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

INTRODUCTION

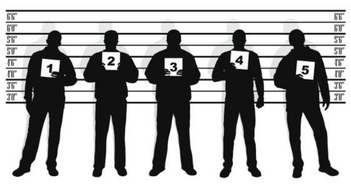
Investigations of **voice similarity** → increasingly important in various fields:

❶ Voice casting (Obin & Roebe 2016)

- To determine the set of target actors that are the most similar to the speech recording of a source actor
- E.g. to transfer a film or video game from a source language to a target language with a small amount of available voices for each language


Dietmar Wunder is the German voice of Daniel Craig. Check their similarity

❷ Voice parades (forensic application)



- Perceptual equivalent of a visual identification line-up
- Used when a witness to a crime could not see the perpetrator's face but could hear him/her speak
- Design by forensic phoneticians:
  - to ensure that ear witness evidence is conducted fairly, the forensic expert chooses a set of similar voices to the suspect (foils)
  - earwitnesses are then asked if they can recognize the offender's voice from the selection of voices (de Jong et al. 2015)
  - undesirable effect → if 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

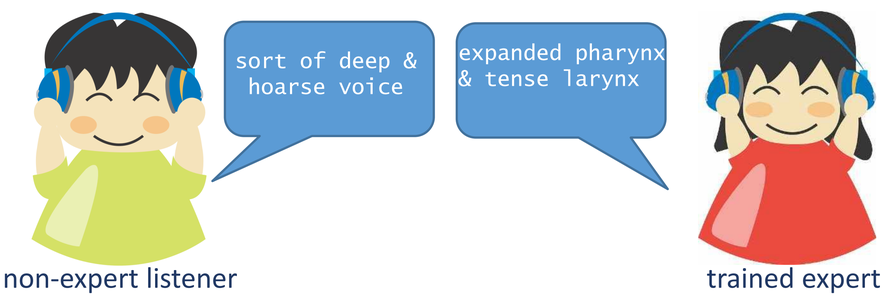
In both ❶ & ❷ the notion of VOICE SIMILARITY is key: the smaller/larger the distance between voices, the closer/farther they are perceived

….but how to measure voice similarity/distances?

OBJECTIVES

✓ Investigate one under-researched aspect: voice quality (VQ) similarity

- some proposals exist on how to measure speaker similarity for voice parades (de Jong et al. 2015), but investigations focusing on VQ are rare
- VQ = quasi-permanent quality of a speaker's voice resulting from a combination of long-term laryngeal and supralaryngeal settings (Laver 1980)



MATERIALS & METHODS

Speaker corpus: DyViS

- 99 male speakers (aged 18-25) Standard Southern British English
- semi-scripted telephone conversation
- 7 min. net speech

VQ protocol: Vocal Profile Analysis (VPA)

- componential approach to the perceptual assessment of VQ
- VQ = emerging from several components or settings (defined in relation to a 'neutral setting')
- 3 point scale: slight > marked> extreme

Perceptual assessment procedure

- 3 trained phoneticians (ESS, PF, PFr)
- 2-stage methodology: (1) pilot assessment of 10 random subject; (2) calibration meeting
- 99 VPA ratings each (blind procedure)
- cross-coder calibration process

INTERRATER AGREEMENT

82.6% absolute agreement
89.1% within 1 scalar degree
Unweigthed Fleiss' kappa = moderate to substantial

(See San Segundo et al. 2017)

Cluster analyses

→ measured distances: squared Euclidean distances

① hierarchical method (Ward's method)

- because non-hierarchical methods require specifying # clusters
- # clusters determined at the step where distance coeff. show a great diff. in agglomeration schedule → checked visually in the scree diagram (here: two clusters)

② non-hierarchical method (k-means)

- used to actually form the clusters
- # clusters is fixed (two) 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

① Hierarchical Ward's method

(−) affected by the order of variables (anatomical progression of VPA settings from lips to larynx)
(+) allow dendrogram representation / tree structure

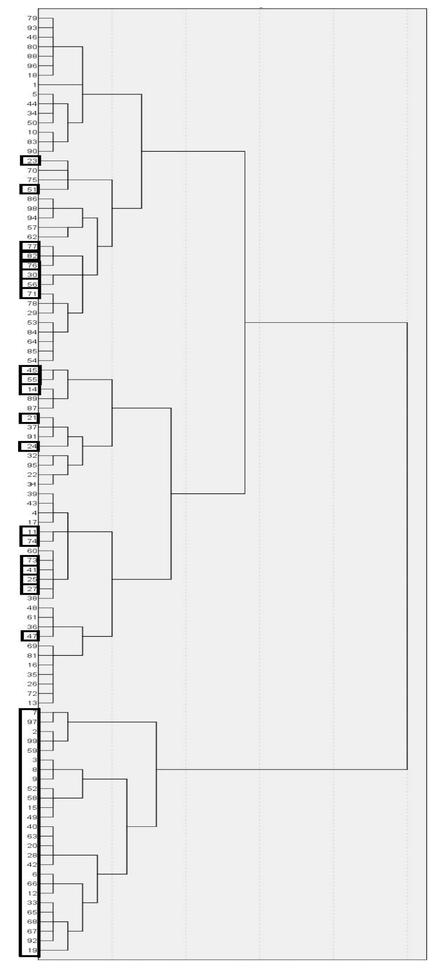


Figure 1: *Dendrogram (Ward's method)*
Bold: Cluster 2 speakers with k-means method

② Non-hierarchical k-means method

- Cluster 1 = 53 speakers ; Cluster 2 = 46 speakers
- VPA settings – assigned to either Cluster 1 or 2

- settings that contribute most to the separation of the clusters because high average % in one cluster vs. the other (ANOVA test)

		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

Table 1: *Final cluster centers*
ANOVA test significance level:
* p<0.01 ** p<0.001

DISCUSSION & CONCLUSIONS

➢ **Cluster analysis** = useful to find patterns in large data sets

- different techniques → slightly diff. results
- hierarchical vs. non-hierarchical → diff. classified speakers

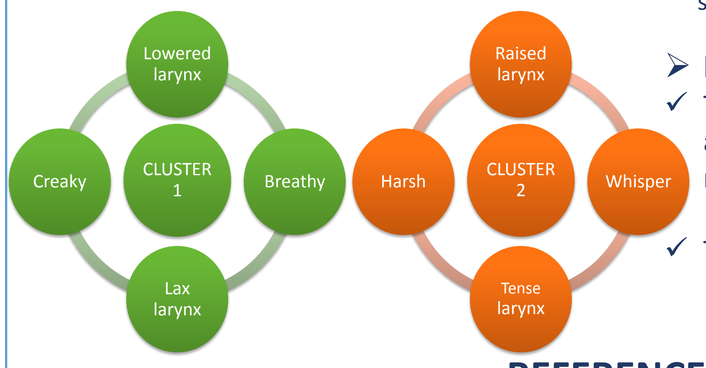
➢ **K-means technique** = better method

✓ not influenced by order of variables (robust method even if some cases are dropped)

✓ meet requirements for robust classification (Eppler & Stoyko 2011):

- simple & clear
- meaningful groupings
- consistent with established theories

(e.g. Laver 1980: creaky & f0 below 100 Hz)



➢ Clustering methods allow to distinguish perceptually similar speakers (basis: VQ ratings)

✓ It was possible to distinguish at least 2 clusters even within a homogeneous population of same-sociolect speakers

✓ Forensic implication → annotate databases (VQ info) prior to voice parade

- automatized search of most similar foils to suspect
- optimize resources by law enforcement agencies (time and money save)
- minimize subjectivity involved in voice selection

➢ **Novelty:** use of perceptual ratings

- Previous studies: acoustic measures (disadvantage: not necessarily perceptually salient for listeners in judging similarity)

➢ **Future:**

✓ To which extent auditory expert ratings are comparable with similarity ratings by naive listeners (San Segundo et al. 2016)?

✓ Testing other data reduction techniques for data mining (e.g. factor analyses)

REFERENCES

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.

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

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

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. & V. Hughes (2016). Holistic perception of voice quality matters more than L1 when judging speaker similarity in short stimuli, *Proceedings of the 16th Australasian International Conference on Speech Science and Technology*, Sydney 6-9 December, 2016

San Segundo, E., Foulkes, P., French, P., Harrison, P., Hughes, V. & C. Kavanagh (2017, under review). The use of the Vocal Profile Analysis for speaker characterisation: methodological proposals.

ACKNOWLEDGEMENTS & CONTACT

This research was funded via the UK AHRC grant *Voice and Identity* (AH/M003396/1)

Email: eugenia.sansegundo@york.ac.uk @sanse_eu Web: eugeniasansegundo.github.io