EFFECTIVENESS OF ELECTRONIC VOICE DISGUISE BETWEEN FRIENDS

MARK HUCKVALE, ANNE-LINN KRISTIANSEN

Department of Speech, Hearing and Phonetic Sciences, University College London, London, U.K.

m.huckvale@ucl.ac.uk, anne-linn.kristiansen.09@ucl.ac.uk

Three experiments were conducted into the identification of speakers from their voices after electronic disguise using pitch scaling and vocal tract length scaling. A cohort of undergraduate students was used as a source of both speakers and listeners. The speed and accuracy with which speakers were identified from their voices was measured in conditions ranging from undisguised to severely distorted. Results show that when listeners know speakers well, identification accuracy can be very high, and it is hard to disguise speakers by pitch and vocal tract length scaling alone. Recognition levels close to chance were only achieved when extreme levels of disguise were applied, corresponding to a pitch increase of 12 semitones together with vocal tract length reduction of 20%. These were also the most unnatural and most distorted conditions. The implications of the study for the use of voice disguise in witness protection are considered.

1 INTRODUCTION

Whenever someone speaks an utterance, they communicate not only a message made up of words and sentences which carry meaning, but also information about themselves as a person [1]. The speaker-specific characteristics in the signal can provide information about the speaker's anatomy, physiology, linguistic experience and mental state. This information can sometimes be exploited by listeners and technological applications to describe and classify speakers, possibly allowing speakers to be categorised by age, gender, accent, language, emotion or health. In circumstances where the speaker is known to the listener, speaker characteristics may be sufficient to select or verify the speaker's identity.

Sometimes, situations arise in which a speaker does not want to be identified from their speech. An important example of this is found in legal proceedings where the recorded testimony of a witness is played in court, and where the safety of the witness would be jeopardised if their identity were revealed. In such situations, the audio recording can be manipulated to mask some of the speaker characteristics present in the speech. A significant problem in this area of *voice disguise for witness protection* is that the disguise should be adequate to protect the identity of the witness even when they are known to the accused. The present study aims to investigate the robustness of the typical means of voice disguise used in witness protection when speakers and listeners are well known to each other.

1.1 Speaker recognition by listeners

Research into the abilities of human listeners to recognize speakers from their voices has a long history. Part of that research has been into the investigation of which properties of the signal are used by human listeners. For example [2] showed that speakers could sometimes be recognized by intonation contours alone. A review of the phonetics of speaker recognition can be found in [3].

Other research has considered the importance to identification of the familiarity of the speaker to the hearer. For example [4] found that a group of 10 speakers were recognized nearly perfectly from their voices alone by speakers who knew them well. However listeners who were trained to recognize the speakers only recognized 40% of the samples on average. A study [5] shows that listener performance on speaker recognition is also well correlated with listener opinion of their own recognition accuracy, with listeners showing a higher confidence and a higher accuracy in the recognition of familiar speakers. Interpreting the results of a clinical study of the recognition abilities of brain-damaged patients, [6] argues that the recognition of familiar and unfamiliar voices shows evidence of dissociation: that is to say that listeners may use different mechanisms for *recognizing* familiar voices compared to *discriminating* between unfamiliar voices. A study into speaker recognition using single vowels [7], suggests that familiar speakers are recognised with respect to a mental prototype of the speaker in the mind of the listener. But nevertheless [7] also shows that listeners use conventional features such as voice quality, fundamental frequency and frequency of the higher formants to help identify speakers.

1.2 Means of disguise

We can broadly divide the means of voice disguise into two types [8]: *non-electronic voice disguise* are methods applied while the person is speaking to change or distort speaker-specific speech characteristics. For example, a speaker may change pitch range, alter voice quality, raise his larynx or constrict his pharynx. He might affect an accent or imitate another individual. He might pinch his nostrils, pull his cheeks, or use mechanical aids such as a bite block. In contrast, *electronic voice disguise* are methods typically applied to audio recordings of normally spoken speech, that use signal processing techniques to modify aspects of pitch, timbre and timing. The goals of electronic voice disguise are to mask speaker identity while maintaining the intelligibility of what was said.

We focus in this paper on electronic voice disguise, although we note that it certainly possible to consider voice disguise based on a combination of non-electronic and electronic means.

The signal processing means by which voices are disguised are relatively straightforward, and tend to involve voice pitch shifting, voice quality changes and spectral warping. These changes address acoustic characteristics found to be useful for identification by human listeners [7] but relate to the long-term spectral properties of a voice rather than to articulatory detail which might expose the speaker's accent. These changes are also relatively easy to apply in real-time using specialised hardware or digital signal processing. It appears that voice disguise systems used by law-enforcement agencies originate in the music and audio effects industries rather than in speech science and technology [9].

1.3 Effects of disguise

Previous work on the effects of voice disguise can be described under three headings. There is research that investigates whether voice disguise impacts forensic speaker identification, that is whether the use of voice disguise would compromise the accuracy with which a forensic phonetician would judge whether two recordings are of the same person. For example, [10] showed that a forensic speaker authentication system was strongly affected by some simple means of non-electronic disguise. A second area of research is into the detection of voice disguise, that is the determination of whether a speech recording has been disguised for speaker identity. For example [11] showed that human listeners were quite good at recognising whether a speaker had changed their normal way of speaking, although [12] found that an automatic speaker verification system was not always able to detect vocal forgery. An acoustic study [9] reports mixed results in detecting the presence of electronic audio manipulation by spectral examination of the signal. Lastly there is research into the impact of disguise on speaker identification by human listeners. For example, [4] reports that recognition accuracy fell from 98% to 79% for 10 speakers after they were allowed to 'freely disguise' their voices. In another study, [13] showed that accuracy in discriminating same-different pairs of speakers fell by 30% when speakers used a hyper-nasal speech quality.

1.4 Effects of electronic disguise on listeners

Surprisingly, there are few studies which investigate the effectiveness of electronic disguise on human listeners. Those that do exist [14, 15] do not address the issues important to the protection of witnesses, namely that speakers should be disguised from identification even by their friends and relatives.

In [14], the effectiveness of fundamental frequency manipulation was assessed using a group of listeners who were trained to recognize 4 voices. Although a significant reduction in identifiability was achieved using a pitch shift of 8 semitones, the listeners only achieved 60% identification accuracy in the unprocessed condition. It cannot be said that the listeners knew the speakers well in this case.

In [15], a voice transformation system was trained for each of five speakers, to transform each person's voice towards (or beyond) a common target voice. In this approach a speaker is disguised only because the alternative candidate voices in the identification experiment are also transformed to the same target. Although the method appeared to work well in informal listening tests, the approach does not necessarily address the issues in witness protection, where the identity of the other possible speakers are not known (and where it may be equally dangerous to target the identity of a different real person). Also such a voice transformation system requires recordings of specific utterances from the speaker for training the transform, which may be impractical in a real application.

1.5 Aims of this study

We have seen that there exist means for electronic voice disguise that manipulate the signal characteristics that have been found relevant for speaker identity in experiments with human listeners. How these products are used for witness protection in any given situation can be up to the audio engineer assigned to the disguise, and seem to be rather ad hoc: one engineer told us "if it is a young woman, we'll make them sound like an old man". Thus we believe that the area is in need of a more systematic exploration of the effectiveness of electronic voice disguise. In particular we need to establish whether current schemes of electronic disguise can be relied upon to protect the identity of speakers even from people who know them well.

In this study we have three primary aims: to establish how well a large group of friends can identify each other by their voices alone (without training); to assess the extent to which pitch and vocal tract length scaling affect the identification of a small group of speakers by their friends; and to investigate how much disguise is required to reduce identification to chance levels. These aims are addressed through a series of three experiments described below.

2 EXPERIMENT 1 - BASELINE VOICE IDENTIFICATION

The goals of this experiment were: to establish a baseline identification performance for familiar undisguised voices, to find the amount of speech needed to make reliable identifications, and to find a set of the most readily identified speakers for the subsequent experiments.

2.1 Method

Speakers and listeners were chosen from the cohort of third-year female undergraduates attending the BSc Speech Sciences programme at UCL in 2010/11. These students had worked and studied together for over two years, and because of the use of much small-group teaching in the programme, knew each other well. Of the year group, 28 participants agreed that their voices could be used in the study. The vast majority of speakers were native British English speakers younger than 25 years.

High-quality recordings (16-bit, 44100 samples/sec) were made in a sound-treated booth of each speaker reading a passage "The Natural World" which contains 335 words and begins:

> "By the end of the twentieth century, very few children in Britain will know what 'unspoilt nature' really means. Well-kept urban parks and gardens will be all they know. They'll never see a rich carpet of wild flowers in a woodland glade, nor hear a bird-song at dawn without the disturbing buzz of traffic."

The recordings were made 12 months earlier for a different purpose, so the speakers did not know that these recordings were going to be used in an identification experiment.

In the experiment, 25 listeners were first asked to visually identify the photographs of the 28 speakers from their names - this was done to ensure that all the listeners knew the individuals concerned. In the subsequent listening task, the listeners were played the recording of each speaker in random order and starting from the first sentence of the passage. Recordings were played over a loudspeaker in a sound-treated booth. Listeners were then asked to identify the speaker as quickly as possible from a display of 28 named photographs. If the listener was also one of the speakers, that response was removed from the results.

2.2 Results

Listeners identified an average of 97.8% of speakers' photographs correctly from their names, and all but 6 listeners made no mistakes.

Listeners recognized speakers from their voices on average 91.4% of the time. Three listeners recognized all the speakers correctly, with the median score being 92.6%, and the lowest score 63.0%.

The median time taken to recognize a speaker correctly was 5.0s. 90% of all correct identifications were made within 11.0s, and 95% of all correct identifications were made within 15.3s.

Speakers also varied in how well they were identified. Twelve speakers were always recognized correctly, and the median identification accuracy was 95.8%. The most readily identified speakers were also recognized more quickly, see Fig.1.

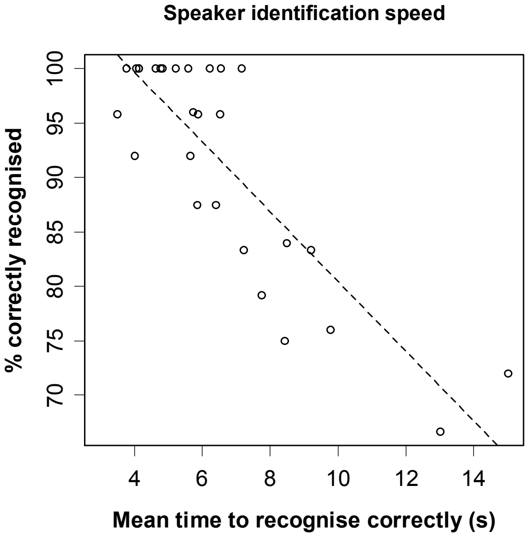

Figure 1. Identification rate of speakers as a function of mean time taken to correctly identify them (r= 0.85).

2.3 Summary

Among this group of 28 young women who know each other well, over 90% of individuals were recognised by their voice alone despite the listeners not having any prior training. This result emphasises the importance of familiarity; it is a much higher rate than is commonly found in studies where listeners are trained to recognise a set of unfamiliar speakers (e.g. [14])

From this study we were also able to identify a sub-group of speakers that were recognised greater than 95% of the time from just 10s of their speech (see Fig.1). These speakers were then candidates for the following experiments in which voice disguise was applied.

3 EXPERIMENT 2 - DISGUISED VOICE IDENTIFICATION BY FRIENDS

The goals of this experiment were to determine how simulated changes in voice pitch and vocal tract length affected the ability of listeners to identify familiar and recognisable voices.

3.1 Method

Five speakers were chosen from the best and fastest recognised speakers in experiment 1. Since some of the best recognised speakers had distinctive regional accents, five speakers were selected on the basis of having similar regional accents, in this case from south-east Britain. Although we expect that accent will make a large contribution to speaker identification, we did not want to confound our results with the effects of regional accent.

The original read passages used in experiment 1 were processed under 11 disguise conditions, see Table 1.

Condition	Vocal Tract Length	Fundamental Frequency
vt100fx+0	Unchanged	Unchanged
vt100fx+4	Unchanged	Increased by 4st
vt100fx-4	Unchanged	Decreased by 4st
vt100fx-8	Unchanged	Decreased by 8st
vt090fx+0	Reduced to 90%	Unchanged
vt090fx+4	Reduced to 90%	Increased by 4st
vt110fx+0	Increased to 110%	Unchanged
vt110fx-4	Increased to 110%	Decreased by 4st
vt110fx-8	Increased to 110%	Decreased by 8st
vt120fx+0	Increased to 120%	Unchanged
vt120fx-4	Increased to 120%	Decreased by 4st
vt120fx-8	Increased to 120%	Decreased by 8st

Table 1. Processing conditions used in listening experiment 2. (st=semitones)

Pitch and vocal tract length changes were implemented using Linear Prediction (LP) vocoding. For pitch changes, the LP residual was first resampled to shift its pitch, then time adjusted to its original duration using the Waveform-Similarity Overlap-Add algorithm (WSOLA, [16]). For vocal tract length changes, the LP coefficients were transformed to line-spectral frequencies [17] which were then warped to new spectral positions before being transformed back for synthesis. The spectral warping functions are shown in Fig 2. The warp functions are linear below 5.5kHz and quadratic above. The processing sample rate was 22050 samples/sec, and a 24th order LP analysis was employed.

Pitch changes of 4 and 8 semitones mirror those used by [14], while vocal tract length changes of 10% correspond roughly to the mean difference between male and female speakers [18]. Since all speakers were female, eight conditions looked at lowering of pitch or increasing VT length while only three conditions looked at raising pitch or VT shortening. The quality of re-synthesis was good, and intelligibility did not seem to be affected by processing. See Fig. 3 for example spectrograms.

Ten listeners were chosen from the original 28 students excluding the chosen speakers. They listened to 3 repetitions of each of the 5 speakers in each of the 12 conditions in random order, indicating their responses against named photographs. Random 10s segments of the processed passages were presented, and the listeners were asked to respond as soon as they were confident about the identity of the speaker.

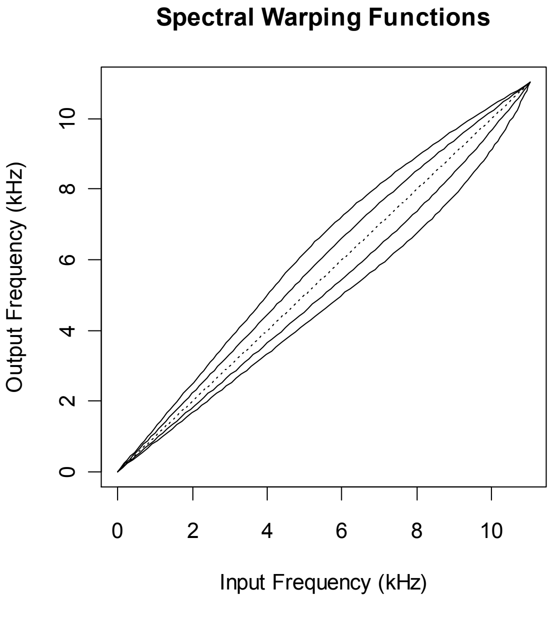
Spectral Warping Functions

Figure 2. Spectral warping functions applied to obtain 80%, 90%, 110% and 120% change in vocal tract length.

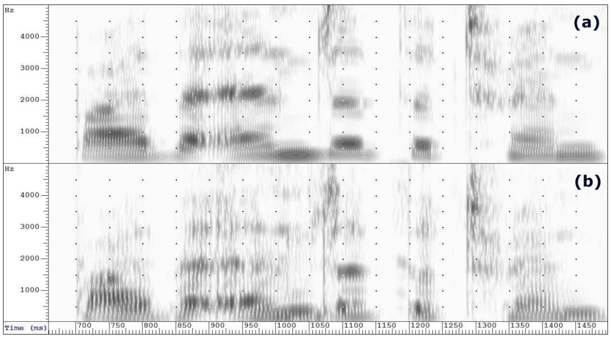

Figure 3. Example electronic disguise. (a) original utterance by female speaker, (b) after pitch decreased by 8st and vocal tract length increased by 20%.

3.2 Results

Overall mean identification accuracy in each of the disguise conditions is listed in Table 2, while distributions of mean identification scores across listeners are shown with boxplots. The boxplots show the 0th, 25th, 50th, 75th and 100th quantiles of the distribution. Outliers are plotted as circles. Recognition accuracy as a function of pitch change without vocal tract length change is shown in Fig 3. Recognition accuracy as a function of vocal tract length change without pitch change is shown in Fig 4. Recognition accuracy with a combination of pitch change and vocal tract length change is shown in Fig 5.

	Pitch Scaling			
VT scaling	-8 st	-4 st	0 st	+4 st
90%	-	-	**86.7**	**66.7**
100%	**84.7**	96.7	98.0	96.7
110%	**81.3**	94.0	98.0	-
120%	**72.0**	**83.3**	**82.0**	-

Table 2. Mean percentage identification scores of 10 listeners across disguise conditions. Scores that are significantly worse ($p<0.05$) than the unprocessed condition are marked in bold.

Statistical analysis of listener judgments was performed using a mixed effects logistic regression model [19], using the lme4 package [20]. Each listener judgment is predicted as a logistic function of a linear combination of the effects of CONDITION (12 levels), SPEAKER (5 levels) and REPETITION (3 levels), with LISTENER as a random effect. Significant reductions ($p<0.05$) in recognition accuracy compared to the unprocessed condition were observed for changes of -8st pitch change, 90% vocal tract length change, and 120% VT length change.

The regression analysis also showed significant differences in the recognisability of speakers, see Fig 7. Speakers 2 and 5 stood out as being more recognisable.

A learning effect was observed, with a significant increase in performance with repetition. Comparison of the first and third trials of each listener showed a significant increase in accuracy (McNemar test, χ^2=19.4, df=1, $p<0.001$). For example the performance in condition vt120fx-8 improved from 66% to 80% between the first and third presentation of each speaker-condition combination.

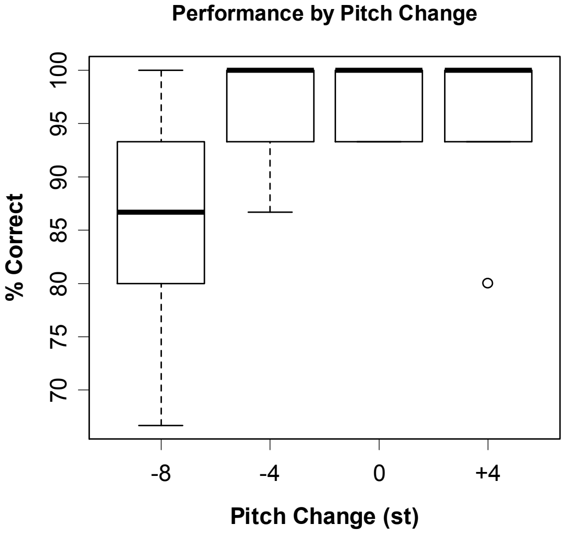

Figure 4. Speaker identification after fundamental frequency shift alone (N=10).

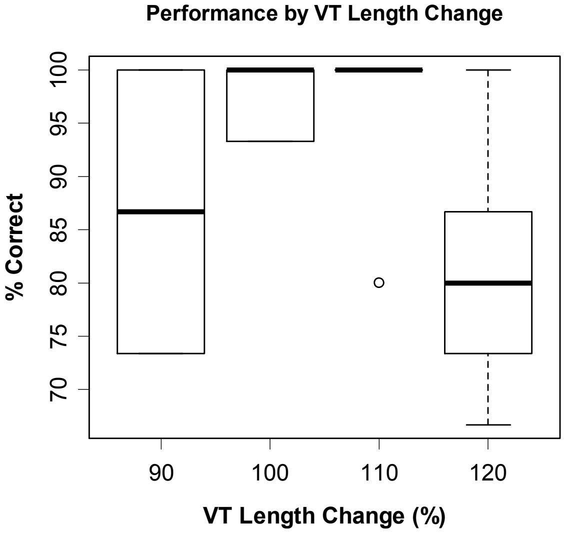

Figure 5. Speaker identification after vocal tract length change alone (N=10).

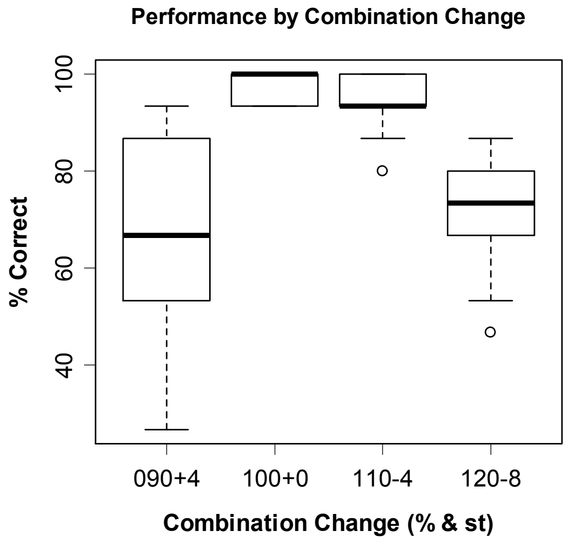

Figure 6. Speaker identification after combined pitch and vocal tract length changes (N=10).

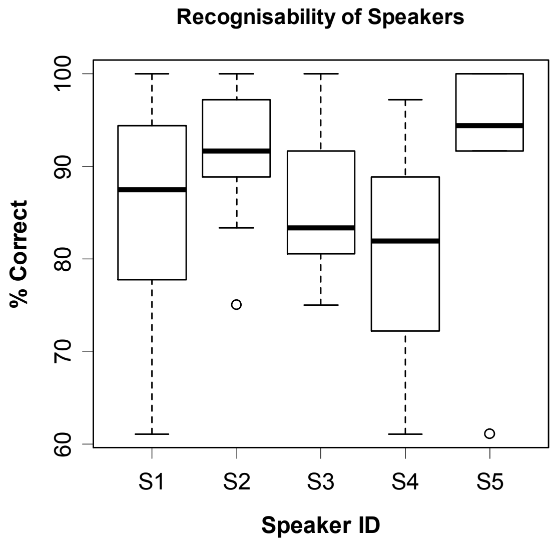

Figure 7. Mean identification rates of speakers for all conditions by all listeners.

3.3 Summary

With pitch changes alone, only a change of 8st caused a significant reduction in identification accuracy. This mirrors the result reported in [14].

Changes in vocal tract length alone were only effective if (for these female speakers) the vocal tract length was shortened by 10% or lengthened by 20%.

When changes to pitch and VT length are combined, significant reductions in identification accuracy were seen for the 90% +4st and the 120%-8st conditions.

Speakers were recognised in over 60% of trials even in the most disguised condition and on the first presentation.

4 EXPERIMENT 3 - EXTREME VOICE DISGUISE

The goals of this experiment were to investigate how large pitch and vocal tract scaling needed to be to reduce speaker identification performance to chance levels.

4.1 Method

The same five speakers were used as in experiment 2. The original read passages were processed under 16 disguise conditions, listed in Table 3.

Condition	Vocal Tract Length	Fundamental Frequency
vt100fx+0	Unchanged	Unchanged
vt080fx+12	Reduced to 80%	Increased by 12st
vt080fx-12	Reduced to 80%	Decreased by 12st
vt080fx-16	Reduced to 80%	Decreased by 16st
vt080fx-20	Reduced to 80%	Decreased by 20st
vt090fx+12	Reduced to 90%	Increased by 12st
vt090fx-12	Reduced to 90%	Decreased by 12st
vt090fx-16	Reduced to 90%	Decreased by 16st
v090fx-20	Reduced to 90%	Decreased by 20st
vt110fx+12	Increased to 110%	Increased by 12st
vt110fx-12	Increased to 110%	Decreased by 12st
vt110fx-16	Increased to 110%	Decreased by 16st
vt110fx-20	Increased to 110%	Decreased by 20st
vt120fx+12	Increased to 120%	Increased by 12st
vt120fx-12	Increased to 120%	Decreased by 12st
vt120fx-16	Increased to 120%	Decreased by 16st
vt120fx-20	Increased to 120%	Decreased by 20st

Table 3. Processing condition used in listening experiment 3.

Scaling values were chosen to extend those used in experiment 2 to extreme values. Since all the source speakers were female, more conditions involved lowering of pitch and lengthening of vocal tract than vice versa.

Unfortunately we could not use listeners from the undergraduate cohort for this experiment. So instead we used 15 new listeners who were first trained on the undisguised voices of the 5 speakers until they could recognise them 100% of the time. Training was performed by allowing the listener to listen to up to 20s of each speaker using material not used in the listening experiment. Then each listener was tested using new 7s fragments of each speaker. The process of train & test was repeated until each listener was able to recognise all speakers twice without error. This change in procedure does mean however, that the results here are not directly comparable with those in experiment 2, and likely to be of worse performance.

Because of the increase in the number of conditions compared to experiment 2, the amount of speech presented to each listener in the disguise experiment, was reduced from 10s to 7s. This was to ensure that the same section of audio was not played more than once. In addition, each voice was played only once in each disguised condition without feedback to avoid any learning effect. Finally, at every third stimulus an undisguised condition was played, and after listener judgment, feedback of the correct answer was given. This was to maintain the attention of the listener and their memory for the voices of the speakers.

4.2 Results

Mean percentage identification accuracy by the 15 listeners across the audio conditions are shown in Table 4. In the undisguised condition, which was presented multiple times in the test with feedback, listeners achieved over 90% accuracy. Otherwise recognition accuracy generally fell with increasingly extreme disguise.

	Pitch Scaling				
VT scaling	-20 st	-16 st	-12 st	0 st	12 st
80%	34.7	34.7	33.3	-	**25.3**
90%	42.7	52.0	48.0	-	48.0
100%	-	-	-	91.7	-
110%	37.3	45.3	38.7	-	40.0
120%	36.7	40.0	42.7	-	33.3

Table 4. Mean percentage identification scores of 15 listeners across disguise conditions. One condition, where performance is not significantly greater than chance is indicated in bold

The distribution of accuracy scores across conditions is shown in Fig 8. The difference between the performance scores in each condition and the chance level of 20% was analyzed using a one-sample t-test. Only condition vt080fx+12 was not significantly different to chance at the p=0.05 level (t=1.07, df=14, p=0.30).

4.3 Summary

The listeners achieved high recognition accuracy in the undisguised condition despite the speakers being unfamiliar. Extreme amounts of pitch scaling and vocal tract length scaling reduce recognition accuracies, but above chance performance is still found in all but one audio condition.

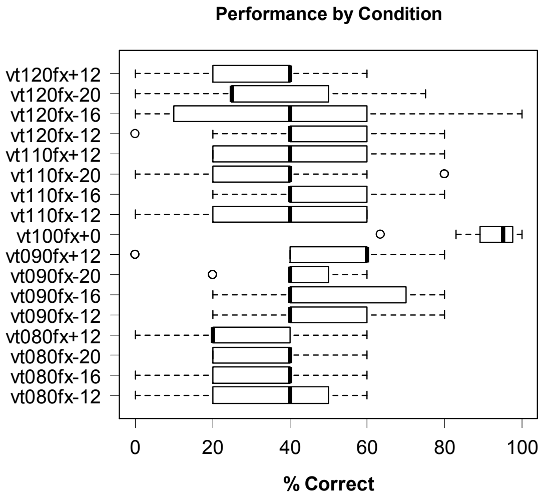

Figure 8. Distribution of recognition performance across disguise conditions (N=15).

5 DISCUSSION

5.1 Aims of study

The first aim of the study was to determine a baseline speaker identification performance for a group of friends. Using a cohort of 28 undergraduate students we showed that their ability to recognize each other from their voices alone was very high (91.4%) and only a little less than their ability to recognize each other from photographs (97.8%). We also found that 95% of the correct recognitions were made from less than 15.3s of speech. This performance can be compared to other studies in which speakers were familiar to listeners: for example [4] in which 10 speakers were recognized 98% of times from 2.5min of speech, and with [7] in which 20 speakers were recognized 49% of times from a single vowel. It is clear that familiarity is a potent factor in the identification of speakers by listeners from voice alone.

The second aim of the study was to look at effects of pitch and vocal tract length scaling on identification by friends. We investigated the recognition of 5 similar speakers in a range of disguise conditions: pitch scaling from -8st to +4st and vocal tract length scaling from -10% to +20%. We found that 10 listeners who were familiar with the speakers showed good performance in recognizing them even after disguise. Performance on the undisguised condition was 98%, and performance on the most disguised condition (90% VT scaling, +4st) still showed a recognition performance of 66%. In contrast to the study [14], we did not observe chance level performance in any disguise condition at the levels of disguise tested.

This experiment also demonstrated a clear learning effect. Recognition performance was significantly higher on the third presentation of each speaker+disguise condition than on the first. This may have been because the listeners became more familiar with the speakers' voices or because they had become more familiar with the effects of disguise. However, even on the first trial of the most disguised condition, recognition accuracy was greater than 60%.

The third aim of the study was to determine the level of disguise necessary to achieve chance performance. Unfortunately we had to use trained listeners rather than the friends used in the second experiment. We used a training and reinforcement protocol to try and mitigate the problem, and we found that our trained listeners were able to show good recognition performance in the undisguised condition. We explored a wide range of extreme disguise conditions, from -20st to +12st in pitch and from -20% to +20% in vocal tract length. However we found that our listeners, while strongly affected by disguise, were still operating above chance in all but one of the disguise conditions.

5.2 Limitations of study

In attempting to consider the generalizations one might take from this study it is important to realize that listener performance will be affected by factors not considered in our experiment. For example our speakers were all of the same sex, of similar age and similar accent. Disguise would likely be less effective if the group of speakers were of different sexes, ages or accents. Our speakers were chosen because they were well identified in our original group of 28 speakers. Disguise may be more effective if applied to a group of speakers with less distinctive voices. In our experiment, listeners were only presented with short fragments of speech (10s in experiment 2, and 7s in experiment 3). Disguise may be less effective if listeners have longer amounts of speech upon which to make judgments

Our signal processing technique relied on LPC vocoding to scale pitch and vocal tract length. Although this technique works well for small changes, the speech signal can become significantly distorted for large changes such as those used in experiment 3. By the time the pitch has been changed by more than one octave, the resulting signal, while still intelligible, becomes rather unnatural. It seems likely that this distortion also plays a part in disguising the identity of the speaker in combination with the change in pitch and VT length.

5.3 Consequences for witness protection

To obtain chance performance in a 5-way speaker identification experiment we had to employ extreme levels of voice disguise, restrict utterances to 7s in duration and play them to listeners who were not originally familiar with the speakers.

In real witness protection situations, there will always be differences in the circumstances of the speaker and listeners compared to our experiments. The cohort of potential speakers within which the witness wants anonymity may contain speakers of different sexes, ages and accents. In which case the results here may actually over-estimate the effectiveness of disguise. In other situations, the cohort of potential speakers may be more similar or less well known to the listeners and hence it may be easier to disguise identity than we have found here.

If voice disguise is to be applied, the best we can say from our experiments is that it should use pitch scaling and vocal tract length scaling in combination, it should use vocal tract length shifts of not less than 20%, and pitch shifts of more than one octave. Even then there are clear risks that a speaker will be recognized by their friends at better than chance levels.

In any case, voice disguise cannot protect a witness from being identified through information available in the content of their spoken utterances.

6 ACKNOWLEDGEMENTS

Thanks to the 3rd year BSc Speech Sciences students at UCL for agreeing to let their voices and their ears be used in this experiment.

7 REFERENCES

[1] Dellwo, V., Huckvale, M., Ashby, M., "How is individuality expressed in voice? An introduction to speech production & description for speaker classification", in Müller, C. (Ed.). *Speaker Classification I,* Berlin: Springer Verlag (2007) 1-20.

[2] Abberton, E., Fourcin, A., "Intonation and speaker identification", Language and Speech 21 (1978) 305-318.

[3] Nolan, F., *The Phonetic Bases of Speaker Recognition.* Cambridge: CUP. (2009).

[4] Hollien, H., Majewski, W., Doherty, T., "Perceptual identification of voices under normal, stress, and disguised speaking conditions", Journal of Phonetics 10 (1982) 139-148.

[5] Yarmey, A., Yarmey, A., Yarmey, M., Parliament, L., "Commonsense beliefs and the identification of familiar voices", Applied Cognitive Psychology 15 (2001) 283-299.

[6] van Lancker, D., Kreiman, J., "Voice discrimination and recognition are separate abilities", Neurophysiologica 25 (1987) 829-834.

[7] Lavner, Y., Rosenhouse, J., Gath, I., "The prototype model in speaker identification by human listeners", International Journal of Speech Technology 4 (2001) 63-74.

[8] Rodman, R., "Speaker Recognition of Disguised Voices", in M. Demirekler, A. Saranli, H. Altincay, and A. Paoloni (eds), Proceedings of the Consortium on Speech Technology Conference on Speaker Recognition by Man and Machine: Directions for Forensic Applications, Ankara, Turkey, April 1998, pp9-22.

[9] Brixen, E., "Digitally disguised voices", AES 39th Conference on Audio Forensics, Copenhagen, June 2010, 35-46.

[10] Zhang, C., Tan, T., "Voice disguise and automatic speaker recognition" Forensic Science International 175 (2008) 118-122.

[11] Reich, A., "Detecting the presence of vocal disguise in the male voice", J.Acoust.Soc.Am. 69 (1981) 1458-1461.

[12] Perrot, P., Aversano, G., Chollet, G., " Voice Disguise and Automatic Detection: Review and Perspectives", Progress in Nonlinear Speech Processing, Springer Lecture Notes in Computer Science, 4391 (2007) 101-117.

[13] Reich, A., Duke, J., "Effects of selected vocal disguises upon speaker identification by listening", J.Acoust.Soc.Am. 66 (1979) 1023-1028.

[14] Clark, J., Foulkes, P., "Identification of voices in electronically disguised speech", The International Journal of Speech, Language and the Law, 14 (2007) 195-221.

[15] Jin, Q., Toth, A., Schultz, T., Black, A., "Speaker De-Identification via Voice Transformation", IEEE workshop on Automatic Speech Recognition and Understanding (ASRU 2009), Merano, Italy.

[16] Verhelst, W, & Roelands, M., "An overlap-add technique based on waveform similarity (WSOLA) for high-quality time-scale modification of speech", IEEE Conference Acoustics, Speech and Signal Processing (1993) 554-557.

[17] Kabal, P., Ramachandran, R., "The computation of line spectral frequencies using Chebyshev polynomials", IEEE Trans. Acoustics, Speech and Signal Processing, ASSP-34 (1986) 1419-1426.

[18] Xue, S., Hao, J., "Normative standards for vocal tract dimensions by race as measured by acoustic pharyngometry", Journal of Voice 20 (2006) 391-400.

[19] Jaeger, T., "Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models", J. Mem. Lang. 59 (2008) 434-446.

[20] http://lme4.r-forge.r-project.org/