Earwitness evidence and the question of voice similarity

KIRSTY McDOUGALL

Consider the following scenario from a real criminal case. A person's house is broken into while the occupant is at home. The intruder is wielding a weapon and wearing a balaclava which conceals his face. The intruder remains in the house for over an hour threatening the occupant, then escapes, having stolen some of the occupant's possessions. The stolen property is later recovered and a suspect identified. The occupant believes that he can remember the voice of the assailant and agrees to participate in a 'voice parade'.

A voice parade can be used to provide 'earwitness' evidence in cases where a voice was heard at the scene of a crime, but the voice was not recorded. Analogous to a visual identity parade, in a voice parade the witness is asked whether he or she can pick out the voice of the speaker heard at the crime scene from a line-up of recordings which includes the suspect's voice and a number of foil voices.

In the United Kingdom, specifically in England, voice parades have been used to collect earwitness evidence in several cases in recent years. In current practice, voice parades are constructed according to the guidelines published in 2003 in the Home Office Circular 'Advice on the use of Voice Identification Parades', prepared by DS (now DCI) John McFarlane of the Metropolitan Police and Professor Francis Nolan, a phonetician at the University of Cambridge. The procedure recommended in the guidelines was developed by extending the existing police procedure for visual identity parades to the aural domain, taking into account findings from the available literature on earwitness performance. This procedure has been successfully implemented on a number of occasions, but it is still evolving as technology improves and research develops.

A voice parade constructed according to the guidelines consists of nine voice samples played to the witness via a video or *PowerPoint* presentation which displays the number of the sample while the audio file is playing. Each voice sample is prepared by a phonetician and contains a 45-60 second 'collage' of short utterances of spontaneous speech representative of the speaker. The utterances are spliced together in a randomised order to give an overall impression of the speaker's voice without the distraction of the topic of discussion or any sense of narrative.

Fairness of the voice parade dictates that the suspect and foil samples are as comparable as possible. Choosing the voices to serve as foils is one of the most difficult aspects of constructing a voice parade as it is not well understood what makes voices sound similar. Whereas in the visual domain the foils for an identity parade may be selected on the basis of a description such as "short black hair, beard and glasses", there is no straightforward equivalent auditory profile available for describing and comparing voices. My research has therefore been tackling the question of voice similarity and its phonetic description. I am interested in determining why listeners perceive some voices to sound more similar to each other than others, and using this knowledge to develop a phonetically principled technique for selecting foil voices for a voice parade. During my British Academy Postdoctoral Fellowship, I have been carrying out an experimental study of the phonetic underpinnings of perceived voice similarity,[1] considering the roles played by aspects of speech such as pitch, resonances, voice quality and speaking rate.

When a listener judges two voices as sounding similar, there are two sources of similarity contributing to this judgement, linguistic factors and personal factors. *Linguistic*

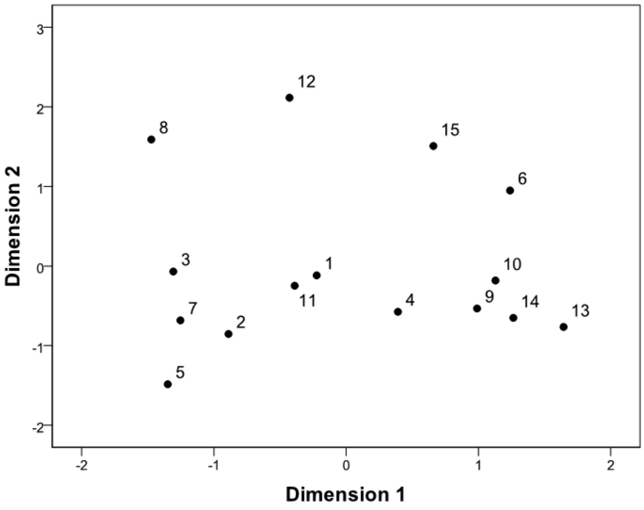


Figure 1. *Plot of the first two dimensions (of five) produced by the Multidimensional Scaling analysis, showing perceived distances among the 15 speakers, labelled 1-15. The closer together the datapoints of a given pair of speakers, the more similar-sounding the pair were judged by the listeners. For example, speakers 1 and 11 were judged as sounding very similar to each other, while speakers 8 and 13 were judged rather different-sounding.*

[1] This work has grown out of an earlier project carried out in collaboration with Francis Nolan and Toby Hudson, 'Voice similarity and the effect of the telephone: a study of the implications for earwitness evidence' (*VoiceSim*) at the University of Cambridge, ESRC Award no. RES-000-22-2582. The author is grateful to Francis Nolan and Toby Hudson for their ongoing interest in the present study and to Geoffrey Potter for technical assistance.

factors include the language or dialect spoken and the accent used. For example, two speakers of Australian English may sound more similar to each other than an Australian English speaker and a British English speaker (all other things being equal). *Personal* factors relate to the speaker's anatomy and physiology and his or her individual way of using them to produce speech. For example, two speakers with large vocal tracts and hence deep voices may sound more similar to each other than to a speaker with a smaller vocal tract and a higher-sounding voice (all other things being equal). In order to probe the notion of voice similarity, it is necessary to tease apart these two underlying components, the linguistic and the personal, and examine their workings – as well as how the two interrelate. In the present study, linguistic factors are held constant and personal factors investigated, by focusing on judgements of the similarity of voices within a group of speakers of the same accent, age and sex. This study was made possible by the recent appearance of the *DyViS* database,[2] a forensic phonetic database of speech recordings of 100 speakers matched for accent (Standard Southern British English, also known as 'modern Received Pronunciation'), age (18-25 years), and sex (male). By presenting listeners with voices of the same demographic profile, judgements can thus be made about the extent of similarity or difference among the voices specifically due to individual variation within the group.

The experiment

Fifteen speakers were chosen from the *DyViS* database, whose recordings were used to construct the voice stimuli for presentation to listeners. Two short samples of spontaneous speech were chosen from each of the test speakers. The speech samples were paired in every possible combination, so that each speaker could be compared with all other speakers, including himself. A group of 20 listeners (10 male, 10 female, all native British English speakers) were recruited and given the task of judging the similarity of the two voices in each stimulus pair on a distance scale from 1 (very similar) to 9 (very different).

The idea behind the study was to investigate whether the listener judgements of voice similarity correlated with a number of phonetic properties of the test voices, in particular pitch, speaking rate, resonance features and voice quality. The voice similarity judgements were therefore subjected to a data reduction technique called Multidimensional Scaling which characterised each speaker in a set of five perceptual dimensions, creating a perceptual space in which the voice similarity relationships among all pairs of speakers could be interpreted. A plot of the 15 speakers' locations on the first two of these dimensions is given in Figure 1. Measures of the phonetic features of interest for each of the 15 speakers could then be tested for their extent of correlation with these perceptual dimensions.

The four types of phonetic feature investigated are explained below with reference to the computer-generated acoustic display of the speech signal, a spectrogram and pitch trace, shown in Figure 2.

1. Pitch

Pitch, how high or low a voice sounds, is intuitively likely to be important in listeners' judgements of the extent to which two voices sound similar. Pitch corresponds to the

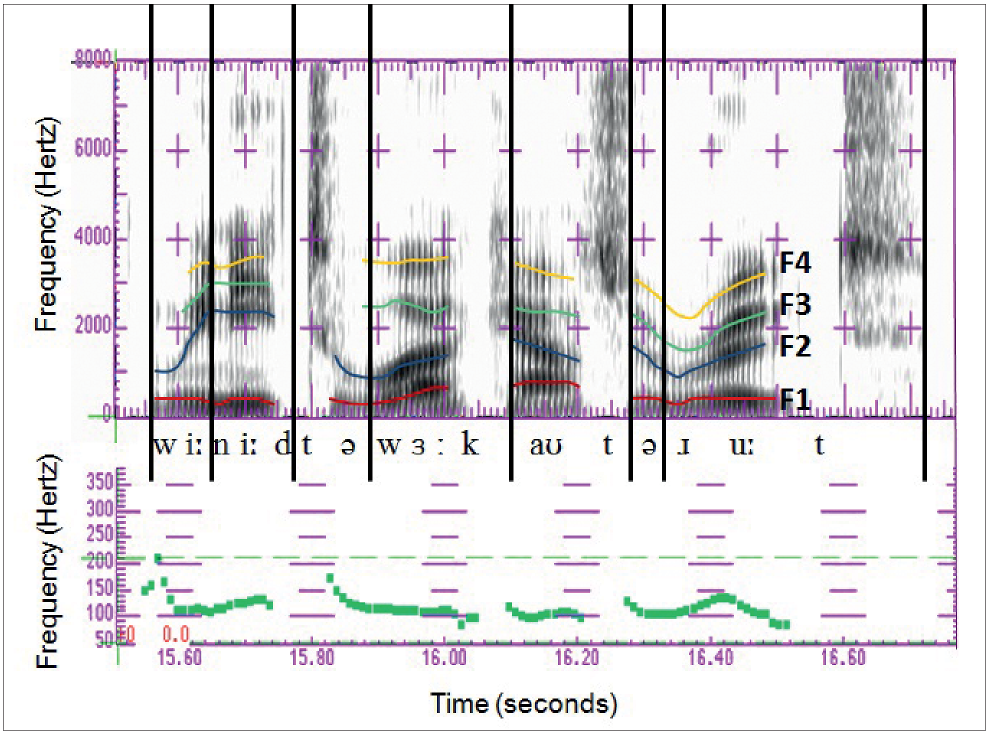


Figure 2. *Spectrogram (upper panel) and pitch trace (lower panel) of a male speaker of Standard Southern British English producing the utterance 'We need to work out a route'. A transcription of the component sounds of the utterance is given below the spectrogram using the International Phonetic Alphabet. The first four formant frequencies are labelled F1, F2, F3 and F4, and shown in red, blue, green and yellow respectively. Approximate syllable boundaries are indicated by vertical lines.*

[2] F. Nolan, K. McDougall, G. de Jong and T. Hudson, 'The *DyViS* database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research', *International Journal of Speech, Language and the Law* 16:1 (2009), 31-57. The *DyViS* Database is available via the UK Economic and Social Data Service, www.esds.ac.uk/findingData/snDescription.asp?sn=6790.

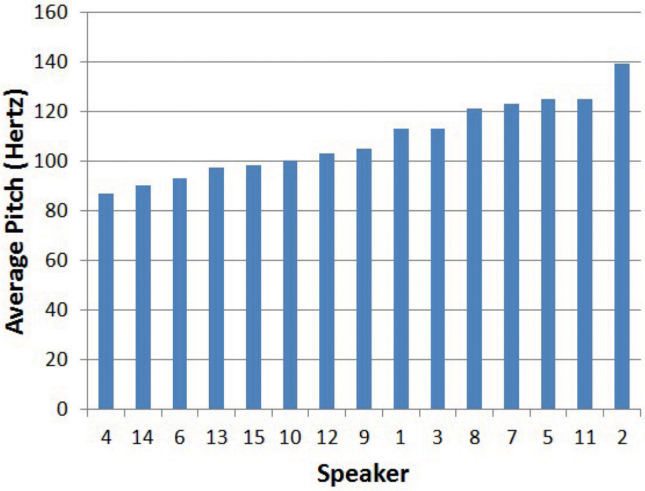


Figure 3. *Average pitch values in Hertz for each of the 15 test speakers, ordered from lowest to highest.*

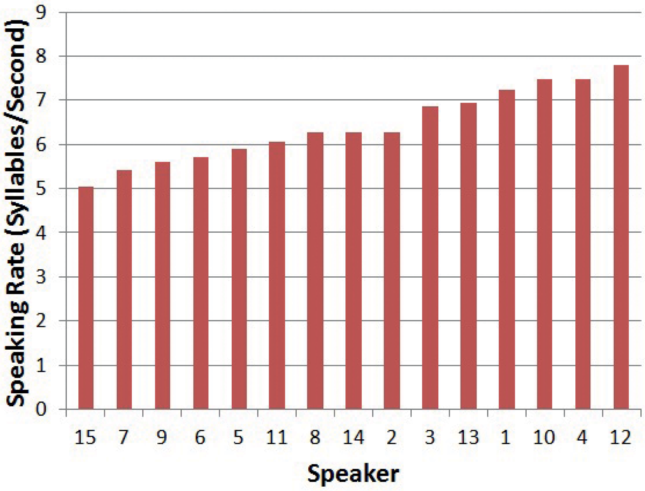


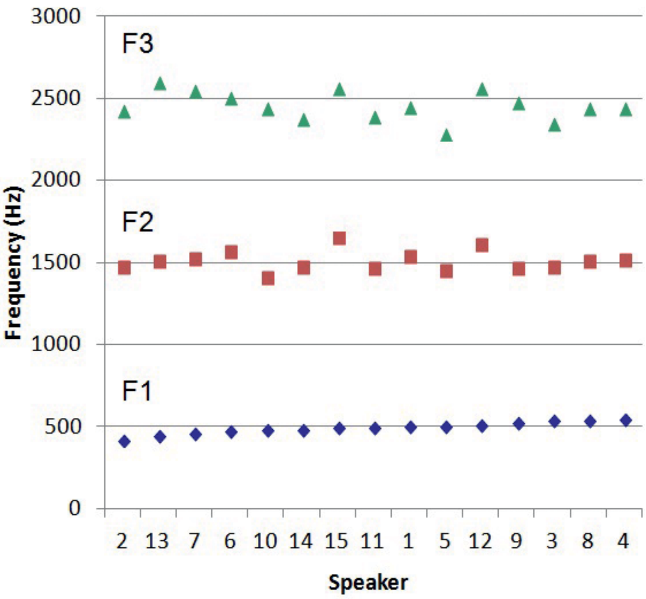
Figure 4. *Average speaking rate in syllables per second for each of the 15 test speakers, ordered from slowest to fastest.*

rate of vibration of the vocal folds, and is measured by phoneticians using the fundamental frequency trace as shown in the lower panel of Figure 2. The average pitch for each of the 15 test speakers is given in Figure 3.

2. Speaking rate

Speaking rate, how fast or slow a person speaks, can be measured by calculating the average number of syllables he or she articulates per second, and also intuitively might be considered relevant to voice similarity. Syllable durations can be seen on the spectrogram in Figure 2. The average speaking rate for each of the 15 test speakers is given in Figure 4.[3]

Figure 5. *Average Long-Term Formant measures of F1, F2 and F3 frequencies for each of the 15 test speakers, ordered by F1.*



3. Resonances

Further phonetic features contributing to the impression a voice makes are its resonances or ‘formant frequencies’, that is, the frequencies at which vibrations of air are at maximum amplitude in the vocal tract during the production of vowels and certain consonants. Formant frequencies appear as dark, roughly horizontal bands on a spectrogram and vary over time as shown in Figure 2. As well as determining the quality of different speech sounds, the patterns of these formant frequencies vary between speakers depending on the shape and dimensions of their individual vocal tracts. Formant frequencies are therefore also of interest when investigating the perception of similarities and differences among individual voices. For the present study, an average measure of each speaker’s first three formant frequencies was calculated using a technique called Long-Term Formant analysis. These values are shown in the graph in Figure 5.[4]

4. Voice quality

Another feature likely to be important in judgements of voice similarity is voice quality, that is, elements of the timbre of the voice such as breathy, creaky, nasalised, falsetto, whispery, and so on. The seminal framework for describing voice quality is John Laver’s *The Phonetic Description of Voice Quality* (CUP, 1980). For the present study, voice quality features of the 15 speakers were analysed using a system involving 33 voice quality settings derived from Laver’s framework.[5]

[3] The author is grateful to Erica Gold, University of York, for making the measurements of speaking rate used in this study.

[4] The author is grateful to Philip Harrison and Christin Kirchhübel, University of York, for their assistance with this part of the study.

[5] The author is grateful to Louisa Stevens and Peter French, University of York, for their collaboration on the part of this research relating to voice quality.

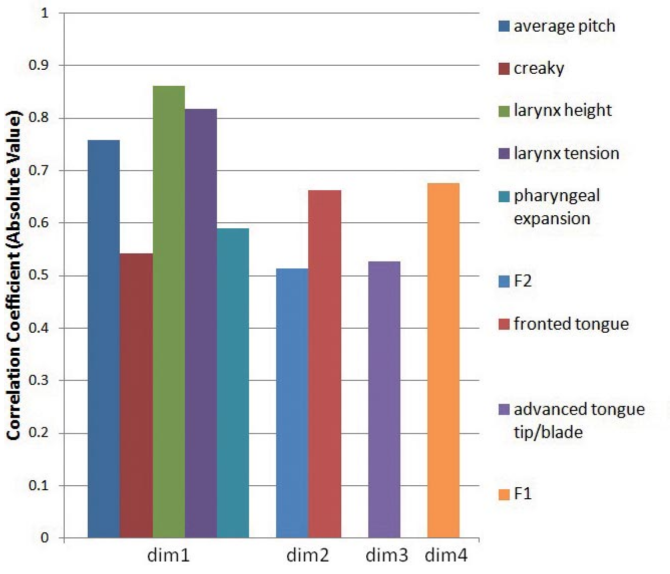


Figure 6. *Significant correlations ($p < 0.05$) between the phonetic features tested and the five perceptual dimensions from the voice similarity judgements, Dimensions 1 to 5, labelled dim1, dim2, ... Note that there were no significant correlations with Dimension 5.*

Linking voice similarity to phonetic features

The extent of correlation between the five perceptual dimensions from the voice similarity judgements and the set of phonetic features measured was tested using Spearman correlation. The results are given in Figure 6 which shows each phonetic feature that achieved a significant correlation with a given perceptual dimension. For this group of Standard Southern British English speakers, the features most important for voice similarity are those correlating with Dimension 1, namely average pitch and several voice quality features linked to the larynx and pharynx. Since pitch itself is a product of larynx behaviour (vocal fold vibration), these results suggest that one major dimension of voice similarity centres on the larynx/pharynx. Dimensions 2, 3 and 4 yield significant correlations with the F2 frequency, voice quality features related to tongue movement and the F1 frequency. Since F1 and F2 are related to tongue posture, as are the significant voice quality features here, a front/back-of-the-mouth factor could more tentatively be posited for Dimensions 2/3/4. It is interesting to note that speaking rate did not achieve a significant correlation with any of the perceptual dimensions. This could perhaps be due to the short duration (3 seconds) of the voice samples the listeners compared in the experiment, which may have been insufficient for speaking rate to establish a clear role in the judgements of voice similarity.

Future directions

The experimental work presented here focused on a single accent, Standard Southern British English. However, investigating the role of accent differences is also crucial in improving our understanding of the perception of voice similarity. Experiments of a similar kind are needed to establish whether the roles for pitch, voice quality features and formant frequencies found here for Standard Southern British English apply in other accents. Investigation of the effect of including more than one accent among the voices to be compared is also required. For instance (extending from the examples initially given to illustrate linguistic versus personal factors), how would an Australian English speaker with a large vocal tract, an Australian English speaker with a smaller vocal tract and a British English speaker with a large vocal tract compare in terms of similarity? Further, research into the role of the accent background of the listener is needed – if the listener judging voice similarity speaks with a different accent from the speakers being compared, how does this impact on the similarity judgements? Does a listener whose accent is more different from that of the speakers under comparison make very different judgments of voice similarity from a listener whose accent is close to the speakers'? I am grateful to the British Academy for the recent award of a Small Research Grant which will enable me to pursue these questions further, initially through collection and analysis of a new database of York English.

As well as its importance for phonetic theory, an improved understanding of voice similarity will have a crucial practical impact on forensic phonetic casework. Developing a more comprehensive model of voice similarity will help us better understand how earwitness speaker identification works, and help us to construct better, more scientifically informed, voice parades.

Dr Kirsty McDougall is a British Academy Postdoctoral Fellow in the Department of Theoretical and Applied Linguistics, University of Cambridge.