PERCEPTION OF VOCAL TRACT TENSION: EXPLORING POSSIBLE PROSODIC CORRELATES

Eugenia San Segundo[1], Sandra Schwab[2], Volker Dellwo[2], Lei He[2] & José Mompeán[3]

[1]Department of Language & Linguistic Science, University of York, UK
[2]Institute of Computational Linguistics, University of Zurich, Switzerland
[3] University of Murcia, Spain
eugenia.sansegundo@york.ac.uk

ABSTRACT

A recent study involving the perceptual analysis of 24 speakers by two raters (San Segundo & Mompeán 2017) revealed a slight inter-rater agreement in the assessment of vocal tract tension (VTT). In the current investigation several prosodic measures related to intensity and durational variability have been extracted per speaker with the aim of testing whether they correlate with the perceptual ratings for VTT provided by the two trained raters. The correlation test showed a significant positive correlation between the ratings of Rater 1 and the variable *varcoM* (mean intensity variability across syllables). In contrast, the ratings of Rater 2 correlated positive and significantly with two rhythmic measures related to mean consonant duration. These results suggest that the acoustic cues playing a role in each rater's auditory judgements are not the same. The different salience of intensity and durational characteristics should be taken into account in future studies on voice quality perception.

Keywords: voice quality, perception, vocal tract tension, prosodic correlates, forensic phonetics

1. INTRODUCTION

Voice quality (VQ) is the characteristic timbre or quasi-permanent quality resulting from a combination of long-term laryngeal and supralaryngeal features which make a speaker's voice recognizably different from others (Laver 1980). A considerable number of protocols have been proposed for describing a speaker's VQ; some of the best known are the GRBAS scale (Hirano 1981), the Buffalo III Voice Profile (Wilson 1987), the Consensus Auditory Perceptual Evaluation (CAPE-V) (Kempster et al. 2009) or the Stockholm Voice Evaluation Approach (SVEA) (Hammarberg 2000). These protocols require the listener to rate several VQ features using different types of ratings, categorical interval scales or visual analogue scales. Clinical voice therapy has extensively applied such formal protocols by expert/trained listeners. Other phonetic applications include forensic phonetics, a discipline that applies phonetic knowledge to legal issues, for instance in Forensic Speaker Comparison. This consists in the comparison of voice samples belonging to an offender and a suspect in order to assist courts in determining speaker identity. In this research area the most commonly used protocol is the Vocal Profile Analysis (VPA). Some recent forensic investigations delving into the potential of the VPA for speaker characterization are French et al. (2015), Hughes et al. (2017), San Segundo & Mompeán (2017) or San Segundo et al. (under revision).

In the VPA scheme the analytic unit defining a speaker's VQ is the 'setting', or long-term articulatory, phonatory and muscular tendency. In the version of the protocol described in San Segundo et al. (under revision), there are 32 settings: 21 describe vocal tract (supralaryngeal) features, seven describe phonation features and four describe overall muscular (laryngeal and vocal tract) tension features. As far as the rating of settings is concerned, each VPA setting is described as a deviation from a clearly defined 'neutral' or standard condition.

The present investigation has focused on one setting in particular: vocal tract tension (VTT), as previous studies have shown that experts frequently disagree on their ratings when assessing this voice feature perceptually. In San Segundo & Mompeán (2017), inter-rater agreement between two independent analysts (judges or raters) for this setting was only 42% (raw percent agreement), 0.13 κ using unweighted Cohen's kappa and 0.21 κ using linear-weighted kappa. In contrast, the same study shows very good intra-rater agreement results; according to these, internal consistency in VTT ratings is very high, comparing two rating sessions by the same judge: Rater 1 achieved 87.5% agreement (unweighted κ = 0.81) and Rater 2 achieved 95.83% agreement (unweighted κ = 0.91). These figures suggest that this perceptual dimension is salient – internal standards for the setting are clear within a rater– but different raters seem to pay attention to different cues in order to inform their auditory ratings, as they do not always converge in their ratings. A search for precisely defined acoustic correlates for VTT is therefore justified, as they could help trained

experts achieve better agreement in their auditory evaluations.

2. RESEARCH OBJECTIVE

Our objective was to find if there are prosodic correlates of lax and tense vocal tract perceptual ratings. In particular, we aimed to test the following impressionistic correlations (described in Beck 2007 and San Segundo et al. under revision, but not tested empirically yet, to the best of our knowledge):

- *lax vocal tract* is associated with slower tempo; phonetic undershoot; both consonantal and vocalic reduction; and higher incidence of unstressed syllables.

- *tense vocal tract* is associated with faster tempo; precise/full consonantal articulation; and lower incidence of unstressed syllables.

3. MATERIALS & METHODS

3.1. Materials

Two types of materials were necessary:

(a) The voice recordings of 24 male speakers (aged 20-36, speakers of Standard Peninsular Spanish), belonging to the Twin Corpus described in San Segundo (2013, 2014). The voice samples are spontaneous conversations held between the participant and the first author (ca. 10 min).

(b) The perceptual ratings provided by two expert raters who assessed the 24 speakers using a simplified version of the VPA protocol (San Segundo & Mompeán 2017) on a scale from more lax (-1) to more tense (1), with a midpoint for neutral (0).

3.2. Methods

3.2.1. Transcription and semi-automatic alignment

The speech samples of the 24 speakers were transcribed and aligned using *EasyAlign* (Goldman 2011; Goldman & Schwab 2014), with minimal manual correction. The acoustic analyses that followed required that each voice sample be transcribed and segmented at the phonetic and syllable level. A CV tier was necessary where each phone was classified as C (consonant) or V (vowel). For the elongation-hesitation variables, a further manual detection of elongated consonants and syllables was carried out.

3.2.2. Acoustic analyses

(a) Articulation rate: It was measured in syllables per second using a script developed ad hoc by Sandra Schwab. As in Spanish a syllable must contain one of these vowels: [a, e, i, o, u], the script counted the number of such vowels per inter-pause (IP) stretch, which was divided by the duration of the IP stretch. This gives the average articulation rate per speaker.

(b) Rhythmic measures: These basically relate to the variability and proportion of duration between consonant and vocalic segments (*%V*, *nPVI-V*, *VarcoC*, *VarcoV*, among others) and were extracted using a script developed by Volker Dellwo (see Dellwo et al. 2015).

(c) Intensity measures: These relate to variability in the average intensity (*meanM*, *stdevM* and *varcoM*) and in the peak intensity (*meanP*, *stdevP* and *varcoP*) of each syllable. They were extracted using a script developed by Lei He (He & Dellwo 2016).

(d) Elongation/Hesitation variables: These measures were envisaged ad hoc for this study and extracted using a script developed by Eugenia San Segundo. They refer to the total number of elongated sounds (vowels and consonants) per second, as well as their relative duration within the corresponding IP stretch.

3.2.3. Statistical analyses

After the extraction of the acoustic features, we carried out a correlation test (*IBM SPSS Statistics v.24*) aimed at testing whether the individual ratings of Rater 1 and Rater 2 correlate with some or any of the prosodic-acoustic features previously extracted.

4. RESULTS

We analysed whether there was a relationship between any of the prosodic measures described above and the VTT ratings of each rater. Table 1 summarizes the results of the correlation test using Spearman's rho ($n = 24$). Only significant results are reported here.

The results show that there is a positive correlation between the ratings of Rater 1 and *varcoM* (mean intensity variability across syllables), $r = 0.344$, $p < 0.05$ (1-tailed).

As for Rater 2, the results show that there is a positive correlation between his ratings and two variables: *meanConLn* (mean duration of Ln normalized consonant durations), $r = 0.362$, $p < 0.05$ (1-tailed), and *meanCLn* (mean duration of Ln normalized C interval durations), $r = 0.395$, $p < 0.05$ (1-tailed).

We also found a negative correlation between the ratings of Rater 2 and the following variables: *nCV* (number of C or V intervals) $r = -0.358$, $p < 0.05$ (1-tailed); *varcoSyL* (coefficient of variation of *deltaSyl*), $r = -0.396$, $p < 0.05$ (1-tailed); *deltaSylLn* (standard deviation of ln normalized syllable

durations), r = -0.464, p < 0.05 (1-tailed); and *nPVI-Syl* (normalized Pairwise Variability Index of Syllable durations), r = -0.490, p < 0.01 (1-tailed).

Figure 1 shows a boxplot representation for the variables that yielded the strongest and most significant correlation: *nPVI-Syl* and the perceptual ratings of Rater 2. Despite the fact that fewer speakers fall within the 'tense' category, we can still observe that there is a correspondence between 'lax' ratings ('-1' in Figure 1) and higher variability of (normalized pairwise) syllable durations.

Table 1: Significant correlation coefficients (r) between the perceptual ratings and the prosodic measures.

Prosodic measures	Ratings Rater 1	Ratings Rater 2
varcoM	0.344*	
meanConLn		0.362*
meanCLn		0.395*
nCV		-0.358*
varcoSyL		-0.396*
deltaSylLn		-0.464*
nPVI-Syl		-0.490**

Note: * p<.05, ** p<.01

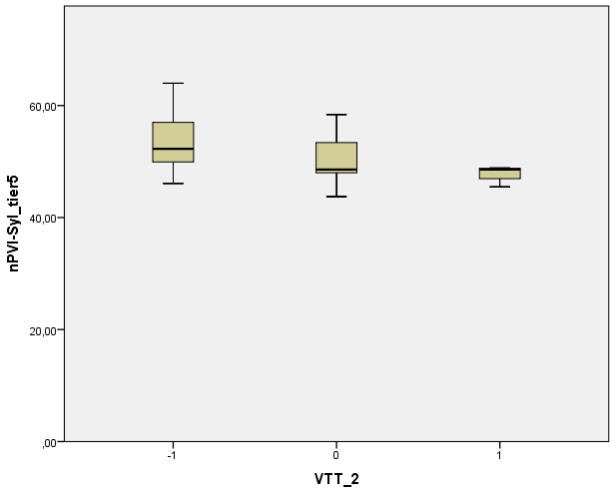

Figure 1: Boxplot showing the distribution of values for *nPVI-Syl* along the ratings of Rater 2 for vocal tract tension (VTT): -1 (lax), 0 (neutral) and 1 (tense).

5. CONCLUSIONS

The obtained results obtained represent a first approach to the quantitative characterization of the voice quality parameter VTT from a prosodic-acoustic point of view. The fact that different correlations were found between a range of prosodic measures and the perceptual ratings of two raters suggest that different listeners pay attention to different cues when evaluating a speaker's VQ. These results seem to explain the low inter-rater agreement found between Rater 1 and Rater 2 in a previous study (San Segundo & Mompeán 2017). The auditory ratings of the former –for this particular setting– seem to be more dependent on intensity variability across syllables than on any other rhythmic variables. The perception of tension by Rater 2, however, is associated with durational variability across syllables. This points to different listening strategies or diverse internal standards as regards how to define this VQ setting and with which type of voices to associate it.

Future studies will explore whether non-prosodic acoustic cues (e.g. those related to the long term average spectrum of the speaker) may also play a role in the auditory ratings given by both raters. All in all, the results imply that better inter-rater agreement could be reached in future perceptual studies if raters take into account this lack of shared acoustic relevance and the fact that the same auditory dimension can mean different things to different listeners. Through preliminary calibration meetings with a small number of voices, aimed at testing these aspects, a redefinition of VTT could be attempted before undertaking the analysis of a larger set of voices. This is one of the methodological proposals described in San Segundo et al. (under revision).

The potential of temporal variables for the study of VQ has been revealed throughout this study. This is an area of research that has seldom been explored, as indicated by Freitas et al (2015). Commonly analyzed measures include jitter, shitter or HNR dB. However, these authors point out that "there are timing and spectral peculiarities of the signal that should not be ignored" (Freitas et al. 2015: 5). Notwithstanding, it is well-known in VQ studies of this sort that direct and unique perceptual-acoustic relations cannot be established.

6. ACKNOWLEDGEMENTS

This research was funded partly via an International Short Visit Grant of the Swiss National Science Foundation (IZK0Z1_173307) and partly via the UK AHRC grant Voice and Identity – Source, Filter, Biometric (AH/M003396/1). A special thanks to Paul Foulkes and Peter French (University of York) for extensive discussions about the VPA settings.

7. REFERENCES

Beck, J. M. 2007. *Vocal profile analysis scheme: a user's manual*. Edinburgh: Queen Margaret University College-QMUC, Speech Science Research Centre.

Dellwo, V.; Leemann, A. & M-J. Kolly. 2015. Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *J. Acoust Soc Am*., 137(3):1513-1528.

Freitas, S. V.; Pestana, P. M.; Almeida, V., & A. Ferreira. 2015. Integrating voice evaluation: correlation between acoustic and audio-perceptual measures. *Journal of Voice* 29(3), 390-e1.

French, P.; Foulkes, P.; Harrison, P.; Hughes, V.; San Segundo, E. & L. Stevens. 2015. The vocal tract as a biometric: output measures, interrelationships, and efficacy. *Proceedings of the 8th International Congress of Phonetic Sciences (ICPhS)*, Glasgow, Scotland.

Goldman, J-P. 2011. EasyAlign: an automatic phonetic alignment tool under Praat. *Proceedings of Interspeech*, September 2011, Firenze, Italy.

Goldman, J.-P. & S. Schwab, S. 2014. EasyAlign Spanish: an (semi-)automatic segmentation tool under Praat. In Y. Congosto, M. L. Montero & A. Salvador (Eds.), *Fonética experimental, educación superior e investigación* (Vol. 1, pp. 629-640). Madrid: Arco/Libros.

Hammarberg, B. 2000. Voice research and clinical needs. *Folia Phoniatrica et Logopaedica* 52, 93-102.

He, L. & V. Dellwo. 2016. The role of syllable intensity in between-speaker rhythmic variability. *International Journal of Speech, Language and the Law* 23: 245-275.

Hirano, M. 1981. *Clinical examination of voice*. Vienna/New York: Springer Verlag.

Hughes, V.; Harrison, P. T.; Foulkes, P.; French, P.; Kavanagh, C. & E. San Segundo. 2017. Mapping across feature spaces in forensic voice comparison: the contribution of auditory-based voice quality to (semi-) automatic system testing. In *Proceedings of Interspeech*, Stockholm, pp. 3892-3896.

Kempster, G. B.; Gerratt, B. R.; Abbott, K. V.; Barkmeier-Kraemer, J. & E. H. Robert. 2009. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *American Journal of Speech-Language Pathology*, 18(2), 124-132.

Laver, J. 1980. *The phonetic description of voice quality*. Cambridge: Cambridge University Press.

San Segundo, E. 2013. A phonetic corpus of Spanish male twins and siblings: Corpus design and forensic application. *Procedia-Social and Behavioral Sciences*, 95, 59-67.

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*. PhD dissertation. Alicante: Biblioteca Virtual Miguel de Cervantes. http://www.cervantesvirtual.com/nd/ark:/59851/bmcm9293

San Segundo, E. & J.A. Mompeán. 2017. A Simplified Vocal Profile Analysis Protocol for the Assessment of Voice Quality and Speaker Similarity. *Journal of Voice* 31 (5), 644.e11 - 644.e27.

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

Wilson, D. K. 1987. *Voice problems of children*. Baltimore: Williams & Wilkins.

CITE THIS ARTICLE AS:

San Segundo, E.; Schwab, S.; Dellwo, V.; He, L. & Mompeán, J. (2017). Perception of vocal tract tension: Exploring possible prosodic correlates. In V. Marrero & E. Estebas (Coords.), *Current Trends in Experimental Phonetics: Cross-disciplines in the Hundredth Anniversary of "Manual de Pronunciación Española" (Tomás Navarro Tomás), Proceedings of the VII Congreso Internacional de Fonética Experimental* (pp. 79-82). Madrid: UNED.