http://dx.doi.org/10.15448/1984-7726.2017.1.26659

# Speaker Identification in Whisper[1]

*Identificação de falante a partir de fala sussurrada*

India Smith
Paul Foulkes
Márton Sóskuthy
University of York – York – Yorkshire – United Kingdom

**Abstract:** Sociophonetic methods and findings have value in application to real-life issues, including providing expert forensic evidence in legal cases. Forensic cases often involve voices which differ markedly from those typically encountered in laboratory or field studies. We assess the ability of people to identify familiar voices produced in whisper, a commonly used form of disguise. Members of a pre-existing social network were recorded speaking normally and in whisper. Speakers found it difficult to maintain whisper beyond 30 seconds. They and other members of the group listened to extracts that were (i) short and whispered, (ii) long and whispered, and (iii) short and normal (non-whispered). Foils were also included. Performance was well above chance, and improved significantly in conditions (ii) and (iii). Differences were found across listeners and voices. The study emphasises how important is it not to overgeneralise from experimental data to a witness's ability under forensic conditions.

**Keywords:** Sociophonetics; Forensic phonetics; Whisper

**Resumo:** Os métodos e achados sociofonéticos são valiosos para aplicação a questões da vida real, como no fornecimento de evidência forense pericial em casos legais. Os casos forenses envolvem vozes que se diferem marcadamente daquelas tipicamente encontradas em laboratório ou estudos de campo. Avaliamos a habilidade de pessoas de identificar vozes familiares produzidas de forma sussurrada, uma estratégia de disfarce comumente utilizada. Membros de uma rede social pré-existente foram gravados falando normalmente e de forma sussurrada. Os falantes consideraram difícil manter o sussurro por mais do que 30 segundos. Esses falantes e outros membros do grupo ouviram trechos que foram (i) curtos e sussurrados; (ii) longos e sussurrados e (iii) curtos e normais (não sussurrados). Distratores foram incluídos. A performance foi bem acima do acaso e melhorou significativamente nas condições (ii) e (iii). Diferenças foram encontradas entre falantes e vozes. O estudo enfatiza o quanto é importante não supergeneralizar a partir de dados experimentais quanto à habilidade da testemunha sob condições forenses.

**Palavras-chave:** Sociofonética; Fonética Forense; Sussurro

## 1 Introduction: applications of sociophonetics in the forensic domain

Sociophonetics, begat by the union of sociolinguistics and phonetics, is a sub-field of linguistics emerging from its adolescence. While heavily influenced by its heritage, like most adolescents it is now finding its own identity and challenging its parents. Sociophonetics seeks to observe and explain the vast range of 'fine phonetic detail' that people produce and understand in the full range of social contexts in which speech is used. Speaking and listening are emphasised as collaborative and goal-driven human activities, interwoven with other modes of behaviour including physical gesture. The rich data set provided by sociophonetic studies provides robust challenges to theoretical models based on idealised, hypothesised or experimentally-controlled data. As a consequence

[1] **Acknowledgments:** We record our thanks to the participants for taking part in this study and for allowing use to use the work premises to carry out recordings. We are also grateful to Colleen Kavanagh, anonymous reviewers, and colleagues at the 2009 IAFPA conference held in Cambridge, UK, for feedback on earlier versions of the work.

we are now developing theoretical models of speech production, perception, and phonological representation that are in some ways radically different from those in the structuralist-generative tradition (FOULKES, SCOBBIE & WATT, 2010, DOCHERTY & MENDOZA-DENTON, 2012, FOULKES & HAY, 2015).

One particular route by which sociophonetic work departs from the mainstream is through its application to forensic phonetic cases, i.e. legal cases where analysis of speech is produced as evidence. Such analysis may be called for under various circumstances and for various reasons (FOULKES & FRENCH, 2012). Most frequent is the comparison of a voice recorded during the commission of a crime (e.g. a threatening phone call or kidnap demand) with that of a suspect accused of having committed that crime. Sociophonetic details may offer crucial help in such cases, for example to demonstrate similarity or difference in the regional accent being spoken.

In the analysis of forensic cases we are often asked to apply the principles and methods of sociophonetics to voices recorded or heard in unusual circumstances, quite different from the types of work typically conducted in the laboratory or the field. In this way we gain insights into speaking and listening towards the peripheries of human experience. For example, the speech to be analysed may involve intoxication by alcohol or drugs (CHIN & PISONI, 1997, PAPP, 2008), and/or heightened emotions, including severe distress (ROBERTS, 2012). Furthermore, the recordings of the crime and suspect might be separated by many years, and the analysis has to take into account processes of ageing on the voice (see RHODES, 2012, and for an account of a real case FRENCH, HARRISON & WINDSOR LEWIS, 2007).

Other types of forensic case involve witnesses to a crime whose memory of a voice or the words spoken might provide critical evidence. In such cases there is usually no recording of the voice in question. The (socio)phonetician's role might be to construct or vet a test of the witness's recall of the voice (NOLAN, 2003). Experimental research in this domain has shown that witness performance can be affected by many factors (see the review by BULL & CLIFFORD, 1984). These include whether the voice was encountered in direct interaction or heard passively (HAMMERSLEY & READ, 1985), the length of the samples heard (LADEFOGED & LADEFOGED, 1980), the time delay between original exposure and subsequent testing (MCGEHEE, 1937), the degree of familiarity of the voice (HOLLIEN et al., 1982), and whether and how it was disguised (HOLLIEN et al., 1982, FIGUEIREDO & DE SOUZA BRITTO, 1996).

Our focus in this study is the effect of whisper on the ability of people to identify familiar voices. Whisper is a commonly used form of disguise, masking the voice by removing fundamental frequency (f0, or pitch). Nonetheless, whisper has rarely been addressed in experimental work, especially with a forensic interest. We turn in the next section to a brief review of research on whisper, before describing our study in sections 3 (methods) and 4 (results). We conclude (section 5) with a discussion of the results relative to previous work, and with some comments on the forensic value of the work.

## 2 Whisper

In whisper, pulmonic airflow creates turbulence at the open but narrowed glottis. It is often generated with a posterior triangular opening of the glottis and with the anterior portion adducted, although the glottis may be narrowed across most or all of its length (LAVER, 1980, p. 121, LAVER, 1994, p. 190). From MRI data, Tsunoda et al. (1997) suggest that supralaryngeal structures may be lowered to contact the vocal folds and thus prevent phonation during whisper. The physiological and aerodynamic bases of whisper are further discussed by e.g. Sundberg et al. (2010), and Gick, Wilson & Derrick (2013). The acoustic effects of whisper are documented by Schwartz (1970, 1972), Kallail & Emanuel (1984a,b), and Swerdlin et al. (2010). Not surprisingly, whisper impairs speech perception, although most studies have been limited to investigation of individual segments (e.g. STURM & JAKIMIK, 1984, TARTTER, 1989). In automatic speech/speaker recognition (ASR), systems have been developed for automatic detection of whisper (ZHANG & HANSEN, 2010), and to reconstruct whisper as phonated speech (AHMADI et al., 2008).

From the forensic perspective, whisper is one of the easiest ways to disguise the voice (ORCHARD & YARMEY, 1995), and so is frequently reported in forensic cases (MASTHOFF, 1996, KÜNZEL, 2000, YARMEY et al., 2001, p. 297). However, relatively little research has been conducted on the effects of whisper on speaker identification. Pollack, Pickett & Sumby (1954, p. 405) conclude that listeners need three times the duration of whisper to reach the same level of performance in speaker identification as with normal speech. In a forced choice test examining pairs of voice samples, Bartle & Dellwo (2015) show that phoneticians perform better and more cautiously than lay (linguistically-untrained) listeners in judging whispered samples.

The largest experimental investigation of whisper is that by Yarmey et al. (2001), who conducted open speaker identification tests with a large group of lay listeners exposed to voices of varying degrees of familiarity and with samples of increasing lengths (from the single word *hello* to two minutes of spontaneous speech). They report severely degraded performance in tasks involving whisper, but better overall identification rates with more familiar voices and with longer samples. Listeners gave 77% correct responses overall with two minute samples. They also observed a very high rate of false hits, i.e. voice samples wrongly attributed to other known speakers. The effects of whisper have also been explored in ASR, with serious or even catastrophic effects reported on the success of ASR systems (ALEXANDER, 2007, ZHANG & TAN, 2008). Although speaker identity may be difficult to judge, speaker sex appears relatively easy, even with very short samples. Schwartz & Rine (1968) report 97.5% accuracy with 3 second isolated vowels, while Lass et al. (1976) obtained 75% accuracy with even shorter vowels (0.5 seconds). This effect is readily explained: whisper maintains the overall acoustic effects of the vocal tract transfer function, reflecting the large differences in vocal tract size that separate male and female voices (NOLAN, 1997). These differences are perceptible through vowel formant patterning.

We now turn to the experiment we conducted. The rationale for the study was primarily to extend the very limited set of experimental research on voice disguise in general, and whisper in particular. Our aim was to test the conclusions of Yarmey et al. (2001) using a different experimental design, and also to explore further the variability in the predicted results. Yarmey et al. (2001) used a between-subjects design, comparing the performance of two groups of listeners who heard either normal speech or whispered speech. In our study the same group hear both. The subject pool in the Yarmey et al. experiment was large and diverse, and the degree of familiarity with the voices used was classified impressionistically by self-report. We contrast this approach by focusing on a considerably smaller but socially coherent group of participants. This enables us to consider more carefully the social connections between them, and potential reasons for variation in their performance in the speaker identification tests. Finally, by using British participants, whose regional accents show marked diversity, we can explore whether sociophonetic (specifically, regional accent) cues assist in the identification process. The details of the experiment are described in further detail in the next section.

## 3 Method

### 3.1 *Design*

The study followed the general design of previous investigations of speaker identification by lay people (LADEFOGED & LADEFOGED, 1980, YARMEY et al., 2001, FOULKES & BARRON, 2000, BLATCHFORD & FOULKES, 2006). A pre-existing social network was chosen for the study: eleven women, including the first author, who worked together at a cosmetics store in York. Henceforth we refer to them as ‘the group’. They had known each other through work and social activities for between eight months and four years. They were not paid for their participation. Members of the group acted as both speakers and listeners, as explained below.

Working with a pre-existing network has both disadvantages and advantages. On the one hand we cannot control for or quantify the degree of familiarity of the participants with each voice, although this problem is less marked than in experiments involving large and diverse groups of participants. We were nevertheless satisfied that all members of the group knew each other well enough that they were highly familiar with the voices used in the experiment. On the other hand, an existing social group enables us to test voice identification skills in an ecologically valid context. The alternative would be to train a group of participants with a set of novel voices through experimental materials (see e.g. CLARK & FOULKES, 2007 for a similar study with artificially disguised voices). It is unclear whether memory of a voice acquired in artificial settings, and the ability to identify that voice in subsequent experiments, mirrors these processes in real life (cf. VAN LANCKER et al., 1985, who suggest that familiar and unfamiliar voices are recognised through different cognitive processes). It should also be borne in mind that for some types of crime the majority of incidents are committed by perpetrators known to the victim.

### 3.2 *Voice recordings*

Six of the women (including the first author) were recorded reading a set of stimuli both in normal voice and whisper. Three foils were also recorded, i.e. people unknown to members of the group. Two of the foils were women, the third a man. The male voice was included as a reference point against which to judge other stimuli, as it was predicted to be easy for the participants to reject as a member of the all-female group. All participants were aged between 20 and 30. The speakers’ regional origins and key accent features are summarised in Table 1.

British dialects are very diverse (see e.g. WELLS, 1982), and we provide here only brief comments on a small set of features well known to typify these accents and to which we expected listeners to be sensitive. Three accent differences are particularly salient in the UK. First, accents of northern England differ from those in the south by virtue of not contrasting the vowels of the FOOT and STRUT lexical sets (using the framework devised by WELLS, 1982). That is, *look* and *luck*, which contrast in southern and standard accents (/lʊk/ versus /lʌk/), are homophones in the north (both are /lʊk/). The contrastive pattern is also found in Scotland and the standard accent. Second, in northern England BATH words have short /a/, and thus pattern with TRAP words (*pass* /pas/ ≈ *gas* /gas/), whereas in the south and standard accent BATH has long /ɑː/, just as in PALM/START (*pass* /pɑːs/ ≈ *palm* /pɑːm/). Third, accents of Scotland and south west England differ from others, again including the standard, through *rhoticity*, i.e. /r/ is pronounced in all syllable positions. The standard accent permits /r/ only in pre-vocalic contexts (thus *red* /red/ and *very* /veri/, but *farm* /fɑːm/ and *far* /fɑː/; in word-final position /r/ is variably produced if the following word begins with a vowel, thus *far out* /fɑːr aʊt/). A final point to note is that not all speakers from a given region necessarily display the stereotypical patterns for the local accent. We therefore report in Table 1 the actual patterns used by our speakers.

**Table 1.** Speakers

| Participant code | Regional origin | FOOT=STRUT | BATH=TRAP | Rhoticity |
|---|---|---|---|---|
| P1 | Yorkshire | ✓ | ✓ | × |
| P4 | South West | × | ✓ | × |
| P5 | Yorkshire | ✓ | ✓ | × |
| P6 | Yorkshire | ✓ | ✓ | × |
| P8 | Leicester | ✓ | ✓ | × |
| P9 | Edinburgh | × | ✓ | ✓ |
| F1 (female foil) | South West | × | × | ✓ |
| F2 (female foil) | North West | ✓ | ✓ | × |
| F3 (male foil) | London | × | × | × |

The speakers read three texts, each of approximately 100 words. Each text was recorded first in normal voice and then in whisper. The first text was a sample from the cosmetic company's advertising brochure, the second a news story on a well-known local crime, and the third a technical extract from an engineering and technology magazine. The range of texts was designed to test whether the participants would be better able to identify familiar voices talking about familiar issues. The speakers also read a list of short phrases taken from the same texts.

In pilot work we originally asked speakers to produce a text of 250 words. However, this length proved to be very strenuous for the speaker when whispering. Whispering for such a sustained period led to difficulties with breath control, the need for deep and audible in-breaths to be taken, and periods of inadvertent phonation. As example of the latter is shown as Figure 1. The texts were thus shortened to 100 words, although for some participants even this proved difficult for consistent whisper. We ensured speakers had water to hand during the recordings. We discuss the forensic implications of these difficulties below.

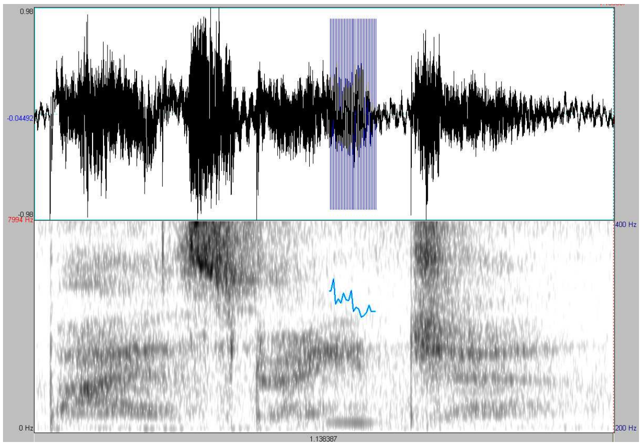

**Figure 1.** Waveform (top, including pulses to mark periodic section) and spectrogram (bottom, including F0 trace) of the phrase *by a stranger,* showing brief period of phonation during whisper.

The recordings were made using a Zoom Handy Recorder H4 device, set at 44.1 kHz sampling frequency with 16-bit resolution to a .wav file output. Recording sessions were held in a quiet room at the cosmetics store. The speaker sat approximately 30 cm from the microphone. Participants were asked to read through each text before recording began in order to familiarise themselves with it, and to query any unfamiliar words. The format of the task was then described and the participants were invited to comment on how they were finding the task between recordings. Before the first recording, the participants were encouraged to practice a loud ('stage') whisper, to ensure that the signal was loud enough for the device to record. Participants were then asked to complete a brief demographic questionnaire.

## 3.3 *Experimental materials*

The recordings were edited via Sony Sound Forge (version 9.0) to provide stimuli for three listening tests. The 'normalise' tool was used to adjust the average RMS level (loudness) for each sample to -6dB in order to overcome variation in the amplitude of the recordings. Some of the whispered samples were so quiet that they were not detected by the default scan settings. These were duly adjusted via the menu option 'ignore below 0db' to ensure that all sections were amplified to -6dB.

The designs of each listening test are summarised in Table 2. In Test 1 the stimuli were short (4 syllables) and whispered, in Test 2 they were also whispered but longer

(16 syllables), and in Test 3 the stimuli were short and spoken in modal voice. The decision to work with stimuli of these lengths was based on pilot testing, to ensure that tests were neither too easy nor too difficult. Tests 1 and 3 consisted of 108 stimuli (9 speakers × 12 extracts). Test 2 comprised 27 stimuli (3 extracts per speaker), both to limit the overall duration of the experiment and because it was predicted to be a relatively easy task. Test 1 was predicted to be the most difficult. Tests 2 and 3 were predicted to be approximately equally difficult, bearing in mind the experimental claim that around three times the duration of whispered material is required for a listener to identify the speaker as well as in normal speech (POLLACK, PICKETT & SUMBY, 1954). The list of stimuli is provided in the **Appendix**.

For each test the stimuli were entered in random order into a PowerPoint presentation. The first slide informed the participant of the nature of the test and how it was to be conducted. Each slide thereafter contained a speaker number (so the participant could navigate the answer sheet) and a sound file. For Tests 1 and 3 (the short samples) a transcription of what was being said was also included. The tests were automated to play the sound files and change slides, with an 8 second delay between each slide (following BLATCHFORD & FOULKES, 2006).

**Table 2.** Test designs

| Test | Stimuli length | Voice type | N stimuli | Approx. test duration |
|---|---|---|---|---|
| 1 | short (4 syllables) | whispered | 108 | 17 minutes |
| 2 | long (16 syllables) | whispered | 27 | 5 minutes |
| 3 | short (4 syllables) | modal | 108 | 17 minutes |

### 3.4 *Listeners*

Ten of the women from the social group (excluding the author, for obvious reasons) acted as listeners for a series of three closed set tests.

### 3.5 *Listening tests*

The three tests were presented in the same order (1-2-3) to all listeners because of the predicted difference in difficulty, and potential learning effects as the experiment progressed.

Listening tests were presented in sequence in a quiet room at the home of the listener or experimenter. The PowerPoint files were presented through an Acer Aspire 3000 laptop and Sennheiser HD 280 Pro headphones. Listeners were given paper questionnaires to record their judgments of the speakers in the tests. The questionnaires offered a closed set of candidates, i.e. the names of the six potential speakers from the group plus a seventh option of 'stranger'. Listeners were asked to place a tick in the appropriate box on the answer sheet. They were asked to provide an answer for each voice and told that they would hear each sample only once. The experiment lasted around 45 minutes in total.

### 3.6 *Statistical analysis*

We used logistic mixed effects regression (JAEGER, 2008) to evaluate our findings statistically. While traditional regression models only incorporate *fixed effects* (capturing systematic or planned contrasts and differences), mixed effects models also include *random effects* which allow the model to capture random variation across individuals, items or different levels of some other grouping factor. Our models are *logistic* regression model, which means that they model the probability of a given event as a function of fixed and random predictors. In the current case, this event – or the outcome variable – is correct identification.

We included the following fixed effects in our models:

- **test** (*short whispered* vs *long whispered* vs *short normal*);
- **type of trial** (*in group* vs *foil*);
- **text** (*cosmetic* vs *local crime* vs *technical*);
- **same speaker-listener** (*same* vs *different*): whether the listener judged a sample produced by herself or by a different speaker;
- **order** (an integer between 1-108): the position of the experimental stimulus within a given test.

As for the last predictor, the effect of presentation order is tested in a separate model that only included the *short whispered* and *short norma*l conditions, but not the *long whispered* condition. This is because the former two tests both included 108 stimuli, while the *long whispered* condition only included 27 stimuli, which means that it is not easily comparable with the other conditions.

We also included two sets of random *intercepts*, which capture overall differences in identification accuracy across speakers and listeners; and two sets of random *slopes*, which capture potential differences in the influence of **test** on identification accuracy across speakers and listeners.

In order to test the significance of our fixed effects, we performed model comparisons between a full model including all terms and a nested model where the fixed effect of interest is excluded (see e.g. SEYFARTH, 2014). We used likelihood-ratio tests to evaluate whether the full model yielded a better fit than the nested ones. In addition, we performed post-hoc pairwise comparisons among the test conditions using the Bonferroni-Holm correction for multiple comparisons.

Note that the male speaker was excluded from the statistical analysis, as listeners were nearly 100% accurate at identifying this voice regardless of other variables.

## 4 Results

### 4.1 *Overall descriptive results*

Table 3 summarises the overall performance of the listeners in the three tests. As predicted, Test 1 proved the most difficult (64% correct responses for all voices, 71% for familiar voices). Identification rates were higher in Test 2, and slightly higher again in Test 3 (87% and 93% respectively for familiar voices). The male foil was rejected almost without fail; listeners only failed to reject him as a stranger in 2 of the 2430 trials (both in Test 1). Listeners struggled, however, to reject the female foils in all three tests (26-36%).

**Table 3.** Correct identification by test and speaker category (average % correct, st. dev. in parentheses)

| Voice category | Test 1<br>short whisper | Test 2<br>long whisper | Test 3<br>short normal |
|---|---|---|---|
| all | 64.1 (30.2) | 76.7 (28.5) | 81.0 (26.2) |
| familiar | 71.1 (24.0) | 87.2 (17.6) | 92.8 (6.3) |
| female foils | 25.8 (2.4) | 33.3 (0) | 36.3 (8.8) |
| male foil | 98.3 (0.4) | 100 (0) | 100 (0) |

### 4.2 *Statistical analysis*

Our first full model focuses on differences in identification accuracy across the three test conditions. The dependent variable was response (a binary variable, correct/incorrect). This model included **test**, **type of trial**, **text** and **same speaker-listener** as fixed effects and all the random effects described in section 3.6. We also coded separately those tokens where the listener was presented with recordings of her own voice (**same speaker-listener**), since it is well known that listeners often find identifying their own voice difficult. Table 4 presents a summary of the fixed effects.

As the model summary shows, only the effect of **test** is significant in this model. Post-hoc comparisons based on Wald tests with Bonferroni-Holm correction for multiple comparisons show that identification accuracy is significantly higher for long whisper than for short whisper (*long whisper – short whisper* = 1.34, SE = 0.50, $z$ = 2.67, $p$ = 0.015), and also significantly higher for short normal speech than for short whisper (*short normal – short whisper* = 1.86, SE = 0.60, $z$ = 3.11, $p$ = 0.006). There was no significant difference between long whisper and short normal speech (*short normal – short whisper* = 0.51, SE = 0.51, $z$ = 1.00, $p$ = 0.317). These findings are summarised in Figure 2, which shows predictions from the regression model as probabilities.

**Table 4.** Summary for a logistic regression model including all three test conditions and the predictors **test**, **type of trial**, **text** and **same speaker-listener**

| Term | ESTIMATE | $\chi^2$ | DF | $p$ ($\chi^2$) |
|---|---|---|---|---|
| **test** = *long whispered* | 1.34 | 7.56 | 2 | 0.029 |
| **test** = *short normal* | 1.86 | – | – | – |
| **type of trial** = *in group* | 1.97 | 2.97 | 1 | 0.084 |
| **text** = *local crime* | –0.07 | 3.87 | 2 | 0.145 |
| **text** = *technical* | 0.20 | – | – | – |
| **same speaker-listener** = *same* | –0.13 | 0.37 | 1 | 0.542 |

**Note:** The estimates represent comparisons against a reference value (*short whispered* for **test**, *foil* for **type of trial**, *cosmetic* for **text** and *different* for **same speaker-listener**). The $\chi^2$ values, degrees of freedom and $p$-values are taken from likelihood-ratio tests that compare the full model against a nested model with the relevant predictor removed. Since each predictor is removed as a whole (not value by value), only a single set of figures is shown for each predictor, even when it has more than a single corresponding estimate.

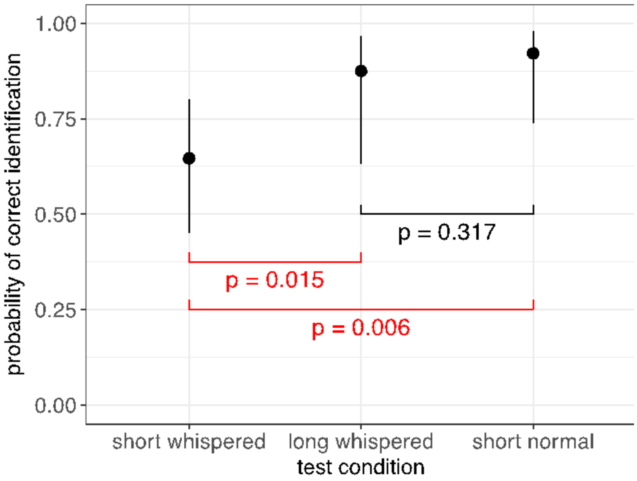

**Figure 2.** Model predictions (solid dots) along with 95% confidence intervals (vertical lines) across different test conditions and the results of post-hoc comparisons (horizontal connectors and $p$-values).

Note that **type of trial** was not significant at the 5% level in this model. This is likely because this predictor only varies across speakers, not within them, and there were only 8 different speakers, of whom just two were in the *foil* group. Although the estimated difference between foils and speakers from the group is large (1.97 in Table 4, and cf. the raw data shown in Table 3), it is not clear whether we can make a reliable generalisation from such a small sample. The relatively high $p$-value for this predictor ($p$ = 0.084) reflects the model's uncertainty about the generalisability of this estimate.

Since significance testing of random effects is a controversial area (e.g. HURLBERT, 1984), we do not report significance values for differences across listeners or speakers. However, the estimates of the random variance components allow us to calculate the ranges within which the listener- and speaker-specific identification accuracies are predicted to vary. In turn this enables us to compare the relative importance of these two grouping factors. Thus, the model predicts that the average accuracy values for different listeners in the short whisper condition will vary between 48-78% (this corresponds to a 95% predictive interval); the average accuracy values for different speakers in the same condition will vary between 18-94% even after we control for differences between foils and within-group samples. In other words, the identity of the speaker who produced a given sample has a much stronger influence on identification accuracy than the identity of the listener. Both of these estimates are based on relatively small samples and are sensitive to outliers, which means that they should be treated with a certain amount of caution.

For the sake of exposition we present raw data by listener and speaker in Figures 3 and 4.

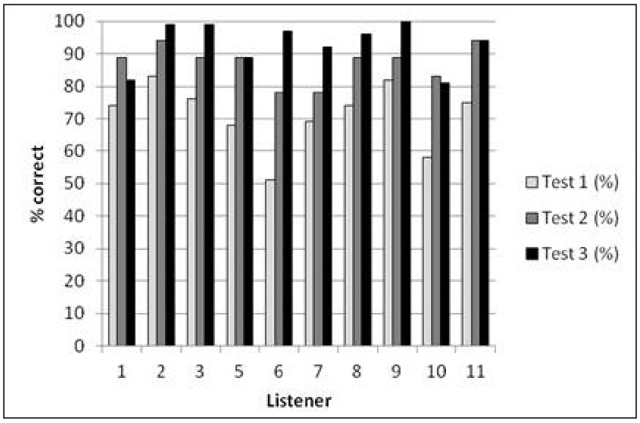

**Figure 3:** Results by listener, familiar voices only

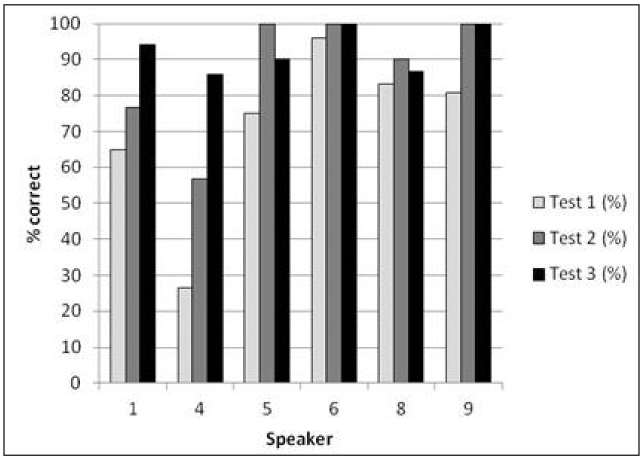

**Figure 4.** Results by speaker, familiar voices only

Figure 3 shows the range in listener performance across tests. It also shows that the same trend for improvement across tests is apparent for most listeners. Note that the only perfect performance was from listener 9 in test 3. Figure 4 supports the conclusion from the statistical analysis that there is greater variation by speaker. In particular, speaker 4 (who was in fact the first author and experimenter) was identified least well in all three tests, and was particularly poorly identified in the short whisper condition. One of the foils, who shared the same regional accent, was regularly misidentified as speaker 4.

Improvement across tests was clear in respect of all voices, although some were approaching ceiling in tests 2 and 3. Three speakers were identified perfectly in test 2, and two of those (6 and 9) were also identified perfectly in test 3. One of these (6) was even identified extremely well in the short whisper condition. Thus we can conclude that certain voices are more and less difficult to identify relative to this group.

A second regression model was fit to the data to test for order of presentation effects. This model includes all of the fixed and random effects from the previous model as well as **order** and an interaction between **order** and **test**, which captures potential differences in order effects between conditions. As explained in section 3.6, only the *short whisper* and *short normal* conditions were tested in this model. Since the estimates for the predictors that were also included in the previous model are largely unchanged, we do not present a separate summary for this model. The effect of **order** has a significant positive effect on identification accuracy (estimate = 0.008, $\chi^2$=9.12, df=2, $p$=0.010), which suggests that listeners' performance improved as they got accustomed to the task. The interaction between order and test is not significant (estimate = -0.006, $\chi^2$ = 1.4956, df = 1, $p$ = 0.2213). Thus we can infer that listeners improved as both tests progressed. Figure 5 shows model predictions as a function of order (labelled "position in experiment" in the graph) and test condition.

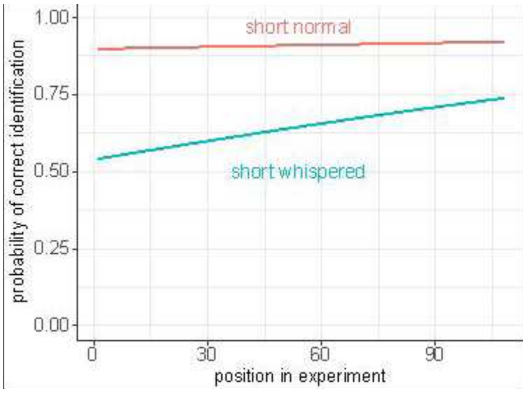

**Figure 5.** Model predictions as a function of position in experiment (**order**; shown along the horizontal axis) and **test** condition (blue = *short whispered*, red = *short normal*).

The increase in identification accuracy seems far greater in the short whispered condition than in the short normal condition, which may make the lack of a significant interaction seem surprising. However, this is simply an artefact of the scale chosen for the vertical axis: the model predictions are plotted as percentages, while the actual model estimates (which serve as the basis of the significance values) are calculated in logits; when viewed on the logit scale, the difference between the slopes of the two lines is smaller. From the perspective of a logistic model, a small increase in probability near the ceiling of probability values (where there is little room for further increase) is more or less equivalent to a much larger increase in probability near the middle of the probability scale (where there is plenty of room for increase / decrease). This should not preclude us from concluding that the improvement in the short whispered condition seems more pronounced – at least when viewed in the form of raw probabilities.

## 5 Discussion and conclusion

Perhaps one of the most surprising findings is the high overall identification rate, even in Test 1 (64% correct overall and 71% identification of familiar voices). This is well above chance (around 14% taking account of all seven response options). This is impressive considering that the stimuli in test 1 consisted of just four syllables, and lasted approximately half a second each. However, listeners made errors in identifying familiar voices in all but one of the 30 tests (Figure 3). It is important to reiterate the conclusion of previous experiments (e.g. LADEFOGED & LADEFOGED, 1980, FOULKES & BARRON, 2000, YARMEY et al., 2001) that speaker identification is certainly not automatic or necessarily straightforward. It is commonplace among the general public, and often assumed or alleged in court, that a witness is infallible when it comes to recognising the voice of someone they know well. This is a myth that needs to be dispelled. All listeners make errors, even with the voices of their closest family. Many factors can affect voice identification at a given instance. It is equally essential that we do not casually extrapolate experimental results to real life scenarios. Some voices are easier than others, while some listeners are better than others and can identify a voice very well even under very testing conditions. Again this finding mirrors those of previous experiments. Despite the short samples it seems likely that certain voices were readily identified because of their distinctive accent features relative to this particular group (cf. FOULKES & BARRON, 2000), notably the one Scottish speaker (speaker 9). In a forensic case involving a witness to a voice it is therefore imperative to consider in detail the facts of the case, and to ask the witness to undergo formal testing if appropriate (NOLAN, 2003).

The identification rates in our experiments might in part be so high because of the optimal conditions under which listeners heard the voices: the recordings were clear and involved 'stage whisper', and they were heard through high quality headphones in a relaxed setting. In other words, the experimental results might not reflect the way in which whisper is encountered as a disguise in forensic cases, where the disguise might be strengthened through low volume or additional voice quality features, and further degraded if transmitted through a telephone line and/or heard in conditions of emotion or stress.

As predicted, Test 1 proved more difficult than the other two tests. This finding is line with those of Yarmey et al. (2001) and Bartle & Dellwo (2015). We tentatively predicted little or no difference between Tests 2 and 3, based on Pollack, Pickett & Sumby (1954), and indeed this prediction was borne out. The longer whispered samples in Test 2 yielded identification rates slightly lower, but not significantly so, than those of the short normal samples in Test 3. It also appears that listeners improved throughout Tests 1 and 3, suggesting that more material assisted them in identifying the speakers relative to each other.

The male foil was rejected with almost no trouble, confirming that broad vocal tract resonance characteristics are robust to loss of phonation (SCHWARTZ & RINE, 1968, LASS et al., 1976, NOLAN, 1997). However, listeners struggled to reject the female foils in all three tests. Yarmey et al. (2001) also report a high rate of false hits, i.e. unfamiliar voices wrongly attributed to familiar speakers. Although we hesitate to confirm this pattern statistically, it does seem that the trend in cases of whisper is that unknown speakers are wrongly identified as someone known to the listener more often than familiar speakers are not identified. Perhaps this trend applies to voice disguise more generally. The forensic implication of such a finding is that courts should exercise caution over the testimony of any ear-witness, but especially so if the accused is someone familiar to the witness.

The final observation worth repeating here is that the speakers found it difficult to maintain whisper beyond about 30 seconds. Gick et al. (2013, p. 90) note how quickly a speaker will run out of air in whisper. Several speakers also lapsed into very brief periods of phonation (Figure 1). The partial glottal closure and high airflow probably explains why phonation may occur inadvertently (LAVER, 1980, 1994). The implication for ongoing forensic cases is clear: if possible, try to keep a whispering perpetrator talking. His disguise will probably fail from time to time, making identification of a familiar voice even more likely.

To conclude, we have presented experimental data to test the performance of listeners hearing whisper, in comparison with normal speech. The study as a whole illustrates how sociophonetic methods and resources can be applied in the forensic domain. It also confirms the value of investigating speaking and listening under circumstances not typical in laboratory or field studies. The production, understanding and cognitive representation of phonetic detail in the full repertoire of human experience is a rich and complex, but rewarding, field of enquiry.

## References

AHMADI, Farzaneh; MCLOUGHLIN, Ian V.; SHARIFZADEH, Hamid R. Analysis-by-synthesis method for whisper-speech reconstruction. *Proceedings of IEEE Asia Pacific Conference Circuits and Systems*. 2008. p. 1280-1283.

ALEXANDER, Anil. Forensic automatic speaker recognition using Bayesian interpretation and statistical compensation for mismatched conditions. *International Journal of Speech, Language and the Law*, v. 14, n. 1, p. 145-156, 2007.

BARTLE, Anna; DELLWO, Volker. Auditory speaker discrimination by forensic phoneticians and naive listeners in voiced and whispered speech. *International Journal of Speech, Language and the Law*, v. 22, n. 2, p. 229-248, 2015.

BLATCHFORD, Helen; FOULKES, Paul. Identification of voices in shouting. *International Journal of Speech, Language and the Law*, v. 13, n. 2, p. 241-254, 2006.

BULL, Ray; CLIFFORD Brian R. Earwitness voice recognition accuracy. In: WELLS, Gary L.; LOFTUS, Elizabeth F. (Ed.). *Eyewitness testimony: psychological perspectives.* Cambridge: Cambridge University Press, 1984. p. 92-123.

CHIN, Steven B.; PISONI, David B. *Alcohol and speech*. New York: Academic Press, 1997.

CLARK, Jessica; FOULKES, Paul. Identification of voices in electronically disguised speech. *International Journal of Speech, Language and the Law*, v. 14, n. 2, p. 195-221, 2007.

DE FIGUEIREDO, Ricardo M.; DE SOUZA BRITTO, Helena. A report on the acoustic effects of one type of disguise. *Forensic Linguistics*, v. 3, n. 1, p. 168-175, 1996.

DOCHERTY, Gerard; MENDOZA-DENTON, Norma. Speaker-related variation–sociophonetic factors. In: COHN, Abigail C.; FOUGERON, Cécile; HUFFMAN, Marie K. (Ed.). *Oxford handbook of laboratory phonology*. Oxford: Oxford University Press, 2012. p. 44-60.

FOULKES, Paul; BARRON, Anthony. Telephone speaker recognition amongst members of a close social network. *Forensic Linguistics*, v. 7, n. 2, p. 180-198, 2000.

FOULKES, Paul; FRENCH, Peter. Forensic speaker comparison: a linguistic-acoustic perspective. In: TIERSMA, Peter; SOLAN, Larry (Ed.). *Oxford handbook of language and law*. Oxford: Oxford University Press, 2012. p. 557-572.

FOULKES, Paul; HAY, Jennifer. The emergence of sociophonetic structure. In: MACWHINNEY, Brian; O'GRADY, William (Ed.). *Handbook of language emergence*. Oxford: Blackwell, 2015. p. 292-313.

FOULKES, Paul; SCOBBIE, James M.; WATT, Dominic J.L. Sociophonetics. In: HARDCASTLE, William; LAVER, John; GIBBON, Fiona (Eds.). *Handbook of phonetic sciences*. 2. ed. Oxford: Blackwell, 2010. p. 703-754.

FRENCH, Peter; HARRISON, Philip; WINDSOR LEWIS, Jack. R v John Samuel humble: The Yorkshire ripper hoaxer trial. *International Journal of Speech Language and the Law*, v. 13, n. 2, p. 255-273, 2007.

GICK, Bryan; WILSON, Ian; DERRICK, Donald. *Articulatory phonetics*. New York: John Wiley & Sons, 2013.

HAMMERSLEY, Richard; READ, J. Don. The effect of participation in a conversation on recognition and identification of the speakers' voices. *Law and Human Behavior*, v. 9, n. 1, p. 71-81, 1985.

HOLLIEN, Harry; MAJEWSKI, Wojciech; DOHERTY, E. Thomas. Perceptual identification of voices under normal, stress and disguise speaking conditions. *Journal of Phonetics*, v. 10, p. 139-148, 1982.

HURLBERT, Stuart H. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, v. 54, n. 2, p. 187-211, 1984.

JAEGER, T. Florian. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, v. 59, n. 4, p. 434-446, set. 2008.

KALLAIL, Ken J.; EMANUEL, Floyd W. An acoustic comparison of isolated whispered and phonated vowel samples produced by adult male subjects. *Journal of Phonetics*, v. 12, n. 1, p. 175-186, 1984a.

KALLAIL, Ken J.; EMANUEL, Floyd W. Formant-frequency differences between isolated whispered and phonated vowel samples produced by adult female subjects. *Journal of Speech and Hearing Research*, v. 27, n. 1, p. 245-251, 1984b.

KÜNZEL, Hermann J. Effects of voice disguise on speaking fundamental frequency. *International Journal of Speech Language and the Law*, v. 7, n. 2, p. 150-179, 2000.

LADEFOGED, Peter; LADEFOGED, Jenny. The ability of listeners to identify voices. *UCLA Working Papers in Phonetics*, v. 49, p. 43-51, 1980.

LASS, Norman J.; HUGHES, Karen R.; BOWYER, Melanie D.; WATERS, Lucille T.; BOURNE, Victoria T. Speaker sex identification from voiced, whispered, and filtered isolated vowels. *Journal of the Acoustical Society of America*, v. 59, n. 3, p. 675-678, 1976.

LAVER John. *The phonetic description of voice quality*. Cambridge: Cambridge University Press, 1980.

LAVER, John. *Principles of phonetics*. Cambridge: Cambridge University Press, 1994.

MASTHOFF, Herbert. A report on a voice disguise experiment. *Forensic Linguistics*, v. 3, n. 1, p. 160-167, 1996.

MCGEHEE, Frances. The reliability of the identification of the human voice. *Journal of General Psychology*, v. 17, n. 2, p. 249-271, 1937.

NOLAN Francis. Speaker recognition and forensic phonetics. In: HARDCASTLE, William; LAVER, John (Ed.). *Handbook of phonetic sciences*. Oxford: Blackwell, 1997. p. 744-767.

NOLAN, Francis. A recent voice parade. *International Journal of Speech, Language and the Law*, v. 10, n. 2, p. 277-291, 2003.

ORCHARD, T. L.; YARMEY, A. D. The effects of whispers, voice-sample duration, and voice distinctiveness on criminal speaker identification. *Applied Cognitive Psychology*, v. 9, n. 3, p. 249-260, 1995.

PAPP, Viktória. *The effects of heroin on speech and voice*. MSc dissertation, University of York, 2008.

POLLACK, I.; PICKETT, J. M.; SUMBY, W. H. On the identification of speakers by voice. *Journal of the Acoustical Society of America*, v. 26, n. 3, p. 403-406, 1954.

RHODES, Richard. *Assessing the strength of non-contemporaneous forensic speech evidence*. PhD dissertation, University of York, 2012.

ROBERTS, Lisa. *Vocal responses of individuals in distress*. PhD dissertation, University of York, 2012.

SCHWARTZ, Martin F. Power spectral density measurements of oral and whispered speech. *Journal of Speech and Hearing Research*, v. 13, p. 445-446, 1970.

SCHWARTZ, Martin F. Bilabial closure durations for /p/, /b/, and /m/ in voiced and whispered vowel environments. *Journal of the Acoustical Society of America*, v. 51, p. 2025-2029, 1972.

SCHWARTZ, Martin F.; RINE, Helen E. Identification of speaker sex from isolated, whispered vowels. *Journal of the Acoustical Society of America*, v. 44, n. 6, p. 1736-1737, 1968.

SEYFARTH, Scott. Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, v. 133, n. 1, p. 140-155, 2014.

STURM, Ruth; JAKIMIK, Jola. The perception of whispered speech. *Journal of the Acoustical Society of America*, v. 76, p. 29, 1984.

SUNDBERG, Johan; SCHERER, Ronald; HESS, Markus; MÜLLER, Frank. Whispering—a single-subject study of glottal configuration and aerodynamics. *Journal of Voice*, v. 24, n. 5, p. 574-584, 2010.

SWERDLIN, Yoni; SMITH, John; WOLFE, Joe. The effect of whisper and creak vocal mechanisms on vocal tract resonances. *Journal of the Acoustical Society of America*, v. 127, n. 4, p. 2590-2598, 2010.

TARTTER, Vivien C. What's in a whisper? *Journal of the Acoustical Society of America*, v. 86, n. 5, p. 1678-1683, 1989.

TSUNODA, Koichi; OHTA, Yasushi; NIIMI, Seiji; SODA, Yasushi; HIROSE, Hajime. Laryngeal adjustment in whispering: magnetic resonance imaging study. *Annals of Otology, Rhinology & Laryngology*, v. 106, n. 1, p. 41-43, 1997.

VAN LANCKER, Diana; KREIMAN, Jody; EMMOREY, Karen. Familiar voice recognition: patterns and parameters. *Journal of Phonetics*, v. 13, n. 5, p. 19-38, 1985.

WELLS, John C. *Accents of English*. Cambridge: Cambridge University Press, 1982. 3 v.

YARMEY A. Daniel; YARMEY A. Linda; YARMEY Meagan J.; PARLIAMENT Lisa. Commonsense beliefs and the identification of familiar voices. *Applied Cognitive Psychology*, v. 15, n. 5, p. 283-299, 2001.

ZHANG, Chi; HANSEN, John H. Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing. *IEEE Transactions on Audio, Speech, and Language Processing*, v. 19, n. 4, p. 883-894, 2011.

ZHANG, Cuiling; TAN, Tiejun. Voice disguise and automatic speaker recognition. *Forensic Science International*, v. 175, n. 2, p. 118-122, 2008.

## Appendix

### *List of items used as stimuli*

**Long samples – 16 syllables**

- *We believe in long candlelit baths, sharing showers and massage.*
- *She sent her last text message a few minutes later to a friend.*
- *ultra high-frequency device that is in the terahertz range.*

**Short samples – 4 syllables**

- *fruity products*
- *candlelit baths*
- *you've been mangoed*
- *I love juicy*
- *later that week*
- *an acquaintance*
- *in good spirits*
- *by a stranger*
- *carbon atom*
- *terahertz range*
- *oscillator*
- *a Pentium*

Recebido: 30 de setembro de 2016
Aprovado: 06 de março de 2017
Contatos: marton.soskuthy@york.ac.uk
paul.foulkes@york.ac.uk