to the individual wishing to be verified. If the average output value was greater than a threshold, the unknown speaker would be accepted [54].

Two other strategies for using neural networks to perform speaker recognition were presented by Rudasi and Zahorian [62]. The first was to use one large network with one output per known speaker. The second was to use binary networks, small networks for distinguishing between two speakers. Although there would be many more networks, the training time for each would be very short. Since each network is responsible for only a small portion of the overall classification, the binary networks can be very specialized and have much better performance than a large network.

Finally, Time-Delay Neural Networks (TDNNs) were developed to capture transient information using a connectionist approach [4].

Text-independent speaker recognition is based on the notion that acoustic parameter measurements of individual speakers speaking any speech may be used to characterize the speaker uniquely. In [23], for example, the authors describe a method of comparing speech utterances to determine whether or not the underlying probability

density functions are the same, hence likely to have been spoken by the same person. Using the King telephone database, accuracies varying between 35 and 90 percent were reported.

Segregating systems treat the text-independent speaker recognition task as a two-step process. First, all input vectors are segregated into categories based on their acoustic phonetic properties. Then, for each category, the vectors from each individual are compared with vectors from the unknown. The weighted summation of the scores from each category is used for making decisions. The conceptual underpinning of this approach is that specific sounds produced by the unknown speaker are compared with the same sounds produced by the known speakers.

Several segregating approaches for speaker recognition appear in the literature. Wang described a system in which feature vectors were segregated using VQ. Features in each category were weighted by the variance within each category [79]. Savic presented a system that used Ergotic HMMs in which each state represented a different broad phonetic category. A Bayes classifier was used for determining the identity of vectors in each category [71]. Matsui and Furui described a similar system in which HMMs were used for segregation and VQ was used within each category. Thus, each known speaker was represented by several codebooks, one for each category [52].

Several other institutions, such as the Oregon Graduate Institute (OGI), MIT Lincoln Laboratory, Nagoya University, AT&T, BBN, ITT and NTT are also working in the speaker recognition field. Many of these current research efforts have focused on new pattern matching techniques and channel variation compensation [21] [22] [23] [25] [32] [44] [58] [66].

At North Carolina State University we have developed a speaker recognition system that uses a different approach to segregation. We use coarse grained features for the initial phase of segregation, and fine grained features for the second stage of comparison. For segregation our system uses only features that model the rough structure of the spectral envelope, viz. formant frequencies. For classification our system uses features that model the fine structure such as the inverse filter spectral coefficients. This combination has yielded very low error rates. The methods and test results are reported in [36].

Relatively few studies have dealt with the issues of disguised voices in speaker recognition. We have compiled an extensive bibliography of works concerned with both human and computer recognition of disguised voices. Rather than list them all here, they can be identified as the starred entries in the References.

Predecessory work in voice disguise is represented in [18] [19]. This pre-World War II research addressed the issue of whether people could recognize various emotions expressed in the voice of actors. Different emotional states are known to affect voice

quality, and have long been problematic in speech and speaker recognition. In effect, an extreme emotional state is a form of voice disguise.

Direct studies of disguised voice speaker recognition began in the 1970s. There tended to be two types of research. One type was non-electronic and attempted to measure the ability of non-expert humans to identify other humans who were disguising their voice in a variety of ways [56] [57] [59]. The second type was electronic, often involving speech spectrograms, or so-called "voiceprints." The ability of experts in voiceprint interpretation to perform speaker recognition of disguised speech was measured [15] [26] [75] [76] [80].

Most of the early studies concerned themselves with the deliberate disguising of the voice, such as speaking in falsetto or feigning a speech defect or foreign accent. Studies complementary to these consider the effect on speaker identification of non-deliberate voice distortion such as those that occur due to aging, intoxication, illness, or emotional stress [8] [15] [27] [28] [65] [78] [82].

Some research concerned itself with voice mimicry. This not only disguises the voice of the speaker, but also has the additional intent of representing a different speaker. The ability of both humans alone, and humans using electronic aids, to detect mimicry has been studied, though not extensively [31] [42] [55].

There has been a recent upsurge of interest in criminal voice disguise, especially that associated with acts of terrorism. This interest has given rise to the field of *Forensic Phonetics* [3] [29]. It has also, in part, motivated the founding in 1994 of the journal *Forensic Linguistics*, in which much of the recent work in voice disguise has been circulated. Many of these papers have already been cited and others can be found in the References.

4. APPLICATION AREAS OF SPEAKER RECOGNITION

The ability to identify people uniquely through speech has spawned several application areas. Voice disguise is significant in two of the most important ones.

4.1. Access restriction

Access restriction is the area in which speaker recognition technology has had the greatest impact. While access to secured areas can be restricted with the use of keys, magnetic cards, and lock combinations, all three can be lost or stolen. Speaker recognition can provide an alternative or supplemental means of entry.

Although voices cannot be stolen, they can be copied with recording devices. Thus, voice-based security systems must protect themselves against this ploy. This can be achieved by varying the text to be spoken by the person wishing access to the secure

area, which requires combining speaker recognition with speech recognition. Both the identity of the speaker and the linguistic content of the speech must be verified.

Another security concern is access to computer systems via terminals, phone lines, automatic teller machines, etc. Currently, access to such systems is restricted by the use of passwords or personal identification numbers. Again, these numbers can be lost, stolen, or copied. In a similar manner to physical access security, speaker recognition could provide security for computer systems.

Voice-based entry systems of all types are vulnerable to disguise in two ways. First, an impostor may gain illicit entry through a voice disguise that mimics a valid voice. Conversely, a valid person may be denied entry because of an unintentional voice disguise that accompanies an illness, emotional stress and similar factors.

4.2. Forensics

The use of speaker recognition in law enforcement is becoming commonplace where evidence is in the form of voice recordings of the suspects [3] [5] [17] [20] [29] [35] [36] [39] [46] [77]. Such cases might include bomb threats, ransom negotiations, undercover tape recordings, wire taps, etc. Results are not always definitive, but they often direct the investigation away from unlikely suspects and toward likely ones. The results of speaker recognition analysis are not freely admitted as evidence in the courtroom, but with improved techniques, and with judges now beginning to understand the significance of probabilistic findings, the situation is expected to change in the future.

The use of voice disguise by criminals is not uncommon, and ranges from the simplistic, Hollywood inspired "handkerchief over the telephone mouthpiece," to sophisticated electronic techniques [29] [45] [51] [56] [59] [60]]64]. Voices may be disguised in the carrying out of any speech specific criminal acts. These include annoying or threatening phone calls, bomb threats, extortion, blackmail, etc.

5. RESEARCH IN DISGUISED VOICE SPEAKER RECOGNITION

5.1. Types of disguises.

One of the striking features of the limited literature in voice disguise is how unsystematically the studies of disguise are treated. Studies permit speakers to "disguise their speech as completely as they could," [26] or to disguise their speech "in a manner which [the speaker] felt would conceal his identity most effectively," [57] or "obscure your identity to the best of your knowledge by disguising your voice while still clearly delivering the meaning of the prescribed sentence" [45].

For the sake of discussion we would like to define *voice disguise* to mean "any alteration, distortion or deviation from the normal voice, irrespective of the cause." The definition is imperfect in several respects, including the lack of a good definition of

normal voice — for example, it is normal for voices to taper off to creaky, or for syllables in falsetto to occur. Nonetheless, such a definition allows us to create a taxonomy of voice disguises that permits research to be more sharply focused.

We further define *disguise* along two independent dimensions: Deliberate versus nondeliberate, and electronic versus nonelectronic. Deliberate-electronic would be the use of electronic scrambling devices to alter the voice. This is often done by radio stations to conceal the identity of a person being interviewed. Nondeliberate-electronic would include, for example, all of the distortions and alterations introduced by voice channel properties such as the bandwidth limitations of telephones, telephone systems, and recording devices. Deliberate-nonelectronic is what is usually thought of as disguise. It includes use of falsetto, teeth clenching, etc. Nondeliberate-nonelectronic are those alterations that result from some involuntary state of the individual such as illness, use of alcohol or drugs (the effects are involuntary), or emotional feelings. Please refer to Table 1.

Broad taxonomy of voice disguise:	**DELIBERATE**	**NONDELIBERATE**
ELECTRONIC	Electronic scrambling, etc.	Channel distortions, etc.
NONELECTRONIC	Speaking in a falsetto, etc.	Hoarseness, intoxication, etc.

Table 1: Type of Disguises

We propose to focus on a single cell in the above table: Nonelectronic-deliberate. Electronic-deliberate disguise is relatively uncommon, occurring in only one to ten percent of voice disguise situations [45]. Electronic-nondeliberate disguise concerns itself mainly with channel distortions, both wire and wireless, and is a well-studied area of research.

Nonelectronic-nondeliberate disguise is of interest, and is a poorly researched area, but such studies are best left to researchers with access to medical personnel in the case of illness, or psychological personnel in the case of emotions.

Even within the nonelectronic-deliberate voice disguise area (to be called henceforth simply "disguise"), there is extreme richness and variety. Please refer to Table 2 for some of the kinds of disguises that have been used and/or studied. The table is not complete, and is in principle not completable given human ingenuity.

PHONATION	**PHONEMIC**	**PROSODIC**	**DEFORMATION**
Raised pitch (falsetto)	Use of dialect	Intonation	Pinched nostrils
Lowered pitch	Foreign accent	Stress placement	Clenched Jaw
Creaky voice (glottal fry)	Speech defect (e.g., feigning a lisp)	Segment lengthening or shortening	Use of bite blocks (Pipe-smoker speech
Whisper	Mimicry	Speech tempo	Lip protrusion
Inspiratory	Hyper-nasal (velum		Pulled cheeks

	lowered throughout)		
Raised or lowered larynx			Tongue holding
			Objects in mouth
			Objects over mouth

Table 2: Table of nonelectronic-deliberate voice disguise

The division into four main types is our own, based loosely on work found in [3] [29] [45] [64]. *Phonation* refers to abnormal glottal activity; *phonemic* refers to the use of abnormal allophones; *prosodic* concerns matters of intonation, stress segment length and speech rate; and *deformation* refers to forced physical changes in the vocal tract. The taxonomy is not really a partition though we have presented it that way for clarity. For example, "raised or lowered larynx" could be considered deformation, especially if it is held in position from the outside by a finger. "Mimicry" involves not only copying the allophonic pronunciation of the person mimicked, but also the glottal and prosodic characteristics, so its placement under *phonemic* is somewhat arbitrary. Surely one area of research would be to improve the taxonomy presented here.

5.2. Motivation for the research

There are two challenges. First, disguised voice is often used in the committal of a crime where the criminal has reason to expect to be recorded. [45] [60]. Often, it is necessary to identify or verify a suspect based on the disguised voice. Some means is needed to (1) determine that a voice has been disguised on a voice recording, (2) determine the method of disguise and (3) perform computer speaker identification or verification despite the disguise.

The second challenge is an academic one. It is stated in [26] that ". . speaker identification essentially is incapable of accurately determining the identity of a speaker when a test sample of his disguised speech is compared to a reference based on his normal speaking mode."

To date, and to the best of our knowledge, the above quoted passage remains true. One goal of forensic speaker recognition is to undertake research to reverse that situation, at least for a large and useful subset of disguise types.

5.3. Research to be carried out

5.3.1. Data Collection

There is not, to our knowledge, any standardized databases of voice disguises. Creating such a database is the first goal of the proposed research.

The data collection should follow the specifications and standards set out in [2] and [12], and by publications from the Linguistic Data Consortium (LDC) and the United States National Institute of Standards and Technology (NIST). Initially we recommend collecting data from 30-40 speakers, with multiple sessions per speaker. The recordings should be digital, sampled at 22kHz, 16 bit quantization, in a low noise environment using high quality components in a consistent manner. Data should be permanently stored on CD ROMs, or other superior media that may appear in the future.

Clearly, to attempt to capture data for all the disguises mentioned in the above table is unrealistic. Furthermore, some of the disguises — inspiratory and tongue-holding in particular — are mentioned in the literature as producing unintelligible speech [45]. We tentatively propose to record subjects speaking normally, in a whisper, in a falsetto, in creaky voice (glottal fry), with pinched nostrils, and with the use of bite-blocks. These choices are made for the following reasons: Some of these forms of disguise have been discussed in the literature [24] [63], and these forms of disguise are among the ones most commonly found in forensic casework [45].

5.3.2. Methods

Since the amount of research methodology on computer processing of disguised speech is small compared to research on speaker recognition in general, we suggest carrying out many preliminary experiments to determine the methods with the greatest potential. From our other experiences with voice processing, we have identified some areas that seem promising.

5.3.2.1. Investigate effects of disguise on conventional speaker recognition systems.

We suggest testing speaker recognition systems based on various modeling techniques — VQ, HMMs, segregation, etc. — against standardized data bases of various speech disguises. The results of these experiments will be summarized as a table in the form shown in Table 3 (q.v.). The table will allow us to predict the performance of each recognition technique for a given disguise type. From these results, we will be able to determine which disguise techniques pose the greatest risks to successful recognition.

Disguise Method	Recognition Performance			
	Segregating	VQ	HMM	…
Normal				
Whispered				
Falsetto				
Glottal Fry				
Pinched Nostrils				
Bite Blocked				

Table 3: Format of Disguise Effect Results

5.3.2.2. Automatic disguise detection

Before carrying out a speaker identification procedure, it is necessary to know if disguise is being used. People can usually tell when someone is disguising their voice; we suggest research to determine whether a computer can be programmed to recognize when disguise is being used and which type of disguise it is. As with automatic speaker recognition, we may discover that we can write computer programs that can detect disguises better than humans in certain situations.

We recommend two approaches, one based on comparing speaker models of normal and disguised speech; the other based directly on interpreting parametric information extracted from the speech signal.

5.3.2.2.1. Investigate whether different modes of disguise can be automatically detected from speaker models

We propose research on methods for identifying systematic differences between speaker models of undisguised speech and speaker models of disguised speech for various disguises and recognition techniques. For example we might find that for systems based on VQ, disguise by pinching the nostrils causes the centroids in only certain regions of the parameter space to differ, and that these differences are consistent across speakers.

To perform such analyses, we will have to compare the disguised and undisguised speech models of many speakers. Statistical techniques can be used to perform the comparison of models and evaluate the results.

From these experiments, we may be able to discover both a detection and compensation mechanism (to be discussed later) for certain types of disguise using certain recognition techniques.

5.3.2.2.2. Investigate whether different modes of disguise can be automatically detected directly from parametric information.

This stage of research would have scientists training several different types of recognizers to make a binary decision: disguised versus undisguised. Hidden Markov Models (HMMs), Neural Networks, and segregating systems could all be examined.

We propose using the following types of parameters as input to the recognizers: pitch, formant frequencies, formant widths, LPC coefficients, LPC cepstral coefficients, inverse spectral coefficients, relative spectral based coefficients (RASTA), and spectral moments. The latter has been central to our research in lip-synching and may turn out to provide useful information regarding disguised voices [37] [47] [48] [49].

During training, exemplars of many known types of disguised speech from many speakers would be used as positive reinforcement, while exemplars from undisguised speech would be used as negative reinforcement.

Since having one recognizer that is able to detect all disguises may be impractical due to poor generalization or excessive training, we also suggest investigating the uses of separate binary-decision recognizers for each known disguise mechanism. For example, once could train one multi-layer feed-forward neural network to distinguish whispered from normal, whispered from creaky, whispered from falsetto, etc. A second neural net would be trained to distinguish pinched nostril speech, a third to distinguish falsetto, and so on.

During training, each "expert" recognizer would be given positive exemplars of speech from many speakers using one specific disguise technique. As negative examples, the recognizer would be given both examples of normal speech and examples of speech using other disguise techniques. Since each recognizer would be more specialized, each would have greater accuracy. An important advantage of this methodology is that new disguise techniques can be brought into the picture simply, without necessitating major changes to already existing systems.

During recognition an unknown exemplar would be used as input to the series of disguise technique recognizers and would be deemed "disguised" if any one of the recognizers yielded a positive result. This collection of experts would have the added benefit of determining that speech was disguised using a specific disguise technique.

5.3.2.3. Automatic disguise compensation

I suggest research to determine a method for compensating for the effect of disguise during speaker recognition. This compensation could take the form of a transformation applied to disguised speech to make it comparable to undisguised speech, as spoken by a particular speaker. For example, if the system detects that the unknown speaker is using excessive nasality to disguise his voice, the system could disregard all sound segments except those that normally contain nasality. If raised pitch is detected, the system could resynthesize the speech at a lower pitch before performing a comparison.

The converse of this technique may also be effective. This would consist in modifying the undisguised speech reference patterns by disguising them in a manner similar to the disguised exemplars. For example, if the disguise is whispered speech, one could extract all the voiceless segments from the training data to create new speaker models more appropriate for comparison. This idea is embodied in the common practice of training reference patterns for recognizing telephone speech using telephone speech over the same or similar channels.

The detection of disguise may also be useful for adjusting confidence values of recognition decisions. First, experiments would be needed to determine the accuracy of the system for various forms of disguised speech. Then, depending on the type of disguise detected, the system would adjust the confidence value of recognition based on its past performance on similarly disguised speech.

5.3.2.4. Use of multiple disguises.

A study reported in [45] found that when speakers are free to disguise their voice to obscure personal identity, but retain intelligibility, 55% chose a single disguise method such as mimicking a foreign accent or altering their natural pitch. The remaining 45% chose multiple disguise methods. For example 15% chose a phonation change and a prosodic change; another 15% chose a phonation change and a phonemic change; and another 15% chose a prosodic change and a phonemic change.

The literature suggests that there is some hope of detecting and eliminating the effect of single disguises such as use of foreign accent or dialect [15] [53], the presence of hoarseness [78], the use of creaky voice (glottal fry) [24], pitch changes [63], nostril pinching [63], and whispering [our own as yet unpublished research].

When multiple disguises are used the entire enterprise becomes more challenging. One approach worth researching is to consider each multiple disguise individually. Thus whispered, bite-block speech would be consider a separate disguise, to be treated similarly to unary disguises. The number of possible disguise techniques proliferate exponentially, but the number that might actually be used is likely to be manageable.

6. SIGNIFICANCE OF RESEARCH

The significance of all research falls into two broad categories. The first category is laying the groundwork for further research, either by initiating a new research area or by advancing knowledge in an already established one. The second category is the opening up of new applications that can be developed based on the advanced knowledge revealed by the research.

The research proposed here is concerned with a relatively immature area of investigation within speaker identification: how to handle disguised voices. Completion of the research will be of two-fold benefit. It will lay the groundwork for other investigators to move forward — such groundwork barely exists presently. And it will extend the usefulness of speaker identification as it is currently practiced.

6.1. Application areas

6.1.1. Law enforcement

The problem of matching the voice of a suspect with a recorded voice, or of matching two recorded voices, is of interest to law enforcement agencies. In [45] it is noted that for 1989-1994 there was ". . an overall occurrence of voice disguise in 52 percent of the cases where the offender used his/her voice and may have expected to have it recorded during the criminal action. This percentage includes cases of blackmailing, where the specific percentage was as high as 69 percent." The latter figures are based on crimes in Germany. Regarding Brazil, the authors of [10] state: "Disguised speech is typically found in situations in which the criminal thinks he is being recorded. This situation is very common in cases of kidnapping, a kind of crime whose incidence has increased considerably in the past years in Brazil." Similar figures are not available for the United States because the individual states and the federal government tend to keep separate records, but there is no reason to believe that the numbers are different than in Germany or Brazil.

6.1.2. Application areas normally requiring speaker verification

While the most immediate application of research into the speaker recognition of intentionally disguised voices is in the forensic field, successful research is likely to establish methodologies for research into unintentionally disguised voice speaker recognition.

As noted earlier, speaker verification is becoming increasingly used to secure access to physical and electronic sites. Moreover, verification is starting to be used to control cellular phone fraud — well in excess of one billion dollars per year [41]. Its use is also incipient in house arrest enforcement, which is becoming more widely used due to the overcrowded condition of conventional prisons. In all such applications, a major problem is false negatives: a legitimate user is rejected, most often because something is affecting the person's usual voice quality, the one for which the system is trained. In effect, the person is speaking with a disguised voice. Though we have not specifically discussed dealing with unintentional disguises for reasons mentioned previously, we expect the methodological approach outlined above to affect research in that arena.

7. APPROXIMATE TIMETABLE OF GOALS TO BE ACHIEVED IN A 36 MONTH PERIOD

The Table 4 suggests a specific timetable for achieving the research goals mentioned in this paper. The goals could, I believe, be accomplished by two leaders — perhaps faculty members or corporate scientists — and two "followers." The followers might be graduate students or research assistants. The availability of sufficient infrastructure for carrying out the prescribed activities is presupposed.

MONTHS	**DESCRIPTION OF RESEARCH**
1 - 12	**Data collection (э5.3.1)**

12 - 18	**Effects of disguise on conventional speaker recognition systems (ə5.3.2.1)**
18 - 24	**Automatic disguise detection (ə5.3.2.2)**
24 - 36	**Automatic disguise compensation. Multiple disguises (ə5.3.2.3, 5.3.2.4)**

Table 4: A time-table for accomplishing the research suggested in this paper.

8. REFERENCES
(starred entries directly relevant to disguise)

[1] B.S. Atal. Automatic speaker recognition based on pitch contours. *Journal of the Acoustical Society of America*, 52:1687-1697. 1972.

[2] L. Boves, T. Bogaart, L. Bos. Design and recording of large data bases for use in speaker verification and identification. *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification.* Pp. 43-46. Martigny, Switzerland. April 1994.

*[3] J.R. Baldwin and P. French. *Forensic Phonetics.* London:Pinter Publishers. 1990.

[4] Y. Bennani and P. Gallinari. On the use of TDNN-extracted features information in talker identification. In *The 1991 International Conference on Acoustics, Speech, and Signal Processing*, pp. 385-388, 1991.

[5] R.H. Bolt et al. Identification of a speaker by speech spectrograms. *Science*, Vol. 166, Oct., 1969.

*[6] D. Boss. The problem of F0 and real-life speaker identification: a case study. *Forensic Linguistics*, 3(1): 155-159. 1996.

*[7] R.H. Bahr and K.J. Pass. The influence of style-shifting on voice identification. *Forensic Linguistics*, 3(1): 24-38. 1996.

*[8] A. Braun. The effect of cigarette smoking on vocal parameters. *ESCA Workshop on Automatic Speaker Recognition , Identification, and Verification*, Martigny, Switzerland, pp. 161-4, April 1994.

[9] J. P. Campbell. Speaker recognition: a tutorial. *Proceedings of the IEEE,* Vol. 85, No. 9, September, 1997.

*[10] R.M. de Figueiredo and H. de Souza Britto. A report on the acoustic effects of one type of disguise. *Forensic Linguistics,* 3(1): 168-175. 1996.

*[11] E.T. Doherty and H. Hollien. Multiple factor speaker identification of normal and distorted speech. *Journal of Phonetics,* 6:1-8. 1978.

[12] A. di Carlo, M Falcone, A. Paoloni. Corpus design for speaker recognition assessment. *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification.* Pp. 47-50. Martigny, Switzerland. April 1994.

[13] G. R. Doddington. Speaker recognition - identifying people by their voices. *Proceedings of the IEEE,* Vol. 73, No. 11, pp. 1651-1663, November, 1985.

[14] J.R. Deller, Jr., John G. Proakis and John H. L. Hansen. *Discrete-Time Processing of Speech Signals.* New York:Macmillan. 1993.

*[15] W. Endres, W. Bambach and G. Flosser. Voice spectrograms as a function of age, voice disguise and voice imitation. J*ournal of the Acoustical Society of America,* 49:1842-1848. 1971.

*[16] A.J. Friedhoff, M. Alpert and R.L. Kurtzberg. An electroacoustical analysis of the effects of stress on voice. *Journal of Neuropsychiatry,* 5:266-272. 1964.

*[17] M. Falcone and N. de Sario. A PC speaker identification system for forensic use: IDEM. *ESCA Workshop on Automatic Speaker Recognition , Identification, and Verification,* Martigny, Switzerland, pp. 169-72, April, 1994.

*[18] G. Fairbanks and L.W. Hoaglin. An experimental study of the durational characteristics of the voice during the expression of emotion. *Speech Monographs,* 8:85-90. 1941.

*[19] G. Fairbanks and W. Pronovost. An experimental study of the pitch characteristics of the voice during the expression of emotion. *Speech Monographs,* 6:87-104. 1939.

*[20] P. French. An overview of forensic phonetics with particular reference to speaker identification. *Forensic Linguistics,* 1(2): 169-181. 1994.

[21] S. Furui. An overview of speaker recognition technology. In [43], pp. 31-56. 1996.

[22] H. Hattori. Text-independent speaker verification using neural networks. *ESCA Workshop on Automatic Speaker Recognition , Identification, and Verification*, Martigny, Switzerland, pp. 103-6, April, 1994.

[23] A. Higgins, L. Bahler and J. Porter. Voice identification using nonparametric density matching. In [43], pp. 211-232. 1996.

[24] A. Hirson and M. Duckworth. Glottal fry and voice disguise: a case study in forensic phonetics. Journal of Biomedical Engineering, Vol. 15, Pp. 193-200. May, 1993.

[25] S. Hayakawa and F. Itakura. The influence of noise on the speaker recognition performance using the higher frequency band. *The 1995 International Conference on Acoustics, Speech, and Signal Processing*, Detroit, MI, pp. 321-4, May, 1995.

*[26] H. Hollien and W. Majewski. Speaker identification by long-term spectra under normal and distorted speech conditions. *Journal of the Acoustical Society of America*, 62(4): 975-980. 1977.

*[27] H. Hollien and C.A. Martin. Conducting research on the effects of intoxication on speech. *Forensic Linguistics*, 3(1): 107-128. 1996.

*[28] H. Hollien, W. Majewski and E.T. Doherty. Perceptual identification of voices under normal, stress and disguised speaking conditions. *Journal of Phonetics*, 10:139-148. 1982.

*[29] H. Hollien. *The Acoustics of Crime: The New Science of Forensic Phonetics.* New York:Plenum Press. 1990.

*[30] M.H.L. Hecker, K. Stevens, G. von Bismark and C. Williams. Manifestations of task-induced stress in the acoustic speech signal. *Journal of the Acoustical Society of America*, 44: 993-1001. 1968.

*[31] M. Hall and O. Tosi. Spectrographic and aural examination of professionally mimicked voices. *Journal of the Acoustical Society of America*, 58: S107A. 1975.

[32] C.R. Jankowski et al. Measuring fine structure in speech: application to speaker identification. *The 1995 International Conference on Acoustics, Speech, and Signal Processing*, Detroit, MI, pp. 325-8, May, 1995.

*[33] M. Jessen. Speaker-specific information in voice quality parameters. *Forensic Linguistics*, 41): 84-104.1997

[34] L.G. Kersta. Voiceprint identification. *Nature,* Vol. 196, No. 4861, pp. 1253-1257, December, 1962.

*[35] B.E. Koenig. Spectrographic voice identification: a forensic survey. *Journal of the Acoustical Society of America,* 79:2088-2090. 1986.

[36] R.L. Klevans and R.D. Rodman. *Voice Recognition.* Boston,MA:Artech House, Inc. 1997.

[37] B. Koster, R. Rodman and D. Bitzer. Direct translation of speech-sound to mouth shape. *Proceedings of the 28th Annual Asilomar Conference on Signals, Systems and Computers.* IEEE:1994, Pp 36-46.

*[38] H.J. Kunzel. Current approaches to forensic speaker recognition. *ESCA Workshop on Automatic Speaker Recognition , Identification, and Verification*, Martigny, Switzerland, pp. 135-41, April, 1994.

*[39] H.J. Kunzel. Identifying Dr. Schneider's voice: an adventure in forensic speaker identification. *Forensic Linguistics,* 3(1): 146-154. 1996.

[40] Y. Linde, A. Buzo, and R. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, vol. 28, No. 1, pp. 84-95, 1980.

[41] J.A. Levine. C.E.O. of e.Scape USA, LLC. Personal Communication. May, 1997.

*[42] R.C. Lummis and A.E. Rosenberg. Test of an automatic speaker verification method with intensively trained professional mimics. *Journal of the Acoustical Society of America,* 51:131-132(A). 1972

[43] C-H Lee, F. K. Soong and K.K. Paliwal. *Automatic Speech and Speaker Recognition: Advanced Topics.* Norwell, MA:Kluwer Academic Publishers. 1996.

[44] C-S Lui et al. An orthogonal polynomial representation of speech signals and its probabilistic model for text independent speaker verification. *The 1995 International Conference on Acoustics, Speech, and Signal Processing*, Detroit, MI, pp. 345-8, May, 1995.

*[45] H. Masthoff. A report on a voice disguise experiment. *Forensic Linguistics,* 3(1): 160-167. 1996.

*[46] W. Majewski and C. Basztura. Integrated approach to speaker recognition in forensic applications. *Forensic Linguistics,* 3(1): 50-64. 1996.

[47]. D. McAllister, R. Rodman and D. Bitzer. Lip synchronization for animation. *SIGGRAPH 97 Visual Proceedings,* Pp 225-226. Los Angeles, CA, August, 1997.

[48] D. McAllister, R. Rodman, D. Bitzer and A. Freeman. Lip synchronization as an aid to the hearing impaired. *Proceedings of AVIOS 97.* Pp 233-248. San Jose, CA, September 1997.

[49] D. McAllister, R. Rodman, D. Bitzer and A. Freeman. Lip synchronization of speech. *Proceedings of AVSP 97.* Pp 133-136. Rhodes, Greece, September 1997.

[50] J.D. Markel and S.B. Davis. Text-independent speaker identification from a large linguistically unconstrained time-spaced data base. *The 1978 International Conference on Acoustics, Speech, and Signal Processing,* pp. 287-289, 1978.

*[51] R. McGlone, H. Hollien and P. Hollien. Acoustic analysis of voice disguise related to voice identification. *Proceedings of the International Conference on Crime Countermeasures,* pp. 31-35. (University of Kentucky Bulletin 113 — generally available in university libraries.) July, 1977.

[52] T. Matsui and S. Furui. Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs. *The 1992 International Conference on Acoustics, Speech, and Signal Processing,* Vol. 2, pp. 157-160, 1992.

*[53] S. Moosmuller. Phonological variation in speaker identification. *Forensic Linguistics,* 3(1): 29-47. 1997.

[54] J. Oglesby and J.S. Mason. Optimization of neural models for speaker identification. *The 1990 International Conference on Acoustics, Speech, and Signal Processing,* 1990.

*[55] G. Papcun and J. Krix. What do mimics do when they imitate a voice? *Journal of the Acoustical Society of America,* 84:S114. 1986

*[56] A.R. Reich and J.E. Duke. Effects of selected vocal disguise upon speaker identification by listening. *Journal of the Acoustical Society of America,* 66:1023-1028. 1979.

*[57] A.R. Reich. Detecting the presence of vocal disguise in the male voice. *J. Acoust. Soc. Am.* 69(5), pp. 1458-60, May, 1981.

[58] D.A. Reynolds et al. The effect of telephone transmission degradations on speaker recognition performance. *The 1995 International Conference on Acoustics, Speech, and Signal Processing,* Detroit, MI, pp. 329-332, May, 1995.

*[59] A. Reich, K. Moll and J. Curtis. Effects of selected vocal disguises upon spectrographic speaker identification. . *Journal of the Acoustical Society of America,* 60:919-925. 1976.

[60] M. D. Robertson, Special Agent. Personal communication. North Carolina Department of Justice, State Bureau of Investigation. 16 E. Rowan St. Suite 500. Raleigh, NC 27609. October, 1995.

[61] L.R. Rabiner and R.W. Schafer. *Digital Processing of Speech Signal,* Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1978.

[62] L. Rudasi and S.A. Zahorian. Text-independent talker identification with neural networks. *The 1991 International Conference on Acoustics, Speech, and Signal Processing,* pp. 389-392, 1991.

*[63] L. Shinan and A. Almeida. The effects of voice disguise upon formant transitions. *The 1986 International Conference on Acoustics, Speech, and Signal Processing,* pp. 885-888, 1986.

*[64] J. Sample. *Methods of Disguise.* Port Townsend, WA:Loompanics Unlimited. Pp 64-68. 1984.

*[65] K.R. Scherer. Effects of stress on fundamental frequency of the voice. *Journal of the Acoustical Society of America,* 62:S25-26(A). 1977.

[66] M. Schmidt et al. Covariance estimation methods for channel robust text-independent speaker identification. *The 1995 International Conference on Acoustics, Speech, and Signal Processing,* Detroit, MI, pp. 333-6, May, 1995.

[67] M. Shridhar et al. Text-independent speaker recognition using orthogonal linear prediction. *The 1981 International Conference on Acoustics, Speech, and Signal Processing,* pp. 197-200, 1981.

[68] L.S. Su, K.P. Li and K.S. Fu. Identification of speakers by nasal coarticulation. *Journal of the Acoustical Society of America,* 56:1876-1882. 1974.

[69] F.K. Soong et al. A vector quantization approach to speaker recognition. *The 1985 International Conference on Acoustics, Speech, and Signal Processing*, pp. 387-390, 1985.

[70] F.K. Soong, A.E. Rosenberg, L.R. Rabiner and B.H. Juang. A vector quantization approach to speaker recognition. *The 1985 International Conference on Acoustics, Speech, and Signal Processing*, pp. 387-390, 1985.

[71] M. Savic and J. Sorensen. Phoneme based speaker verification. *The 1992 International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 165-168, 1992.

*[72] F. Schlichting and K.P.H. Sullivan. The imitated voice — a problem for voice line-ups? *Forensic Linguistics*, 4(1): 148-165. 1997.

[73] K.N. Stevens, C.E. Williams, J.R. Carbonelli and B. Woods. Speaker authentication and identification: a comparison of spectrographic and auditory presentations of speech material. *Journal of the Acoustical Society of America*, 43:1596-1607. 1968.

[74] N. Tishby. On the application of mixture AR hidden Markov models to text independent speaker recognition. *IEEE Trans. Signal Processing*, SP-39, pp. 563-570, 1991.

*[75] T. Takagi and H. Kuwabara. Contributions of pitch, formant frequency and bandwidth to the perception of voice personality. *The 1986 International Conference on Acoustics, Speech, and Signal Processing*, pp. 889-892, 1986.

[76] O. Tosi, H. Over, W. Lashbrook, C. Pedrev, J. Nicol and E. Nash. Experiment on voice identification. *Journal of the Acoustical Society of America*, 51:2030-2043. 1972

[77] O. Tosi. *Voice Identification: Theory and Legal Applications.* Baltimore:University Park Press. 1979.

*[78] I. Wagner. A new jitter-algorithm to quantify hoarseness: an exploratory study. *Forensic Linguistics*, 21): 18-27. 1995.

[79] R-h. Wang, L-s. He and H. Fuisaka. A weighted distance measure based on the fine structure of feature space: application to speaker recognition. *The 1990 International Conference on Acoustics, Speech, and Signal Processing*, pp. 273-276, 1990.

[80] J.J. Wolf. Efficient acoustic parameters for speaker recognition. *Journal of the Acoustical Society of America*, 51:2044-2056. 1972.

[81] E.H. Wrench. A real time implementation of a text independent speaker recognition system. *The 1981 International Conference on Acoustics, Speech, and Signal Processing*, pp. 193-196, 1981.

*[82] C.E. Williams and K.N. Stevens. Emotions and speech: some acoustical correlates. *Journal of the Acoustical Society of America,* 52:1238-1250. 1972.

[83] M.A. Young and R.A. Campbell. Effects of context on talker recognition. *Journal of the Acoustical Society of America*, 42:1250-1254. 1967.