Computer Recognition of Speakers Who Disguise Their Voice

Robert D. Rodman
Michael S. Powell
Department of Computer Science
North Carolina State University
Raleigh, North Carolina, 27695-8206
U.S.A.

1. Introduction

Speaker identification is the process of determining who spoke a recorded utterance through computer analysis of the recorded speech. More specifically, given a set of speakers $\Sigma=\{S_1 \; . \; . \; S_N\}$, a set of utterance sets $Y=\{U_1 \; . \; . \; U_N\}$ made by those speakers, and an utterance u_X made by an unknown speaker: *closed set* speaker identification determines a value for X in {1 . . N}; *open set* speaker identification determines a value for X in {0, 1 . . N}, where X=0 means "the unknown speaker $S_X \notin \Sigma$."

During the process, acoustic feature sets $\{F_1 \; .. \; F_N\}$ are extracted from the utterances in $\{U_1 \; . \; .U_N\}$. In the same manner, a feature set F_X is extracted from u_X. A matching algorithm determines which, if any, of $\{F_1 \; . \; . \; F_N\}$ sufficiently resembles F_X. The identification is based on the match (or lack of same), and is generally proffered with a probability-of-error coefficient.

A common scenario is to have an utterance associated with a crime such as a recorded bomb threat. This would be u_X. The usual suspects are rounded up (Σ), sets of utterances are elicited from them (Y), and an analysis carried out to determine the likelihood that one of the suspects was the speaker, or that none of them was.

In practice, the criminal often disguises his or her voice. The effect of the disguise is that F_X, the acoustic features of the criminal exemplar, is altered to become less similar to F_{actual}, the acoustic features of the actual criminal's undisguised utterances.

2. Types of disguises.

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 and sound "creaky," or for syllables in falsetto to occur. Nonetheless, such a definition allows us to classify voice disguises so that our research can 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 [8]. 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 drug-related alterations, or psychological personnel in the case of emotional disturbance.

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 lowered throughout)		Pulled cheeks
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 [2] [7] [9] [13]. *Phonation* refers to abnormal glottal activity; *phonemic* refers to the use of abnormal allophones; *prosodic* concerns matters of intonation, stress placement, 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,* rather than *phonation,* 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. Part of the on-going research is to improve and refine the taxonomy presented here.

3. Motivation for studying disguised voices

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. [9] [10]. 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 despite the disguise.

The second challenge is an academic one. It is stated in [6] 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.

4. Methodology for studying disguised voices

4.1. Data Collection

There are not, to our knowledge, any standardized databases of voice disguises. Creating such a database is a natural beginning to a systematic study of disguise.

The data collection should follow the specifications and standards set out in [1] and [4], 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 Table 2 is unrealistic. Furthermore, some of the disguises — inspiratory and tongue-holding in particular — are mentioned in the literature as producing unintelligible speech [9]. 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. Some of these forms of disguise have been discussed in the literature [5] [12], and they are among the ones most commonly found in forensic casework [9].

4.2. Choosing among different types of speaker identification systems

Since the amount of research on the 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 type of system — VQ-based, HMM-based, etc. [11] — with the greatest potential. From our other experiences with voice processing, we have identified some areas that seem promising.

4.2.1. 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 databases of various speech disguises. The results of these experiments will be summarized as a table in the form shown in Table 3. The table will allow us to predict the performance of each recognition system for a given disguise type.

Disguise Method/Type of Recognition System	Recognition Performance			
	Segregating	VQ	HMM	…
Normal				
Whispered				
Falsetto				
Glottal Fry				
Pinched Nostrils				
Bite Block				

Table 3: Format of Disguise Effect Results

4.2.2. 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 are investigating whether a computer can be programmed to recognize when disguise is being used and which type of disguise it is. It is conceivable that 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. Application areas

5.1. 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 [9] 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 [3] 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.

5.2. 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.

Speaker verification, as opposed to speaker identification [11], 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 [8]. 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.

6. References

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

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

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

[10] 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.

[11] R.D. Rodman. *Computer Speech Technology.* Norwood, MA:Artech House

[12] 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.

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

7. Contacts

rodman@csc.ncsu.edu

mspowell@unity.ncsu.edu