SOME ACOUSTIC CORRELATES OF PERCEIVED (DIS) SIMILARITY BETWEEN SAME-ACCENT VOICES

Francis Nolan, Kirsty McDougall & Toby Hudson

Department of Theoretical and Applied Linguistics, University of Cambridge, UK
fjn1@cam.ac.uk; kem37@cam.ac.uk; toh22@cam.ac.uk

ABSTRACT

Subjects rated the (dis)similarity of paired voice samples on a nine-point scale. The short voice samples were taken from the *DyViS* database of young male speakers with ‘Standard Southern British’ pronunciation. Accent was thus controlled, and ratings can be presumed to tap perceived personal voice quality differences. Multidimensional scaling (MDS) was applied to the ratings to derive five pseudo-perceptual dimensions.

These were then correlated with measures of f0 and the first three formants. Significant correlations were found with all measures. The first MDS dimension correlated with f0, confirming f0’s key role in voice similarity, followed in order of importance by F3, F2, and F1.

Keywords: personal voice quality, voice similarity, acoustic correlates, voice parades

1. INTRODUCTION

A fair identity parade is one in which the suspect does not ‘stick out’. In constructing a voice parade, part of the role of the phonetician as forensic expert is to examine the voice samples which are candidates to serve as foils in the parade ‘to ensure that the accent, inflection, pitch, tone and speed of the speech used provides a fair example for comparison against the suspect’ [5] point 15. A quantitative method cf. [11] for testing the similarity of the suspect’s and foils’ voices with the help of mock witnesses has been successfully applied by the second author in the preparation of real-world voice parades; but voice (dis)similarity is not well understood in phonetic terms, and research into how the acoustic properties of speech contribute to the perceived similarity of voices is relatively sparse. As a result, phoneticians have yet to establish a framework for the description of voice similarity [8], and there is no established set of procedures for quantifying the degree of similarity between two speakers. Such a model would be central to the construction of fair voice parades, but also helpful in forensic speaker comparison cases with similar voices.

When experiments have required similar-sounding voices, researchers have tended to choose speakers on the basis of anecdotally reported similarity such as family members and/or speakers whose voices have been confused over the telephone e.g. [8, 12, 13]. One possible quantitative approach is suggested by [14], who chooses pairs of father and son, twins, or brothers whose long-term spectra have similar properties, although he acknowledges that speakers with similar-sounding voices may not necessarily furnish similar-looking spectrograms. [17] demonstrates some correlation of fundamental frequency (f0) and word duration with perceived talker similarity. However, this experiment examined a read single-word utterance only. The importance of f0 is confirmed indirectly by studies on imitation, e.g. [3]. [10] compares judgments of talker similarity made on natural utterances and sine-wave replicas of the same utterances which preserved broad patterns of formant dynamics but lacked other detail provided by the glottal source. It concludes that this formant dynamic information plays a major role in perception of voice similarity since listeners’ judgments were much the same for the natural and sine-wave conditions. However, this study is again limited to read speech, and the test speakers were heterogeneous: a group of only 10 included males and females and American and British English dialects.

The present study, which was part of the *VoiceSim* project [16], examines correlations between the listener-assessed similarity of a number of speakers and acoustic properties of their voices to determine which acoustic features are most important in the perception of voice similarity and to lay the foundations for the development of a framework for its description. In its methodology it has some aspects in common with [1], which reports highest correlations between two derived perceptual similarity dimensions and, from among a large pool of acoustic dimensions, f0 and F1 respectively (female speakers), and f0 and the mean difference between F4 and F5 (male speakers) – the latter surprising, perhaps, given the low energy of these formants and the difficulties inherent in their accurate estimation. Unlike the present experiment it used isolated vowels, from Canadian French (accent not specified). Crucially our experiment uses voices from a population homogeneous for accent, so that the perception of personal voice similarity can be studied untrammelled by linguistic variation.

2. METHOD

2.1. Stimuli

Fifteen male speakers of Standard Southern British English, aged 18-25, were selected from the *DyViS* database [2, 9] for construction of the stimuli. The speakers were selected on a random basis (using the random number generator at <*http://www.random.org*>); however, speakers whose voices sounded impressionistically relatively unusual (e.g. extremely high or low pitched) were excluded.

Recordings from *DyViS* Task 2, a telephone conversation recorded at both studio quality and at the remote end of a telephone landline, were used. For each speaker, two short audio clips (labelled 'utterance 1 (U1)') and 'utterance 2 (U2)') of approximately three seconds were selected from the recordings. All U1 speech pertained to the subject denying knowledge of a man named Robert Freeman; U2 speech involved the subject denying having been at the Yewtree Reservoir on Wednesday evening. Each speaker was matched with all other speakers and with himself to form 120 pairings. Each pairing of speakers was represented by a U1 sample and a U2 sample randomly assigned. The order in which the two utterances were presented was determined at random by *Praat*. The present study is drawn from a larger experiment in which judgments were made on both studio- and telephone-recorded utterance pairings; analysis of studio-only pairings is presented here.

The 'ExperimentMFC' (Multiple Forced Choice) facility in *Praat* was used to present the stimuli to listeners. The playlist for each set was generated in random order on each occasion by *Praat*, and the order in which the sets were presented was reversed for half of the listeners.

2.2. Listeners

Twenty listeners (10 male, 10 female), all native speakers of British English aged 17-42 years, were recruited to participate in the experiment. Listeners had no known speaking or hearing impairments and were paid for their time.

2.3. Procedure

The experiment was conducted in a silent room, the stimuli being played via headphones. Each listener was asked to compare the voice pairings and assess the degree of similarity of the voices in each pairing. The listeners were instructed to take into account voice quality and accent, but as far as possible to ignore the meaningful content of the speech. Each listener undertook a practice test to familiarise him- or herself with the experimental set-up. For each voice pairing, the question 'How similar are these voices?' was displayed on the screen with, below, buttons showing the numbers 1 (very similar) to 9 (very different) for the listener to click on in order to select his or her response and move to the next trial Before each pairing there was a silence of 1.5 seconds, and between the two speech samples in each pair there was a silence of 1 second. The listener was asked to give 'snap' reactions and not agonise over particular comparisons, but nevertheless the timing between the pairs was in his or her control.

2.4. Acoustic analysis

For each of the 15 test speakers the first to third formant frequencies of six tokens of the vowels /iː/, /æ/, /ɑː/, /ɔː/, /ʊ/ and /uː/ in /hVd/ contexts in read speech were measured, using the *DyViS* database Task 4 (see [7] for details of the Praat-based measurement and manual validation procedures). Formant means (F1 to F3) were calculated for each of a speaker's vowels, and a 'global' mean across the six vowels. Mean and mode fundamental frequency measurements were made for each speaker using both the stretch of speech from the experimental stimuli (approximately 6 seconds) and a longer stretch of speech from *DyViS* Task 1 (spontaneous interview speech) of 3-5 minutes (see [6] for further details of the measurement procedure). Only the f0 results from the stimuli are reported here since these results were very similar to those from the longer recordings, indicating that the f0 variation exhibited by the short samples can be assumed to be approximately representative for each speaker.

3. RESULTS

The similarity judgments on each pairing of voices were subjected to Multidimensional Scaling (MDS). This data reduction technique, widely used in psychological research, derives a small number of pseudo-perceptual dimensions which enable the perceived distance or similarity amongst all of the objects to be inferred [15]. The analysis with five perceptual dimensions was chosen (cf. Giguère's [4] p. 35 guideline thresholds for stress); this yielded a stress-value of 0.18596, RSQ = 0.16006. Each speaker was thus characterised by a set of five coordinates on five perceptual dimensions of the form (dim1, dim2, dim3, dim4, dim5).

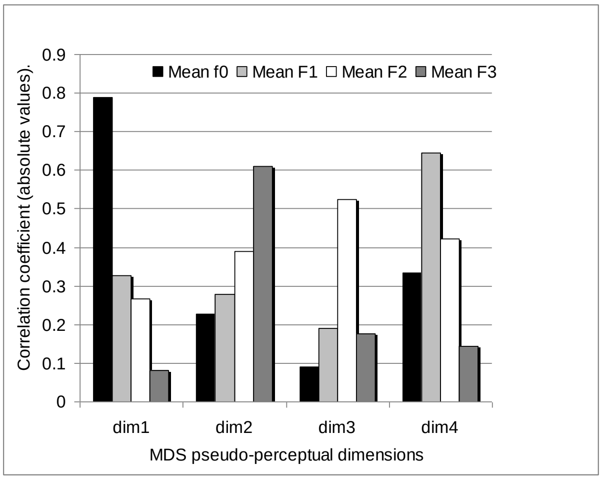
Correlations between the set of acoustic variables and the five perceptual dimensions were calculated using Pearson's formula, and are shown in Table 1.

Table 1: Correlations between the acoustic variables and MDS perceptual dimensions (Pearson's *r*). Significant ($p < 0.05$) correlation coefficients are in bold type. Mean F1/F2/F3 refers to the average of the respective formant across the six different vowels.

	dim1	dim2	dim3	dim4	dim5
Mean f0	**-0.783**	-0.270	0.186	-0.434	0.028
Mean F1	-0.327	0.278	-0.191	**0.646**	0.025
Mean F2	-0.267	0.390	**0.525**	0.423	0.234
Mean F3	-0.081	**0.611**	0.176	0.144	0.070
/iː/ F1	-0.191	-0.080	-0.258	**0.539**	0.033
/iː/ F2	-0.242	**0.553**	0.382	0.246	**0.009**
/iː/ F3	0.214	0.462	0.312	0.133	-0.066
/æ/ F1	-0.311	0.373	0.095	0.400	0.061
/æ/ F2	0.077	0.344	0.160	-0.484	-0.113
/æ/ F3	-0.266	0.406	0.158	0.148	0.004
/ɑː/ F1	-0.233	**0.534**	-0.281	0.292	0.011
/ɑː/ F2	-0.266	0.382	-0.196	0.304	-0.063
/ɑː/ F3	-0.091	0.357	-0.474	-0.359	0.060
/ɔː/ F1	-0.336	0.025	-0.231	0.504	-0.200
/ɔː/ F2	-0.179	-0.218	0.219	0.374	-0.137
/ɔː/ F3	0.048	**0.557**	0.264	0.276	0.038
/ʊ/ F1	-0.181	-0.092	-0.293	**0.581**	0.159
/ʊ/ F2	-0.218	0.125	0.494	0.465	0.380
/ʊ/ F3	-0.091	0.433	0.318	0.441	0.278
/uː/ F1	-0.036	0.114	-0.001	**0.590**	-0.040
/uː/ F2	-0.083	0.050	0.347	0.405	0.409
/uː/ F3	-0.042	**0.590**	0.487	0.201	0.029

Whilst the formants of individual vowels show sporadic correlations, the strongest and most interpretable correlations are with the means. Mean f0 will reflect laryngeal anatomy, and the formant means will reflect the individual's vocal tract and articulatory setting independent of specific vowel qualities. Figure 1 shows the correlation of these four acoustic variables (mean f0, and the global means of F1, F2 and F3) with each of the first four MDS dimensions. Unsurprisingly, f0 is dominant (intuitively, the 'pitch' of a voice is salient), while F3 may be relatively stable within a speaker, and reflect vocal tract size. F2 and F1, on the other hand, vary greatly with vowel quality, and correlate only with lower ranked MDS dimensions. F2 might, though, be (for instance) distinctively high in a speaker with a markedly palatalised articulatory setting.

Figure 1: Correlations between four acoustic dimensions and the first four MDS dimensions. The highest correlation with a dimension is significant ($p < 0.05$) in each case.



Correlations among the MDS dimensions and formant frequencies of the individual vowels are lower overall. Certain formants of some vowels, specifically F1 of /iː, ɑː, ʊ, uː/, F2 of /iː/, and F3 of /ɔː, uː/ bear some relationship with the second and fourth perceptual dimensions. However, the complex non-linear dependency of formants on vocal tract configurations and dimensions makes robust generalisations unlikely.

4. CONCLUSIONS

The findings reported here provide the foundation for a model of perceived speaker similarity. The foundation can be augmented by considering further acoustic measures such as measures of rhythm, f0 dynamics, and other spectral parameters. Note, however, that in order to disentangle the perception of personal voice quality from linguistic factors, this experiment used speakers carefully matched for accent. A question for future research will be the interaction of these two aspects of speech in the perception of voice similarity. The present experiment is a first step in the construction of a comprehensive model of voice similarity, a model that would be able to predict how similar two voices would sound on the basis of their acoustic and linguistic properties.

Accent and personal voice quality will normally both be relevant in forensic casework. A witness describing a bomb threat might describe the voice as ‘northern sounding, deep, and a bit nasal’. Any future database which might provide voice samples for voice parades would have to be cross-categorised for features of both accent and personal voice quality. Nevertheless, our experiment is directly relevant to the manual construction of voice parades as it is currently carried out in the UK. Candidate foils will normally already have been largely ‘controlled’ for accent, in that only voices matching the suspect’s accent will have been selected by the phonetician carrying out the work. Acoustic measures of the kind discussed here, appropriately weighted, could in principle be used to check similarity of personal voice quality. At this early stage, however, it would be premature to jettison the usual perceptual pre-test, in which listeners rate similarity to check that the suspect lies within the range defined by the foils. As yet, the human ear must be the final arbiter.

5. ACKNOWLEDGMENTS

We would like to thank Geoffrey Potter for technical assistance, and both him and Gea de Jong for moral support. Phil Harrison generously allowed us to use his f0 software. Two reviewers gave very helpful suggestions. The *VoiceSim* project within which this work was carried out was funded by the UK Economic and Social Research Council (ESRC) Grant RES-000-22-2582. The database from which the speech samples were taken was collected during the *DyViS* project supported by ESRC Grant RES-000-23-1248. The second author is supported by a British Academy Postdoctoral Fellowship.

6. REFERENCES

[1] Baumann, O., Belin, P. 2010. Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychological Research* 74, 110-120.

[2] Dynamic Variability in Speech. *http://www.ling.cam.ac.uk/dyvis/*

[3] Eriksson, A., Wretling, P. 1997. How flexible is the human voice? A case study of mimicry. *Proceedings of Eurospeech-1997*, 1043-1046.

[4] Giguère, G. 2006. Collecting and analyzing data in multidimensional scaling experiments: A guide for psychologists using SPSS. *Tutorial in Quantitative Methods for Psychology* 2(1), 27-38.

[5] Home Office. 2003. Advice on the use of voice identification parades. UK Home Office Circular 057/2003 from the Crime Reduction and Community Safety Group, Police Leadership & Powers Unit. *http://www.homeoffice.gov.uk/about-us/home-office-circulars/circulars-2003/057-2003/*

[6] Hudson, T., de Jong, G., McDougall, K., Harrison, P., Nolan, F. 2007. F0 statistics for 100 young male speakers of Standard Southern British English. *Proc. 16th ICPhS* Saarbrücken, 1809-1812.

[7] de Jong, G., McDougall, K., Nolan, F. 2007. Sound change and speaker identity: an acoustic study. In Müller, C. (ed.), *Speaker Classification II: Selected Papers*. Berlin: Springer, 130-141.

[8] Loakes, D. 2006. *A Forensic Phonetic Investigation into the Speech Patterns of Identical and Non-Identical Twins*. Ph.D. Dissertation, the University of Melbourne.

[9] Nolan, F., McDougall, K., de Jong, G., Hudson, T. 2009. The DyViS database: Style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law* 16(1), 31-57.

[10] Remez, R.E., Fellowes, J.M., Rubin, P.E. 1997. Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance* 23(3), 651-666.

[11] Rietveld, A.C.M., Broeders, A.P.A. 1991. Testing the fairness of voice identity parades: the similarity criterion. *Proc. 12th ICPhS* Aix-en-Provence, 46-49.

[12] Rose, P. 1999. Differences and distinguishability in the acoustic characteristics of *hello* in voices of similar-sounding speakers: A forensic phonetic investigation. *Australian Review of Applied Linguistics* 22, 1-42.

[13] Rose, P., Duncan, S. 1995. Naive auditory identification and discrimination of similar voices by familiar listeners. *Forensic Linguistics* 2(1), 1-17.

[14] Rothman, H.B. 1977. A perceptual (aural) and spectrographic identification of talkers with similar sounding voices. *Proceedings of the International Conference on Crime Countermeasures* Oxford, 37-42.

[15] Schiffman, S.S., Lance Reynolds, M., Young, F.W. 1981. *Introduction to Multidimensional Scaling: Theory, Methods, and Applications*. New York: Academic Press.

[16] Voice Similarity and the Effect of the Telephone. *http://www.ling.cam.ac.uk/voicesim/*

[17] Walden, B.E., Montgomery, A.A., Gibeily, G.J., Prosek, R.A., Schwartz, D.M. 1978. Correlates of psychological dimensions in talker similarity. *Journal of Speech and Hearing Research* 21, 265-275.