Perceptual similarity of identical twins across different L1 listeners: the importance of voice quality in Forensic Phonetics

Eugenia San Segundo[1], Paul Foulkes[1], Peter French[1,2], Vincent Hughes[1] and Olaf Köster[3]
[1]Dept. Language & Linguistic Science, University of York, UK [2]J P French Associates, York, UK
[3]BKA (Federal Criminal Policy Office, Germany)

BACKGROUND & OBJECTIVE

- **Round robin test** recently conducted by the *Bundeskriminalamt* (BKA) to evaluate the performance of experts in speaker identification tasks:
  - **auditory evaluation** since the technical characteristics of the recordings prevented most experts from carrying out specific acoustic analyses.
  - speakers for comparison = **pair of female German twins**.
  - widely assumed that twins' voices are similar, and thus recognition of voices is especially difficult (e.g. [1,2]).
- **Results:**
  - **lack of native knowledge** of the language spoken by the twins was not a disadvantage for telling the twins apart.
  - informal feedback from participants suggested that **voice quality (VQ) –approached holistically rather than analytically– was the main cue** used by non-native listeners to distinguish the twins.
- **Limitations:**
  - **limited and idiosyncratic** data set (the twins were of advanced age and had lived in different dialectal regions)

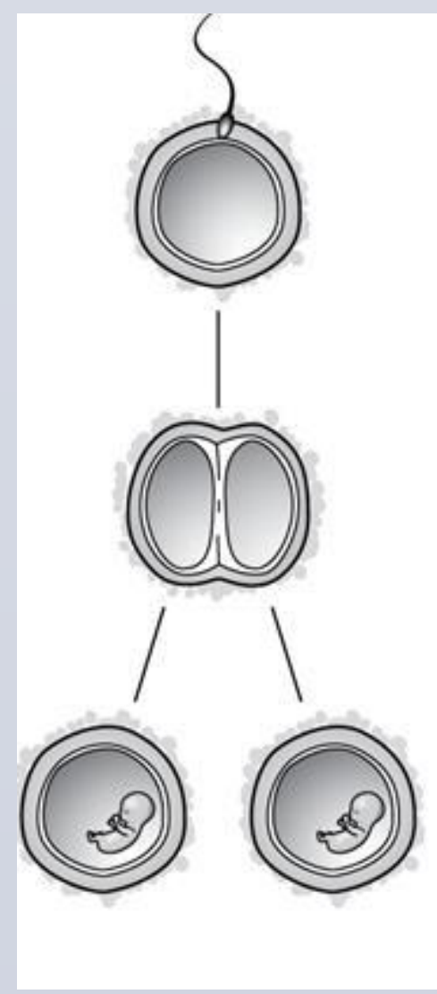
SO WHAT NOW?

→ The BKA test called for the design of a perceptual experiment of a different nature which could shed light on how listeners of different L1s perform when assessing the voice of very similar-sounding speakers.

→ **In this study we have tested, with a larger twin sample and under controlled conditions of age and dialect, whether the different L1 of listeners affect the perceptual distances between speakers.**

WHY IDENTICAL TWINS

- **Monozygotic (MZ) twins** occur when a single ovum is fertilized by a sperm cell to form one zygote, which then divides in two.
- MZ twin pairs share all their genes in common. The fact that they usually share environmental (educational + social) influences makes them examples of extreme similarity, also in voice. Both *organic* (vocal tract anatomy) and *learned* (phonetic choices) variation –which usually explain between-speaker variation – are minimised in MZ twin pairs.



Importance of twins for voice quality research

Different speakers present isomorphic but not identical vocal tracts, this being one of the shortcomings of perceptual protocols for the assessment of voice quality, such as the VPA [3]:

"Laver's framework would not be designed to capture the less linguistic aspects of speech, i.e. the relevant 'phonetic detail'. In other words (…) the small differences in size or shape that two speakers have will make them sound different even if they choose the same articulatory options" [5]

→ **Investigations with twins (identical vocal tract) may then prove useful to assess VQ closeness in very similar-sounding speakers.**

MATERIALS & METHOD

Subjects: 10 speakers selected from the corpus collected in [6], i.e. five pairs of male MZ twins (native speakers of Standard Peninsular Spanish). Original corpus contains 54 speakers (aged 18-54), so for the selection of the 10 speakers of this experiment some criteria were established:

- ❑ Similar age (mean: 21, sd: 3.7)
- ❑ Similar mean f0 (mean: 113 Hz, sd: 13 Hz)
- ❑ Similar Euclidean distance (ED) between each speaker and his twin (as calculated in [7]), in order to select only the most similar-sounding twin pairs. EDs are based on the perceptual assessment of their VQ using a simplified version of the Vocal Profile Analysis (VPA) scheme [3]. The mean ED between twin pairs was 0.62, measured in Similarity Matching Coefficients (SMC), a typical distance measure for categorical data where the number of matches for each variable is divided by the number of variables.

	MAJOR SETTING GROUPS									
Key	**Labial**	**Mandib.**	**Ling. tip**	**Ling. body**	**Pharyng.**	**Velo-pharyng.**	**Larynx Height**	**VT tension**	**L tension**	**Phon. Types**
1a	Lip rounding	Close	Advanced	Front &Raised	Constricted	Audible nasal escape	Raised Larynx	Tense	Tense	Falsetto
1b	Lip spreading	Open	Retracted	Back &Lowered	Expanded	Nasal	Lowered larynx	Lax	Lax	Creak.
1c	Labiodent.	Protr.				Denasal				Whisp.
1d										Harsh.
1e										Tremor

Table 1: Simplified Vocal Profile Analysis Scheme (SVPAS): 10 major setting groups and 26 total settings, with the category key for marking non-neutrality (1a-1e). *VT*: Vocal Tract; *L*: Laryngeal; *Labiodent*: Labiodentalization; *Protr:* Protruded; *Creak*: Creakiness; *Whisp*: Whisperiness/Breathiness.

	Lab	Mand.	Ling tip	Ling body	Pharyng.	Velo-pharyng.	Larynx Height	VT tension	L tension	Phon. Types	
AGF	0	1a	1a	0	0	0	0	1a	1a	0	
SGF	0	1b	1a	0	0	0	1b	1a	0	1c	
Matches	1	0	1	1	1	1	0	1	0	0	0.6
											SMC
AMG	0	1b	0	1b	0	0	0	1a	0	0	
EMG	1a	1b	1a	1b	0	1b	1a	1a	0	0	
Matches	0	1	0	1	1	0	0	1	1	1	0.6
											SMC
ARJ	0	1a	1a	0	0	0	0	1b	1b	1c	
JRJ	0	1a	0	0	0	0	1b	1b	1b	1c	
Matches	1	1	0	1	1	1	0	1	1	1	0.8
											SMC
ASM	1a	0	0	0	1b	0	1b	0	0	1c	
RSM	1a	1a	0	0	1b	0	1b	1a	1b	0	
Matches	1	0	1	1	1	1	0	0	0	0	0.6
											SMC
DCT	1a	0	0	1a	0	0	1a	1a	1a	1d	
JCT	0	0	1a	0	1a	1b	1a	1a	1a	1d	
Matches	0	1	0	0	0	0	1	1	1	1	0.5
											SMC

Table 2: Summary of SMCs for all twin pairs. Mean SMC: 0.62, indicating that around 6 VQ settings were shared on average by the twin pairs.

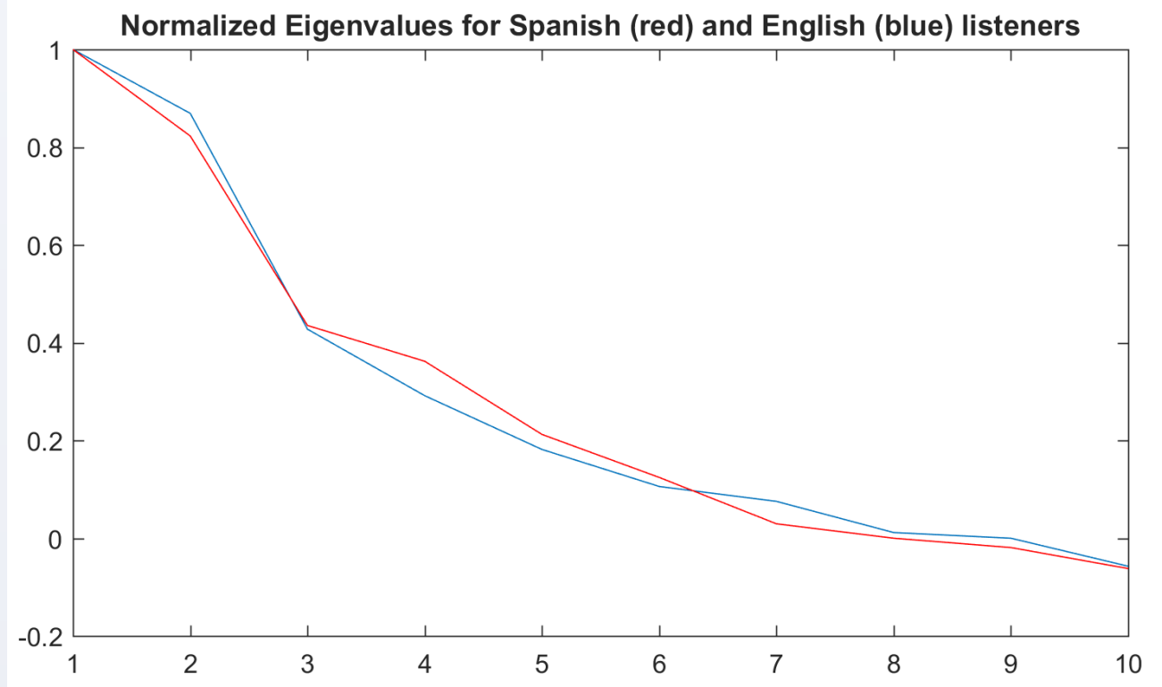
Stimuli: Speech samples (approx. duration 3 seconds) extracted from semi-directed spontaneous conversations [6]. Declarative sentences of different linguistic content (diverse neutral topics).

Listeners: Two different groups of listeners were recruited to take part in the perceptual experiment: native Spanish speakers (N=20; age range 22-51, mean 33) and native English speakers with no knowledge of Spanish (N=20; age range 19-35, mean 25).

Design of perceptual test: A *Multiple Forced Choice* experiment was set up with 90 different-speaker pairings, i.e. each speaker compared with everyone else. Stimuli were presented in random order and listeners had to indicate the degree of similarity of each stimuli pair on a scale 1 to 5. They didn't know that the stimuli included twin pairs. The test was run on a PC with HQ headphones. A short pre-test allowed familiarization with this type of test.

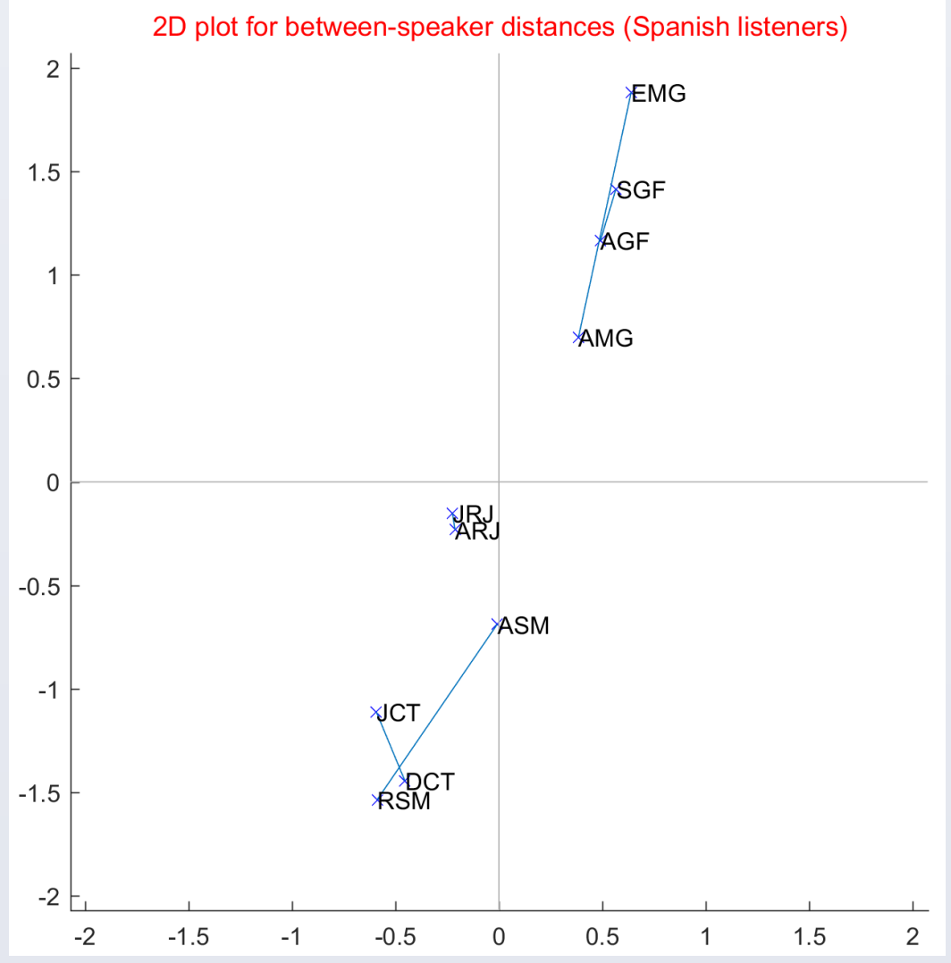
Analysis method: Following [4], the degree of perceived similarity was measured using Multidimensional Scaling (MDS), a means of visualizing the level of similarity of individual cases in a dataset and of detecting meaningful underlying dimensions that explain observed similarities or dissimilarities (distances).

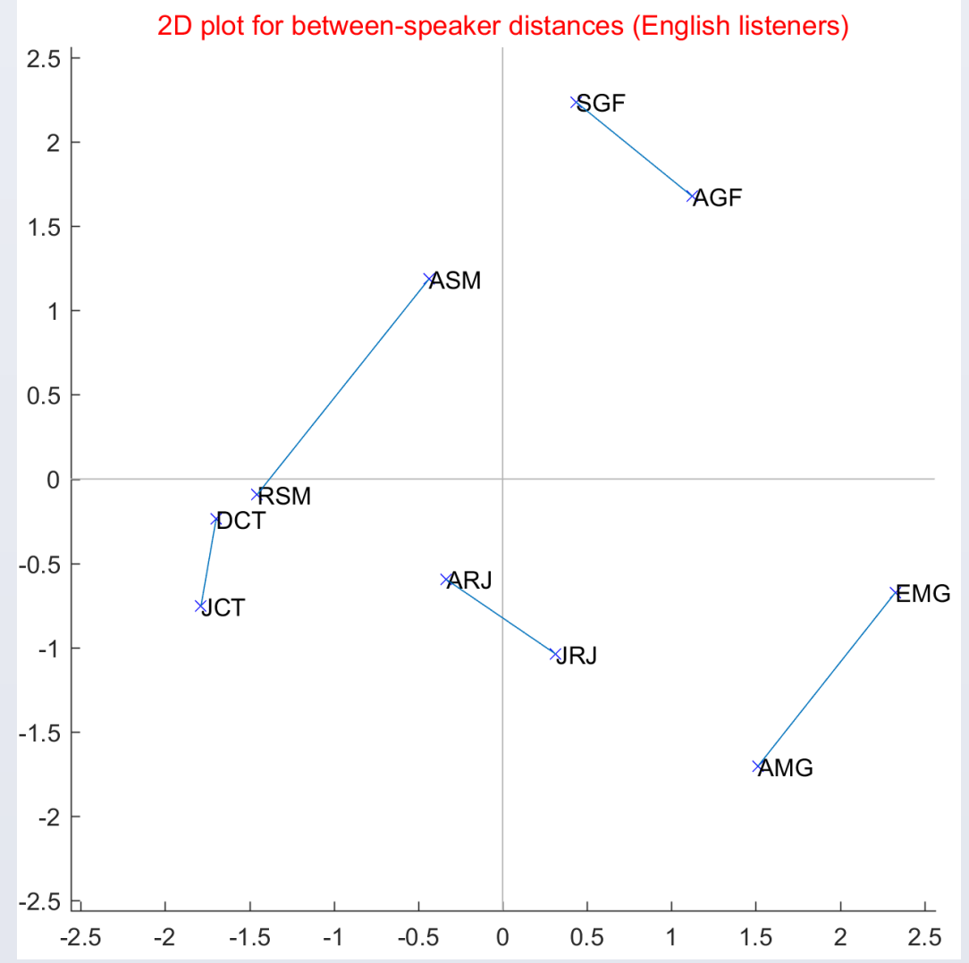
RESULTS

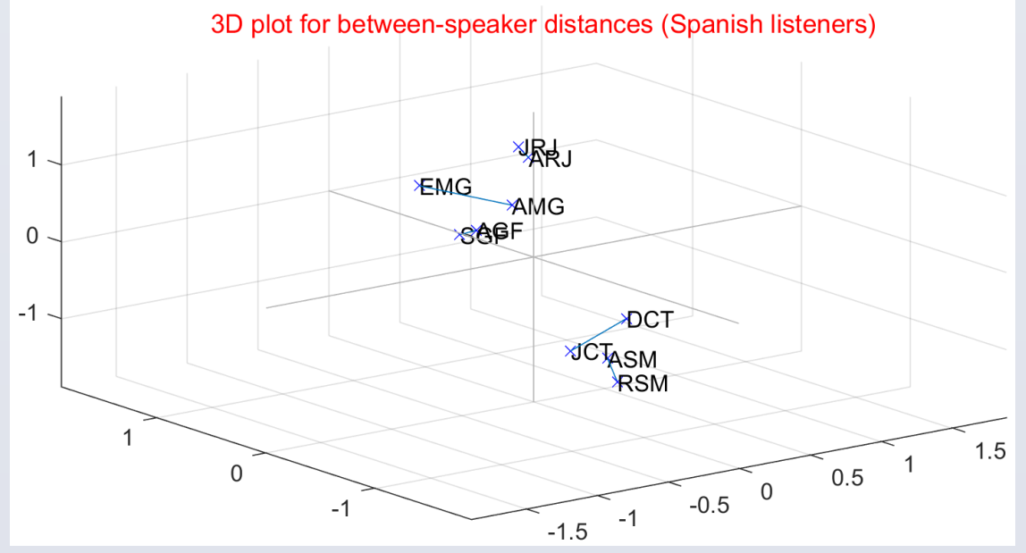


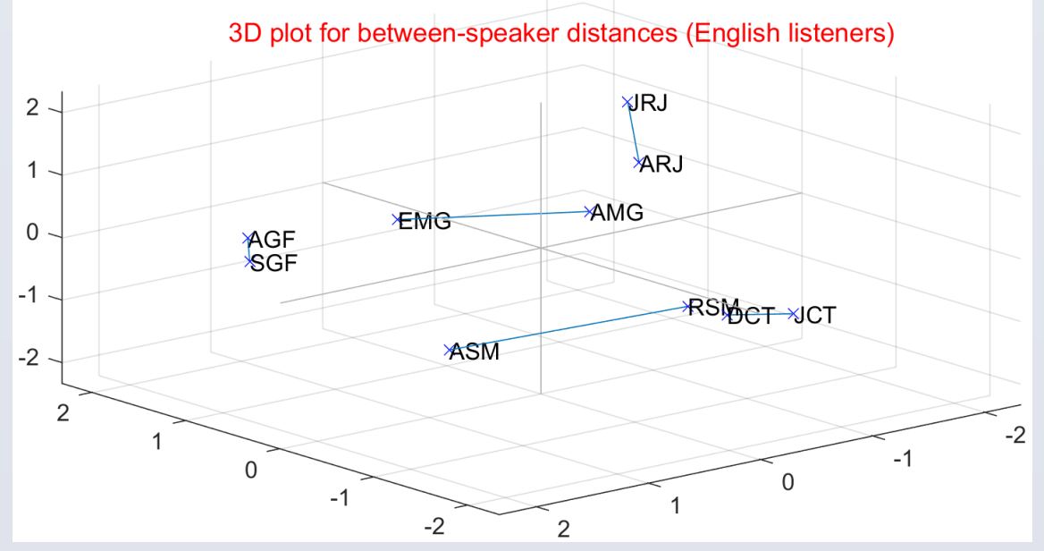
MDS analyses were carried out using similarity scores to construct a perceptual map of the speakers. The scree plot (right) shows the relative magnitude of the sorted Eigenvalues.

7 dimensions necessary to accurately reproduce between-speaker distances in the perceptual space, but MDS results typically visualized using only the first 2 or 3 dimensions.





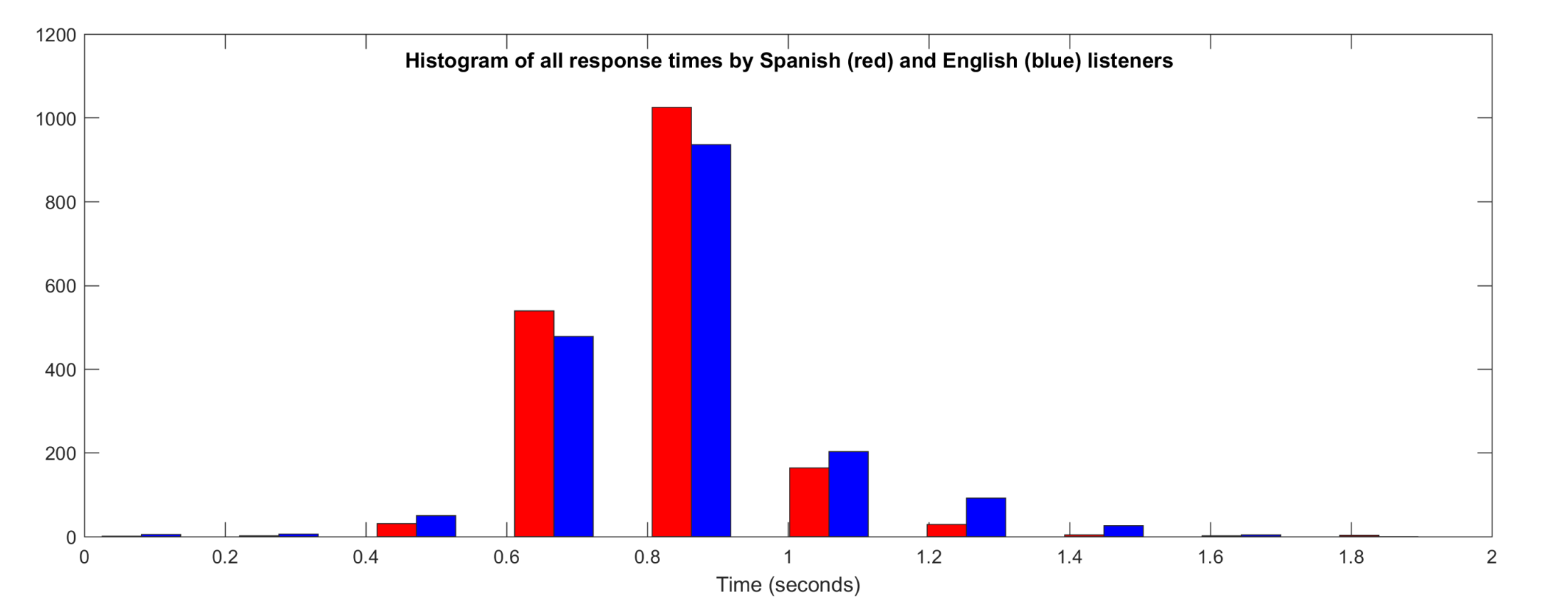




Normalized intra-pair Euclidean distances (7 dimensions):

	DCT_JCT	AGF_SGF	ARJ_JRJ	ASM_RSM	AMG_EMG
ENGLISH	0.219	0.264	0.349	0.435	0.445
SPANISH	0.343	0.341	0.345	0.369	0.607

Reaction times were very similar for Spanish (mean: 0.82 secs; std: 0.14) and English listeners (mean: 0.84; std: 0.18).



DISCUSSION

SPANISH

- All speakers are closer in the perceptual space. Does this imply that knowledge of the linguistic content makes the task more difficult? Distraction effect of the message?

- Better detection of very similar speakers (i.e. twins). Smaller distances between these in comparison with English listeners. Note also the different magnitude of the plots.

- Most similar twin pair: AGF-SGF and ARJ-JRJ . Most different twin pair: AMG-EMG.

ENGLISH

- All speakers are more spread in the perceptual space. Some twin pairs are very far apart, which even makes them have an unrelated speaker as their closest speaker.

- Most similar twin pair: DCT-JCT. Most different: AMG-EMG.

CONCLUSIONS

- Eigen-decomposition: 7 main dimensions explain similarity decisions by listeners (both English and Spanish).
  - ✓ voice is highly multidimensional; reducing the perceptual space to 2D or 3D may be misleading.
  - ✓ similarity of the relative magnitude of the sorted Eigenvalues suggests that similar perceptual strategies operate for both listener groups.
- Almost the same ranking of twin similarity for both listener groups could indicate the same cue prominence, i.e. regardless of familiarity with the language spoken or understanding of the linguistic content, both groups seem to rate the same twin pairs as most similar and the same twin pairs are most dissimilar. **Exception:** AGF-SGF most similar for Spaniards (VQ analysis: *tense VT* & *advanced tongue tip*) while DCT-JCT most similar for English (VQ analysis: *harshness* & *raised larynx*). Different perceptual salience of VQ settings?
- Equivalent reaction times point to similar listening strategies ('gut' impressions; holistic VQ perception).
  - ✓ However, qualitative feedback from participants also point to other cues: mainly rhythmic aspects but also segmental features.
- Besides: your twin may not necessarily be your best impostor (e.g. RSM: closer perceptual distance with unrelated speaker)
- **Future work:** Correlate Euclidean distances obtained from VQ holistic perception with componential-featural analysis of VQ
- **Other future work**: (1) Divide listeners in musical and non-musical training; (2) Test with listeners of other languages (e.g. Germans - would these obtain similar results as English?)

REFERENCES

[1] **Decoster, W., Van Gysel, A., Vercammen, J., & Debruyne, F**. (2000). Voice similarity in identical twins. *Acta Oto-Rhino-Laryngologica Belgica*, *55*(1), 49-55.

[2] **Künzel, H. J.** (2010). Automatic speaker recognition of identical twins. *International Journal of Speech, Language and the Law*, *17*(2), 251-277.

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

[4] **McDougall, K.** (2013). Assessing perceived voice similarity using Multidimensional Scaling for the construction of voice parades, *International Journal of Speech Language and the Law*, 20 (2): 163-172.

[5] **Nolan F.** (2005). Forensic speaker identification and the phonetic description of voice quality. In W. J. Hardcastle & J. Mackenzie Beck (Eds.) *A Figure of Speech: A Festschrift for John Laver*. 385–411. London/Mahwah, NJJ: Laurence Erlbaum Associates.

[6] **San Segundo, E.** (2014). *Forensic speaker comparison of Spanish twins and non-twin siblings: A phonetic-acoustic analysis of formant trajectories in vocalic sequences, glottal source parameters and cepstral characteristics*, Doctoral dissertation, Menéndez Pelayo International University & Spanish National Research Council.

[7] **San Segundo, E. & Mompeán, J.A**. (2016). Voice quality similarity based on a simplified version of the Vocal Profile Analysis: A preliminary approach with Spanish speakers including identical twin pairs, *Sociolinguistics Symposium 21*, University of Murcia, Spain, 15-18 June.

ACKNOWLEDGEMENTS

Thanks to Andrew MacFarlane and Duncan Robertson for references about measuring reaction time. A special thanks to Juana Gil for letting ESS run the test with Spanish listeners at the Phonetics Lab (CSIC, Madrid).