Cluster analysis of voice quality ratings: Identifying groups of perceptually similar speakers

Eugenia San Segundo, Paul Foulkes, Peter French, Philip Harrison, Vincent Hughes and Colleen Kavanagh

Department of Language and Linguistic Science, University of York, UK

{eugenia.sansegundo|paul.foulkes|peter.french|philip.harrison|vincent.hughes|colleen.kavanagh}@york.ac.uk

Abstract

Cluster analysis is a way of classifying individual cases into groups on the basis of their similarity in respect of a defined set of variables. In this investigation, our cases are 99 male speakers of Southern Standard British English, matched for age and occupation. The classifying variables are the perceptual ratings given by three phoneticians to these speakers on 32 settings of the Vocal Profile Analysis scheme, one of the most widely used protocols to assess voice quality perceptually. The results reveal that it is possible to distinguish at least two speaker clusters. The forensic implications of these results are discussed.

Introduction

Investigations of voice similarity have become increasingly important in research fields such as voice casting (Obin & Roebe 2016). Different methodologies have been proposed for the development of computer-aided voice casting to determine the target actors that are most similar to the voice of a source actor. There are also forensic applications of voice similarity research, such as the design and validation of voice parades, i.e. the perceptual equivalent of a visual identification line-up (de Jong et al. 2015). These are used when someone has been witness to a crime wherein she/he could not see the perpetrator's face but could hear him/her speak. In such situations, forensic phoneticians may be requested to construct a voice parade. To ensure that ear witness evidence is conducted fairly, the forensic expert must choose a set of similar voices to the suspect (foils). After the line-up design and validation, earwitnesses are asked if they can recognize the offender's voice from the selection of voices.

Ensuring a good selection of foils is vital to ensure the parade is fair for both the witness and the suspect. One undesirable effect occurs when the suspect's voice stands out because it is not similar enough to the foils, and consequently it is too easy for the earwitness to pick out – even if the suspect were not in fact the offender. This may lead to a miscarriage of justice.

Although some proposals exist on how to measure speaker similarity for voice parades (de Jong et al. 2015), in this investigation we tackle the under-researched question of voice quality (VQ) similarity. The VQ of a speaker might be described by a non-expert listener as 'deep' and 'hoarse'. A phonetician might describe the same voice by using terms such as 'expanded pharynx' and 'tense larynx'. VQ is thus defined as the quasi-permanent quality of a speaker's voice resulting from a combination of long-term laryngeal and supralaryngeal settings. The Vocal Profile Analysis (VPA) scheme in particular is a componential approach to the perceptual assessment of VQ. In this protocol, VQ is seen as emerging from various components or settings, defined in relation to a 'neutral setting'. The version used here comprises 32 settings: 21 supralaryngeal, 7 laryngeal and 4 referring to muscular tension (San Segundo et al. 2017). A 3 point scale was used to judge the degree of deviation from neutral. The points were defined as 'slight' (degree 1), 'marked' (2) and 'extreme (but non-pathological)' (3).

Materials and methods

Speaker corpus

Data for analysis were extracted from Task 2 of the DyViS corpus (Nolan et al. 2009), which contains recordings of 100 non-pathological male speakers of Standard Southern British English (SSBE) aged 18-25 in a semi-scripted telephone conversation. These are high-quality recordings (44.1 kHz sample rate, 16-bit resolution) with around 7 minutes net speech. One speaker was excluded because of technical problems with his recording.

Perceptual assessment

Three phoneticians (the three first authors) – all trained in the VPA protocol– conducted the perceptual evaluation of the voices. They followed a two-stage methodology: a pilot assessment of 10 randomly selected subjects, followed by a calibration meeting (San Segundo et al 2017). Analysts then produced the 99 VPA ratings independently (i.e. a blind procedure) and a cross-coder calibration process produced agreed ratings for each speaker. Overall inter-rater agreement was high (82.6% absolute agreement and 89.1% agreement within 1 scalar degree). Unweighted Fleiss' kappa ranged between moderate and substantial agreement (San Segundo et al. 2017).

Cluster analyses

Selecting squared Euclidean distances, we followed two cluster methods (using *IBM SPSS Statistics 24*):

(1) *Hierarchical method:* : Since non-hierarchical clustering requires specifying the number of clusters (i.e. an arbitrary decision), we first implemented a hierarchical method to define the number of clusters (Ward's method). This computes the sum of squared distances within clusters and then aggregates clusters with the minimum increase in the overall sum of squares.The number of clusters was determined at the step where the distance coefficients showed a greater difference in the agglomeration schedule; this was checked visually in the scree diagram. In this instance, the number of cases (99) minus the step where the greatest coefficient difference was found (97th) gave two clusters.

(2) *Non-hierarchical method:* A non-hierarchical procedure, *k-means,* was used to properly form the clusters. Under this approach, the number *k* of clusters is fixed (here: two clusters) and an initial set of *k* 'seeds' (aggregation centers) is provided. Given a certain threshold, all units are assigned to the nearest cluster seed. New seeds are computed until no reclassification is necessary.

Results

Hierarchical Ward's method

This method revealed that the speakers split into two main clusters. On the one hand, hierarchical clustering presents the disadvantage that results are affected by the way in which the variables are ordered. In this case, the 32 VPA settings were ordered following anatomical progression from lips to larynx ('lip rounding' to phonation settings such as 'creak'; see Table 1). On the other hand, this type of clustering has the advantage of allowing dendrogram representations, a type of tree structure that fosters the understanding of data relations by placing similar cases together and positioning relatively unrelated cases at a greater distance (Figure 1).

Non-hierarchical k-means method

This analysis showed that 53 speakers belonged to Cluster 1 and 46 speakers to Cluster 2. Table 1 shows the final cluster centers, where all variables are assigned either to Cluster 1 or Cluster 2. The settings that contribute most to the separation of the clusters present higher average values in one cluster than in the other. The following settings (p-value) belong to Cluster 1: 'lowered larynx' (4.1×10^{-15}), 'lax larynx' (1.7×10^{-13}), 'creaky' (5×10^{-6}) and 'breathy' (1.3×10^{-5}). Cluster 2 contains the following settings (p-value): 'raised larynx' (9.1×10^{-13}), 'tense larynx' (6.6×10^{-11}), 'harsh' (2×10^{-6}) and 'whispery' (3×10^{-3}).

*Table 1. Final cluster centers (supralaryngeal settings 1-21; muscular tension settings 22-23; laryngeal settings 24-32). ANOVA test significance level: * p <0.01 **p <0.001*

	Cluster 1	Cluster 2
1. Lip rounding	.02	.00
2. Lip spreading	.08	.02
3. Labiodentalisation	.00	.00
4. Extensive labial range	.00	.00
5. Minimised labial range	.00	.00
6. Close jaw	.00	.02
7. Open jaw	.00	.00
8. Extensive mandibular range	.00	.00
9. Minimised mandibular range	.04	.09
10. Advanced tongue tip	.70	.96
11. Retracted tongue tip	.06	.00
12. Fronted/raised tongue body	1.26	1.28
13. Backed/lowered tongue body	.00	.00
14. Ext. lingual range	.04	.02
15. Min. lingual range	.00	.04
16. Pharyngeal constriction	.00	.07
17. Pharyngeal expansion	.06	.00
18. Nasal	1.08	1.41
19. Denasal	.09	.04
20. Raised larynx **	.04	.96

21. Lowered larynx **	1.09	.04
22. Tense vocal tract	.62	.65
23. Lax vocal tract	.64	.59
24. Tense Larynx **	.09	.93
25. Lax Larynx **	1.06	.11
26. Falsetto	.00	.00
27. Creaky **	1.53	.85
28. Whispery *	.02	.30
29. Breathy **	1.49	.74
30. Murmur	.08	.00
31. Harsh **	.11	.70
32. Tremor	.00	.00

The settings mentioned above are the ones that contribute most to the separation of the two clusters. They point to very different VQ configurations from an articulatory point of view; mutually exclusive configurations in the case of ‘lowered larynx’ vs. ‘raised larynx’, and ‘lax larynx’ vs. ‘tense larynx’. Since the observed significance level for the remaining VPA settings was large ($p > 0.01$), they cannot be considered to contribute much to the separation of the clusters.

Figure 1 is the dendrogram representation following the Ward’s cluster method. Here, the speakers highlighted in bold were the ones classified as belonging to Cluster 2 by the k-means method. As it can be seen, most of these classifications are coincident (large bold box at the bottom right-hand corner). However, 20 speakers were classified differently by each method: speakers in individual bold boxes were classified as belonging to Cluster 1 by the Ward’s method but belonging to Cluster 2 by the k-means method.

Discussion

Cluster analysis allows us to find patterns in large data sets. However, the choice of one clustering technique over others leads to slightly different results, as shown in this investigation. Although the number of differently classified speakers in the hierarchical Ward’s method in comparison with the non-hierarchical k-means method is relatively small, it serves to raise the question of which method is more suitable for this type of data. We suggest that k-means is a better method for drawing meaningful distinctions between speakers. On the one hand, the fact that this method is not influenced by the order of the variables makes it more robust (the analysis is stable even if cases are dropped). On the other hand, the k-means classification fulfils the requirements for a robust classification (Eppler & Stoyko 2011): it must be simple and clear, contain meaningful groupings, and above all, it should be consistent with established theories. The fact that ‘raised larynx’ and ‘lowered larynx’, together with their constellation of related settings, differentiate the two clusters is in line with phonetic theory. For instance, ‘harsh’ phonation (Cluster 2) is achieved with strong adductive glottal tension and fundamental frequencies (f0) consistently above 100 Hz. ‘Creaky’ phonation (Cluster 1) occurs with f0 consistently below 100 Hz and is clustered together with ‘lax larynx’ (cf. articulatory description of ‘creaky’ in Laver 1980). Note, however, that different types of ‘creaky’ may exist with varying degrees of laryngealization (Keating & Garrellek 2015). This could explain that many speakers evaluated as ‘creaky‘ can be found in both clusters.

Conclusion

The aim of this study was to explore clustering methods to distinguish perceptually similar speakers on the basis of VQ ratings. The fact that two main clusters were found even within a homogeneous population of same-sociolect speakers has forensic implications. For instance, it suggests the importance of annotating speaker databases with VPA information prior to a new voice parade design. This methodology would enable an automatized search of the most similar set of foils for each suspect. This would allow for optimization of resources by law enforcement agencies, with a considerable reduction in the time and costs currently involved in the design of voice parades. It would also minimize the subjectivity involved in the selection of voices.

Cluster analysis is not novel in forensic/biometric investigations (Fong 2012) or sociophonetics (Ferragne & Pellegrino 2010). However, while these techniques have been widely used with a range of acoustic features, to our knowledge they have not yet been explored with perceptual ratings or discussed in relation to voice parades. The potential of *perceptual* ratings to identity cohorts of *perceptually* similar speakers seems clear. Previous studies have been largely confined to acoustic measures such as vowel formants and f0 (Kelly et al. 2016) or MFCCs (Adachi et al. 2009). The shortcoming of purely acoustic methods is that they disregard the fact that acoustic similarities may not be perceptually salient for a naïve listener in judging speaker similarity.

Future investigations will examine the extent to which auditory expert ratings are comparable with the similarity ratings of naïve listeners, thereby further testing the appropriateness of the method in voice line-up construction.

References

Adachi, S. Kawamoto, T. Yotsukura, S. Morishima, & S. Nakamura (2009). Automatic voice assignment tool for Instant Casting movie System, *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.1897-1900.

Beck, J. (2007). *Vocal profile analysis scheme: a user's manual*. Edinburgh: Queen Margaret University College-QMUC.

Eppler, M.J. & Stoyko, P. (2011). Drawing distinctions: The visualization of classification. MCM working paper, no. 2, Univ. St. Gallen.

Ferragne, E. & Pellegrino, F. (2010). Vowel systems and accent similarity in the British Isles: Exploiting multidimensional acoustic distances in phonetics. *Journal of Phonetics*, *38* (4), 526-539.

Fong, S. (2012). Using hierarchical time series clustering algorithm and wavelet classifier for biometric voice classification. *BioMed Research International*. Article ID 215019

de Jong, G., F. Nolan, K. McDougall & T. Hudson. (2015). Voice lineups: a practical guide. In *ICPhS 2015 – 18th International Congress of Phonetic Sciences*, August 10–14, Glasgow, UK. Paper number 1041.1-9.

Keating, P. & Garrellek, M. (2015). Acoustic analysis of creaky voice, LSA Annual Meeting.

Kelly, F., Alexander, A., Forth, O., Kent, S., Lindh, J. & Åkesson, J. (2016). Identifying Perceptually Similar Voices with a Speaker Recognition System Using Auto-Phonetic Features. In *INTERSPEECH 2016*, September 8–12, San Francisco, USA, pp. 1567-1568.

Laver, J. (1980). *The phonetic description of voice quality*, Cambridge: CUP.

Nolan, F. K. McDougall, G. de Jong & T. Hudson (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *IJSLL*, *16*, 31-57.

Obin, N. & A. Roebe (2016). Similarity search of acted voices for automatic voice casting," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24* (9), 1642-1651.

San Segundo, E., Foulkes, P., French, P., Harrison, P., Hughes, V. & Kavanagh, C. (under review). The use of the Vocal Profile Analysis for speaker characterisation: methodological proposals.

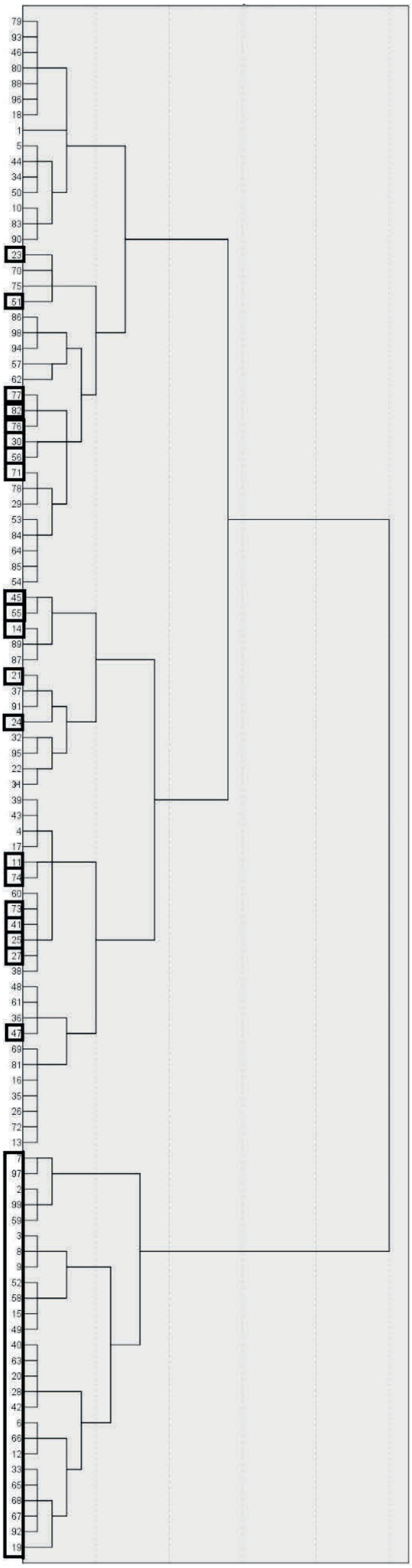

Figure 1. Dendrogram (Ward's method). Bold: Cluster 2 speakers with k-means.