Assessing the abilities of phonetically untrained listeners to determine pitch and speaker accent in unfamiliar voices

James Tompkinson & Dominic Watt

University of York, UK

Abstract. *It is sometimes the case that a victim of a crime will never see the perpetrator's face, but will be exposed to his or her voice. This could occur in situations such as masked robberies, telephone fraud, or the receipt of bomb threats via phone or voicemail. In such cases, attempts can be made by the police or intelligence services to get the witness or victim to describe the voice of the offender. However, there is a high likelihood that a given earwitness will lack the linguistic expertise and technical vocabulary of the kind used by trained phoneticians when they describe voices. One question that arises from this problem is whether phonetically untrained listeners have the ability, using verbal means, to accurately capture different aspects of speakers' voices. This paper presents an experiment in which a group of listeners were tasked with assessing how 'high-pitched' the voices of 12 speakers were, along with providing a description of each speaker's accent. These assessments were then compared to measured Fundamental Frequency (F0) values and prior knowledge of the speakers' accents in order to assess listener performance. The results suggest that while some listeners have the ability to make reliable judgements of relative vocal pitch, the overall correlations between measured F0 and perceived pitch were weak. With regard to accent, the results suggest that the more unfamiliar a speaker's accent is to the listener, perhaps owing to the geographical distance of the area where the accent is spoken from the listener's place of origin, the harder it will be for the listener to accurately describe that accent. We argue that testing the abilities of earwitnesses to assess aspects of speakers' voices before their descriptions are used further would be a useful safeguard against the use of potentially inaccurate or erroneous earwitness evidence in police investigations.*

Keywords: *Keywords: Voice identification, vocal pitch, speaker accent, earwitness evidence.*

Resumo. *Acontece com frequência a vítima de um crime não ver o rosto do perpetrador, mas ser exposta à sua voz. Pode acontecer em casos como roubos por assaltantes encapuzados, fraudes por telefone ou ameaças de bomba por telefone ou voicemail. Nesses casos, as forças policiais ou a polícia judiciária podem tentar que a testemunha ou a vítima descrevam a voz do(a) criminoso(a). Contudo, existe uma probabilidade elevada de uma testemunha auditiva não possuir os* *conhecimentos linguísticos e o vocabulário técnico utilizado por foneticistas especializados ao descrever as vozes. Uma questão, decorrente deste problema, é se ouvintes sem formação fonética possuem a capacidade de, usando meios verbais, captar com precisão diferentes aspetos das vozes dos falantes. Este artigo reporta uma experiência na qual se pediu a um grupo de ouvintes para determinar quão agudas eram as vozes dos 12 informantes, para além de fornecer uma descrição da pronúncia de cada falante. Estas avaliações foram, depois, comparadas com os valores de Frequência Fundamental (F0) medidos e com o conhecimento prévio da pronúncia dos falantes para avaliar o desempenho dos ouvintes. Os resultados sugerem que, embora alguns ouvintes tenham a capacidade de fazer julgamentos fiáveis do tom vocal relativo, as correlações gerais entre os valores F0 medidos e o tom percebido eram fracos. Relativamente à pronúncia, os resultados sugerem que, quanto menos familiar for a pronúncia de um falante para o ouvinte, talvez devido à distância geográfica da área onde se fala a pronúncia do local de origem do falante, mais difícil é para o ouvinte descrever com precisão a pronúncia. Defendemos que, testar as capacidades das testemunhas para avaliar aspetos das vozes dos falantes antes de utilizar descrições mais detalhadas, seria uma boa salvaguarda contra o uso de provas testemunhais potencialmente imprecisas ou errónas em investigações policiais.*

Palavras-chave: *Identificação de voz, tom vocal, pronúncia do falante, prova testemunhal.*

Introduction

In January 2018, media reports broke of a series of violent burglaries taking place in southeast England, during which a masked intruder physically assaulted and robbed victims of high-value possessions such as jewellery (BBC News, 2018). When asked to provide a description of the perpetrator, one victim described him as follows: “I would say he spoke well, he had no accent, he didn’t have bad grammar, he’s an intelligent man, he knows how to assess the situation and carry this out.” Examples of this kind illustrate some of the difficulties that witnesses may have when asked to provide linguistically precise descriptions of the speech of a criminal who provided few or no other useful clues to identity, e.g. from his face, while the offence was in progress. Under such circumstances, the witness’s description of the offender’s voice may become the most useful evidence available. The potential importance of such earwitness testimony is highlighted by Nolan and Grabe (1996: 74), who point out that victims of certain crimes – verbal threats delivered via the telephone, masked robberies, sexual assaults, or instances where criminal activity has been overheard but not seen, and so on – may have been exposed only to the voice of the culprit, and not his or her face.

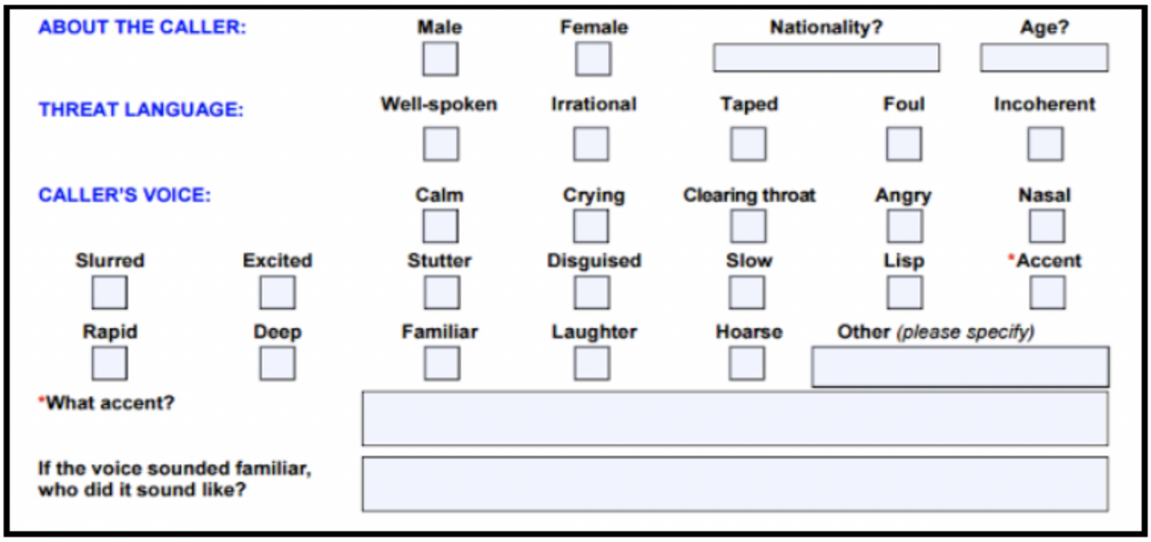
However, the great majority of earwitnesses to crimes will have had little or no formal linguistic training (Griffiths, 2012), and, according to Shuy (1993), will almost always lack both the ability and the vocabulary needed to give adequately detailed descriptions of other speakers’ language behaviour. Furthermore, Sherrin (2015) documents two examples of cases in Canada in which unreliable earwitness voice identification led to wrongful convictions, and also cites 17 US cases of wrongful imprisonment that were based, at least in part, on faulty earwitness testimony. These issues present an ongoing problem to police officers and security personnel, who from time to time will wish to elicit meaningful descriptions of the voices of criminals from earwitnesses.

Although speaker identification by earwitnesses and earwitnesses' descriptions of offenders' voices are not equivalent, dependent as they are upon different sorts of memory recall, they are closely related. Broeders and van Amelsvoort (2001) state that the foil (non-suspect) samples in a voice identity parade should match as closely as possible with the verbal description given by the witness, although they also point out that such descriptions do not necessarily form a solid foundation upon which foil selection should take place. Furthermore, the UK guidelines on constructing voice lineups (Nolan, 2003: 288) explicitly state that "the identification officer in charge should obtain a detailed statement from the witness" which "should contain as much detail and description of the [offender's] voice as is possible". This emphasises the need for voice descriptions to be promoted as best practice in the UK as a part of eliciting earwitness evidence.

It has also been argued that the process by which phonetically untrained listeners identify voices operates below the level of consciousness (Broeders and van Amelsvoort, 2001; Watt, 2010), making it difficult for an earwitness to introspect about and then verbally externalise what can essentially be viewed as an automatic process. The problem is further compounded by the often highly technical nature of the terminology used by expert phoneticians to capture aspects of a speaker's voice, much of which – in spite of the relative transparency of labels like 'creaky', 'whispery' or 'breathy' for certain voice quality attributes – is unlikely to form a part of the non-linguist's lexicon. Watt and Burns (2012) highlight that it is unlikely that the majority of earwitnesses will have voice description skills comparable to those who have received specialised training in phonetics or linguistics. This issue was earlier commented on by Yarmey (2001), who obtained voice descriptions of unfamiliar speakers using an open-ended question format in which listeners were free to provide as many or as few descriptors as they considered appropriate. Yarmey (2001) observed that listeners provided, on average, between 4 and 5 descriptors, but that these were often non-technical and somewhat limited in their usefulness.

However, despite warnings from researchers that phonetically untrained listeners perform poorly when tasked with describing the voices of speakers, some research has shown that listeners appear to be able to identify certain aspects of speakers' voices with relative accuracy. In an investigation of listener accent attribution, Griffiths (2012) found that lay listeners were able to label speakers' accents reasonably accurately, although descriptions of the voices of speakers with localisable accents, such as that of Cardiff, Wales, were more accurate than those for speakers with less region-specific accents like Standard Southern British English (SSBE). Additionally, Watt and Burns (2012) found that listeners were able to provide phonetically interpretable descriptions of voice quality with a tolerable degree of accuracy and consistency, and in a way that was compatible with expert terminology. Both Griffiths (2012) and Watt and Burns (2012) stress the importance of further research on how non-linguists describe voices in forensically relevant contexts. Griffiths (2012: 76) specifically warns that this research is needed because "non-linguist members of the general public [by which here he means police officers] are appointed to elicit the best possible linguistic evidence, from other non-linguist members of the general public [witnesses], which other non-linguists [legal counsels such as barristers] then represent in law courts."

At present, voice descriptions elicited from witnesses by law enforcement officers in the UK are still unlikely to be systematically collected, not least because no standardised protocol to structure the task has yet been developed in this country (Watt and Burns, 2012; Watt, 2010). By contrast, the Netherlands Forensic Institute (NFI) have for over a decade been making use of a questionnaire for the elicitation of earwitness voice descriptions (Watt and Burns, 2012). There do exist, however, UK government documents such as the National Counter Terrorism Security Office bomb threat checklist (National Counter Terrorism Security Office, 2016), which incorporates questions about speakers' voices in the context of lay-listener evaluations of telephoned bomb threats. The relevant section of this document is reproduced in Figure 1. It invites the user to comment on a range of vocal features using predominantly non-technical terms such as 'slurred', 'lisp', and 'deep', alongside a range of labels for presumed emotional states such as 'angry' and 'calm'. Whereas the NCTSO document uses simple 'present/absent' tick-boxes to elicit earwitness descriptions, the NFI questionnaire uses scales ranging from one extreme to another as a means of getting witnesses to describe the voices of speakers they have heard. The latter technique mirrors the method used by Handkins and Cross (1985), which elicits scalar judgements of rate of speech, rate variation, pitch variation, 'expressive style', 'enunciation', 'inflection', tremor, pauses, and nasality. While some of these terms, such as nasality, have clear phonetic or linguistic correlates, the precise meaning of others, such as enunciation (the scale for which ranges from 'very poor' to 'very good') or inflection (from 'none, flat' to 'very much'), is harder to pin down.



ABOUT THE CALLER:		Male □	Female □	Nationality? □		Age? □
THREAT LANGUAGE:		Well-spoken □	Irrational □	Taped □	Foul □	Incoherent □
CALLER'S VOICE:		Calm □	Crying □	Clearing throat □	Angry □	Nasal □
Slurred □	Excited □	Stutter □	Disguised □	Slow □	Lisp □	*Accent □
Rapid □	Deep □	Familiar □	Laughter □	Hoarse □	Other *(please specify)* □	

*What accent? □

If the voice sounded familiar, who did it sound like? □

Figure 1. Extract from UK National Counter Terrorism Security Office bomb threat checklist (full form available at http://bit.ly/2o8UDBq).

Research aims

Following the work of Griffiths (2012) and Watt and Burns (2012), the goal of this paper is to assess how accurate listeners are at making judgements of specific aspects of speakers' voices under particular conditions. The research presented focuses on two particular aspects of voice: *pitch* and *regional accent*. This research addresses whether listeners' judgements of the high-pitchedness of a speaker's voice align with acoustic measurements of average Fundamental Frequency (F0), the key acoustic correlate of vocal pitch (Laver, 1994). It also builds on Griffiths' (2012) work by eliciting and examining listeners' descriptions of the accents of speakers of three different varieties of English.

These two aspects of voice were chosen owing to the fact that neither parameter relies on lay-listeners' ability to interpret specialised linguistic or phonetic terminology, and on account of both types of voice characteristics being used in documents such as the NCTSO bomb threat evaluation checklist (Figure 1).

Method

Stimuli

Twelve speakers (6 male) provided informed consent to take part in a recording session during which they were asked to produce the utterances "There's a bomb at York Station. It will go off this afternoon" and "I'm warning you about a bomb at York Station, which will go off this afternoon". Given that the NCTSO document is used to evaluate bomb threats, we sought to mirror this context when designing the stimuli for the experiments. Each speaker was instructed to produce each utterance twice, once with extra emphasis on the word 'will' and once with emphasis on the word 'this'. This yielded 48 recordings to be used as experimental stimuli. All speakers were students at the University of York or Newcastle University, UK. Recordings were made in a quiet recording environment using a Zoom H4N handheld recorder with the microphone positioned on a table approximately 30cm from each speaker. Among the group of speakers, four were speakers of Standard Southern British English (SSBE), four were speakers of Northern Irish English, and four were L2 speakers of English having 'Middle Eastern' languages as an L1 (three Arabic speakers, one Persian speaker) [1]. Each accent group contained an equal number of male and female speakers. The SSBE and Northern Irish accent samples were checked by the researchers to ensure they were appropriately representative of the target accents.

The median F0 measurements for each voice were extracted using the ProsodyPro script (Xu, 2013) in Praat (Boersma and Weenink, 2016), with pitch trace errors being manually corrected before the script was used. Additionally, measures were taken for the *F0 range, formant dispersion* (measured as the "average distance between adjacent formants up to F3" (Xu, 2013)), *jitter* (an index of variability in glottal cycle duration), *shimmer* (glottal cycle amplitude variation), and *harmonic-to-noise ratio* of each speaker's voice in each utterance. These additional measurements were also extracted using the ProsodyPro Praat script (Xu, 2013). These variables were used to capture a range of information about each speaker's vocal tract resonances and phonation qualities.

Finally, in order to make the experimental stimuli sound more like real-world telephone calls, all recordings were band-pass filtered between 300 and 3400Hz to simulate a landline telephone channel (Künzel, 2001; Nolan *et al.*, 2013). A 0.5-second period of silence was also added to the end of each utterance, and this was followed by a 1-second long 175Hz tone which was designed to resemble the 'hangup tone' ending of a telephone call.

Participants

85 student participants (9 male, mean age = 20, age range = 18-55) received payment or course credit to take part in an experiment in which they were tasked with evaluating a subset of the recordings created for the experiment. No participant in the study had received advanced-level formal phonetic training. Each participant heard a different subset of the total number of voices, presented in a computer-generated randomised

order. The mean number of voices evaluated per listener was 11, and the mean number of times each utterance was evaluated was 20. All participants were tested in either the Department of Psychology or the Department of Language and Linguistic Science at the University of York, and all participants were native English speakers who self-identified as having a British English accent. Furthermore, no participants were from Northern Ireland, and no listeners reported that they themselves had an accent with any Arabic, Persian or Middle Eastern influence.

Procedure

Participants wore closed-cup headphones in a quiet environment, and were instructed to listen to each voice and then to answer a series of questions about the speaker they had heard. Listeners were not limited as to the number of times they could hear a given recording, but were not able to amend previous answers when progressing through the experiment. As part of the questionnaire, listeners were instructed to say how high-pitched they thought the voice of each of the speakers they heard was. These evaluations were provided on a scale ranging from 0-100, where 0 represented 'very low-pitched' and 100 represented 'very high-pitched'. The scale was the same for male and female voices, and listeners were not given any instructions to provide ratings in accordance with gender norms. Listeners were also instructed to say what accent they thought each speaker had. This was done using an open-answer format, in response to the question "'*What accent do you think this speaker has? Leave the box blank if you are unsure*". We also collected information about how similar listeners thought their own accents were to a range of different UK accents. This list included *Oxford/Cambridge* (designed to reflect SSBE), *Newcastle, Yorkshire, Manchester, Liverpool, Belfast* and *Glasgow*. This information was collected in order to assess perceived similarity between listeners' accents and two of the target varieties in the experiment (SSBE and Northern Irish English). The other accents were included as distractors so as not to focus listeners' attention entirely on the target varieties.

Given that the research was not concerned with listeners' abilities to remember speakers' voices, the evaluation of each voice sample took place immediately after exposure to that sample. It is also acknowledged that the experimental design of this study created a more favourable environment for voice evaluations to take place than would be expected in real-world forensic scenarios. For example, evaluations took place in a stress-free environment, listeners were aware that they would be evaluating each voice using a repeated evaluation format, listeners had the option to listen to each recording as many times as they wished to, and the evaluative questions were asked immediately after exposure to a given vocal stimulus.

Results and discussion

Pitch perception

To assess how accurate listeners' pitch judgements were, their perceived pitch scores were compared to the corresponding measured median F0 values for the voices of the speakers in the experiment. Figure 2 plots listeners' pitch judgements against the measured median F0 values, separated in accordance with speaker sex, given that F0 is a sexually dimorphic aspect of voice (e.g. Puts *et al.*, 2006).

Figure 2 shows a weak positive correlation between listeners' pitch judgements and the measured median F0 values for both male (Pearson's $r = 0.33$, $df = 492$, $p < 0.001$)

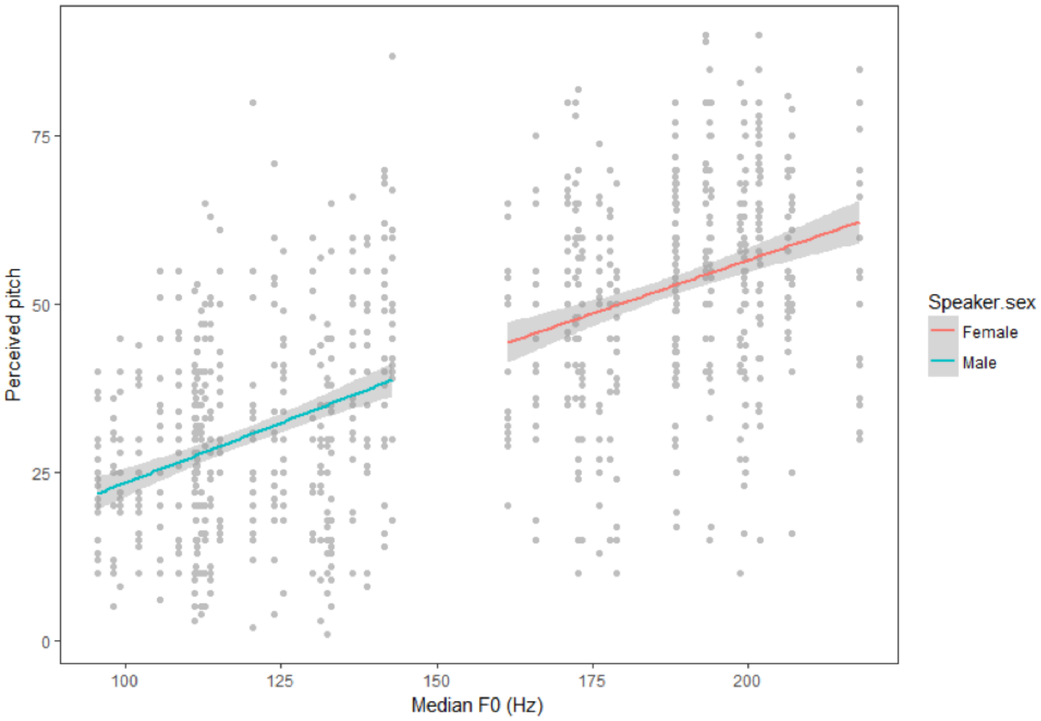


Figure 2. Relationship between listeners' perceived pitch scores and median F0. The axis units are Hz (x axis) and listeners' subjective pitch ratings on a scale between 0 ('very low-pitched') and 100 ('very high-pitched') (y axis). Each dot represents a single listener judgement, and each column of dots represents an individual recording (four produced by each speaker; male and female speakers are treated separately).

and female (Pearson's r = 0.28, df = 479, p < 0.001) speakers. We use Cohen's (1992) approach to the estimation of the magnitude of these effects, where r=0.10 equates to a small effect size, r=0.30 is the threshold for a medium effect size (classified by Cohen (1992: 156) as an effect which would "represent an effect likely to be visible to the naked eye of a careful observer"), and r=0.50 is said to represent a large effect size. We can therefore posit small to medium effect sizes for the relationship between median F0 and perceived pitch for female and male speakers in the present experiment. Additionally, although the relationship is reported by our models to be statistically significant for both male and female speakers, the graph in Figure 2 shows that a high level of variation exists between listeners' perceptual pitch scores and the corresponding measured F0 values. The relatively high level of variation in the sample is also evident in the r^2 values for the relationship between measured mean F0 and perceived pitch. For the male speakers, 10% of the variation ($r^2 = 0.10$) in the sample was accounted for by the relationship between measured mean F0 and perceived pitch. For the female speakers, 7% ($r^2 = 0.07$) of the total variation was accounted for by this relationship. Figure 2 also illustrates that while male voices were, overall, perceived to be lower-pitched than the female voices, there was a relatively high degree of overlap between the perceived pitch judgements for the male and female voices in the experiment. This was, however, not mirrored in the measured median F0 values, which showed complete separation between male and female speakers.

Three potential explanations can be proposed to explain the results seen in Figure 2. The first is that individual listeners interpreted the perceived pitch scale differently, and that a given value on the scale did not, therefore, map onto the perceived pitch scale equivalently for each listener. Secondly, it could be the case that other aspects of speakers' voices, such as voice quality or the relative distribution of formants, could also influence pitch perception. This would mean that making direct comparisons between average F0 and perceived pitch is a rather crude one-dimensional measure of the accuracy of listeners' pitch judgements. Thirdly, it could be the case that listeners are both inconsistent and inaccurate when tasked with gauging how high-pitched the voice of a given speaker is.

In an attempt to reduce the influence of individual differences in how listeners interpreted the scale used to elicit perceived pitch judgements, standardised scores were calculated for each listener's judgements of the high-pitchedness of speakers' voices using the scale() function in R (Baayen, 2008: 61). Figure 3 plots the standardised perceived pitch scores against the corresponding measured median F0 values. The figure reveals that there was some reduction in variation when standardised scores were used, in comparison to the raw data displayed in Figure 2. Analysis of the correlation coefficients showed a slightly tighter positive correlation and increased effect size between perceived pitch and median F0 for both male (Pearson's $r = 0.40$, $df = 492$, $p < 0.001$) and female (Pearson's $r = 0.32$, $df = 479$, $p < 0.001$) speakers when standardised scores were used. However, the r^2 values for both male ($r^2 = 0.16$) and female ($r^2 = 0.10$) speakers showed that only a limited amount of variation was accounted for by the relationship between standardised perceived pitch scores and measured mean F0.

The second reason that was proposed above for the weakness of the relationship between perceived pitch and measured mean F0 was that other variables, such as voice quality or the dispersion of formants across the frequency spectrum, could influence listener pitch perception alongside average F0. In order to assess the relationship between multiple acoustic phonetic variables and listeners' pitch judgements, multiple linear regression models were constructed using the lm() function in R. These contained listeners' perceived pitch scores as the dependent variable, and measurements of *median F0, F0 range, formant dispersion, jitter, shimmer* and *harmonic-to-noise ratio* as independent variables. Separate models were constructed for male and female speakers. Analysis of the r^2 values from the models for both male ($r^2 = 0.15$) and female ($r^2 = 0.13$) speakers showed that a greater proportion of variance was accounted for when the additional acoustic measures were considered, although the respective models still only accounted for 15% and 13% of the variation in the data. The proportion of variance accounted for in the relationship between perceived pitch and acoustic aspects of voice was further enlarged by using the standardised pitch judgement scores instead of the raw judgement scores, with 20% of the variation being accounted for by the model for male speakers ($r^2 = 0.20$), and 19% by the model for female speakers ($r^2 = 0.19$). However, in order to capture this level of variation, multiple judgements made by the same listener were required in order to calculate the standardised pitch judgement scores. To some degree this could be considered unrealistic for users of documents such as the NCTSO bomb threat checklist, which is designed to obtain earwitness evaluations from a single listener about a single speaker on a single occasion.

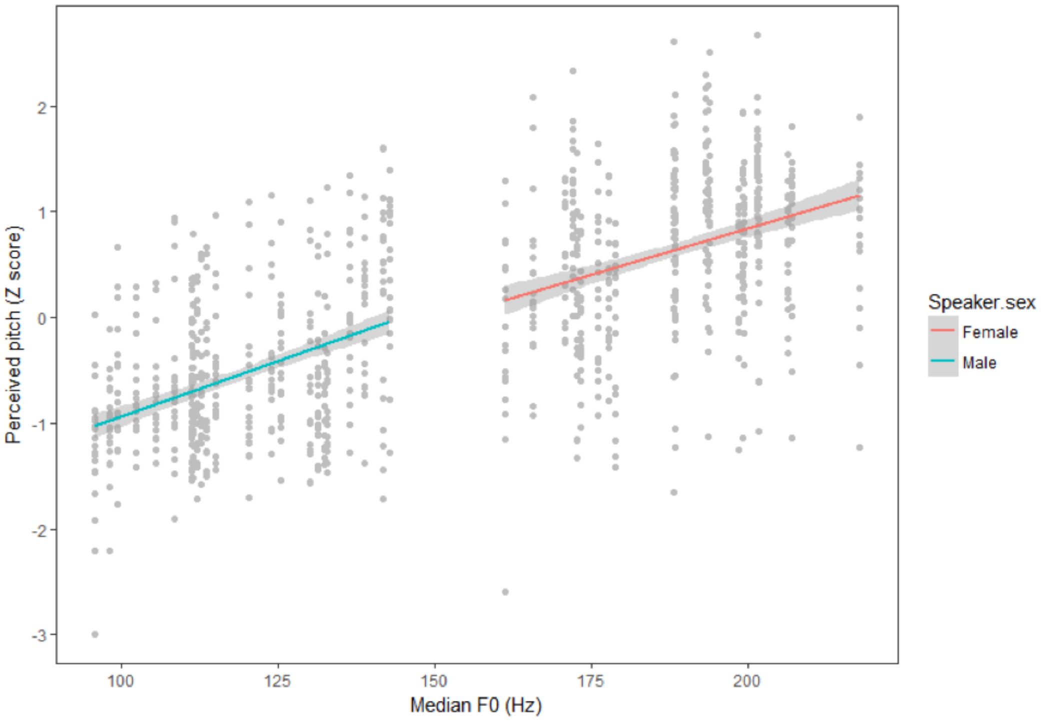
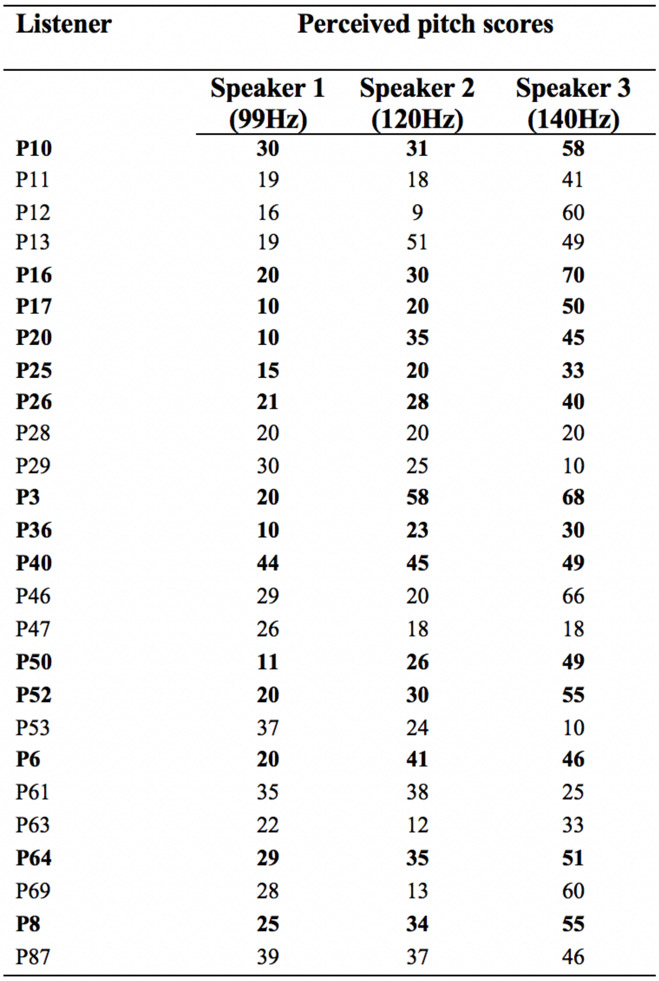


Figure 3. Relationship between listeners' standardised perceived pitch scores and median F0. The axis units are Hz (x axis) and listeners' standardised subjective pitch ratings on a scale between 0 ('very low-pitched') and 100 ('very high-pitched') (y axis). Each dot represents a single listener judgement, and each column of dots represents an individual recording (four produced by each speaker; male and female speakers are treated separately).

Further questions arise from this analysis relating to the role of the individual listener in the pitch judgement task. Specifically, is it simply the case that some listeners are good at the task, and some are not? If this were indeed the case, then there might be some merit in testing the ability of an earwitness to distinguish aspects of voice, for instance pitch, before his/her earwitness evidence is further used. In order to address this question using the data from the current experiment, a subset was created containing the responses of all those participants who provided a pitch judgement for three of the male speakers in the experiment, hereafter labelled Speaker 1, Speaker 2 and Speaker 3. These speakers were chosen because their average median F0 values (a) spanned the range found in the data for male speakers, (b) reflected the population statistics for English speakers' average F0 reported by Hudson *et al.* (2007), and (c) were almost equally spaced from each other along the pitch continuum (Speaker 1 - 99Hz, Speaker 2 - 120Hz and Speaker 3 - 140Hz). Given that the question randomisation process meant that not all listeners evaluated the voices of all speakers, some listeners were excluded from this analysis.

In total, 26 listeners provided at least one perceived pitch judgement for utterances produced by the three speakers described above. Table 1 shows the perceived pitch scores for each listener. The interest in this analysis is not in the absolute values, but rather in the relative pitch judgements provided by listeners. Given the 20Hz gaps between the three speakers' average median F0s, it was expected that listeners would provide a lower perceived pitch score for Speaker 1 (99Hz) than for Speaker 2 (120Hz), and that the score applied to Speaker 2 would, in turn, be lower than the score for Speaker 3 (140Hz). If a listener met these criteria, they were classified as an accurate listener, shown in bold type in Table 1. This analysis showed there were 14 accurate listeners within the subset. This would support the view that some listeners are simply unable to judge pitch accurately according to the present criterion, while other listeners are capable of performing the task adequately or well.



Listener	Perceived pitch scores		
	Speaker 1 (99Hz)	Speaker 2 (120Hz)	Speaker 3 (140Hz)
P10	**30**	**31**	**58**
P11	19	18	41
P12	16	9	60
P13	19	51	49
P16	**20**	**30**	**70**
P17	**10**	**20**	**50**
P20	**10**	**35**	**45**
P25	**15**	**20**	**33**
P26	**21**	**28**	**40**
P28	20	20	20
P29	30	25	10
P3	**20**	**58**	**68**
P36	**10**	**23**	**30**
P40	**44**	**45**	**49**
P46	29	20	66
P47	26	18	18
P50	**11**	**26**	**49**
P52	**20**	**30**	**55**
P53	37	24	10
P6	**20**	**41**	**46**
P61	35	38	25
P63	22	12	33
P64	**29**	**35**	**51**
P69	28	13	60
P8	**25**	**34**	**55**
P87	39	37	46

Table 1. Listeners' perceived pitch scores for Speaker 1, Speaker 2 and Speaker 3. Bold type denotes listeners who assigned the 'correct' ranking of the three speakers from low to high pitch, irrespective of the spacing on the perceptual scale between Speakers 1 and 2 and Speakers 2 and 3, or the placement of the scores on the 0-100 scale.

Accent perception

The data in this study also permitted an assessment of how accurately listeners could describe a speaker's accent via the responses to the question "What accent do you think this speaker has? Leave the box blank if you are unsure". The experiment made use of three different accents: Standard Southern British English (SSBE), Northern Irish English, and Middle Eastern-accented English. Listeners were also asked to state how similar they felt their own accent was to a series of other UK accents using a 0-100 scale (very different - very similar). The list of accents included *Oxford/Cambridge* and *Belfast* so as to facilitate an assessment of how closely aligned listeners thought their own accents were to the British target varieties in the experiment.

Responses to the question which obtainened listeners' assessments of how similar they thought their own accent was to the accents of both *Oxford/Cambridge* (SSBE) and *Belfast* (Northern Irish English) showed that listeners in the experiment aligned their own accents much more closely to SSBE than to Northern Irish English. The mean similarity score across the sample for SSBE was 40.4 (range 0-100), whereas the mean similarity score for Northern Irish English was 4.6 (range 0-47). Additionally, 51 listeners provided a similarity score of 0 for *Belfast*, in contrast to 17 listeners who gave a similarity score of 0 for *Oxford/Cambridge*.

A summary of the accent attributions for the SSBE speakers is shown in Figure 4. The results shown in Figure 4 reveal that listeners appeared to describe the accent of the SSBE speakers relatively accurately when they opted to describe it, although the most commonly chosen option was to leave the box blank to indicate uncertainty. When labelling SSBE, the most common way of answering besides selecting Unsure (Blank) was to choose one of the set of accent labels relating to SSBE or RP. These answers included "Southern", "Southern accent", "Southern England", "Southern English", "SSBE", "Standard Southern British", "RP", "Roughly RP" and "RP but grew up in London/'Estuary English' area" (Estuary English being the relatively newly-emerged 'hybrid' of RP and working-class London English; see Altendorf, 2011). An association between the SSBE speakers and the prestigious university towns of Oxford and Cambridge was also found in the data, as was a link between the SSBE accent and the capital city of London. More specific places in southeast England, including Kent, Chelsea and Surrey, were occasionally listed by listeners. More general terms such as "British" and "English" were also used, possibly owing to the generalisable nature of the accent, or to the presence of other accents in the experiment which were non-English.

Given the lack of a fixed geographical location for SSBE, and the position of Received Pronunciation (RP) as a social rather than a geographical accent of the UK (Hughes *et al.*, 2012), it can be argued that an association with any location within the south or south east of England could validly be considered an accurate attribute of an SSBE accent. It can also be argued that if a listener was unsure about a speaker's accent, then providing no answer rather than risking an inaccurate description was an appropriate strategy. Furthermore, it could be contended that the explicit instruction to provide no answer when the listener was unsure about a speaker's accent was a useful means of allowing listeners to express their uncertainty with confidence, instead of providing instructions which could implicitly encourage listeners to provide an accent label solely because the question asks for one. It is also possible that a listener's decision not to provide an answer was based on the perception that speaking with an SSBE accent means the talker has 'no' accent, a belief which is commonly held among laypeople in the UK (Mugglestone, 2003), or that because SSBE is not confined to a specific locality in Southern England, it was not possible for listeners to define the speaker's accent to a specific town, city or region. Indeed, one participant in the experiment (P65) described her own accent as "no accent – plain southern but not posh", which further illustrates these possible explanations.

Figure 4 also shows that a small number of more inaccurate labels, including "Yorkshire", "Manchester", "York" and "Lancashire" were provided by listeners. While it is certainly true that some people from these places speak with RP/SSBE accents, or accents

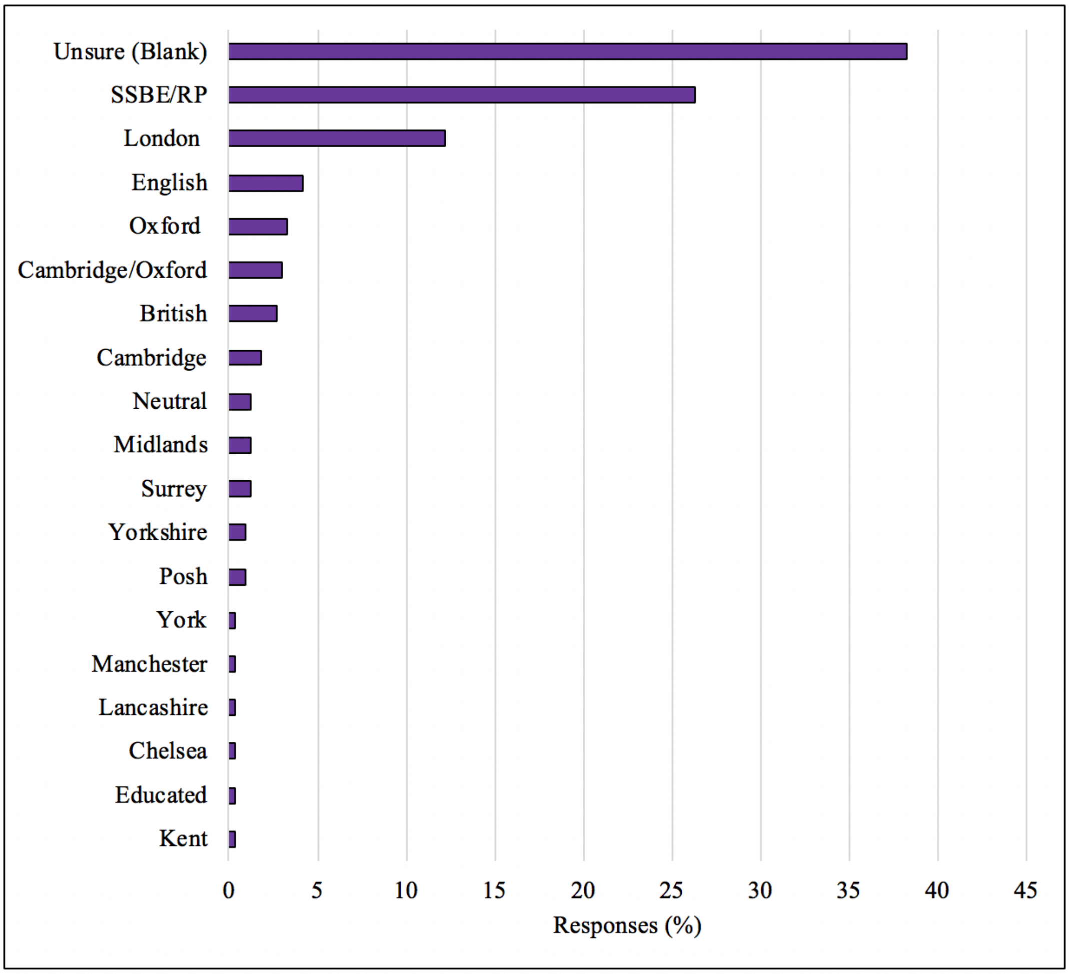


Figure 4. Percentage distribution of responses to the question "*What accent do you think this speaker has? Leave the box blank if you are unsure*" for the Standard Southern British English speakers.

phonetically very close to the standard model, they are not areas where the majority of speakers would have such an accent. The descriptors "York" and "Yorkshire" could also be attributable to the fact that participants in the experiment were students at the University of York, an institution attracting significant numbers of students – many of whom have SSBE accents – from southern England and/or from affluent middle-class backgrounds. For a northern English city, York and its surrounding area is home to an unusually high proportion of university graduates and people in professional occupations, and otherwise has a demographic profile that is markedly different from those of other urban areas of Yorkshire (Dorling, 2010). These factors mean that students have numerous opportunities to be exposed to SSBE accents within their university city.

Figure 5 shows the responses provided by listeners when they were asked to describe the accent of the Northern Irish speakers in the experiment.

In contrast to the SSBE accent description, the "Unsure (Blank)" classification was not the most popular label provided by listeners for the Northern Irish-accented speakers. There was a much greater proportion of "Irish" labels compared with the number of "Northern Irish" and "Southern Irish" labels. This suggests that many speakers either could not,

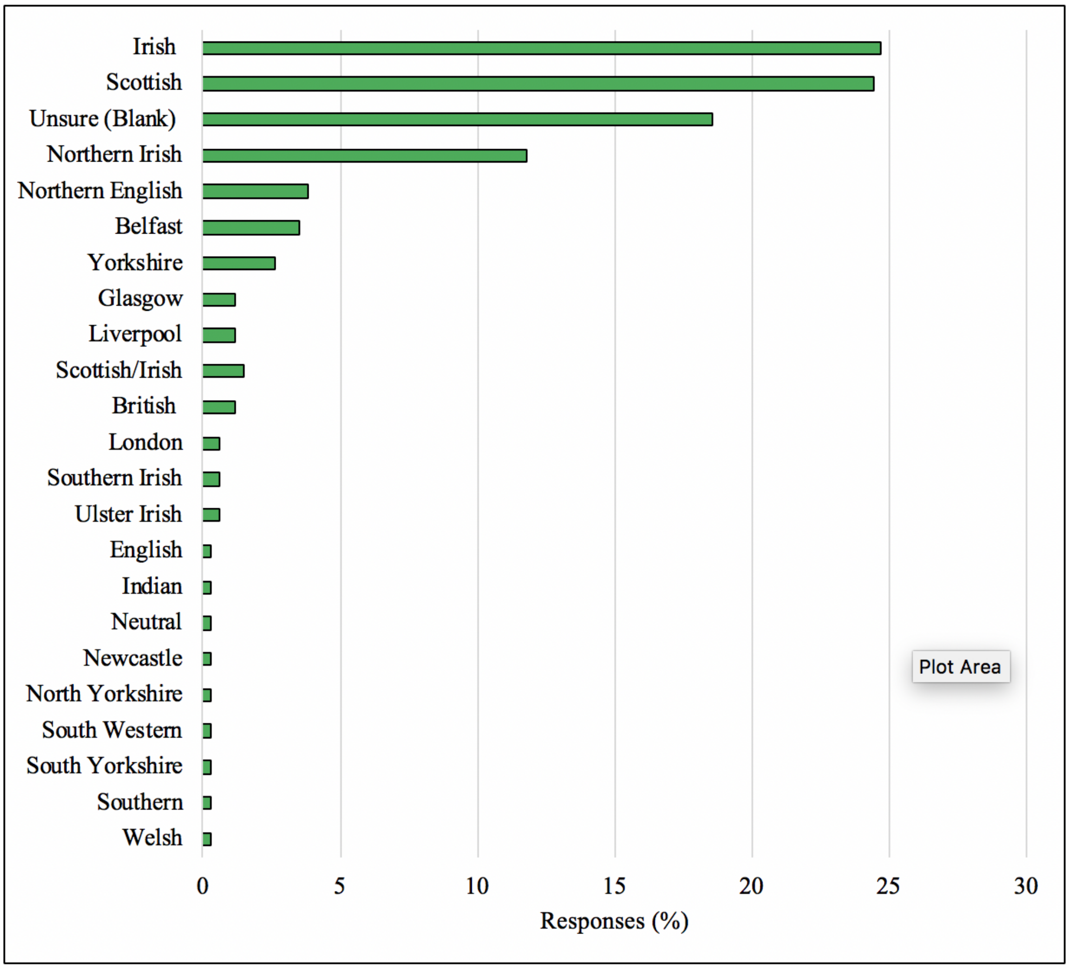


Figure 5. Percentage distribution of responses to the question "'*What accent do you think this speaker has? Leave the box blank if you are unsure*" for the Northern Irish speakers.

or were unwilling to, determine the speaker's accent more precisely than to say he/she had an Irish accent. They may alternatively have thought the "Northern" qualifier to be superfluous, just as listeners from outside England might not think it necessary to specify whether an evidently English speaker is from the north or south of England. The results in Figure 5 also show that there appears to be confusion between listeners' perceptions of Northern Irish and Scottish accents. The Northern Irish speakers in the sample were frequently reported to have a Scottish accent, which was either indicated using a generic "Scottish" label or a more specific label such as "Glasgow". Given that the trend in the data was for listeners to say that their own accent was dissimilar to both Northern Irish English (mean similarity score to *Belfast* = 4.6/100) and Scottish English (mean similarity score to *Glasgow* = 5.2/100), the confusion is perhaps explainable by the relative lack of perceived similarity to and/or familiarity with, the target varieties.

Subsequent analysis was conducted to assess whether the confusion between the Northern Irish and Scottish accents was either speaker-specific, or listener-specific, or both. Figure 6 shows the number of 'Scottish' and 'Irish' labels assigned to each of the four Northern Irish speakers in the sample. For the purposes of this analysis, labels were grouped so that the "Northern Irish", "Southern Irish", "Ulster Irish", "Irish" and "Belfast" descriptors were all grouped into the 'Irish' category, while the "Scottish" and "Glasgow" labels were grouped into the 'Scottish' category. Figure 6 shows that while the proportions of 'Scottish' and 'Irish' classifications were not the same for each speaker, no single speaker was consistently misidentified as sounding particularly Scottish by the listener group. This suggests that the confusion between the two accents seen in Figure 5 was not the consequence of one or two speakers in the study being frequently mistaken for Scottish speakers, but rather that the misidentification applied across all the speakers in the study.

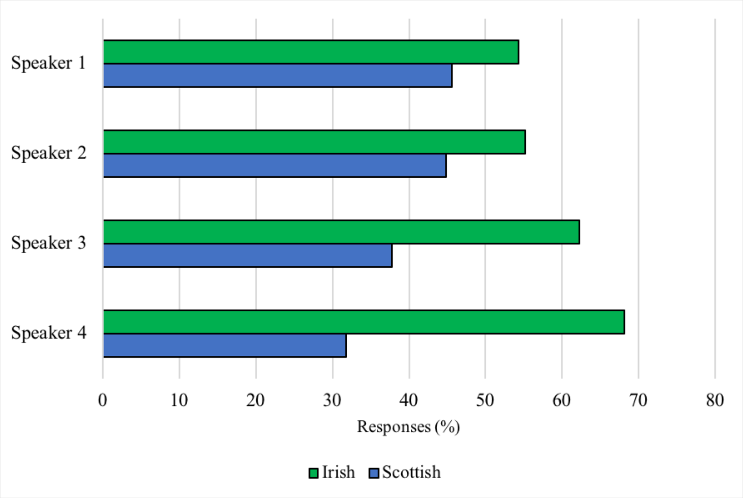


Figure 6. Percentage of Scottish and Irish accent labels assigned to each of the four Northern Irish speakers in the sample.

Given that the confusion of the Northern Irish and Scottish accents was not closely associated with any particular speaker, analysis was also conducted to assess how good individual listeners were at attributing the relevant accent labels correctly. The responses of each individual listener were assessed, with a count taken for the number of 'Irish' labels assigned to the Northern Irish voices. These results are shown in Table 2. Due to the automatic question randomisation process, results are displayed as percentages, as different listeners heard different numbers of the Northern Irish recordings (range = 1-8; mean = 3.87). Listeners were grouped according to the extent to which they assigned an 'Irish' label to the voices of the Northern Irish speakers (in percent).

Percentage of 'Irish' attributions	Number of listeners
0-20	33
21-40	12
41-60	9
61-80	9
81-100	20

Table 2. Percentages for the number of listeners who provided 'Irish' labels for the Northern Irish speakers' accents.

The data in Table 2 show that 20 listeners classified the Northern Irish accent using ‘Irish’ labels between 81 and 100% of the time. Conversely, 33 listeners classified the Northern Irish accent using ‘Irish’ labels between 0 and 20% of the time. This shows that the majority of listeners within the sample performed either very accurately or very inaccurately when assigning accent labels to the Northern Irish voices, and it suggests that labelling inaccuracies within the data shown in Figure 5 were the result of some listeners being consistently unable to provide a correct label.

The third accent included in this experiment was Middle Eastern-accented English. Figure 7 shows the accent labels provided for the Middle Eastern speaker samples. Given the large number of accent labels used to describe the Middle Eastern speakers’ voices, Figure 7 excludes labels which were used on just one occasion. These excluded labels were *African, American, British Arabic, Central European, Central Asian, Korean, Automated, Greek, Hispanic, South American, Leeds, Northern British, Malaysian, Non-regional, Welsh, Swedish, Scandinavian, Scottish, South Africa, South East, Turkish*, and *Thai.*

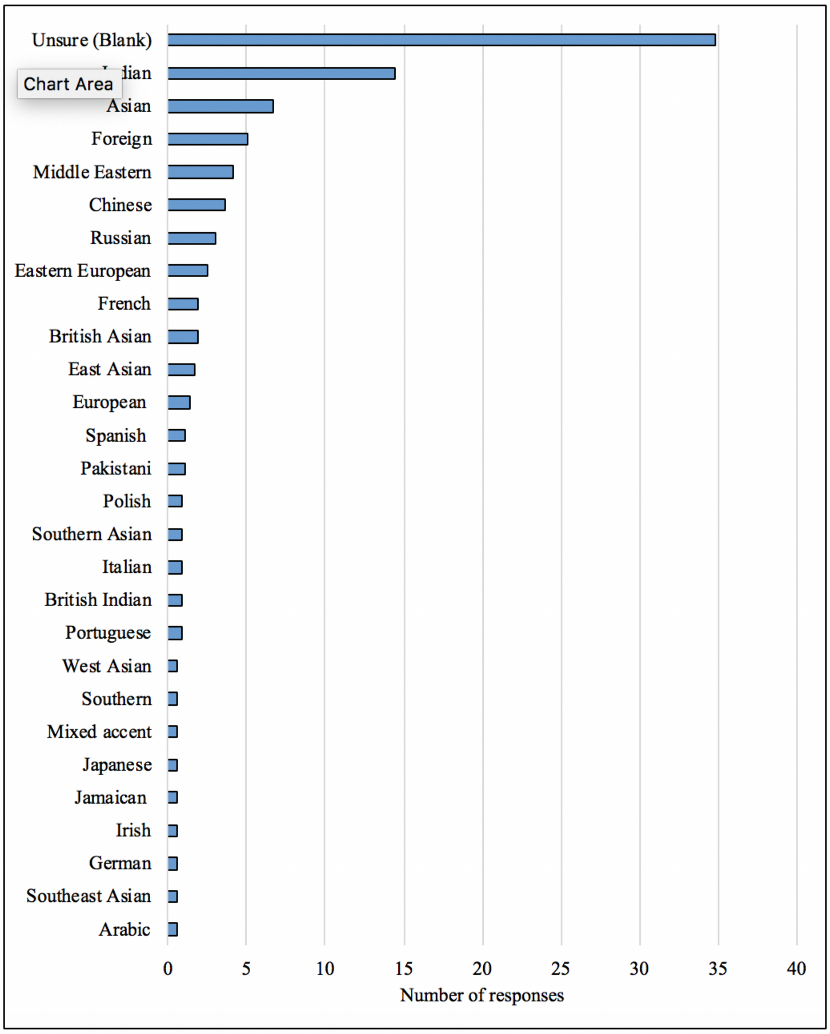


Figure 7. Percentage distribution of responses to the question “’*What accent do you think this speaker has? Leave the box blank if you are unsure*” for the Middle Eastern speakers.

Figure 7 shows that, like the SSBE accent, the most popular accent label assigned to the Middle Eastern speakers was "'*Unsure (Blank)*". There was also a greater number of different labels assigned to the Middle Eastern speakers (n=40) than to the SSBE speakers (n=19) or the Northern Irish speakers (n=23), suggesting a greater level of inconsistency among listeners when assigning accent labels to the foreign accent compared to the British accents in the experiment. The overwhelming majority of responses named a non-British location for the accent of the Middle Eastern speakers in this experiment, but there was a high level of inconsistency within the labels assigned, which made reference to 35 different countries spanning five continents. Additionally, relatively few responses (n=17) pinpointed the Middle Eastern speakers' accents as having any Arabic, Persian or Middle Eastern origin, with *Indian, Asian* and *foreign* the most commonly assigned labels. While the *foreign* descriptor is non-specific, it can be considered an accurate description in so far as listeners were able to say that the speakers in the Middle Eastern recordings were not native British English speakers. It may also not be unreasonable to suggest that Asian is a relatively accurate accent label for the Middle Eastern speakers, given the proximity of parts of the Middle East to the Asian subcontinent[2]. Again, however, the descriptor is relatively broad and arguably would be of rather limited use within a forensic investigation.

Conclusion

The goal of the research presented in this paper was to assess how accurately a group of listeners could perceive different aspects of a speaker's voice within an experimental setting, with a view to evaluating the usefulness of such a practice in certain forensic contexts. With respect to assessments of vocal pitch in line with measured average F0, the analysis showed small to medium-sized correlations in the data between median F0 and listeners' judgements of how high-pitched speakers' voices were. The coefficients improved when standardised pitch scores were used, and when other acoustic measurements relating to pitch and voice quality were included alongside average F0 measurements. However, the best-performing model – that for male speakers – still only accounted for 20% of the total variation present. The analysis also illustrated how some listeners within the sample seemed unable to correctly appraise the relative differences between three speakers' voices with average median F0 values of 99Hz (Speaker 1), 120Hz (Speaker 2) and 140Hz (Speaker 3). Of the 26 listeners who evaluated these speakers' voices, only 14 assigned relative pitch judgements in accordance with the increase in average F0 across the three speakers' voices. This suggests that some listeners lack the ability to reliably judge how high-pitched a speaker's voice is, while some listeners are able to accurately estimate vocal pitch in line with measured acoustic correlates.

As for the description of accents, the analysis suggests that listeners' abilities to describe accents decreases as the degree of unfamiliarity or geographical distance increases. There were relatively few inaccurate labels used to describe the accents of the SSBE speakers, with a higher number of confusions shown when listeners were asked to describe the Northern Irish speakers' accents, and further confusion when listeners were asked to describe the accents of the Middle Eastern speakers. Given the trend in the data for speakers to identify the SSBE accents as being more similar to their own in comparison to the Northern Irish English, these data would support the idea that the more geographically distant or unfamiliar an accent is, the greater the scope for confusion or otherwise inaccurate accent labelling (the L1 accents of English spoken in Australia and

New Zealand are obvious likely counterexamples to this generalisation, but they do not invalidate the general ‘proximity effect’ patterns observed in numerous perceptual dialectology studies. See e.g. Montgomery, 2015; Shen and Watt, 2015; Preston, 2018. Had this study tested listeners from Glasgow, Edinburgh or Belfast, then the misclassification of the Northern Irish-accented speakers as Scottish speakers would not have been expected as the listeners would have presumably been more familiar with the tested varieties.

Additionally, given the wide variety of answers provided by listeners in this study when they were asked to describe the accents of the Middle Eastern speakers, the data lend support to the view that there is limited value in asking phonetically untrained non-native listeners to assess the geographical provenance/nationalities of speakers based on vocal information alone. As listeners rarely assigned a “British” label to the Middle Eastern accents, it could be argued that while listeners were adequately equipped to assess whether a speaker was a native or non-native speaker of English, any information beyond this was unreliable. This generalisation is especially important in view of the fact that the NCTSO bomb threat checklist encourages users to give an opinion concerning a speaker’s possible nationality. With respect to asking untrained listeners to determine the accent of a given speaker, the data in this study suggest that there may be some usefulness in asking this question. The results nevertheless also urge caution, owing to the poor performance of some listeners in the Northern Irish accent classification task, and suggest that factors such as the background of the listener and their general accent classification ability should also be considered.

It can also be contended that the use of information about a speaker’s accent obtained through asking non-linguist earwitnesses to describe the voice of a given speaker should also be used in conjunction with the knowledge that not all accents have a well-defined corresponding geographical location. For example, the spread of geographical locations that listeners associated with the SSBE speakers in this study spanned much of the south of England, and yet it cannot be considered ‘inaccurate’ to suggest that SSBE speakers could come from any of those places. We argue, therefore, that the use of speech-based evidence in the form of phonetically untrained listeners’ descriptions of voices and accents should be treated with due scepticism by default, and that such information should be used in conjunction with empirically verified data about UK and international varieties of English that have been collected by professional linguists.

One possible improvement to the practice of eliciting information from earwitnesses would be the development of a set of materials designed to test a listener’s abilities to identify various aspects of speakers’ voices. Given that the evidence recorded in documents such as the NCTSO bomb threat checklist would, in many cases, be based only on the perceptions of a single listener, it would potentially be useful to assess the capability of that earwitness to make reliable observations of different aspects of speaker’s voices. This would allow the police and other investigative agencies to verify whether the checklist user can consistently and accurately identify different aspects of voice before any use is made – either in court or for the purposes of further investigative work – of subsequent checklist evidence he or she might produce (cf. the recommendations laid out in Nolan (2003) concerning testing of earwitness reliability using the voice parade paradigm). However, such a recommendation would require more research to be implemented in practise, specifically regarding the finer details of how such a test could be standardised and implemented by those working in the criminal justice sphere.

There is plentiful scope for expansion of the design of this study in future work, which could focus on other aspects of voice such as speech rate, variation in the F0 contour as a cue to how 'monotonous' or 'lively' a speaker's utterances are perceived to be, nasality, disfluency features (e.g. hesitations, filled pauses, etc.), and the use of paralinguistic markers such as clicks. As we referred to earlier, it is also acknowledged that the experimental conditions in this study created a more favourable earwitness environment than would be expected in certain real-world scenarios, such as the handling of bomb threats in emergency service control rooms, hospitals or schools. However, the aim of the work in this study has been to generate empirical data as a basis upon which to make recommendations about how earwitness evidence can be better collected and later deployed by those tasked with gathering such information. It is hoped that this approach could be helpful in guarding against the use of erroneous, redundant, vague or otherwise low-value earwitness testimony in the sphere of criminal investigation. At the very least, we hope that the availability of systematically-collected data of the sort described above will serve to encourage more discriminating, better-informed evaluations of the utility of earwitnesses' voice descriptions on the part of members of the law enforcement and intelligence communities.

Notes

[1] Arabic and Persian are of course languages with highly distinct phonologies, but we take the view that in the present context the differences in the way these participants speak English are not large enough to create significant disparities in terms of the listeners' evaluations of the speakers' accents

[2] The term Asian in the UK tends to be used to denote people with origins in the countries of the Asian subcontinent – chiefly India, Pakistan, and Bangladesh – rather than people of East Asian ancestry (China, Korea, Japan, Vietnam, etc.). We recognise also that the Middle Eastern countries, including those of the Arabian Peninsula, are conventionally said to be part of the continent of Asia.

References

Altendorf, U. (2011). *Estuary English: Levelling at the Interface of RP and South-Eastern British English.* Tübingen: Gunter Narr Verlag.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

BBC News, (2018). 'Ex-soldier' raiding Home Counties houses at gunpoint.

Boersma, P. and Weenink, D. (2016). Praat: Doing Phonetics by Computer.

Broeders, A. and van Amelsvoort, A. (2001). A practical approach to forensic earwitness identification: constructing a voice line-up. *Problems of Forensic Sciences*, 47, 237–245.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155.

Dorling, D. (2010). Persistent North-South divides. In N. M. Coe and A. Jones, Eds., *The Economic Geography of the UK.* London: Sage, 12–28.

Griffiths, M. (2012). "'Did he have an accent?': Forensic speaker descriptions of unknown voices". In S. Tomblin, N. MacLeod, R. Sousa-Silva and M. Coulthard, Eds., *Proceedings of the International Association of Forensic Linguists' Tenth Biennial Conference*, 245–261, Birmingham, UK: Aston University Centre for Forensic Linguistics (Online).

Handkins, R. E. and Cross, J. F. (1985). Can a voice lineup be too fair? In *Paper presented at Symposium on Forensic and Scientific Issues in Voice Recognition at the Meeting of the Midwestern Psychological Association*, Chicago, IL.

Hudson, T., De Jong, G., McDougall, K., Harrison, P. and Nolan, F. (2007). F0 statistics for 100 young male speakers of Standard Southern British English. In *Proceedings of the 16th international congress of phonetic sciences (Vol. 6, No. 10)*, Saarbrücken, Germany.

Hughes, A., Trudgill, P. and Watt, D. (2012). *English Accents and Dialects: An Introduction to Social and Regional Varieties of English in the British Isles.* London: Hodder Education, 5th ed.

Künzel, H. J. (2001). 'Beware of the "telephone effect": the influence of telephone transmission on the measurement of formant frequencies'. *International Journal of Speech, Language and the Law*, 8(1), 80–99.

Laver, J. (1994). *Principles of Phonetics.* Cambridge: Cambridge University Press.

Montgomery, C. (2015). Borders and boundaries in the north of England. In R. Hickey, Ed., *Researching Northern English.* Amsterdam: John Benjamins, 345–368.

Mugglestone, L. (2003). *'Talking Proper': The Rise of Accent as a Social Symbol.* Oxford: Oxford University Press, 2 ed.

National Counter Terrorism Security Office, (2016). Bomb threat checklist.

Nolan, F. (2003). A recent voice parade. *International Journal of Speech Language and the Law*, 10(2), 277–291.

Nolan, F. and Grabe, E. (1996). Preparing a voice lineup. *International Journal of Speech Language and the Law*, 3(1), 74–94.

Nolan, F., McDougall, K. and Hudson, T. (2013). Effects of the telephone on perceived voice similarity: implications for voice line-ups. *International Journal of Speech, Language & the Law*, 20(2), 229–246.

Preston, D. C. (2018). Perceptual dialectology. In C. Boberg, J. Nerbonne and D. Watt, Eds., *The Handbook of Dialectology.* Oxford: Wiley-Blackwell, 173–207.

Puts, D. A., Gaulin, S. J. C. and Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior*, 27, 283–296.

Shen, C. and Watt, D. (2015). Accent categorisation by lay listeners: Which type of "native ear" works better? *York Papers in Linguistics Series*, 2(14), 106–131.

Sherrin, C. (2015). Earwitness evidence: the reliability of voice identifications.

Shuy, R. W. (1993). *Language Crimes: The Use and Abuse of Language Evidence in the Courtroom.* Oxford: Blackwell.

Watt, D. (2010). The identification of the individual through speech. In C. Llamas and D. Watt, Eds., *Language and Identities.* Edinburgh: Edinburgh University Press, 76–85.

Watt, D. and Burns, J. (2012). Verbal descriptions of voice quality differences among untrained listeners. *York Papers in Linguistics*, Series 2, 1–28.

Xu, Y. (2013). ProsodyPro - A Tool for Large-scale Systematic Prosody Analysis. In *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*, 7–10, Aix-en-Provence, France.

Yarmey, A. D. (2001). Earwitness descriptions and speaker identification. *International Journal of Speech Language and the Law*, 8(1), 113–122.