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

Neville Ryant and Mark Liberman

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 *La Casa de los Espiritus*, 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, 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, frication strength, and laryngeal coarticulation of /s/, with aspiration, deletion and voicing as continuum endpoints.

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.[1–4] 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 [ɦ] 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[3, 5–8] 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.[9–12] Males lenite more than females[13–15] and urban speakers more than rural speakers.[3, 16] 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: [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 and confirmation bias.[20] 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.[21–25]

This new work is welcome, but for the most part it suffers from lack of scalability, due to reliance 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 Fox[21] and Torreira and Ernestus,[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

A. CORPORA

The present study examines speech from a preliminary corpus covering four varieties of Spanish from seven audiobooks:

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.

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

	Hours	Speakers	Words	/s/
Peninsular	24.24	3	204,448	85,645
Chilean	16.98	2	165,620	66,726
Argentinian	28.72	2	226,489	88,246
Mexican	10.15	1	92,811	37,493
TOTAL	80.09	8	689,368	278,110

Table 1: Corpus composition.

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. We note that the LibriVox collection of English-language audiobooks now comprises more than 50,000 hours of speech from several thousand speakers, and it is reasonable to expect the available set of Spanish-language audiobooks to grow to similar proportions in coming years.

B. 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.[27] The aligner was trained with the Kaldi ASR toolkit,[28] using the CALLHOME Spanish pronunciation dictionary[29] with pronunciations for out-of-vocabulary (OOV) words generated from the Sequitur G2P toolkit.[30] 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.[31] 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

A. ACOUSTIC CHARACTERISTICS OF /S/ SEGMENTS

Following existing work on gradient /s/-lenition,[18, 21–23] we began by extracting three acoustic features previously observed to correlate with /s/-lenition and compared their distribution within /s/ segments across phonological contexts:

- Spectral centroid as computed from the decibel power spectrum using a 25 ms Hamming window and excluding spectral components below 1,000 Hz
- Probability-of-voicing (POV) derived from the output of the KALDI pitch tracker[32]
- Duration of the /s/ segment in seconds as derived from the forced alignment boundaries

For spectral centroid and POV, features were extracted every 5 ms and and averaged across all frames within each /s/ segment for a total of 278,100 measurements (one per underlying /s/). Five phonological contexts were then examined: word-final before a pause, word-final before a vowel, before a voiced stop, before a voiceless stop, and before a nasal. These contexts were chosen for consistency with previous work by Fox,[21] who observed the following ranking in terms of deletion percentage for Spanish broadcast news: word-final before a vowel < word-final before a pause < before a voiceless stop < before a voiced stop < before a nasal.

If weakening represents a decrease in vocal effort and/or blending and overlap of articulatory gestures, then we would expect the longest /s/ durations, highest spectral centroids, and lowest POV values in contexts encouraging retention, such as word-final before a vowel or pause, and lowest durations and spectral centroid and highest POV values in contexts typically associated with frequent weakening, such as before a nasal or voiceless stop. And this is in fact what we observe as is apparent in Figure 1, though more pronounced for the dialects typically considered to be leniting (Chilean and Argentinian).

Figure 1 clearly shows the expects context-dependent and dialect-dependent variation in frication strength, voicing assimilation, and duration. Figure 2 and Figure 3 present the frame-level

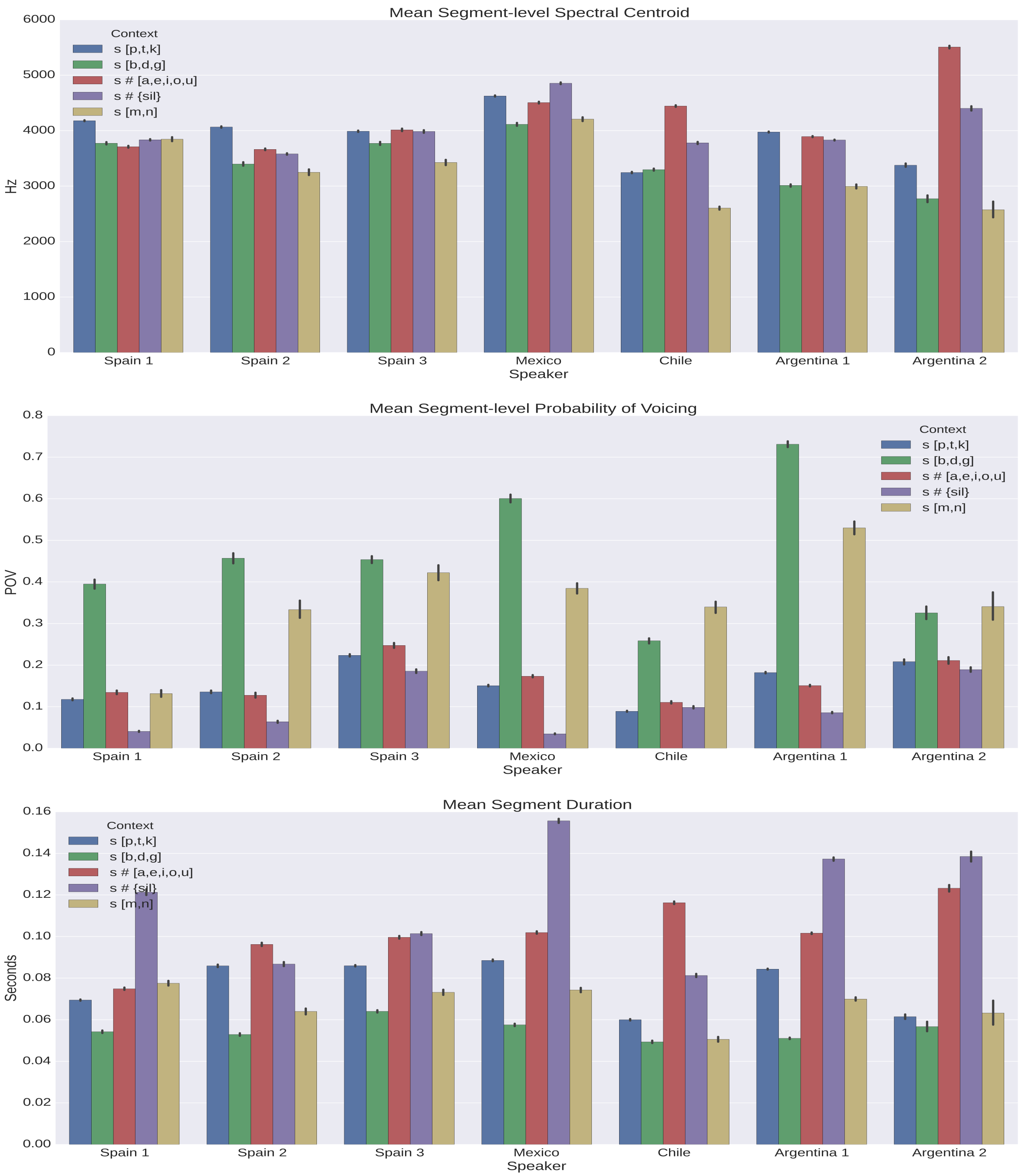

Figure 1: Means of acoustic measurements by speaker and following phonological context. Top: spectral centroid, Center: probability of voicing (POV), Bottom: Duration. Phonological contexts: word-final before a pause: s # {sil}; word-final before a vowel: s # [a,e,i,o,u]; before a voiced stop: s [b,d,g]; before a voiceless stop: s [p,t,k]; before a nasal: s [m,n].

distribution of spectral centroid measures for the Mexican audiobook and the Chilean audiobook. These density plots help to visualize the greater contextual variability of /s/ for the Chilean speakers, as well as the multi-modal nature of some of the distributions, for example the pre-nasal and

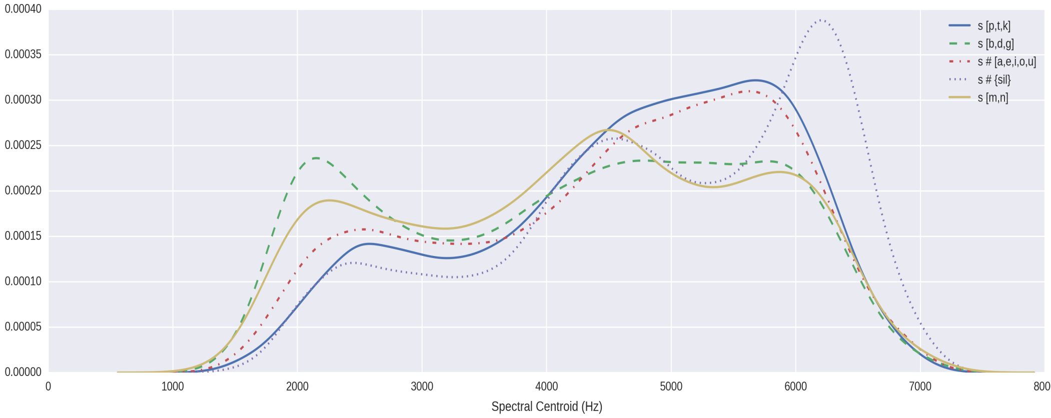

Figure 2: Frame-level density of spectral centroid for Mexican speaker (Angelina) by context. Phonological contexts: word-final before a pause: s # {sil}; word-final before a vowel: s # [a,e,i,o,u]; before a voiced stop: s [b,d,g]; before a voiceless stop: s [p,t,k]; before a nasal: s [m,n].

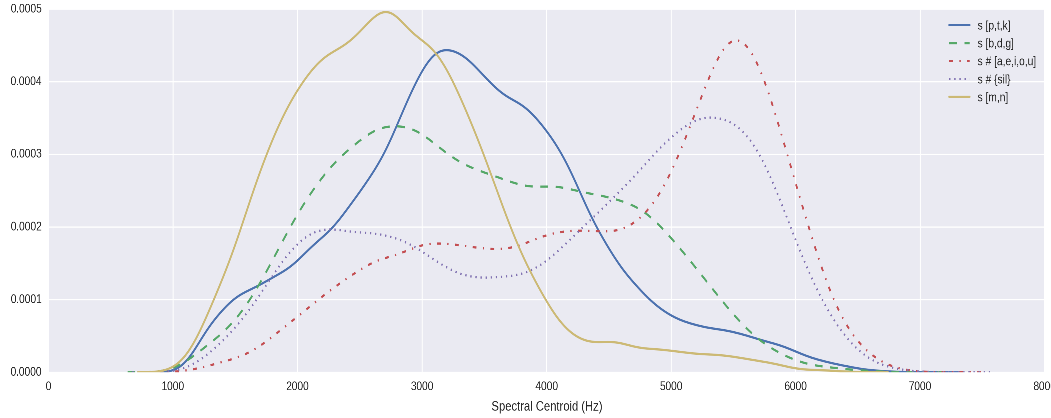

Figure 3: Frame-level density of spectral centroid for Chilean speakers (La Casa de los Esperitus) by context. Phonological contexts: word-final before a pause: s # {sil}; word-final before a vowel: s # [a,e,i,o,u]; before a voiced stop: s [b,d,g]; before a voiceless stop: s [p,t,k]; before a nasal: s [m,n].

pre-voiced stop instances of /s/ for the Mexican speaker. There are at least three possible explanations: (a) additional subdivision of phonological or phonetic contexts, leading to different phonetic targets; (b) non-linearities in articulatory-to-acoustic mapping, of the type discussed in Stevens (1989);[33] (c) phonological assignment of distinct symbolic categories such as [ɦ], [z], [s].

It’s clear that quantal effects of type (b) do exist in the articulatory space associated with /s/: lenition of the lingual gesture, beyond a certain threshold, turns a fricative into an approximant; lenition of the laryngeal devoicing gesture produces a partly-voiced segment and beyond a certain threshold may yield a completely voiced segment; and so on. Whether explanations of types (a) and (c) are also relevant in explaining these distributions is a matter for further research.

B. DEFINING AND EXPLORING NEW ACOUSTIC DIMENSIONS

Datasets like these make it easy to explore some non-traditional approaches to phonetic analysis. One promising idea is to use statistical or machine-learning methods to find relevant acoustic-phonetic dimensions. In this section, we'll display some results from a maximally simple 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.

Note that in contrast to an approach that simply applies SVD or other dimensionality-reduction techniques directly to spectral parameters (or waveform time-series), this approach focuses attention on those aspects of acoustic variation that are related to phonetic identity.

As an example of this approach, consider just the distributions in the dimension corresponding to the largest singular value of the /s/ segments in the audiobook *Angelina*, which was read by a speaker of Mexican Spanish from Texas. Figure 4 shows the distribution of values on this dimension for /s/ segments in general (all 798,079 frames) and in a number of phonological contexts. It is immediately apparent that the distribution of /s/ overall is strongly bimodal, as is the distribution of /s/ before voiceless stops and in word-initial position. Moreover, these three distributions are very similar.

We can see just what these modes are by examining the distributions for /s/ in additional contexts; specifically, before voiced stops, before nasals, and word-final before a pause. The distributions for /s/ before a voiced stop and before a nasal are strongly unimodal, with that mode near one of the modes of the bimodal distributions. The distribution of word-final /s/ before a pause is unimodal, with its single mode located near the other mode of the bimodal distributions. 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.

Speakers from dialect regions with other allophonic regularities show other patterns in such plots. Thus Figure 5 shows distributions for the same contexts in the Audible version of Isabel Allende's *La Casa de los Espiritus*, read by two Chilean actors. 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 Figure 1 indicates, it is much shorter).

The other books in our preliminary dataset similarly reflect the different allophonic patterns of their speakers – and SVD dimensions 2 and 3 add additional useful differentiation. Figure 7 through Figure 6 plot the joint density for the first two SVD dimensions, for all /s/ segments in each of seven books, with the plotting characters 1, 2, 3 at the mean values for prevocalic /s/, /s/ before voiceless stops, and /s/ before voiced stops. The horizontal dimension is the one corresponding to the largest singular value, and the vertical dimension is the next one in order.

The three Peninsular speakers in Figure 7 show a relatively unimodal overall distribution in

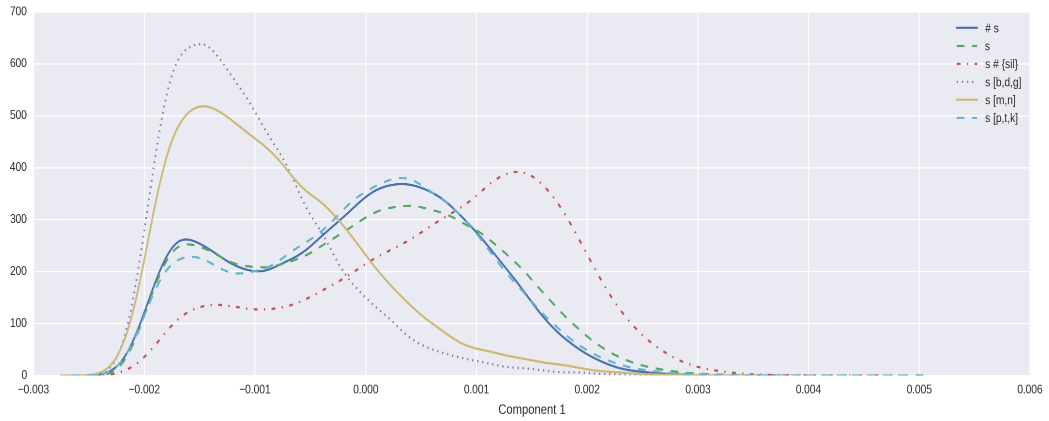

Figure 4: Frame-level density of most important singular component for Mexican speaker (Angelina). Phonological contexts: word-initial: # s; all /s/: s; word-final before a pause: s # {sil}; before a voiced stop: s [b,d,g]; before a voiceless stop: s [p,t,k]; before a nasal: s [m,n].

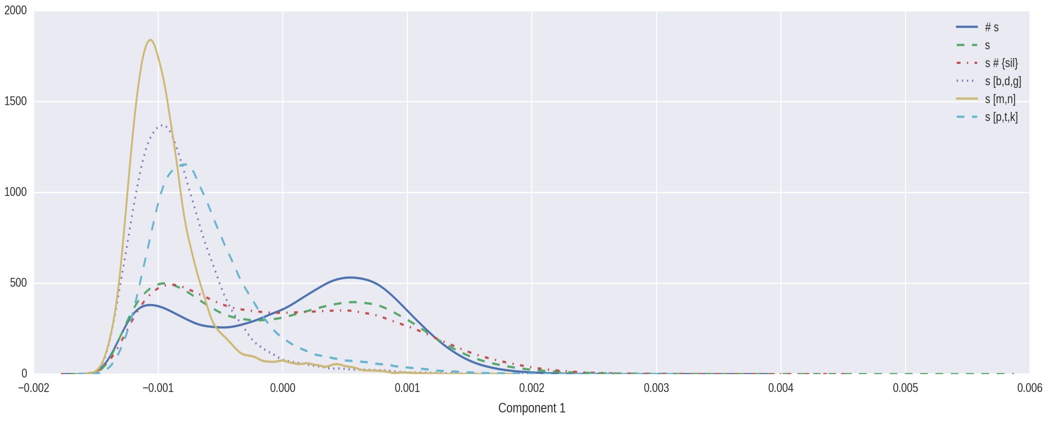

Figure 5: Frame-level density of most important singular component for Chilean speakers (La Casa de los Esperitus). Phonological contexts: word-initial: # s; all /s/: s; word-final before a pause: s # {sil}; before a voiced stop: s [b,d,g]; before a voiceless stop: s [p,t,k]; before a nasal: s [m,n].

this space, with prevocalic /s/ and /s/ before voiceless stops near one another, and /s/ before voiced stops somewhat further away – always in a consistent order in the horizontal dimension.

The Mexican speaker in Figure 9 shows a generally similar distribution, except that there is the beginning of a second mode associated with voiced variants, and /s/ before voiced stops is substantially further away from /s/ in the other two contexts, in the voiced-allophone region.

For the Chilean speakers in Figure 8, the overall distribution is clearly bimodal – and this time, the instances of /s/ before voiceless stops are closer to the pre-voiced-stop group, reflecting the strong lenition of these preconsonantal cases.

The two Argentinian speakers in Figure 6 also show clearly bimodal distributions. But Argentinian Speaker 1 has /s/ before voiceless stops close to word-final /s/ before vowels, whereas

Argentinian Speaker 2 has /s/ before voiceless stops closer to /s/ before voiced stops. As Figure 1 shows, these two speakers also differ strikingly in their spectral centroid and POV patterns for these contexts.

These 2D-density figures offer a suggestive picture of the (frame-wise) spectral aspects of allophonic variation in /s/ realization across contexts for these varieties of Spanish. Ideally one would like to capture contextual variation in spectral trajectories and timing as well, and also take account of the latent structure of higher-order interaction terms in these patterns of spectro-temporal variation. Those goals motivate experimentation with the non-linear dimensionality-reduction capabilities of modern "neural" models – but we feel that this simple linear version suggests the potential of the approach.

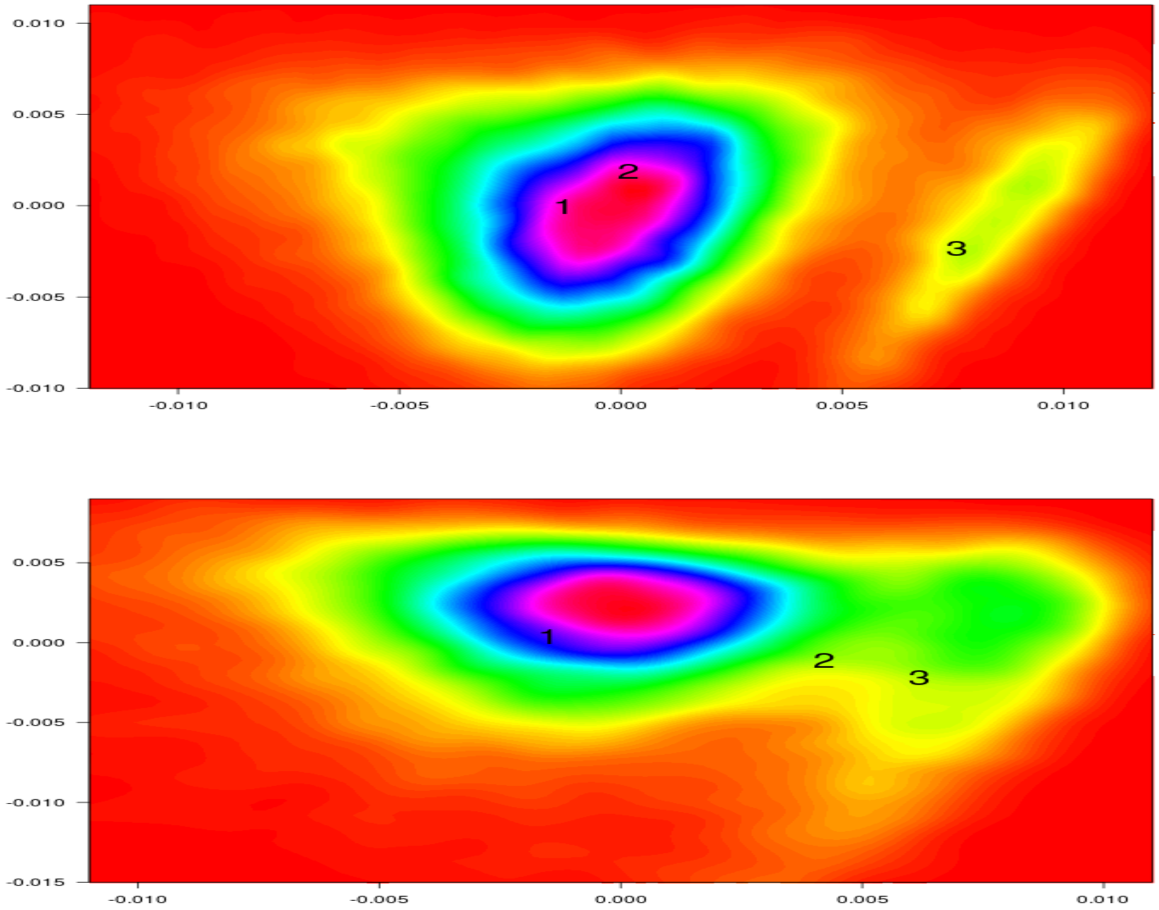

Figure 6: Joint density of the two most important singular components for the two Argentinian speakers. Top: Speaker 1 (Cien Años de Soledad), Bottom: Speaker 2 (La Isla Del Tesoro). Numbers indicate centroids of /s/ in different phonological contexts: 1: word-final before a vowel; 2: before a voiceless stop; 3: before a voiced stop.

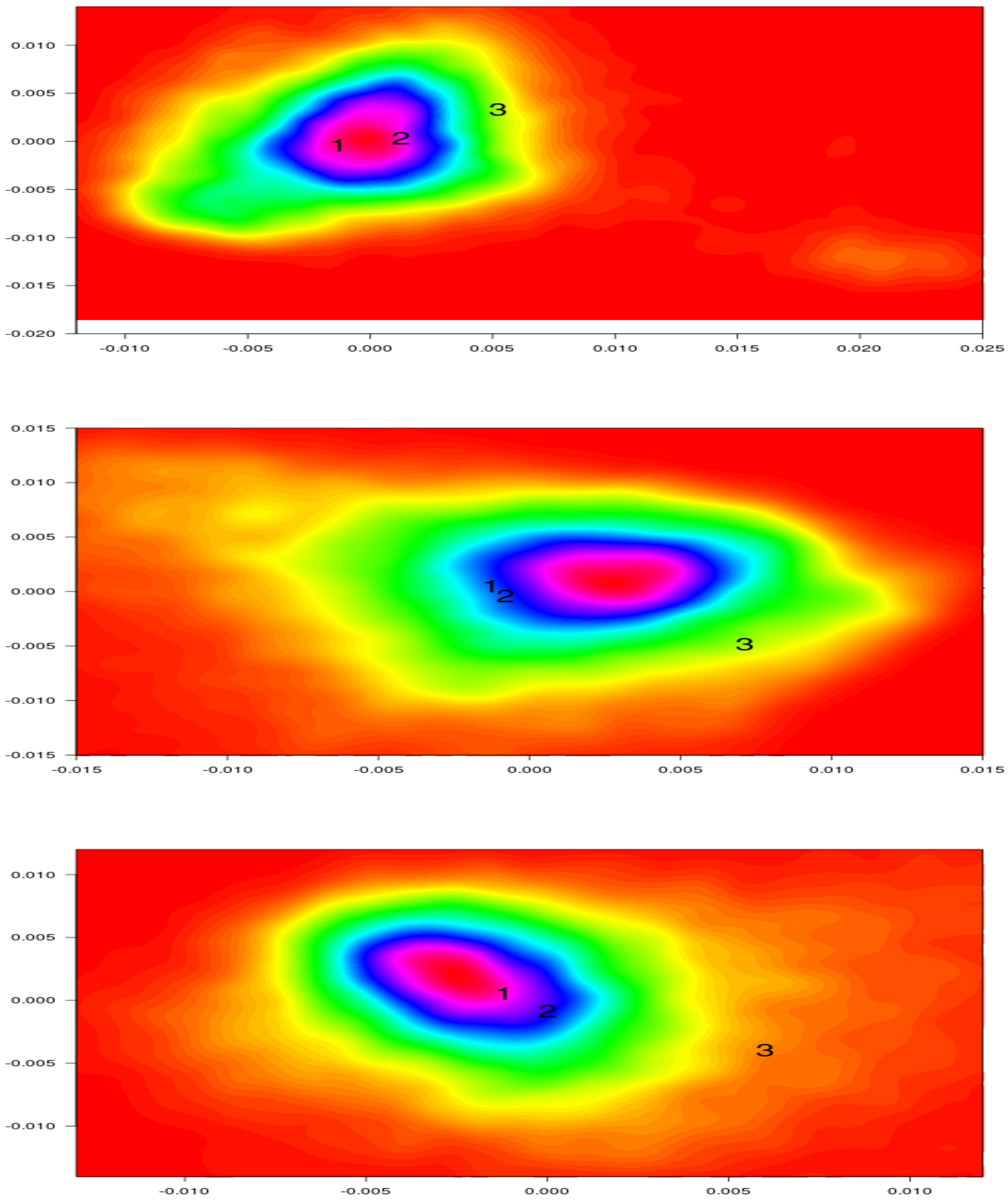

Figure 7: Joint density of the two most important singular components for the three Spanish speakers. Top: Speaker 1 (Los Pazos de Ulloa), Center: Speaker 2 (Historietas Nacionales), Bottom: Speaker 3 (El 19 de Marzo y el 2 de Mayo). Numbers indicate centroids of /s/ in different phonological contexts: 1: word-final before a vowel; 2: before a voiceless stop; 3: before a voiced stop.

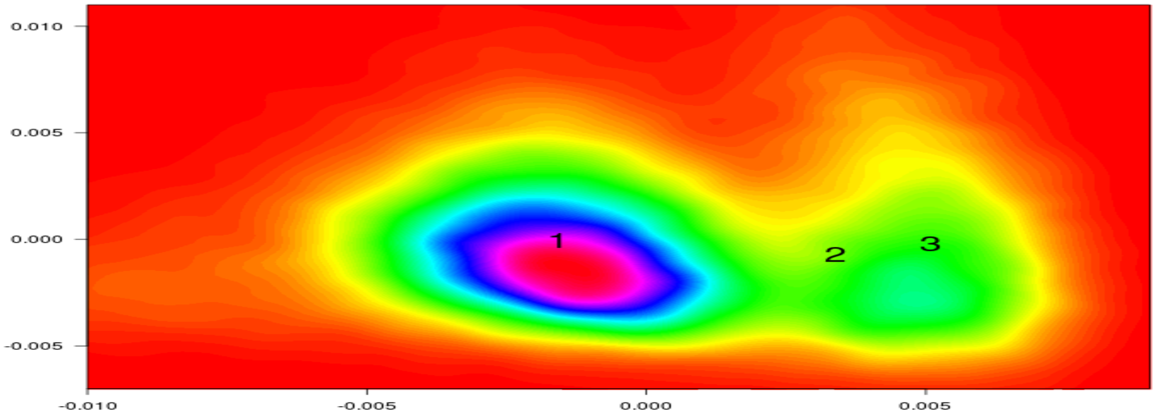

Figure 8: Joint density of the two most important singular components for the Chilean speakers (La Casa de los Esperitus). Numbers indicate centroids of /s/ in different phonological contexts: 1: word-final before a vowel; 2: before a voiceless stop; 3: before a voiced stop.

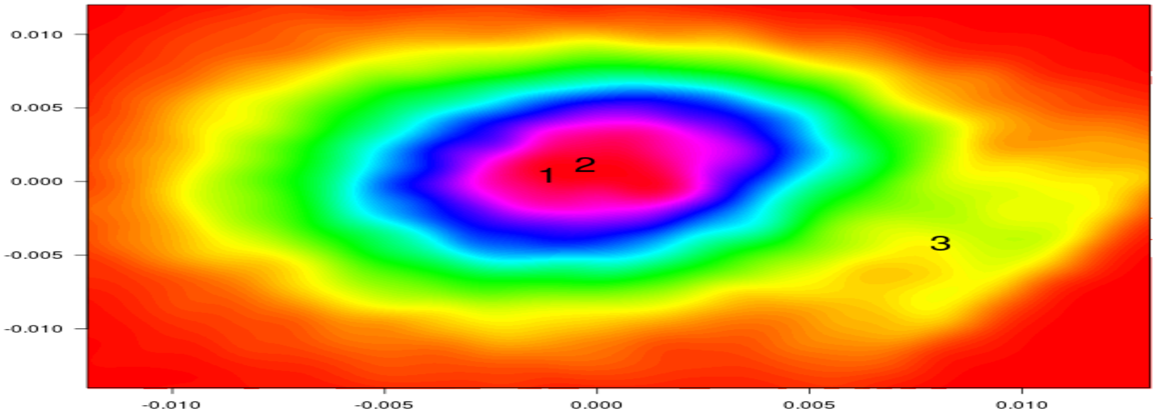

Figure 9: Joint density of the two most important singular components for the Mexican speaker (Angelina). Numbers indicate centroids of /s/ in different phonological contexts: 1: word-final before a vowel; 2: before a voiceless stop; 3: before a voiced stop.

C. 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. However, 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 2 shows that many other segments are also relatively well separated by this simple-minded method.

Thus the goal of focusing a dimensionality-reducing technique on phonetic sources of variation can be accomplished without any data-independent knowledge of the language. How far this approach will take us remains to be seen.

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

Table 2: *Angelina*: ***Partial segment-wise confusion matrix (percentages) for k=100 clusters***

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[34] 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.[35] 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

[1] R. Lapesa and R. M. Pidal, "Historia de la Lengua española", (1942).

[2] T. Navarro Tomás, *El Español en Puerto Rico* (Río Piedras: Editorial Universitaria) (1948).

[3] E. G. Cotton and J. M. Sharp, *Spanish in the Americas* (Georgetown University Press) (1988).

[4] J. M. Lipski, *Latin American Spanish* (Longman Pub Group) (1994).

[5] D. L. Canfield, *La pronunciación del Español en América: ensayo histórico descriptivo*, 17 (Instituto Caro y Cuervo) (1963).

[6] D. L. Canfield, *Spanish Pronunciation in the Americas* (University of Chicago Press) (1981).

[7] J. M. Lipski, "On the weakening of/s/in Latin American Spanish", Zeitschrift für Dialektologie und Linguistik 31–43 (1984).

[8] J. M. Lipski, "/s/ in Central American Spanish", Hispania **68**, 143–149 (1985).

[9] H. C. J. Cedergren, "The interplay of social and linguistic factors in Panama", Ph.D. thesis, Cornell University (1973).

[10] S. Poplack, "Function and process in a variable phonology", (1979).

[11] R. G. Sutil, "Una cuestión de fonosintaxis: realización en Andaluz de la" s" final de palabra seguida de vocal", Anuario de estudios filológicos 135–154 (1992).

[12] G. Cepeda, "Retention and deletion of word-final/s/in Valdivian Spanish (Chile)", Hispanic Linguistics **6**, 329–354 (1995).

[13] M. B. F. D. Weinberg, "Comportamiento ante-" s" de hablantes femeninos y masculinos del español bonaerense", Romance Philology **27**, 50–58 (1973).

[14] T. Terrell, "Diachronic reconstruction by dialect comparison of variable constraints: s-aspiration and deletion in Spanish", Variation omnibus 115–124 (1981).

[15] J. Magaña and H. Valdivieso, "Variación fonética de" s" en el habla espontánea", RLA: Revista de lingüística teórica y aplicada 97–114 (1991).

[16] L. Rodríguez-Castellano and A. Palacio, "El habla de cabra", Revista de dialectología y tradiciones populares **4**, 387 (1948).

[17] O. Alba, *Cómo Hablamos Los Dominicanos: Un Enfoque Sociolinguistico* (2004).

[18] R. J. File-Muriel, “The role of lexical frequency in the weakening of syllable-final lexical /s/ in the Spanish of Barranquilla, Colombia”, Hispania 348–360 (2009).

[19] V. J. Boucher, “Alphabet-related biases in psycholinguistic enquiries: Considerations for direct theories of speech production and perception.”, Journal of Phonetics (1994).

[20] R. J. File-Muriel and M. Díaz-Campos, “Grammatical judgments and phonetic reality: A study of internal constraints in phonological variation”, in *Conference paper. 114th Meeting Acoustical Society of America. Austin* (2003).

[21] M. A. M. Fox, “Usage-based effects in Latin American Spanish syllable-final /s/ lenition”, Ph.D. thesis, University of Pennsylvania (2006).

[22] D. G. Erker, “A subsegmental approach to coda /s/ weakening in dominican spanish”, International Journal of the Sociology of Language **2010**, 9–26 (2010).

[23] R. J. File-Muriel and E. K. Brown, “The gradient nature of s-lenition in Caleño Spanish”, Language Variation and Change **23**, 223–243 (2011).

[24] F. Torreira and M. Ernestus, “Weakening of intervocalic /s/ in the Nijmegen Corpus of Casual Spanish”, Phonetica **69**, 124–148 (2012).

[25] J. I. Hualde and P. Prieto, “Lenition of intervocalic alveolar fricatives in Catalan and Spanish”, Phonetica **71**, 109–127 (2014).

[26] E. K. Brown, M. S. Gradoville, and R. J. File-Muriel, “The variable effect of form and lemma frequencies on phonetic variation: Evidence from /s/ realization in two varieties of Colombian Spanish”, Corpus Linguistics and Linguistic Theory **10**, 213–241 (2014).

[27] J. Morgan, *West Point Heroico Spanish Speech* (Linguistic Data Consortium) (2006).

[28] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, *et al.*, “The Kaldi speech recognition toolkit”, in *IEEE 2011 workshop on automatic speech recognition and understanding* (2011).

[29] S. Garrett, T. Morton, and C. McLemore, *CALLHOME Spanish Lexicon* (Linguistic Data Consortium) (1996).

[30] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, “Automatic speech recognition for under-resourced languages: A survey”, Speech Communication **56**, 85–100 (2014).

[31] J. Yuan, N. Ryant, M. Liberman, A. Stolcke, V. Mitra, and W. Wang, “Automatic phonetic segmentation using boundary models”, in *INTERSPEECH*, 2306–2310 (2013).

[32] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, “A pitch extraction algorithm tuned for automatic speech recognition”, in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2494–2498 (IEEE) (2014).

[33] K. N. Stevens, “On the quantal nature of speech”, Journal of Phonetics **17**, 3–45 (1989).

[34] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an ASR corpus based on public domain audio books”, in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 5206–5210 (IEEE) (2015).

[35] N. Ryant and M. Liberman, “Automatic analysis of phonetic speech style dimensions”, in *INTERSPEECH* (2016).