The role of voice quality 'settings' in perceived voice similarity

Francis Nolan[1], *Peter French*[2], *Kirsty McDougall*[1], *Louisa Stevens*[2], *and Toby Hudson*[3]

[1]*Department of Linguistics* | [3]*RCEAL, University of Cambridge, Cambridge, UK*

{fjn1|kem37|toh22}@cam.ac.uk

[2]*J.P. French Associates and University of York, York, UK*

{jpf|lcs}@jpfrench.com

Background

This paper reports part of a research programme which explores the notion of 'perceived voice similarity'. Results have been previously reported from an experiment in which listeners rated the difference between all pairings of 15 voices (controlled for accent – see Nolan et al. 2009), including pairings of two different 3-second samples from the same speaker. Multidimensional scaling (MDS) was used to reduce the pairwise ratings of (dis)similarity to five pseudo-perceptual dimensions (*dim1 ... dim5*). A strong correlation between the MDS *dim1* and the acoustic parameters of average F_0 confirmed the perceptual salience of pitch in judging voice similarity. *Dim2*, *dim3*, and *dim4* correlated significantly with average F3, F2, and F1 respectively. Whilst such correlations are interesting in themselves, they leave open the question of whether these acoustic descriptors capture voice quality characteristics which are salient for impressionistic observation and description. To explore this, authors 2 and 4 carried out an auditory analysis of the experimental samples within the framework of Laver (1980).

The Laver Voice Quality framework

Laver's descriptive framework provides a means to describe voice quality, understood as 'the characteristic auditory colouring of an individual speaker's voice…a cumulative abstraction over a period of time of a speaker-characterizing quality' Laver (1980:1). The descriptive system is componential; for instance a voice can be described as manifesting 'larynx raising, denasality, and creaky voice'. Such terms are referred to by Laver as 'settings', each reflecting a bias in the cumulative tendency of the speaker's vocal tract away from hypothetical neutral values. The setting labels imply plausible productive mechanisms, but are ultimately defined by auditory impressions; and the framework is learned through training rather in the way Cardinal Vowels are. Authors 2 and 4 have recently undertaken such training in order to implement the framework in forensic casework.

Experiment

For each speaker used in the similarity rating experiment described above authors 2 and 4 completed a 'vocal profile' – a *pro forma* assessment of the degree to which each of 38 principal settings was evident (if at all) in that speaker's sample. For this purpose the two samples from a speaker were concatenated and treated as one. Ratings were on a scale from 0 (no deviation from the hypothetical 'neutral' setting) to 3 (extreme deviation).

Results

A number of settings were not manifested in any of the speakers. *Tremor*, for instance, was absent; it might be more likely in elderly speakers. *Falsetto* and *whispery voice* were also not encountered, perhaps being more likely to be confined to specific paralinguistic functions. However, 26 of the 38 settings considered were observed. A preliminary way to see how important a setting is for discriminating a population of speakers is to look at the standard deviation of the ratings across the 15 speakers, as shown in Fig. 1. It can be seen that the five settings with the highest between-speaker variation in rating are nasality, three phonatory settings, and raised larynx. For comparison, five settings with moderate between-speaker variation are shown; these include another laryngeal setting (harshness), and four 'latitudinal' settings modifying the cross-sectional area of the vocal tract. Further analysis will be presented to relate these findings to the previous naïve-listener assessments of (dis)similarity, and to the acoustic properties of the voices.

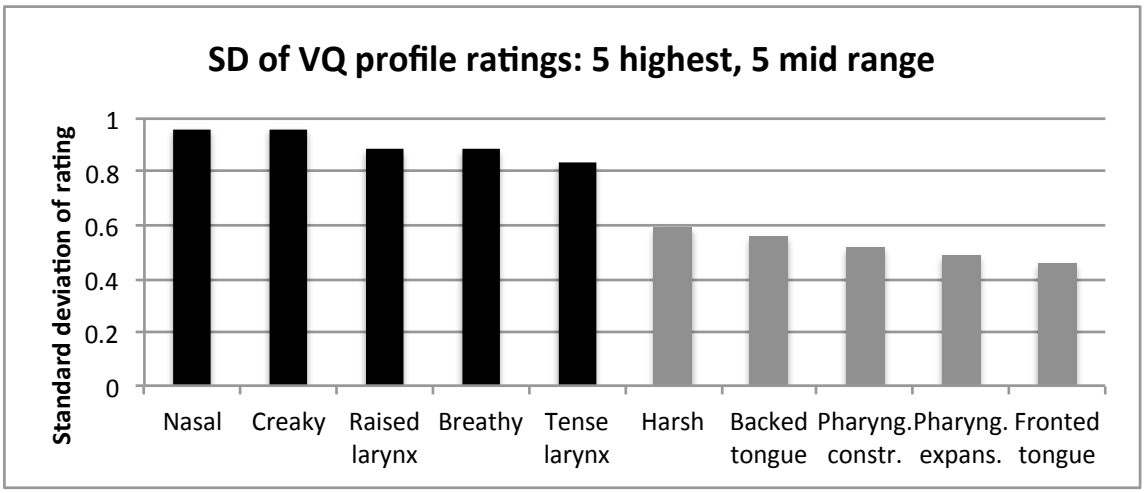

Figure 1 The five voice quality settings (black) rated most variably across the 15 speakers, and five (grey) with intermediate variation.

Conclusion

This study is a preliminary attempt to incorporate systematic impressionistic analysis of voice quality into the search for what makes same-accent voices similar or different. The results are significant because there is increasing interest in the application of systematic voice quality analysis in forensic phonetic casework, and the more that is understood about the relation of voice quality settings to between-speaker (dis)similarity, the more reliable and transparent that application will be. More generally, the findings will contribute to our understanding of perceptual 'speaker space'.

References

Laver, J. (1980). The Phonetic Description of Voice Quality. Cambridge: Cambridge University Press.

F. Nolan, K. McDougall, G. de Jong & T. Hudson (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.