Acoustic [voice] correlate variation by dialect: data from Venezuelan Spanish

Stephanie Lain

The University of Texas at Austin

lain.steph@gmail.com

Abstract

The present study is an investigation of acoustic correlates corresponding to the category [voice] in two dialects of Venezuelan Spanish, one highland (Andean mountain dialect Mérida), the other lowland (Caribbean coastal dialect Margarita). In order to test what repercussions observed differences in consonant articulation may have on the acoustic correlates that encode [voice], a production experiment was run. The materials were 44 CV syllable prompts, analyzed with respect to the following: consonant closure duration, VOT, percent vocal fold vibration (%VF), root mean square amplitude (RMS), preceding vowel duration, CV duration ratio, F1 onset frequency, F0 contour, and burst.

1. Introduction

The feature [voice] is used contrastively in the consonant systems of most of the languages of the world [1]. At a basic level, [voice] describes the state of the vocal folds during the production of a given segment. [+voice] denotes vibration of the vocal folds throughout the duration of the segment. [-voice] is the absence of such vibration. However, the phonetic implementation of these categories is rarely absolute. For example, in perception, many acoustic cues (not just vocal fold vibration) contribute to the listener's categorization of a consonant as voiced or voiceless [2], [3], [4], [5], [6]. Similarly, in consonant production there is a long string of acoustic correlates that correspond to [voice], including (but not limited to): voice onset time (VOT), the presence or absence of a release burst, presence or absence of aspiration, duration of a preceding vowel, consonant vowel duration ratio (CV ratio), F1 onset frequency, F0 contour following closure, and relative amplitude.

Despite the near-universality of the voicing contrast in languages and the biological commonality of the gross gesture (engagement or no of the vocal folds), the initiation, maintenance and cessation of phonation depend on a subtle interplay of articulatory factors. [7] asserts that in this the phonatory system is fundamentally different from other sub-systems used in speech production, for in addition to the muscle-controlled adjustments of the vocal folds, aerodynamic conditions of the glottis (particularly the transglottal pressure), the intrinsic elasticity of the folds and tension held in these by the muscles of the larynx are all contributing factors. Even the slightest change during phonation in any of these parameters has the potential of altering the mode of vibration and hence the auditory quality of the sound produced. Much of this fine phonetic detail will be codified within the particular language system and, along with other nuances in articulation, will influence the quality of voicing, how voiced or voiceless consonants behave in different segmental environments, the nature of the category contrast itself [1]. In 1970, Lisker and Abramson noted: "In many languages some phoneme categories are distinguished by the timing of glottal adjustments relative to supraglottal articulation, and this timing relation determines not only the voicing state as narrowly defined, but the degree of aspiration and certain features associated with the so-called force of articulation as well" [8], p. 563.

Spanish is a language that displays considerable regional variety with regard to stop consonant articulation [9], [10]. In this case, the difference is of a fortis/lenis nature, where in highland (fortis) varieties of Spanish, occlusive realizations of phonological stops (particularly in [-voice]) are common and in lowland (lenis) varieties of Spanish, fricative or approximant realizations prevail, especially in intervocalic contexts.

The present study looks at the acoustic repercussions of dialectal variation in stop consonant voicing in Spanish, with the understanding that: 1) the phonetic realization of voicing is closely tied to the physical gesture and hence perturbed by small adjustments to the tension and timing of articulators, and; 2) the tension and timing of consonant articulation in voicing is language-specific, but not necessarily dialect-specific. The following question is posed: if a language displays substantial regional variation in stop consonant articulation, what will be the acoustic repercussions on the voicing system?

In particular I focus on the acoustic correlates of [voice] for two dialects of Venezuelan Spanish: the Spanish from Margarita Island (situated on the Caribbean coast of Venezuela) and the Spanish from the region of Mérida in the Venezuelan Andes. These dialects exhibit similar characteristics to other highland and lowland dialects that have been described in the literature [9]. The Mérida dialect (henceforth, MER) is known for its strong consonant closure. The Margarita dialect (henceforth, MAR) is a coastal variety of the type known for frequent fricative and/or approximant realizations of phonological stops.

2. Methodology

2.1. Subjects

Subjects were 25 adult monolingual speakers of Spanish between the ages of 20-35, with educational experience ranging between 1st grade and high school. Speakers in Margarita (10 females; 4 males) were recruited from a fishing village, El Tirano, which lies close to Playa el Agua. Speakers in Mérida (7 females; 4 males) were recruited from a town in the Venezuelan highlands, San Rafael de Mucuchíes. An effort was made to choose towns that were small, of roughly the same size, with mostly an indigenous population. All subjects were recruited on-site. No subjects had foreign language experience beyond that required in the public schools. All were paid volunteers. Language background and biographical information were assessed through a questionnaire administered verbally. None of the subjects reported hearing or speech problems.

2.2. Materials and procedure

Materials were 44 CV syllable prompts preceded by the word *son*. *Son* was included for two reasons: 1) to provide a word-initial (but not utterance-initial) context, and; 2) to elicit maximum contrast between the word-initial and word-medial positions. Occlusive pronunciations have been noted for both word-initial environments and after /n r l/ [10].

After the word *son*, each nonsense word that appeared began with a stop consonant and ended with a canonical Spanish vowel /a e i o u/. The distribution of prompts throughout the sample was as follows. All stop consonants appeared an equal number of times in the sample (namely, 15 times) and an equal number of times before each vowel (namely, 3 times). All prompts were randomized, both in the regular and in the training blocks. Prompts were viewed in a PowerPoint slideshow administered via laptop computer. The timing of screen changes was controlled by the investigator. During the recording session, subjects were asked to create an alternation, a wordplay whereby Target Syllable 1 (word-initial; appearing on the screen beside *son*) became the first syllable of a nonsense word. The second syllable of the nonsense word would be comprised of /r/ + the vowel in the first syllable. The last syllable (Target Syllable 2 [word-medial]) would be a repetition of Target Syllable 1. Stress would fall on the penultimate syllable (according to the default stress assignment for vowel-final words in Spanish). Therefore, if the prompt was "Son TO", subjects said, "Son toróto", "Son BU", "Son burúbu", etc. (a written accent was included for clarity. If the words containing target syllables were real words in Spanish instead of nonsense words, they would bear no written accent). In ten out of the fifteen times each stop consonant appeared, Target Syllable 1 (henceforth, TS1) was followed by a color (portrayed as a colored square) or a number. The particular colors and numbers used in the sample were chosen for their status as disyllabic trochee words ending in vowels, the most frequent word type in Spanish. Half began with either /p/, /t/, or /k/. The other half began with either /b/, /d/, or /g/. On screens where a color or number appeared to the right the target syllable, subjects were asked to say "Son toróto verde" or "Son burúbu cinco", etc. There was an even distribution of beginning /p t k/ and /b d g/ with respect to the target syllables.

The prompts were grouped into 4 blocks. Block 1 was a training block consisting of repetitions of *son* plus different target syllables. Block 2 was a second training block that introduced the prompts with colors and numbers. Block 3 was a combination of Blocks 1 and 2, with some prompts containing colors or numbers and others not. The screens in Blocks 1 and 2, as well as the first 10 screens of Block 3 were considered training slides and as such were not included in the measurements. Block 4 was a speeded trial in which subjects were asked to run through the prompts as quickly as possible while maintaining accuracy.

2.3. Acoustic and statistical analysis

Subjects were recorded with a Shure head-mounted dynamic microphone adjusted to the left corner of the mouth, at approximately ½ inch from the lips. Responses were recorded onto a compact flash card using a Marantz PMD 660 steady-state recorder. They were later transferred via a Macintosh G4 PowerBook onto an external hard drive. The sound files were recorded as WAV files at 48 kHz. All recordings took place in the field, in places that the subject and researcher agreed upon as being both amenable and relatively quiet. Most often, this was on the sidewalk outside the subject's residence or place of work.

Word-initial and word-medial target syllables from Blocks 3 and 4 were analyzed separately. Word-initial target syllables were analyzed with respect to the following: consonant duration, VOT, percent vocal fold vibration (%VF), and RMS amplitude (RMS), F1 onset frequency, F0 contour following closure and presence/absence of a release burst. Word-medial target syllables were analyzed for the same measures as in the word-initial contexts, with the inclusion of preceding vowel duration and CV duration ratio. Measurements were taken from a spectrogram display viewed in conjunction with the waveform and pitch track using WaveSurfer speech analysis software [11]. Settings in WaveSurfer were adjusted to view the display in a Hanning window. For duration measurements the bandwidth was set at 250 Hz, a value that is intermediate between the preferred values for male (200 Hz) and female (300 Hz) speakers [12].

Measurement procedures for most of these measures were according to convention and therefore require no special comment. However, I provide a description of two of the less common measures (%VF and RMS).

%VF was obtained following a procedure by [13]. The percentage of voiced frames was quantified by counting the number of glottal pulses throughout the closure gap and dividing this number by the duration in seconds.

RMS (taken from [12], [14], [15]) is a common measure of intensity, dependent on the amplitude of the sound wave as measured in decibels (dB). The RMS value is obtained by squaring individual amplitudes in a given time window, averaging these, then taking the square root of the average, producing a single value that the measure is applied to. By this procedure, the intensity of a sound relative to a given reference sound is calculated not by comparing the relative amplitudes but instead by comparing the relative powers of the two sounds (the power of a sound=the square of its amplitude). For the present study I chose to examine the unstressed CV window for both TS1 and 2. Three points were measured: 1) initial trough signifying amplitude low following closure; 2) peak representing maximum aperture of the vocalic gesture; 3) final trough signifying closure of the gesture and transition to the following segment.

Statistical analysis was performed with SPSS software, using a linear mixed model ANOVA and nesting tokens within subject. The model for the present study tested three fixed effects: *voicing category* ([± voice]), *dialect* (MAR, MER), and *condition* (unspeeded, speeded). The model also tested for interactions between *voicing category*dialect* and *dialect*condition*.

3. Results

In this section I report on the primary findings of the present investigation. A complete account of the results as well as more detailed discussion can be found in [16].

3.1. Approximant realizations

From the outset it is important to note that many of the target “consonants” in this sample might not be defined as such in a traditional sense. That is, formant structure (normally restricted to vowels, or at least to sonorants) appeared throughout a great many of the target consonant realizations. Where there was a visible dip in the waveform (evidence of a closing gesture), the segments were measured as consonants.

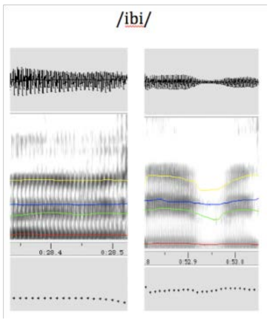

Figure 1: *MAR and MER /ibi/.*

Figure 1 shows an example of a VCV (/ibi/) waveform spectrogram slice for MAR (on the left) and MER (on the right).

3.2. Statistical analysis

3.2.1. Fixed effects

The fixed effect *dialect* showed significance (p≤.05) for RMS in initial and medial contexts (p=.004; p=.042). *Voicing category* was significant for consonant duration, %VF, RMS, F1 onset and burst in initial and medial contexts. In addition, *voicing category* showed significant values for preceding vowel duration and CV duration ratio in medial position. The *voicing category* values for VOT were significant in initial but not medial position (p=.000; p=.092). F0 contour failed to achieve significance in either initial or medial position (p=.996; p=.552). The fixed effect *condition* showed significant values for consonant duration in both initial and medial contexts (p=.025; p=.035). VOT was significant for condition in initial and in medial contexts (p=.045; p=.002). RMS was significant for *condition*, but in initial position only (p=.006).

3.2.2. Means

The following tables show the means for the fixed effect *voicing category* for four key measures in the study: RMS, consonant duration, VOT and %VF. These measures have been chosen to illustrate the main findings of the study.

Mean initial values with (standard error)		
RMS (in dB)	[-voice]	31.491(1.519)
	[+voice]	33.166(1.525)
consonant duration (in sec.)	[-voice]	.087(.005)
	[+voice]	.041(.005)
VOT (in sec.)	[-voice]	.022(.002)
	[+voice]	.007(.002)
%VF	[-voice]	33.626(2.786)
	[+voice]	94.693(2.899)

Table 1: *Mean initial values for RMS, consonant duration, VOT and %VF.*

Table 1 shows the [±voice] means for four measures. These results are consistent with previous findings that show longer durations associated with [-voice] segments and greater vocal fold vibration and amplitude with [+voice] segments. It is to be noted that the difference between [±voice] RMS is negligible. When the MAR and MER data are pooled, the higher RMS values for MAR increase the values for both [±voice].

Mean medial values with (standard error)		
RMS (in dB)	[-voice]	29.046(1.675)
	[+voice]	32.313(1.677)
consonant duration (in sec.)	[-voice]	.099(.005)
	[+voice]	.056(.005)
VOT (in sec.)	[-voice]	.020(.001)
	[+voice]	.017(.002)
%VF	[-voice]	38.036(4.016)
	[+voice]	88.290(4.025)

Table 2: *Mean medial values for RMS, consonant duration, VOT and %VF.*

Table 2 shows the means in medial position for the same measures as in Table 1. Of note is the lack of significance in VOT for [±voice].

3.2.3. Interactions

In initial position, the interaction of *voicing category*dialect* was significant for consonant duration and VOT (p=.000; p=.026).

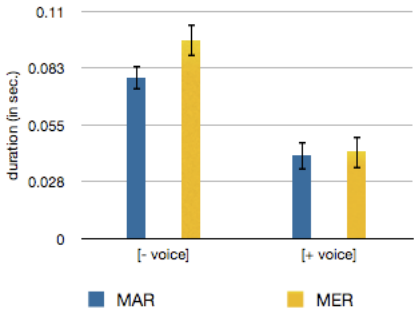

Figure 2: *Initial consonant duration interaction voicing category*dialect.*

Figure 2 shows the overall higher consonant duration values for MER and the greater separation of [±voice] values consistent for this dialect throughout the study. This was reflected in the VOT measure as well. In medial position, *voicing category*dialect* was significant for consonant duration, %VF, RMS, CV duration ratio and burst.

Figure 3 shows the difference in percentage of vocal fold vibration between the two dialects. This difference holds for the [-voice] but not the [+voice] category, where %VF values are roughly equivalent.

Figure 4 shows the elevated values of RMS in [±voice] in MAR by comparison with MER, particularly in the [-voice] category whose values in fact exceed the [+voice] values for MER.

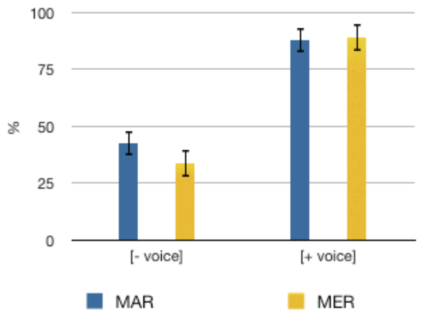

Figure 3: *Medial %VF interaction voicing category*dialect.*

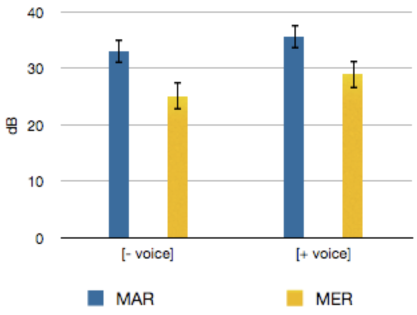

Figure 4: *Initial RMS interaction voicing category*dialect.*

4. Conclusions

The results of the study show that the statistical difference between the two dialects MER and MAR lies in the acoustic correlate RMS, with the MAR values being significantly higher in both [±voice]. All acoustic correlates (with the exception of F0 contour) showed significance for the fixed effect *voicing category*, with a clear separation between [±voice]. Consonant duration and VOT were sensitive to the rate of speech, as was RMS in the initial context.

Several of the acoustic correlates (consonant duration, %VF, RMS, CV duration ratio, burst) displayed significant interactions with respect to *voicing category*dialect*. Most of these observed significant interactions occurred in medial position. With the exception of RMS and %VF, MER values were consistently higher and the differences between [±voice] greater than in MAR. These results seem to indicate that although the acoustic correlate inventories of MER and MAR are not substantially different in content (same correlates correspond to [voice] in both dialects), the way that [±voice] categories relate to one another differs. The difference lies in the range of the [±voice] continuum, with a greater category separation for MER than MAR. [-voice] appears to be more affected than [+voice]. Acoustically, the dialects are differentiated by RMS, an amplitude measure corresponding to sonority and intensity.

One area where further investigation is needed is in determining how duration, sonority and intensity measures work together to cue both voicing and dialect information. In the present study, phonological information pertaining to the voicing contrast resided primarily in the domain of initial position. Sociophonetic information identifying a speaker as a member of a dialectal community emerged in medial context. The experimental design of the study precluded a direct comparison of initial with medial contexts, since the phonetic environments were not identical. While the correlation between the phonetic implementation of phonological contrasts with word position has been well studied in recent years, the relationship between phonological domain and dialectal or sociophonetic contrast remains an open line of inquiry.

5. Acknowledgements

I would like to thank the members of the scientific committee and the participants of the 2010 "Sociophonetics at the Crossroads of Speech Variation, Processing and Communication" conference for their constructive feedback on this project. All errors are my own.

6. References

[1] Ladefoged P. and I. Maddieson. 1996. *The Sounds of the World's Languages*. Malden: Blackwell Publishing.

[2] Denes P. 1955. Effect of duration on the perception of voicing. *Journal of the Acoustical Society of America* 27. 761-764.

[3] Diehl R. and D. Rosenberg 1977. Acoustic feature analysis in the perception of voicing contrasts. *Perception and Psychophysics* 21(5). 418-422.

[4] Lisker L. 1986. 'Voicing' in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech* 29, 1. 3-11.

[5] Haggard M., Q. Summerfield and M. Roberts. 1991. Psychoacoustical and cultural detriments of phoneme boundaries: Evidence from trading F0 cues in the voiced voiceless distinction. *Journal of Phonetics* 9. 49-62.

[6] Whalen D., A. Abramson, L. Lisker and M. Mody. 1993. F0 gives voicing information even with unambiguous voice onset times. *Journal of the Acoustical Society of America* 93 (4). 2152-2159.

[7] Hoole P., C. Gobl and A. Chasaide. 1999. Laryngeal coarticulation. In Hardcastle W. and N. Hewlett (eds.). *Coarticulation: Theory, Data and Techniques*. Cambridge: Cambridge University Press. 105-143.

[8] Lisker L. and A. Abramson. 1970. The voicing dimension: Some experiments in comparative phonetics. In *Proceedings of the 6th International Congress of Phonetic Sciences*. Prague: Academia. 563-567.

[9] Lipski J. 1994. *Latin American Spanish*. New York: Longman.

[10] Hualde J. 2005. *The Sounds of Spanish*. Cambridge: Cambridge University Press.

[11] Sjölander K. and J. Beskow. 2000. WaveSurfer - An open source speech tool. In Yuan B., T. Huang and X. Tang (eds.), *Proceedings of ICSLP 2000, 6th International Conference on Spoken Language Processing*. Beijing. 464-467.

[12] Ladefoged P. 2003. *Phonetic Data Analysis*. Malden: Blackwell Publishing.

[13] Riede T., B. Mitchell, I. Tokuda and M. Owren. 2005. Characterizing noise in nonhuman vocalizations: Acoustic analysis and human perception of barks by coyotes and dogs. *Journal of the Acoustical Society of America* 118 (1). 514-522.

[14] Harrington J. and S. Cassidy. 1999. *Techniques in Speech Acoustics*. Dordrecht: Kluwer Academic Publishers.

[15] Gelfand S. 2001. *Essentials of Audiology*. 2nd ed. New York: Thieme.

[16] Lain S. 2010. *Acoustic correlates of [voice]: Data from two dialects of Venezuelan Spanish*. Saarbrücken: Lambert Academic Publishing.