Holistic perception of voice quality matters more than L1 when judging speaker similarity in short stimuli

Eugenia San Segundo, Paul Foulkes, Vincent Hughes

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

{eugenia.sansegundo|paul.foulkes|vincent.hughes}@york.ac.uk

Abstract

Spanish and English naïve listeners judged the similarity of 5 pairs of Spanish speaking identical twins. Listeners rated speaker similarity in a comparable way irrespective of their L1. This is of forensic relevance in non-native earwitness evidence, as it suggests that similar listening strategies operate (i.e. holistic approach to voice quality) when stimuli are short and no other segmental cues are available for the naïve listener – native or non-native – to judge speaker similarity.

Index Terms: voice quality, perception, twins, forensic phonetics, English, Spanish

1. Introduction

Voice quality (VQ) is defined as the quasi-permanent quality resulting from a combination of long-term laryngeal and supralaryngeal features, which typically makes a speaker's voice different from others [1]. The study of VQ has produced fruitful research in speech pathology and therapy [2], L2 phonology [3], and sociolinguistics, including studies exploring cross-dialectal patterns [4][5]. VQ serves as a social marker to indicate membership of a speech community [6], but it is also idiosyncratic. As such, it has received considerable attention in forensic phonetics, a discipline which applies phonetic knowledge to legal issues. Forensic Speaker Comparison (FSC) tasks, the most frequent in forensic casework [7], consist of the comparison of voice samples belonging to an offender and a suspect in order to assist courts in determining speaker identity.

The study of VQ can be approached from articulatory, acoustic or perceptual perspectives, including hybrid instrumental and perceptual assessment methods. In this investigation we focus on the auditory assessment of VQ, specifically as it is carried out by *naïve* listeners as opposed to *experts* (cf. technical speaker identification [8]). Our hypothesis is that under controlled conditions of speaker similarity (i.e. similar-sounding speakers sharing dialect, approximate age and mean F0), naïve listeners would rely on a holistic VQ perception in order to judge similarity between speakers. Native knowledge of the speaker language would be irrelevant when short stimuli belonging to different voices are very similar in segmental aspects. Only the combination of VQ characteristics (e.g. harsh voice, nasality or close jaw) would be available for the listener to judge speaker similarity. Under this holistic approach of VQ, both native and non-native listeners are expected to perform in a similar way.

To explain the factors which account for similarity ratings we nonetheless consider the perceptual evaluation carried out from a *featural* approach by a trained phonetician (expert listener). The holistic-featural dichotomy has traditionally accompanied the perceptual study of VQ and it continues to be an issue today. Previous studies [9][10] based on neuropsychological evidence suggest that the perception of VQ cannot be explained as the sum of separate features; instead, it involves a component of holistic, gestalt-like pattern processing. However, the different perceptual protocols (e.g. VPA; GRBAS; CAPE-V or SVEA; cf. summary in [11]) that are available for forensic phoneticians rely on the description of a voice in terms of a number of settings or perceptual dimensions: they are thus featural or componential analyses. How to handle the holistic-featural dichotomy is still a challenge, and more investigations are needed to explore in relation to how both perceptual approaches correlate or interact.

Laver's Vocal Profile Analysis protocol (VPA, [1]) is perhaps the most widely used analytic method whose components are referred to as 'settings', defined as long-term tendencies of the vocal apparatus to adopt a particular configuration [12][13]. Recent studies show growing interest in VQ – from an auditory perspective – for forensic purposes [14][15][16]; most using VPA or a simplified version of it. Despite the popularity of featural approaches, much remains to be explored as regards holistic judgments of voice quality made by lay listeners. This paper aims to fill this gap by looking at the role played by non-featural perception of VQ by naïve listeners, and to explore the language independence hypothesis of this holistic approach when judging speaker similarity. Some recent studies have explored lay perceptions of voice similarity, but without a focus on VQ. For instance, using Multidimensional Scaling (MDS) [17] proposes a method for assessing the degree of perceived similarity among a group of speakers for potential inclusion in voice parades. In [18] acoustic correlates are investigated for the perceptual dimensions obtained in the MDS analysis; and in [19] voice similarity judgements are found to depend on the accent background of the listener. Preliminary correlation results seem to show that different phonetic features contribute to the perceived similarity ratings for the two accents.

There are fewer studies focusing on the language dependency of VQ perceptual assessment. Most previous studies on native language effects in voice identification tend to suggest that native listeners have an advantage over non-natives [20][21]. Other investigations, however, fail to support this claim: [22] found that although identification improves the larger the phoneme repertoire in the voice sample, it is still possible to identify voices successfully when stimuli are random phonemes with no meaning and not belonging to any language. It can be then hypothesized that listeners pay attention to cues in a voice which do not require knowledge of the speaker's language, for instance suprasegmental aspects. Ho [23] found no native language effect when comparing British English and Chinese listeners in a speaker identification task where F0 was modified; listeners responded to the stimuli differently regardless of their L1, suggesting that F0 is a language-independent factor for voice identification.

In this study we focus on a different suprasegmental aspect (VQ), but the scope of the investigation differs from the above-mentioned studies in that we are not conducting identification tests or same-different tests. Instead, listeners are asked to rate speaker similarity, so that their ratings can be used as input to a MDS analysis in order to explore listeners' perceptual representations of very similar speakers. That is the reason why we selected a cohort of same-age, same-dialect speakers, with similar F0. Stimuli pairs belonging to monozygotic twins (MZ) were included as they represent extreme examples of voice similarity. Johnson & Azara [24] suggest that twins "serve as a unique control population for studies of the perception of personal identity". An important limitation of [24] is the heterogenous nature of the subjects (5 MZ twins and 1 dizygotic pair) with very different ages (20-67) and dialects. The first two dimensions of MDS solution in [24] correlated with age and dialect correspondingly. In our experiment a larger twin population is used, but most importantly age and dialect are controlled. Perceived speaker similarity is predicted to be explained solely by VQ characteristics, assessed holistically in a very similar way by native and non-native listeners.

2. Materials and method

2.1. Subjects

Five pairs of male MZ twins were selected from the corpus collected in [25]. All were native speakers of Standard Peninsular Spanish, and none reported any voice pathology. Three criteria were established in order to select only the most similar-sounding twin pairs from the corpus: (i) *similar age* (mean: 21, sd: 3.7); (ii) *similar mean F0* (mean: 113 Hz, sd: 13 Hz); and (iii) *similar Euclidean distance (ED)* between each speaker and his twin. EDs were based on the perceptual assessment of their VQ using a simplified version of the VPA scheme [26] by a trained phonetician (Author 1).

Table 1. *VPA speaker evaluation and Similarity Matching Coefficient (SMC) per twin pair. VT: vocal tract*

	VPA settings										
Speaker	**Labial**	**Mandibular**	**Lingual Tip**	**Lingual body**	**Velopharyngeal**	**Pharyngeal**	**Larynx Height**	**VT tension**	**Larynx tension**	**Phonation type**	**SMC**
AGF	0	0	0	0	0	0	2	1	1	1	
SGF	0	1	0	0	1	0	2	1	1	1	
Match	1	0	1	1	0	1	1	1	1	1	*0.8*
AMG	0	1	0	2	2	2	1	1	1	0	
EMG	0	1	0	2	0	2	1	1	0	0	
Match	1	1	1	1	0	1	1	1	0	1	*0.8*
ASM	1	0	0	0	1	0	2	0	1	1	
RSM	1	0	0	0	1	0	2	2	2	0	
Match	1	1	1	1	1	1	1	0	0	0	*0.7*
ARJ	2	2	1	1	0	1	2	2	1	0	
JRJ	1	2	0	0	0	0	2	2	2	0	
Match	0	1	0	0	1	0	1	1	0	1	*0.5*
DCT	0	2	0	2	0	2	0	2	1	1	
JCT	0	0	0	0	2	2	2	0	1	1	
Match	1	0	1	0	0	1	0	0	1	1	*0.5*

The simplified VPA consists in a reduced number of perceptual dimensions (10 settings) where each one is reduced to a continuum with three possibilities: neutral setting (labelled as 0) and two non-neutral possibilities (labelled as 1 and 2), typically going in opposite directions (e.g. labial setting: spread-neutral-round).

The EDs between twins were 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 (Table 2). Mean SMC for all twins was 0.66, indicating that around 6 VQ settings were shared on average by the twin pairs.

2.2. Stimuli and listeners

2.2.1. Stimuli

Voice samples (~3 secs.) were extracted from semi-directed spontaneous conversations ([25]), held by the 10 twins individually with Author 1. The interlocutor is therefore controlled, resulting in the same type of speaking style in all conversations. All utterances were declarative sentences of different linguistic content (diverse neutral topics).

2.2.2. Listeners

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) took part in the perceptual experiment. They were recruited at universities in Spain and England, and none reported any hearing difficulty.

2.3. Design of perceptual test

A Multiple Forced Choice experiment was set up in Praat [26] 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 (very similar) to 5 (very different). Listeners were not told that the stimuli included twin pairs. The test was run on a PC with headphones in a silent room. A short pre-test with four voices (also twins but different ones) allowed familiarization with the test. Reaction times were measured from the end of the second stimulus. The test duration was approximately 15 minutes with a short break every 30 stimuli.

2.4. Analysis method

2.4.1. Multidimensional Scaling

The degree of perceived similarity was visualized using Multidimensional Scaling (MDS), a means of detecting meaningful underlying dimensions that explain observed similarities or dissimilarities (distances).

2.4.2. Mixed-effects modelling

Ordinal mixed effects modelling (MEM) was used to fit models to the similarity ratings using the *Ordinal* package in R [28]. The following fixed effects (predictors) were tested:

- Listener language – Spanish or English
- Similarity matching coefficient (SMC) – between the speakers in the target trial
- Reaction time
- Twins – whether speakers were twins or not

Random intercepts were fitted for listener and trial (i.e. target speaker comparison). The first model tested for the effect of

language on the similarity ratings provided by listeners. A step-up approach was then adopted whereby predictors and interactions were added iteratively and models compared using ANOVAs. Model comparison was conducted in order to assess the best fit to the data.

3. Results

3.1.1. Multidimensional Scaling

MDS analyses were conducted using the similarity scores. The relative magnitude of the sorted Eigenvalues indicates that seven dimensions would be necessary to accurately reproduce between-speaker distances in the perceptual space for both English and Spanish listeners (stress: 0.03 for Spanish listeners; 0.07 for English listeners). However, MDS results are typically visualized using only the first 2 or 3 dimensions. Figures 1-2 show MDS plots for Spanish and English listeners respectively using 2 dimensions (stress: 0.8). Each point represents the location of a speaker in the listeners' perceptual space. 3D models showed an important drop in stress (0.4). Table 2 shows the normalized intra-pair Euclidean distances taking into account the seven dimensions in which listeners' ratings seem to be based.

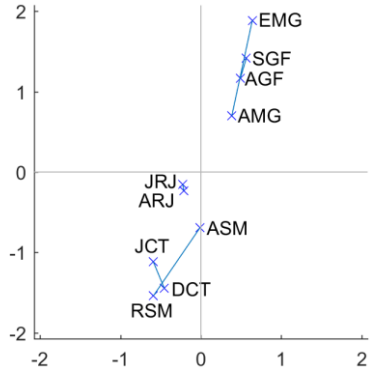

Figure 1: *MDS 2D plot (Spanish listeners)*

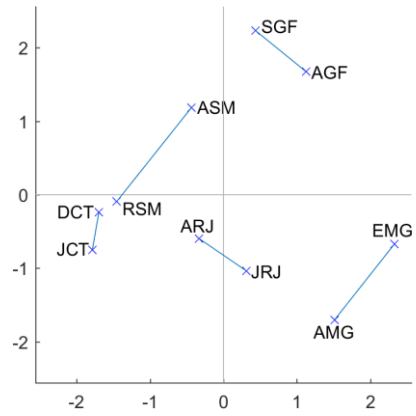

Figure 2: *MDS 2D plot (English listeners)*

Table 2: *Normalized intra-pair Euclidean distances based on seven perceptual dimensions*

speakers → / listeners ↓	AGF SGF	DCT JCT	ARJ JRJ	ASM RSM	AMG EMG
Spanish	0.341	0.343	0.345	0.369	0.607
English	0.264	0.219	0.349	0.435	0.445

3.1.2. Mixed-effects modelling

The best model fitted to the data based on model comparison incorporated all fixed effects and interactions between fixed effects. Significant interactions were found between language and both reaction time and whether the target trial contained a twin pair or not. For English listeners, similarity ratings were not affected by reaction time. That is, listeners were no more likely to judge speaker pairs as being very similar or dissimilar as a function of reaction time. However, Spanish listeners were more likely to give a rating of 5 ('very dissimilar') if reaction time was short, and a rating of 1 ('very similar') if reaction time was longer. This is of special interest if we consider that average reaction times were very similar for Spanish (mean: 0.82 secs; SD: 0.14) and English listeners (mean: 0.84; SD: 0.18).

A number of language independent effects were also found. Across all listeners, twin pairs were rated as being more similar to each other (ratings closer to 1) than non-twin pairs. Twin pairs with low SMC values (i.e. those who are objectively more similar to each other) were also rated as being more similar than twin pairs with high SMC values. For non-twin pairs, listeners did not rate speaker pairs with higher or lower SMC values as being more or less similar. Finally, for twin pairs all listeners were more likely to respond with 1 (i.e. very similar) if reaction time was short. Conversely, for non-twin pairs, all listeners are more likely to respond with 5 if they respond quickly. If they did not respond quickly, reaction time was no predictor of similarity rating.

4. Discussion

MDS analyses show that the optimal configuration to visualize speaker distances would require a 7-dimensional space (lowest possible positive stress value). 2D plots are therefore poorer representation of the data, reflected in a high stress value. This confirms what is well known for VQ: its high multidimensionality. A thorough understanding of perceptual judgements by naïve listeners require other types of analyses, and that is why MEM was conducted. Even though we cannot explain listener decisions with only two dimensions, we still find similar trends in both listener groups, like extreme closeness of speakers DCT and RSM. When we look at their featural VQ analysis (Table 1), the only setting that they share relates to vocal tract tension, possibly evidencing the higher salience of this setting.

Since seven dimensions seem to best represent listeners' perceptual space, we ordered twin pairs from most to least similar and the same ranking appears in both listener groups: AGF-SGF and DCT-JCT being the closest twin pair with slight differences in their normalized intra-pair Euclidean distances; ASM-RSM and AMG-EMG consistently appearing as the least similar pairs for both listener groups. Looking in detail at Table 1, we find that settings shared by AGF-SGF and DCT-JCT relate to the larynx (laryngeal tension and phonation types). Supralaryngeal matches (e.g. labial and lingual tip) are due to shared neutral settings, which should probably not weigh in the same way as matches due to deviations from neutrality in future SMC calculations. In contrast, matches for ASM-RSM and AMG-EMG (discarding matches based on shared neutral settings), are only supralaryngeal matches. These seem not to be so salient for naïve listeners from a holistic perspective, as these twin pairs are consistently far apart in the listeners' perceptual space. This finding seems to point to the same cue prominence by all naïve listeners, i.e. regardless of language familiarity or understanding of the linguistic content.

Equivalent reaction times suggest similar listening strategies ('gut' impressions; holistic VQ perception). However, qualitative feedback from participants suggest that other cues, mostly rhythmic aspects (e.g. speaking rate) may have contributed to perceived similarity as well. These deserve future investigations, as they are also suprasegmental features, apparently also salient even in short stimuli and possibly independent of the listener L1.

Mixed effects modelling revealed a number of effects involving language, although no clear indication of different listening strategies across groups. Significant effects involved reaction time, indicating that, for certain target pairs, similarity ratings are different for English and Spanish listeners depending on how long it took to make the decision. Notably, statistical modelling did suggest a number of language independent factors. Most notably, twin pairs were rated as being more similar to each other than non-twin pairs irrespective of listener language.

5. Conclusions

It is well known that multiple factors affect unfamiliar naïve recognition, from individual listener ability to the distinctiveness of the speaker's voice; the contribution of the latter not being fully understood. In order to explore in which VQ aspects speaker distinctiveness may lie, we have designed a perceptual test where Spanish and English listeners had to rate speaker similarity in pairwise comparisons. Results have shown that when other linguistic cues are suppressed –because of short stimuli– native and non-native listeners rate speaker similarity in a very similar way. Using short speech samples makes it difficult for listeners to base their similarity judgments in other aspects which are not VQ. Although these results should not be extrapolated to earwitness evidence with different characteristics and the native advantage may still hold true in situations where listeners are exposed to longer speech samples, this investigation has aimed to explore the role of VQ holistic perception, which seems to be the resource available for lay listeners to judge speaker similarity at least in a homogenous population of same-accent, same-age, similar-sounding speakers. Future investigations will look further at interrelationships between naïve holistic VQ and the featural decomposition of VQ by expert listeners, as the latter reveals speaker similarities for specific settings that do not appear to be salient in the holistic perception of VQ or at least do not have a strong weight in the similarity ratings made by naïve listeners.

6. Acknowledgements

This research was funded via the UK AHRC grant *Voice and Identity – Source, Filter, Biometric* (AH/M003396/1). We thank Olaf Köster and Jose Antonio Mompean for their helpful insights as well as Juana Gil for her assistance in the recruitment of Spanish listeners.

7. References

[1] Laver, J. The Phonetic Description of Voice Quality, Cambridge University Press, 1980.

[2] Webb, A. L., Carding, P. N., Deary, I. J., Mackenzie , K., Steen, N., Wilson, J. A. "The reliability of three perceptual evaluation scales for dysphonia", European Archives of Otorhinolaryngology, 261:429–434, 2004.

[3] Esling, J. H. & Wong, R.F. "Voice quality settings and the teaching of pronunciation", TESOL Quarterly 17:89–95, 1983.

[4] Stuart-Smith, J. "Glasgow: accent and voice quality", in P. Foulkes & G. Docherty [Eds], Urban Voices, 203-222, Arnold, 1999.

[5] Esling, J. H. Voice quality in Edinburgh: a sociolinguistic and phonetic study. PhD dissertation, University of Edinburgh, 1978.

[6] Laver, J. Principles of Phonetics, CUP, 1994.

[7] Foulkes, P., & French, P. "Forensic speaker comparison: a linguistic-acoustic perspective", in L. Solan & P. Tiersma [Eds], The Oxford Handbook of Language and Law, 418-421, Oxford University Press, 2012.

[8] Rose, P. "Forensic speaker identification", CRC Press, 2003.

[9] Kreiman, J. & Gerrat, B. "Comparing two methods for reducing variability in voice quality measurements", Journal of Speech, Language and Hearing Research, 54:803-812, 2011.

[10] Kreiman, J. & Sidtis, D. Foundations of Voice Studies, Wiley-Blackwell, 2011.

[11] Gil Fernández, J. & San Segundo, E. "La cualidad de voz en fonética judicial", in E. Garayzábal, M. Jiménez & M. Reigosa [Eds], Lingüística forense. La lingüística en el ámbito legal y policial, 154-199, Euphonía Ediciones, 2013.

[12] Honikman, B. "Articulatory settings", in D. Abercrombie, D.B. Fry, P.A.D. MacCarthy, N.C. Scott, & J.L.M. Trim [Eds], In Honour of Daniel Jones, 73–84, Longman, 1964.

[13] Beck, J. "Perceptual analysis of voice quality: the place of Vocal Profile Analysis", in W.J. Hardcastle & J. Mackenzie Beck [Eds], A Figure of Speech: a Festschrift for John Laver, 285–322, Laurence Erlbaum Associates, 2005.

[14] González-Rodríguez, J., Gil, J., Pérez, R., & Franco-Pedroso, J. "What are we missing with i-vectors? A perceptual analysis of i-vector based falsely accepted trials", Proc. Odyssey 2014, 33-40, 2014.

[15] French, P., Foulkes, P., Harrison, P., Hughes, V., San Segundo, E. & Stevens, L. "The vocal tract as a biometric: output measures, interrelationships, and efficacy", Proc. 18th ICPhS, Glasgow, 2015.

[16] San Segundo, E., Hughes, V., French, P., Foulkes, P. & Harrison, P. "Developing the vocal profile analysis scheme for forensic voice comparison", BAAP Colloquium, Lancaster, 2016.

[17] McDougall, K. "Assessing perceived voice similarity using multidimensional scaling for the construction of voice parades", Int. Journal of Speech Language and the Law, 20:163-172, 2013.

[18] Nolan, F., McDougall, K. & Hudson, T. "Some acoustic correlates of perceived (dis)similarity between same-accent voices", Proc. 17th ICPhS, Hong Kong, 2011.

[19] McDougall, K., Hudson, T. & Atkinson, N. "Perceived voice similarity in Standard Southern British English and York English" UKLVC Conference, York, 2015.

[20] Perrachione, T.K., Pierrehumbert, J.B. & Wong, P.C.M. "Differential neural contributions to native- and foreign-language talker identification", Journal of Experimental Psychology: Human Perception and Performance, 35:1950-1960, 2009.

[21] Köster, O. & Schiller, N. O. "Different influences of the native language of a listener on speaker recognition", Forensic Linguistics, 4:8-28, 1997.

[22] Bricker, P. D., & Pruzansky, S. "Effects of stimulus content and duration on talker identification", Journal of the Acoustical Society of America, 40:1441-1449, 1966.

[23] Ho, C.-T. Is pitch a language-independent factor in forensic speaker identification?, MA diss., University of York, 2007.

[24] Johnson, K. & Azara, M. "The perception of personal identity in speech: evidence from the perception of twins' speech" Unpublished manuscript, 2000.

[25] San Segundo, E. Forensic speaker comparison of Spanish twins and non-twin siblings, PhD dissertation, Menéndez Pelayo International University & CSIC, 2014.

[26] San Segundo, E. & Mompean, J. "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, Murcia, 2016.

[27] Boersma, P., & Weenink, D. Praat: doing phonetics by computer [Computer software] (Version 5.3.79), 2012.

[28] Christensen, R.H.B., "Ordinal", R Package [Computer software] (v.3.3.0), 015.