A Study of Intentional Voice Modifications for Evading Automatic Speaker Recognition

Sachin S. Kajarekar, Harry Bratt, Elizabeth Shriberg, Rafael de Leon

Speech Technology and Research Laboratory, SRI International, Menlo Park, CA
{sachin,harry,ees,rdeleon}@speech.sri.com

Abstract

We investigate the effect of intentional voice modifications on a state-of-the-art speaker recognition system. The investigation includes data collection, where normal and changed voices are collected from subjects conversing by telephone. For comparison purposes, it also includes an evaluation framework similar to that for NIST extended-data speaker recognition. Results show that the state-of-the-art system gives nearly perfect recognition performance in a clean condition using normal voices. Using the threshold from this condition, it falsely rejects 39% of subjects who change their voices during testing. However, this can be improved to 9% if a threshold from the changed-voice testing condition is used. We also compare machine performance with human performance from a pilot listening experiment. Results show that machine performance is comparable to human performance when normal voices are used for both training and testing. However, the machine outperforms humans when changed voices are used for testing. In general, the results show vulnerability in both humans and speaker recognition systems to changed voices, and suggest a potential for collaboration between human analysts and automatic speaker recognition systems to address this phenomenon.

1. Introduction

The performance of a speaker recognition system is affected by many factors. These can be divided into two broad groups: variations in communication channel and variations in a person's voice. The communication channel usually has both handset/microphone variations and environment variations. These factors have been extensively studied by researchers who have suggested several normalizing techniques to counter their effects.

Typically, the variations in a person's voice can be divided into those that are unintentional and those that are intentional. Unintentional variations are caused by emotion or by physical condition, such as a cold, a sore throat, or aging. Intentional variations can be further divided into two types.

In the first type of intentional variation, people change their voices to sound like someone else, usually to falsely assume the other's identity. This causes a type II errors, that is, the person is attempting to be identified as someone else, and the system accepts him/her.

Earlier research [1-6] in this area was done by Endres and others [3], who used speech spectrograms in the 1970s to recognize speakers. They observed that disguised voices had different formant structures, and that imitators varied their formant structure but the imitators could not match the structures obtained from the imitated voices. Eriksson and Wretling [1] studied professional mimics and showed that professionals match speaking rate and mean fundamental frequency.

Rosenberg and Sambur [6] were the first to report automatic verification results for variations in people's voices. They found little degradation using professional mimics compared to the nonexperts. Pellom and others [4] did experiments where a speech synthesizer was used to change the voices, and they observed a significant degradation in performance.

The second type of intentional variation is where people are trying to evade detection. This is a type I error, where a person tries to not sound like himself/herself and the system rejects the identity claim. These variations can be made by mimicking other people and other accents, trying to sound like someone else or just trying to not sound like oneself. This interesting case of intentional evasion of an automatic speaker recognition system and the question of how well state-of-the-art systems can detect this evasion is under investigation here.

We describe a comprehensive data collection and experimental design to test the effect of intentional voice changes on the ability of both humans and machines to recognize speakers. We describe a data collection paradigm in which normal and disguised voices were collected. Changes include pitch, duration, and their patterns, as well as phonation type and regular segmental substitutions. We describe the experimental design and the results with an automatic speaker recognition system. We also describe the human listening setup and results. Finally, we compare automatic recognition performance with human performance.

2. Data collection

The data collected was conversations between the subjects and an experimenter. All subjects are native speakers of American English. The same experimenter is used throughout the collection. Each conversation lasts about 5 minutes, and subjects are encouraged to participate equally in the conversation, resulting in approximately 2.5 minutes of speech from the subject. This is done so the resulting data will include common conversational phenomena, such as back-channels, that would not be present in the absence of an interlocutor. In each session, two conversations are recorded with the subject's normal voice. Then, the subject is asked to disguise his/her voice (specifically speaking "style"). They were encouraged them to disguise voices in a variety of ways, such as mimicking an accent, changing pitch, or changing duration, but were not insisted on a particular approach. Subjects are instructed to try anything as long as they do not block their mouths or the microphones or obstruct the path between the mouth and the microphone. Two or more conversations are recorded in different disguised voices. During the recordings, the examiner sits unseen in another

room that is separated by a one-way mirror. This setup, shown in Figure 1, is intended to simulate telephone conversations.

Different topics are used for each recording, selected from the following:

- Movies
- Growing up
- Food
- Neighborhood
- Holidays
- Travel
- Hobbies

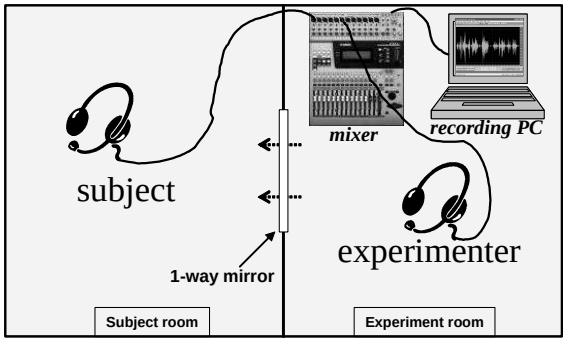

Figure 1: Recording setup

We recorded conversations with 32 speakers, with an average of four conversations per subject. Table 1 shows the different variations the speakers tried with their voices. Note that the labels are chosen to give an idea about the types of disguises, and are typically a name told to the experimenter by the speaker, explaining what kind of voice the speaker was going to use. These are just crude impressions and were not produced by professional or expert mimics. The table shows that the most popular form of voice disguise by these subjects is raising or lowering of pitch. A few speakers also tried changes in vocal energy and/or phonation type by whispering and murmuring. Also common are Southern or other accents. Some speakers also tried more complicated voice disguises, such as a valley girl accent (final pitch rises, frequent use of "like", "oh my god", "you know"), "whale voice" (extremely exaggerated vowel lengthing with highly exaggerated pitch excursions), "Kermit voice" (constant pharyngeal constriction), and stutter.

Table 1: Different voices from the data collection

Speaker	Disguised voices		
	Recording 1	Recording 2	Recording 3
1	Southern	British	Fargo
2	High pitch	Southern	
3	Low pitch	British	Southern
4	Low pitch	Southern	
5	Low pitch	Whisper	French
6	Low pitch	Whisper	
7	French	British	Russian
8	High pitch	Slow, high pitch	Low pitch
9	Southern	French	Whisper
10	Whisper	Low pitch	
11	French	Southern	
13	Whisper	High pitch	
14	Hawaiian	Southern	
15	Low pitch	Valley girl	
16	Low pitch	Nasal	
17	Clenched teeth	Kermit	Murmur
18	Southern	French	Murmur
19	Low pitch	Valley girl	Murmur
20	High pitch	Southern, round	Southern
21	Whale	Stutter	
22	Southern	Tongue out	
23	Indian	Elvis	
24	High pitch, slow	High pitch, nasal	
25	Whisper	Russian	Southern
26	Pirate	Brooklyn	
27	Cockney	German	
28	Indian	Hoarse	
29	High pitch	Stuffy nose	
30	Low pitch	Low pitch, robot	
31	High pitch, nasal	Low pitch, round	Low pitch
32	Israeli	Clenched teeth	High pitch
33	High pitch	High pitch, lisp	

3. Automatic speaker recognition experiments

3.1. Experimental conditions

We used speech from the subject's side of the conversation for the experiments. As mentioned above, each subject (a speaker) spoke for about 2.5 minutes. We used the entire conversation side for training and testing. This is similar to the NIST extended data speaker recognition evaluation paradigm.

Speaker recognition can be performed in four conditions:

1) training and test data from normal voice (normal-normal)
2) training data from normal voice and test data from disguised voice (normal-disguised)
3) training data from disguised voice and test data from normal voice (disguised-normal)
4) training and test data from disguised voice (disguised-disguised)

The first condition is the reference for evaluating other results. We expect almost perfect results for this condition, as the data is collected in a noise-free environment and without significant channel variation. The second condition is of the most interest, and represents situations where only the normal data from speakers is available for training and speakers are trying to deceive the system by changing their voices. The third condition is of little interest, and we expect its performance to be similar to that of the second case. The fourth condition is extreme because the two voices are changed in different ways. In the work reported here, we investigate the first two conditions.

The numbers of speaker models and test trials are obtained as follows. Each conversation side from a speaker is used to train one model, so the number of speaker models is

the same as the number of conversations. Each speaker model is tested against all the conversation sides, except the one used to train the model. Then, the trials are chosen to keep an overall (across all conditions) ratio of true speaker trials to impostor trials of 1:15. All the true speaker trials are kept, and impostor trials are sampled uniformly from a distribution of impostor trials.

Table 2: Recognition setup

Training condition	Test condition	True speaker trials	Impostor trials
Normal	Normal	63	2421
Normal	Disguised	150	2421

3.2. Speaker recognition system

The speaker recognition system is the same baseline system as used in our NIST speaker recognition evaluation submission [7]. This system uses 13 Mel frequency cepstral coefficients estimated by a 300- to 3300-Hz bandwidth front end consisting of 19 Mel filters. Cepstral features are normalized using cepstral mean subtraction and are concatenated with delta, double-delta, and triple-delta coefficients. For channel normalization, the feature transformation described in [8] is applied to the features.

Our baseline uses 2048 Gaussian components for the background model. This Gaussian mixture model (GMM) is trained using gender- and handset-balanced data (electret, carbon-button, cell-phone). We use approximately 300 hours of data from FISHER, part of the NIST 2003 extended Speaker Recognition Evaluation (SRE) data, and the NIST 2002 cellular SRE development data [9].

Target GMMs are adapted from the background GMM using maximum a posteriori (MAP) adaptation of the means of the Gaussian components. Verification is performed using the 5-best Gaussian components per frame selected with respect to the background model scores. No score normalization is used.

3.3. Results and discussion

Table 3 shows recognition results obtained with the GMM speaker recognition system. Two types of results are shown: percent equal error rate (%EER) and the decision cost function (DCF). The %EER is a point on the receiver operating characteristic (ROC) curve of false rejection (FR) versus false acceptance (FA) at which both errors are equally weighted. The DCF shows the point at which a NIST-defined cost function [9] gives the lowest value. For the normal-normal condition, the baseline system performs almost perfectly. This is not surprising because the data was recorded in clean condition without any additive noise or channel distortion. With this performance as a reference, we can see that when people disguise their voices, the performance degrades to an EER of 7.46%.

Table 3: Performance of GMM system

Experiment condition		%EER	DCF (x10)
Training	Test		
Normal	Normal	0.05	0.0041
Normal	Disguised	7.46	0.2992

Note that the EER on the normal-disguised condition is obtained using a "cheating" threshold. That is, it is obtained based on the normal-disguised trials, implying that the recognition system has prior information about disguised voices. A stricter test is without this prior, namely, what would happen if the system did not have access to disguised voices. To simulate this test, we take the threshold for the EER from the normal-normal condition and apply it to the normal-disguised condition. The results in Table 4 show that the false rejection rate increases from 7.33% to 39.3% when the system does not have any access to the disguised voices. This means that about 40% of the speakers can fool the system by changing their voices if the system is trained without disguised voices. This error-rate can be reduced to a fifth of that if system is trained with disguised voices. This clearly shows the effect of disguised voices on this state-of-the-art speaker recognition system.

Table 4: False rejection (FR) and false acceptance (FA) error rates using different thresholds

Experiment	Threshold from	% FR	% FA
Normal-Disguised	Normal-Disguised	7.33	7.60
Normal-Disguised	Normal-Normal	39.3	0

4. Human listening experiments

We compared the machine recognition performance with human performance. Note that we consider the listening experiments to be a pilot study. This is our first attempt to design a listening experiment and we were able to recruit only 25 listeners for this experiment. Both the results and the design are likely to change in the subsequent experiments, based on the results from this study (as mentioned in Section 6 later).

4.1. Experimental design

It is complicated to design a human listening experiment that is comparable to machine recognition for a variety of reasons. First, machines have infinite memory and they can load hundreds of speakers in that memory and perform recognition across thousands of trials. Humans, on the other hand, can not remember that many speakers at once, do not perform recognition across many trials at once, and are limited by real-time constraints (exposing them to two hours of data takes at least two hours). Further, machines can be programmed to make independent decisions from one trial to the next. However, humans tend to remember voices and can use idiosyncratic information about a speaker across different trials.

We designed the listening experiment so that two utterances were played, one after the other, and listeners had to decide whether they were spoken by the same or different speakers. The experiment was performed in four 1o-minute sessions, with a 5-minute break after each session. In each session, the pairs were randomized to reduce the occurrence of duplicate speakers. Although listeners were aware that

utterances might have been spoken in normal or disguisesed voices, no probability was specified.

Another issue, making the human experiment comparable to that of the machine, was the amount of speech. The automatic recognition experiment was performed using 2.5 minutes of speech, which is too long for the human experiment, because human listeners might not pay attention to the entire conversation side and would get bored. There is a question of how much data can be used effectively in a human experiment so that 1) it is long enough to contain useful information and 2) it is short enough so that people remember the characteristics of the first utterance when they listen to the second one. The duration was chosen to be around 5 seconds for each of the two utterances. The listeners were asked to decide whether the utterances were spoken by the same speaker, by a different speaker, or they were not sure. If a listener selected the "not sure" option, 5 more seconds of each utterance were played in addition to the original 5 seconds. Listeners were provided more speech in 5-second increments until 20 seconds had been played. At that point, they had to make a final decision about whether utterances were spoken by the same or different speakers. A screenshot of the presentation tool is shown in Figure 2.

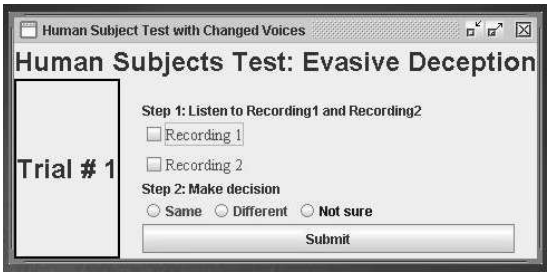

Figure 2: Screenshot of the software tool used for listening experiments

It is important to note that people do not need balanced train and test data (as used in an automatic experiment). Typically, we might need lot more data to get ourselves familiarize with a speaker, but once acquainted we need very little to recognize speakers. Also, we need only a small amount of test data for familiar speakers and more data for unfamiliar speakers. To control this effect, we chose listeners who did not know the subjects used in the recordings.

As mentioned earlier, the automatic recognition experiment was performed using a 1:15 ratio of target to impostor trials. However, a ratio of 1:1 was used for the listener experiment. Listeners were made aware of this ratio at the beginning of a session. Therefore, the worst human performance should have been no worse than 50%.

4.2. Data

Original recordings were preprocessed for experiments with different duration conditions. From each file, we chose contiguous segments of approximately 5 seconds each. In cases where a single segment was not available, we used a collection of segments that were approximately 5 seconds. The inter-segment pause lengths and the turn lengths between different speakers were removed.

4.3. Results and discussion

Human performances were evaluated using a general formula:

$$perf(s) = \frac{\#correct(s)}{\#decided(s)} + 0.5 \times \#undecided(s),$$

where *#decided(s)* is the number of yes/no decisions made for segment length *s*, *#undecided(s)* is the number of "not sure" decisions, and *#correct(s)* is the number of correct decisions. The idea behind this measure is that it accounts for all types of decisions by measuring correct judgments from yes/no decisions and by assuming a 50% chance of correctness for "not sure" decisions. The *#decided* increases as people listen to more data and for s=20sec, *#undecided=0*. Most typically, performance improved with longer segments but longer segments sometimes misled people into making wrong decisions. A total of 900 trials were chosen for the listening experiment. On average, people listened to about 23 trials in a session and about 90 trials in an hour. These trials were randomly chosen from the total trials without replacement.

Performance of the automatic recognition system is evaluated in two ways. First, Table 5 shows the %EER for all the trials for all the different durations. As expected, performance for the normal-normal condition is always better than for the normal-disguised condition, and performance improves with longer segments.[1] A comparison of results between the last two columns of Table 5 shows that performance using 20 seconds of data is still much worse than that obtained with 2.5 minutes (150 seconds).

Table 5: Automatic speaker recognition results for different durations compared with original results (150 seconds) from *Table 3*

Condition	%EER				
	5sec	10sec	15sec	20sec	150sec
Normal-Normal	21.9	10.9	7.8	6.2	0.05
Normal-Disguised	31.8	20.7	20.7	17.5	7.46

Second, the automatic performance is compared to the performance of human listeners follows. We selected the trials that were presented to each listener and obtained scores for those trials using the automatic speaker recognition system. The equal error rate was computed for these trials. This was averaged over the set of the trials presented to all the listeners.

Figures 3 and 4 show this performance comparison, with a box plot for human performance. The box shows lower quartile, median and upper quartile values of the identification error. The circle shows automatic speaker recognition performance. Results indicate that for the normal-normal condition, automatic performance is similar to the lower quartile of human performance. However, for the normal-disguised condition, automatic performance has lower error human performance. Very few people in our experiment achieved performance comparable to that of the machine.

[1] The similar performance for normal-disguised condition using 10 and 15 seconds is partially due to limited data. However, the DCF shows improvement from 0.053 to 0.050 using more data for the 20-second condition.

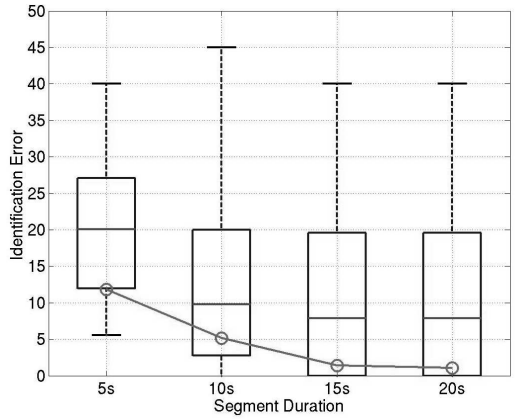

Figure 3: Comparison of human (box plot) and automatic speaker recognition (circles connected with magenta line) performance in the **normal-normal** condition

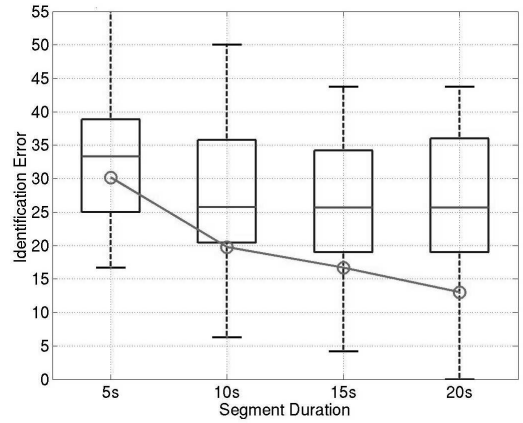

Figure 4: Comparison of human (box plot) and automatic speaker recognition (circles connected with magenta line) performance in the **normal-disguised** condition

5. Summary

We investigated the effect of intentional change of voice on an automatic speaker recognition system. In particular, we looked at intentionally disguised voices. We encouraged subjects to disguise their voices in a variety of ways, such as mimicking an accent, changing pitch, or changing speaking rate, but we did not ask for a particular approach. The most popularly used disguise was a change in pitch, and the second most popular was a southern accent. A total of 32 subjects were recorded, with an average of four conversations between the subject and an examiner. Two of these conversations were recorded in the subject's normal voice and two or more in disguised voices.

An evaluation set was created from this data with two conditions: 1) train and test with normal voice, and 2) train with normal voice and test with disguised voice. A state-of-the-art GMM system using cepstral features was evaluated on this data, and showed significant degradation in performance for disguised voices. The degradation was even worse (about five times) when the system does not use any information about the disguised voices. This indicates vulnerability of existing speaker recognition systems to intentional voice disguises.

Further experiments were designed to test the ability of humans to recognize unknown speakers. In a carefully controlled experiment, human performance was measured and was compared to automatic recognition performance. Results showed that machine performance is comparable to average human performance in the normal-normal condition, but is better than average human performance in the normal-disguised condition. Although this is a preliminary result, given the limited number of subjects used in this experiment, it suggests a potential for collaboration between human analysts and automatic speaker recognition systems.

6. Future work

We hope to expand this work with more subjects and in a more controlled setup. We will ask the subjects to try one or more of pitch, duration, or intensity variations. This will give us more data for each type of variation so that we can study in depth what variations really affect the speaker recognition system and how.

In addition, more data needs to be collected with speakers from other cultures and languages. The choice of how to modify one's voice is likely to be highly culturally dependent. For example, it may not be common in some cultures to imitate other accents or cartoon characters.

After collecting more data, it will probably prove valuable to incorporate high-level stylistic feature-based systems [10-12]. These systems model pitch, energy, and duration patterns. Since these are the patterns the subjects will be changing, it will be interesting to see the effect of disguised voices on these systems.

One issue with our human listening experiment is that people were not allowed to listen to more data after making a decision. This eliminates the possibility that people could have changed their minds after listening to more data, which in turn would have improved human performance.

Another issue involves familiarity with the speaker. Usually, humans use a lot of data from a speaker to characterize that speaker's voice, and need a small amount of additional data to recognize the speaker. In the listening experiments, we required unfamiliarity to avoid contamination of results. But it is worthwhile doing the same experiment with listeners familiar with the speakers and comparing the performance across the experiments [13, 14].

Another type of variation, as mentioned in the introduction, is unintentional variation in a person's voice. This can happen when someone has a cold or a sore throat, or is aging. It would be interesting to collect data, study effects of these variations, and compare results with the results from experiments with intentional variations.

7. References

[1] A. Eriksson and P. Wretling, "How Flexible Is the Human Voice? A case study of mimicry," presented

at European Conference Speech Technology, Rhodes, 1997.
[2] J. Lindh, "Visual Acoustic vs. Aural Perceptual Speaker Identification in a Closed Set of Disguised Voices," presented at FONETIK, 2005.
[3] W. Endres, W. Bambach, and G. Flosser, "Voice Spectrograms as a Function of Age, Voice Disguise, and Voice Imitation," *The Journal of the Acoustical Society of America*, vol. 49, pp. 1842-1848, 1971.
[4] B. L. Pellom and J. H. L. Hansen, "An Experimental Study of Speaker Verification Sensitivity to Computer Voice-Altered Imposters," presented at ICASSP, Phoenix, 1999.
[5] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous Probabilistic Transform for Voice Conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 131-141, 1998.
[6] A. E. Rosenberg and M. R. Sambur, "New Techniques for Automatic Speaker Verification," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 23, pp. 169-176, 1975.
[7] S. Kajarekar, L. Ferrer, E. Shriberg, K. Sonmez, A. Stolcke, A. Venkataraman, and J. Zheng, "SRI's 2004 NIST Speaker Recognition Evaluation System," presented at ICASSP, Philadelphia, 2005.
[8] D. Reynolds, "Channel Robust Speaker Verification via Feature Mapping," presented at ICASSP, Hong Kong, China, 2003.
[9] NIST, "http://www.nist.gov/speech/tests/spk/index.htm."
[10] L. Ferrer, H. Bratt, V. R. R. Gadde, S. Kajarekar, E. Shriberg, K. Sonmez, A. Stolcke, and A. Venkataraman, "Modeling Duration Patterns for Speaker Recognition," presented at Eurospeech, Geneva, 2003.
[11] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling Prosodic Feature Sequences for Speaker Recognition," *Speech Communication*, vol. 46, pp. 455-472, 2005.
[12] S. Kajarekar, L. Ferrer, K. Sonmez, J. Zheng, E. Shriberg, and A. Stolcke, "Modeling NERFs for Speaker Recognition," presented at Odyssey 04 Speaker and Language Recognition Workshop, Toledo, Spain, 2004.
[13] D. Van Lancker, J. Krieman, and K. Emmorey, "Familiar voice recognition: Patterns and parameters - Recognition of backward voices," *Journal of Phonetics*, pp. 19-38, 1985.
[14] D. Van Lancker, J. Krieman, and T. Wickens, "Familiar voice recognition: Patterns and parameters: Part II. Perceptions of rate-altered voices," *Journal of Phonetics*, vol. 13, pp. 39-52, 1985.