VOICING ASSIMILATION IN WHISPERED SPEECH

Martin Kohlberger[1] and Patrycja Strycharczuk[2]

[1]LUCL, Leiden University; [2]CASL, Queen Margaret University
[1]m.kohlberger@hum.leidenuniv.nl; [2]PStrycharczuk@qmu.ac.uk

ABSTRACT

A large body of literature has shown that phonemic voicing contrasts are preserved in the production and perception of whispered speech. Nevertheless, it is unclear to what extent allophonic voicing is also maintained in whisper. The present study investigates whether a non-contrastive voicing distinction in Spanish fricatives – which results from voice assimilation in obstruent clusters – is also acoustically cued in whispered speech. In order to test this, a production experiment was conducted with 11 speakers of Peninsular Spanish. A number of acoustic cues relating to the fricatives in question and their surrounding phonological environment were measured. Four cues were found to be affected by voicing assimilation in normal phonation. Crucially, one cue (preceding vowel duration) was found to be affected by voicing assimilation in both normal and whispered phonation. These results show that non-contrastive voicing distinctions are also maintained in whispered speech.

Keywords: Assimilation, coarticulation, voicing, whisper, Spanish

1. INTRODUCTION

Whispered speech is defined by the absence of vocal fold vibration. Nevertheless, phonemic voicing contrasts have been found to be preserved in both the production [8,11,13,16,17,18,22,25,26] and the perception [4,7,18,25] of whispered speech. The studies cited show that voiced obstruents are differentiated from their voiceless counterparts by a range of cues, including decreased duration of burst/frication, increased duration of the preceding vowel, lowered F_1 following the burst, and decreased burst intensity.

Many of these cues are also used to signal voicing contrasts in normal phonation [5,19,23]. However, in normal speech their relative prominence is decreased due to the presence of other salient voicing cues pertaining to voice onset time and fundamental frequency. Whispered phonation has therefore been treated as a speech perturbation, where the role of various phonetic cues is rearranged but phonemic contrasts are preserved.

In addition to the acoustic investigations cited above, articulatory studies of whispered speech have also provided evidence for the contrast. Endoscopic data by Mills [17] confirm that laryngeal gestures are preserved in whispered speech: English speakers display a glottal aperture difference between voiced and voiceless sounds in both normal and whispered phonation. Interestingly, the same study goes further in showing that speakers differentiate between whispered voiced stops, produced with narrow glottal aperture, and whispered vowels, produced with intermediate glottal aperture, despite the fact that both types of sounds are voiced.

The last observation suggests that fine-grained articulatory distinctions – beyond those necessary to convey contrastive information – are maintained in whispered speech. This raises the question of whether acoustic details pertaining to non-contrastive voicing phenomena are also preserved in whisper. The present study addresses that question by examining various voicing-related cues in cases of allophonic voice assimilation (i.e. in non-contrastive environments) in normal and whispered speech in Spanish.

Spanish presents itself as an interesting test case for two reasons. First, it is a so-called 'true voicing' language in which the voicing contrast is primarily cued by the presence or absence of vocal fold vibrations. Secondly, Spanish exhibits voicing on both stops and fricatives, but the nature of the voicing is different in both types of consonants. In stops, voicing is contrastive. It should be noted that the voicing distinction in stops is also signalled by a change in manner of articulation: voiced stops are lenited to voiced fricatives in most non-initial environments. For fricatives, on the other hand, voicing distinctions are positionally determined: they are phonemically voiceless but they are subject to allophonic voicing in certain environments. When fricatives immediately precede voiced obstruents in a cluster, they undergo voicing themselves. This voicing has been treated as coarticulation rather than categorical assimilation in varieties of Peninsular Spanish because it has been found to be variable and gradient [12,21]. This provides an ideal environment to investigate whether non-contrastive voicing distinctions are maintained in whisper.

In this study, we investigate which acoustic cues signal allophonic voicing in Spanish fricatives in

normal speech, and we then determine whether any of those cues are maintained in whispered speech despite the non-contrastive nature of the voicing. The following section explains how the data were collected and analysed.

2. METHOD

2.1. Data collection

A production experiment was conducted with 11 native speakers of central Peninsular Spanish (10 females). All participants were born and grew up in northern and central regions of Spain, although they were all living abroad at the time of the recording. Participation was voluntary and speakers did not receive remuneration.

The participants were asked to read out test items which contained fricative-stop clusters of four different types, resulting in the four conditions shown in Table 1: two clusters with voiced stops, two with voiceless stops; two clusters across a word boundary and two within a word.

Table 1: Stimuli.

Condition	Example	Translation
V_1s#T	viaj**es p**agados	'paid trips'
V_1s#D	luc**es b**ajadas	'lowered lights'
V_1sT	seis **esp**añoles	'six Spaniards'
V_1sD	qué d**esb**alance	'what a lack of balance'

The critical fricative was always /s/. This fricative was chosen because of its high incidence in Spanish and due to the fact that its high frequency acoustic components make it easier to segment than other fricatives. Word size, stress and the phonological environment of the fricative (preceding vowel, following consonant and subsequent vowel) were controlled for. There were three test items per condition, and they were always embedded in a fixed carrier phrase: *Ahora digo …* 'I now say…'.

The recordings were made in a sound-attenuated room at two different sites. Six speakers were recorded in Leiden University using the Adobe Audition CS6 software, version 5.0.2. and the Roland Quad Capture UA-55 Audio Interface. The microphone was a Sennheiser MKH416T. Five speakers were recorded at Queen Margaret University Edinburgh on an Apple iMac, using Digidesign Pro Tools LE8 software and a Digidesign DIGI003 recording interface. The microphone was a Neumann U89i. The audio data were sampled at 44100Hz with a 16-bit depth.

For all the recordings, the speakers were positioned ca. 30cm away from the microphone. The participants read four repetitions of the experimental material out loud in normal speech and four repetitions in whispered speech. The test items were semi-randomised in blocks for each speaker (excluding immediate repetitions in neighbouring blocks) and presented on a computer screen, one at a time. The experiment was self-timed. The speakers were instructed to speak as naturally as possible. They were also encouraged to correct themselves if they made a mistake, by repeating the entire sentence.

Although 1056 tokens were recorded (12 items × 4 repetitions × 2 phonation types × 11 participants), data from three speakers had to be discarded because they used low-amplitude normal phonation instead of true whisper. In the end, 768 tokens were analysed.

2.2. Segmentation and measurements

The data were analysed using Praat version 5.3.59 [3] on a 5ms Gaussian window. The acoustic signal was segmented using EasyAlign for Spanish [9]. The boundaries for /s/ and its surrounding segments (preceding vowel, following vowel/consonant) were further inspected and adjusted manually by the second author to comply with the following segmentation criteria. We defined the onset and offset of /s/ as the onset and offset of frication visible in the region of 3-5 kHz (and higher). The onset of the vowel preceding /s/ (V_1) was placed at the onset of visible formant structure, and we used intensity transitions as an additional criterion for identifying the vowel onset. In a number of instances, the obstruent preceding the vowel was lenited and formant structure was visible during the obstruent. In such cases, we relied on intensity transitions alone to identify the vowel onset.

Based on the segmentation described above, the following cues were measured:

- V_1 (preceding vowel) duration (ms)
- C_1 (/s/) duration (ms)
- C_1 (/s/) voicing ratio (ms)
- spectral moments of C_1 (/s/):
    - centre of gravity
    - standard deviation
    - skewness
    - kurtosis
- V_1-C_1 intensity (difference between mean intensity of the /s/ and the preceding vowel, measured in dB):
    - in a low frequency band (50-500 Hz)
    - in a high frequency band (500-10000 Hz)

The measurements of the spectral moments of /s/ were based on time-averaged DFTs, using a script developed by Christian diCanio [6]. Although intensity-based measures are not standard in this type of study, we included the last two cues because similar measures have been previously shown to be relevant for voicing in some dialects of Spanish [10,24]. V_1-C_1 intensity was calculated by subtracting the mean intensity of the filtered portion of the consonant from the intensity of the preceding unfiltered vowel.

2.3. Statistical analysis

Statistical analysis was carried out in R [20] version 3.0.1. We analysed the individual measurements using linear mixed-effects regression modelling [2]. We included three main predictors in our modelling procedure, namely context (voiced C_2 vs. voiceless C_2), phonation type (normal vs. whisper) and presence of a word boundary (C_1C_2 vs. $C_1\#C_2$). We then checked for significant interactions between these predictors, using log likelihood comparison of nested models [1]. The interactions were only retained if they were found to significantly improve the model.

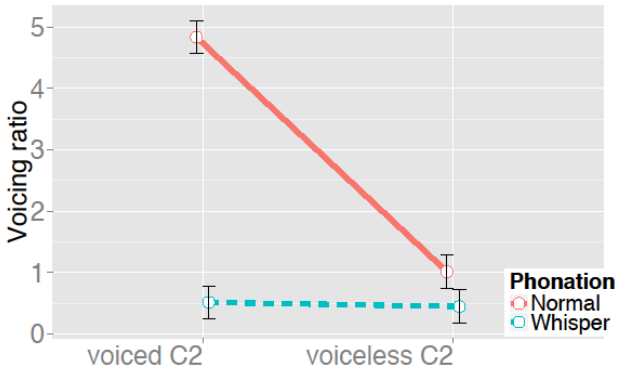
We hypothesise that, at least for normal speech, /s/ will undergo allophonic voicing when preceding a voiced stop. This should result in a significant effect of context (i.e. C_2 voicing). A significant interaction between context and phonation type would reveal that the effect of the voicing assimilation is different in normal speech compared to whispered speech. Finally, a significant interaction between context and the presence of a word boundary would imply that the allophonic voicing of fricatives is affected by the latter, or alternatively, by the position of a fricative within a word. In Section 3 below, we report the best model for each dependent variable selected using this procedure. The *p*-values we report were calculated using Satterthwate's approximations in the lmerTest package [15].

3. RESULTS

The results confirm that there is a significant main effect of C_2 voicing on the preceding fricative, but there was variation in how individual cues were affected. A number of cues (C_1 duration, standard deviation, skewness and kurtosis) did not vary significantly between conditions. Four other cues are indeed affected by C_2 voicing but they show an interaction between that variable and phonation type: they only signal voicing distinctions in normal phonation. These cues are C_1 voicing ratio, V_1-C_1 intensity (in the high and low frequency bands) and centre of gravity. Note that all of these cues are directly related to the presence of vocal fold vibration which increases intensity overall but especially in the lower frequencies, thereby decreasing the centre of gravity. It is thus not surprising that these cues are not affected in whispered speech.

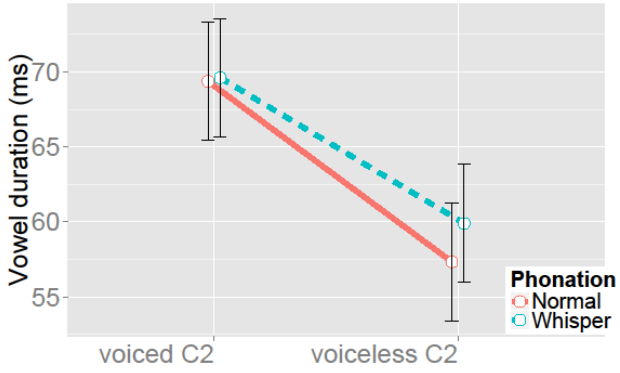
Figure 1 is an interaction plot for the variable voicing ratio. It shows that voicing ratio is only higher in assimilation contexts when the phonation type is normal. As would be expected, there is no change in voicing ratio in whispered phonation because there is no voicing at all in whisper.

Figure 1: Interaction plot for voicing ratio.



Finally, there is one cue which shows a significant effect of C_2 voicing, and no interaction between that and phonation type: V_1 duration is increased when C_2 is voiced (β=-10.88, *SE*=2.36, *t*=-4.61, *p*=0.001), and there is no model improvement for it with an added interaction between C_2 voicing and phonation type (ΔLL=1.2, *p*=0.2). In other words, the duration of the preceding vowel is equally affected by voicing assimilation in both normal and whispered speech. The interaction plot in Figure 2 shows that V_1 duration increases as an effect of C_2 regardless of phonation type.

Figure 2: Interaction plot for V_1 duration.



A summary of the way that all the cues are affected by C_2 voicing can be found in Table 2.

Table 2: Effect of C_2 voicing on individual cues.

No effect	In normal speech only	In normal and whispered speech
C_1 duration	C_1 voicing ratio	V_1 duration
standard deviation	V_1-C_1 intensity (low-band)	
skewness	V_1-C_1 intensity (high-band)	
Kurtosis	centre of gravity	

4. DISCUSSION

With regard to the allophonic voicing of fricatives in Spanish, the results of this study confirm what has been previously stated in the literature [12,21]. We find that C_2 voicing has a significant effect on a number of cues of an immediately preceding fricative: /s/ undergoes voicing in obstruent clusters when it is followed by a voiced stop. Our findings are also consistent with the hypothesis that the voicing of fricatives is variable and gradient in Peninsular Spanish, as evidenced by the relatively low voicing ratio of /s/ in assimilatory environments (0.48, *SD*=0.38).

Returning to our main research question: are acoustic cues of non-contrastive voicing maintained in whispered speech? Our results show that although most of the cues affected by C_2 voicing are found in normal phonation only, one cue, namely the duration of the preceding vowel, is also robustly maintained in whispered speech. This indicates that voicing distinctions are not only preserved when there is a phonemic contrast present (as shown by a large body of literature), but that they are also preserved when voicing is contextual and allophonic. This is in line with findings by Mills [17] that demonstrate that English speakers maintain differences in the articulation of different types of voiced sounds (voiced stops and vowels) in whisper despite the absence of phonological contrastiveness.

The results of this study also suggest that there are two different types of processes at work in signalling voicing distinctions in Spanish. In normal phonation, voicing assimilation was primarily implemented by the presence of vocal fold vibration. This is why acoustic cues that are directly related to voicing (voicing ratio, V_1-C_1 intensity and centre of gravity) were affected in normal speech. In whispered speech, however, true voicing is absent so these cues could not be utilised to mark the contrast.

However, in addition to the cues that are directly linked to the voicing gesture, there seems to be a specific gestural timing relationship in assimilated sequences that results in the extended duration of the preceding vowel of contextually voiced fricatives. This type of cue is unlike the others in that it is not an immediate (physiological and automatic) result of voicing, but rather one which requires fine speaker control. The extended duration of vowels before voiced consonants has been previously observed in the literature and has been explained as an intentional auditory enhancement of the perceptual effect of the voicing gesture [14]. This cue was indeed found to be preserved in whisper and was implemented in a way that is statistically indistinguishable from the way in which it is implemented in normal speech.

Teasing apart the mechanisms which underlie voicing might provide insight to the seemingly contradictory observation that Spanish voice assimilation appears to be phonetically gradient on the one hand, but under fine speaker control on the other. Future research should focus on the relationship between the different articulatory and acoustic signals employed and how they interact with gestural timing.

5. REFERENCES

[1] Baayen, R. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

[2] Bates, D., Maechler, M. 2009. lme4: Linear mixed-effects models using S4 classes. http://CRAN.R-project.org/package=lme4. R package version 0.999375-32.

[3] Boersma, P., Weenink, D. 2009. Praat: doing phonetics by computer. http://www.praat.org

[4] Dannenbring, G. L. 1980. Perceptual discrimination of whispered phoneme pairs. *Perceptual and Motor Skills* 51, 979-985.

[5] Denes, P. 1955. Effect of Duration on the Perception of Voicing. *Journal of the Acoustical Society of America* 27, 761-764.

[6] diCanio, C. 2013. Time-Averaging for Fricatives. Praat script. http://www.acsu.buffalo.edu/~cdicanio/scripts/Time_averaging_for_fricatives.praat

[7] Gilichinskaya, Y., Strange, W. 2011. Perception of final-consonant "voicing" in whispered speech. *The Journal of the Acoustical Society of America* 129, 2420-2420.

[8] Gilichinskaya, Y. D., Strange, W. 2008. Final consonant voicing and vowel height contrasts in whispered speech. *The Journal of the Acoustical Society of America* 123, 3459-3459.

[9] Goldman, J.P. 2011. EasyAlign: An automatic phonetic alignment tool under Praat. *Proceedings of Interspeech*, 3233-3236.

[10] Gradoville, M.S. 2011. Validity in Measurements of Fricative Voicing: Evidence from Argentine Spanish. In: Alvord, S.M. (ed), *Selected Proceedings of the 5th Conference on Laboratory Approaches to Romance Phonology*. Somerville, MA: Cascadilla Proceedings Project, 59-74.
[11] Higashikawa, M., Green, J. R., Moore, C. A., Minifie, F. D. 2003. Lip kinematics for /p/ and /b/ production during whispered and voiced speech. *Folia phoniatrica et logopaedica* 55, 17-27.
[12] Hualde, J. I. 2005. *The sounds of Spanish*. Cambridge: Cambridge University Press.
[13] Jovičić, S. T., Šarić, Z. 2008. Acoustic analysis of consonants in whispered speech. *Journal of Voice* 22, 263-274.
[14] Kluender, K. R., Diehl, R. L., Wright, B. A. 1988. Vowel-length differences before voiced and voiceless consonants: An auditory explanation. *Journal of Phonetics* 16, 153-169.
[15] Kuznetsova, A., Bruun Borckhof, P., Haubo Bojesen Christensen, R. 2013. lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). http://CRAN.R-project.org/package=lmerTest. R package version 2.0-0.
[16] Mills, T. I. P. 2003. *Cues to voicing contrasts in whispered Scottish obstruents*. Master's thesis.
[17] Mills, T. I. P. 2009. *Speech motor control variables in the production of voicing contrasts and emphatic accent*. Ph.D. thesis. Edinburgh: University of Edinburgh.
[18] Munro, M. J. 1990. Perception of `voicing' in whispered stops. *Phonetica* 47, 173-181.
[19] Peterson, G.E., Lehiste, I. 1960. Duration of Syllable Nuclei in English. *Journal of the Acoustical Society of America* 32, 693-703.
[20] R Development Core Team. 2005. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. http://www.R-project.org.
[21] Romero, J. 1999. The effect of voicing assimilation on gestural coordination. *Proc. 14th ICPhS* San Francisco, 1793-1796.
[22] Sharf, D. J. 1964. Vowel duration in whispered and in normal speech. *Language and speech* 7, 89-97.
[23] Stevens, K.N., Klatt, D.H. 1974. Role of formant transitions in the voiced-voiceless distinction for stops. *The Journal of the Acoustical Society of America* 55, 653-659.
[24] Strycharczuk, P., Van 't Veer, M., Bruil, M., Linke, K. 2014. Phonetic evidence on phonology–morphosyntax interactions: Sibilant voicing in Quito Spanish. *Journal of Linguistics* 50, 403-452.
[25] Tartter, V. C. 1989. What's in a whisper? *The Journal of the Acoustical Society of America* 86, 1678-1683.
[26] Zeroual, C., Esling, J. H., Crevier-Buchman, L. 2005. Physiological study of whispered speech in Moroccan Arabic. *Proceedings of Interspeech*, 1069-1072.