SC: Paper ICA2016-721

Large-scale analysis of Spanish /s/-lenition using audiobooks

Neville Ryant\textsuperscript{(a)}, Mark Liberman\textsuperscript{(b)}

\textsuperscript{(a)}Linguistic Data Consortium, USA, nryant@gmail.com
\textsuperscript{(b)}Linguistic Data Consortium, USA, markyliberman@gmail.com

Abstract:

Given forced alignment and accurate automatic phonetic classification and measurement, audiobooks are an important potential source of large-scale evidence about phonetic variation. For example, the audiobook version of the novel \textit{La Casa de los Espíritus}, read by two Chilean actors, presents 17 hours of audio containing nearly 68,000 /s/ segments, distributed in a natural way across a wide variety of prosodic, lexical, morphological, syllabic, and phonetic environments. Thus we believe that this one audiobook offers more /s/ tokens than have been examined in the entire 50-year history of sociolinguistic study of Spanish /s/-lenition — and analysis on this scale allows statistical evaluation of a much larger set of hypotheses about phonetic variation and its conditioning factors. For broad comparison of geographical variants, we can use audiobooks whose readers exhibit a variety of accent types, in this case comparing works read by Chilean, Argentinian, Caribbean, Mexican, and Peninsular speakers. Most of the sociolinguistic literature on variation in Spanish syllable-final /s/ treats it as involving three distinct categories: retained [s], aspirated [h], and deletion. In our data we see coherent gradient variation in the duration and frication strength of /s/, with aspiration and deletion as continuum endpoints. Large-scale data also allows us to argue for cases of allomorphy, i.e. variable lexicalization of particular forms. In addition, we see several types of outcome not usually described, including the variable interpenetration of frication or aspiration with voiced portions of adjacent vowels or with following consonants, often resulting (for example) in breathy-voiced nasals.

\textbf{Keywords:} Acoustic Phonetics; Audiobooks; Spanish /s/-lenition
Large-scale analysis of Spanish /s/-lenition using audiobooks

1 Introduction

The phenomenon of Spanish /s/-lenition, the weakening of syllable-final /s/ to [h] or even to zero, has spawned an extensive literature, due to its large social and geographical variability in the Spanish-speaking world. This process of /s/-lenition is attested to occur at least to some degree in the Spanish of Spain (Andalusia, Madrid, and Castile-Le Mancha), the islands of the Caribbean and coastal parts of Mexico and Central America, as well as Colombia, Chile, the Pacific coast of Peru, parts of Bolivia, Uruguay, Paraguay, and most of Argentina [1, 2, 3, 4]. While the result of weakening is typically described as [h] or zero, other surface forms are also possible. In most dialects /s/ may instead weaken to [z] or [f] when the following syllable begins with a voiced stop or nasal; and in some cases the /s/ gesture may merge with the following consonant gesture to create a voiced fricative or a breathy-voiced nasal. Also, in many regions aspiration may take the form of [x] rather than [h][3]. In parts of Nicaragua and coastal Chile, it may be realized as a glottal stop [3]. And in Puerto Rican Spanish, it may even be realized as a devoiced nasal when it follows a vowel and precedes a nasal [2].

The vast majority of the literature approaches /s/-lenition as a categorical process that applies in syllable-final position and attempts to explicate dialectal differences [5, 6, 7, 8, 3] or to produce a linguistic or sociolinguistic account of its production. Studies have found the rate of /s/-lenition to be conditioned on traditional sociolinguistic variables such as age, gender, education, and socioeconomic status, and of course to vary geographically. The rate of /s/-lenition in many Spanish-speaking cultures shows an especially strong age effect, with young speakers showing more weakening than older speakers [9, 10, 11, 12]. Males lenite more than females [13, 14, 15] and urban speakers more than rural speakers [16, 3]. As with many sociolinguistic variables, /s/-lenition shows effects of genre and speaking style with the rate of retention higher in read speech and formal genres [17, 18], while weakening or deletion are highest in fast, spontaneous speech [8].

Despite the variability of its outcome, /s/-lenition has usually been studied using impressionistic coding into one of three discrete outcomes\(^1\): [s], [h], or zero. While necessary in earlier years due to technological constraints, this has limited the depth and scope of work in this area. Transcription is an inherently subjective decision prone to the biases of the transcriber [10, 19] and, as with all human annotation, prone to error [20] and confirmation bias. Moreover, it is inherently wrong to artificially partition a gradient phenomenon into a small set of classes. These concerns have inspired a spate of recent papers that dispense with impressionistic coding and instead reanalyze /s/-lenition as a gradient process, best understood in terms of continuous acoustic features such as duration, spectral centroid, and voicing [21, 22, 23, 24, 25].

This new work is welcome, but for the most part it suffers from lack of scalability, due to reliance

---

\(^1\)One notable exception is [2], an early analysis in terms of gradient changes in phonetic realization, though Navarro’s insights were lost for half a century.
on manual segmentation and measurement. For example, four of these papers [22, 23, 25, 26] survey between them a total of only 4,007 tokens. Exceptions to this generalization are [21] and [24], where the segmentation and measurements were totally automated. However, even these latter two works suffer from reliance on purpose-transcribed speech recordings, which are expensive to produce in time and money, and therefore limited in availability. We have turned to audiobooks as one source of large quantities of freely available speech with accompanying text.

The cross-product of phonological context, lexical frequency and possible allomorphy is as usual astronomically large, without even allowing for social, geographical, individual, and stylistic variation. This state of affairs is normal in speech research, and motivates the use of large-scale resources and automated methods.

2 Data

2.1 Corpora

The present study examines speech from a preliminary corpus covering five varieties of Spanish from seven audiobooks along with the HUB4 [27] collection of Spanish broadcast news:

**Peninsular** Three speakers reading the books *Los Pazos de Ulloa* by Emilia Pardo Bazán, *Historietas Nacionales* by Pedro Antonion de Alarcón y Ariza, and *El 19 de Marzo y el 2 de Mayo* by Benito Pérez Galsó.

**Chilean** Two speakers reading the novel *La Casa de los Esperitus* by Isabel Allende.

**Argentinian** Two speakers reading the novels *Cien Años de Soledad* by Gabriel García Márquez and *La Isla Del Tesoro* by Robert Louis Stevenson (translated by Manuel Caballero).

**Mexican interior** One speaker reading the novel *Angelina* by Rafael Delgado.

**Caribbean** Because we were unable to locate any audiobooks read by Caribbean speakers, we selected one hundred and five speakers from HUB4 [] who were identified in the transcripts as belonging to the coastal dialect. Primarily, this consists of Caribbean and Cuban speakers on Voice of America radio broadcasts, though also some speakers from coastal Mexico in UNIVISION and ECO broadcasts. Turns containing speech from a non-native speaker, overlapping speech, non-Spanish speech, untranscribable speech, and narrowband audio were filtered out as were turns containing fewer than three words.

Precise figures for the number of words, /s/ segments, and total duration for each corpus are presented in Table 1

Obviously we cannot draw general conclusions about a regional variant on the basis of recordings from one to three speakers. On the other hand, audiobook recordings give us an especially large sample of speech from a single speaker, in a single context, offering an opportunity to explore individual differences in unparalleled detail. And as we’ll see, the individual readers in this collection show the effects that the literature leads us to expect for speakers from their regions. As the number of available sources grows in the future, it will be possible to
characterize both individual and regional variation in similar detail.

2.2 Alignments

For each corpus, segmentations were produced by forced alignment using an aligner trained on all turns from the West Point Heroico corpus of Spanish speech [28]. The aligner was trained with the Kaldi ASR toolkit [29], using the CALLHOME Spanish pronunciation dictionary [30] with pronunciations for out-of-vocabulary (OVV) words generated from the Sequitur G2P toolkit [31]. The acoustic front end consisted of 13 mel frequency cepstral coefficient (MFCC) features extracted every 10 ms using a 25 ms Hamming window plus first and second differences; all features were normalized to zero mean and unit variance on a per-speaker basis. A standard 3-state Bakis model was used for all speech phones and 5-state models allowing forward skips used to model non-speech phones (silence, breaths, coughs, laughter, and lipsmacks) and out-of-vocabulary words (words which were not in the pronunciation dictionary and for which grapheme-to-phoneme transduction failed). To improve segmentation accuracy, special 1-state boundary models were inserted at each phone transition as in [32]. Acoustic modeling was performed using a deep neural network consisting of 4 layers of 512 rectified linear units and an 11-frame context window (5-1-5).

3 Results

3.1 Acoustic characteristics of /s/ segments

Following previous work and common sense, we looked at features such as spectral centroid, probability of voicing, and segment duration for each of the sources in our preliminary dataset. These values are presented for two of the sources, Angelina and La Casa, in Tables 2 and 3 and in Figures 1 and 2 we show the distributions and summary statistics for all /s/, /h/ and /z/ segments in each source, and for /s/ segments in five phonological contexts.

3.2 Defining and exploring new acoustic dimensions

Datasets like these make it easy to explore in a preliminary way some non-traditional approaches to phonetic analysis. One promising idea is to use statistical methods to find relevant phonetic dimensions. In this section, we'll display some results from a maximally simple
Figure 1: Densities for frame-level spectral centroid by segment in *Angelina*.

Figure 2: Densities for frame-level spectral centroid by segment in *La Casa*.

example of this approach.

We begin with the forced-alignment segmentation of one of the audiobooks, relative to 13 MFCC parameters calculated every five milliseconds, and mean-variance normalized per speaker. For each of 28 phone classes, we collect all of the corresponding frames, and calculate the mean vector and the full covariance matrix for all the feature vectors for that class. Then for each analysis frame, we calculate its Mahalanobis distance to each of the 28 phone classes. In the case of the audiobook *Angelina*, for example, this yields a matrix with 7,727,059 rows and 28 columns. After subtracting the column means, we use singular value decomposition (SVD) to find a new coordinate system. The dimensions corresponding to the highest two or three singular vectors are then a good source of information about allophonic variation.

As an example, we'll consider just the distributions in the dimension corresponding to the largest singular value in the case of /s/ segments in the audiobook *Angelina*, which was read by a speaker of Mexican Spanish from Texas. The solid black line in the left-hand panel in Figure 3 shows the distribution of values on this dimension for all of the frames in all of the /s/ segments – 798,079 /s/-segment frames in this book. The dashed red line shows the distribution for word-initial /s/ segments, and the dotted blue line shows the distribution for /s/
segments before the voiceless stops /p/, /t/, /k/. For this speaker, these three distributions are very similar, and all of them are also clearly bimodal.

The right-hand panel in Figure 3 reveals to us what these two modes are. It shows the distribution for all /s/ segments again in the black curve. Then the distributions for /s/ before voiced stops /b/, /d/, /g/, and before nasals /m/, /n/, are shown in the dashed magenta and brown lines respectively, while the distribution for /s/ in pre-silence position is shown in the dashed gold line. If we look at samples of these environments, we find that this speaker voices /s/ to /z/ routinely before voiced consonants, while her pre-silence /s/ segments are thoroughly devoiced.

Figure 3: Angelina: Frame-wise distributions of U column 1 for 5 contexts.
Speakers from dialect regions with other allophonic regularities show other patterns in such plots. Thus Figure 4 shows distributions for the same contexts in the Audible version of Isabel Allende’s *La Casa de los Espíritus*, read by two Chilean actors:

![Figure 4: La Casa: Frame-wise distributions of U column 1 for 5 contexts.](image)

In contrast to the Mexican case, we see here that /s/ before /p/, /t/, /k/ is very different from the word-initial context, while the Chilean pre-silence /s/ is frame-wise similar in this dimension to the word-initial case (though as Table 3 indicates, it is much shorter). The other six parts of our preliminary dataset similarly reflect the different allophonic patterns of their speakers – and SVD dimensions 2 and 3 add additional useful differentiation.

### 3.3 Unsupervised learning of phonetic spaces

In the last section, we used the segmentation induced by forced alignment of text and audio to define a novel phonetic space. This segmentation is far from perfect since we did not try to do any pronunciation modeling, but just forced an alignment of the dictionary pronunciation in every case. But in fact, our general approach doesn’t require a text-based segmentation at all – entirely unsupervised methods can produce similar results.

For example, if we just use k-means clustering with k=100 on mean-variance normalized MFCC vectors, a simple segment-wise majority vote yields 97.9% correct identification of /s/ segments, suggesting that SVD on this trivially-derived set of clusters would produce similar results. Table 4 shows that many other segments are relatively well separated by this simple-minded method.
Table 4: Angelina: Partial segment-wise confusion matrix (percentages) for k=100 clusters

<table>
<thead>
<tr>
<th></th>
<th>/a/</th>
<th>/i/</th>
<th>/o/</th>
<th>/u/</th>
<th>/m/</th>
<th>/n/</th>
<th>/r/</th>
<th>/l/</th>
<th>/s/</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a/</td>
<td>90.2</td>
<td>0.5</td>
<td>2.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.1</td>
<td>0.3</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>/i/</td>
<td>0.0</td>
<td>89.5</td>
<td>0.1</td>
<td>0.1</td>
<td>0.4</td>
<td>1.0</td>
<td>0.0</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>/o/</td>
<td>3.0</td>
<td>0.3</td>
<td>89.0</td>
<td>1.2</td>
<td>0.7</td>
<td>1.2</td>
<td>1.3</td>
<td>2.1</td>
<td>0.2</td>
</tr>
<tr>
<td>/u/</td>
<td>0.1</td>
<td>1.3</td>
<td>9.5</td>
<td>20.5</td>
<td>7.8</td>
<td>4.8</td>
<td>0.1</td>
<td>1.1</td>
<td>0.7</td>
</tr>
<tr>
<td>/m/</td>
<td>0.2</td>
<td>0.4</td>
<td>1.3</td>
<td>0.7</td>
<td>69.7</td>
<td>10.8</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>/n/</td>
<td>2.5</td>
<td>2.5</td>
<td>3.5</td>
<td>2.4</td>
<td>25.8</td>
<td>61.9</td>
<td>1.2</td>
<td>1.3</td>
<td>0.3</td>
</tr>
<tr>
<td>/r/</td>
<td>6.6</td>
<td>3.5</td>
<td>6.8</td>
<td>0.2</td>
<td>1.8</td>
<td>4.3</td>
<td>11.9</td>
<td>1.3</td>
<td>2.6</td>
</tr>
<tr>
<td>/l/</td>
<td>1.7</td>
<td>16.1</td>
<td>4.5</td>
<td>0.2</td>
<td>1.4</td>
<td>6.1</td>
<td>2.4</td>
<td>29.5</td>
<td>0.8</td>
</tr>
<tr>
<td>/s/</td>
<td>0.2</td>
<td>0.3</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
<td>0.4</td>
<td>0.0</td>
<td>0.1</td>
<td>97.9</td>
</tr>
</tbody>
</table>

4 Conclusions and future work

In future work, the simple-minded k-means clustering employed in the previous section will be replaced by hidden layer activations of a recurrent neural network autoencoder. This should give an overall improvement, especially for inherently dynamic segments like stops and strongly coarticulated segments like /r/.

Another important direction for future work is to expand the quantity and range of audiobooks accessible for this type of analysis. For instance [33] presents text-aligned versions of more than 1500 hours of English-language audiobooks, which is only a small portion of the more than 50,000 hours of English audiobooks currently available from LibriVox. The LibriVox catalogue at present offers about 561 hours of Spanish-language audiobooks – as well as more than 3,000 hours of Dutch, 1,000 hours of French, and so on.

These large datasets will allow us to explore allophonic variation on a much larger scale than in the past, effectively modeling not only the effects of phonological context but also prosodic and lexical variation and the individual and stylistic dimensions examined in [34]. By extending and improving new acoustically-defined dimensions of the type described in this paper, we hope to contribute to the phonetic aspects of “Acoustics for the 21st Century.”

References


