INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Automatic Analysis of Phonetic Speech Style Dimensions

Neville Ryant, Mark Liberman

Linguistic Data Consortium
nryant@gmail.com, markyliberman@gmail.com

Abstract

We apply automated analysis methods to create a multidimensional characterization of the prosodic characteristics of a large variety of speech datasets, with the goal of developing a general framework for comparing prosodic styles. Our datasets span styles including conversation, fluent reading, extemporized narratives, political speech, and advertisements; we compare several different languages including English, Spanish, and Chinese; and the features we extract are based on the joint distributions of F0 and amplitude values and sequences, speech and silence segment durations, syllable durations, and modulation spectra. Rather than focus on the acoustic correlates of a small number of discrete and mutually exclusive categories, we aim to characterize the space in which diverse speech styles live.
Index Terms: speaking style, prominence, computational paralinguistics

1. Introduction

Many studies have shown that spontaneous speech and fluent reading differ in a variety of prosodic measures. For instance spontaneous speech has been observed to have both somewhat faster speaking rates [1, 2, 3, 4] and greater variability in pause and speech segment durations [5, 6] than read speech. Measures of pitch also distinguish the two styles with read speech having lower F0 mean, greater F0 standard deviation, and greater F0 range than spontaneous speech [2, 3] as well as more pronounced F0 declination [7, 3]. However, as noted by Laan [3] "none of these acoustic features ... can clearly discriminate between the two speaking styles", because "the performance of the speakers ... varied enormously".

Disfluencies (including filled pauses and self-corrections) are an obvious symptom of extemporized speech, absent by definition in fluent reading. In both German and Mandarin filled pauses can be detected fairly well (A' about 0.90) by listeners who didn't know the languages in question and partial-word disfluencies can be detected reasonably well (A' about 0.80) [8]. Similarly, Chinese speakers can detect disfluencies in Swedish at above chance levels [9]. Automated methods for disfluency detection have been developed [10, 11, 12] and work well for filled pauses, though less well for self-corrections.

In this paper, we follow the advice of Wagner et al. [13] in opposing the dichotomous division of speech styles, instead promoting "generally increased methodological awareness and a higher variety of investigated styles of speech". Our approach is to define a set of prosodic measures that can be applied automatically to speech recordings with and without transcriptions. Application of such measures to speech datasets representing a wide variety of styles, languages, and speakers will allow us to characterize (at least some of) the prosodic differences involved, and should lead us towards an understanding of the latent dimensions of speech planning and production, linguistic systems, and cultural styles that underlie the phenomena under study.

2. Methods

2.1. Data

The present study examines speaking style using data from four large-scale collections of non-laboratory recordings:

Fresh Air: Fourteen radio interviews between Terry Gross and Dave Davies and public figures ranging from Lena Dunham to Stephen King to Gloria Steinem that were conducted as part of National Public Radio's (NPR) Fresh Air program. Recordings and transcripts were downloaded from NPRs website, and the transcripts were "unedited" to include disfluencies and to correct other transcription errors. In the analysis below, Terry Gross is treated separately from the interviewees, though there is little difference between them in the measures used. This collection is being prepared for publication at the Linguistic Data Consortium.

YouthPoint: YouthPoint was a radio program produced by students at the University of Pennsylvania in the late 1970s containing interviews with opinion leaders of the era. The broadcast versions, edited for radio, are all 30 minutes in duration though the original interviews may be much longer. Our data set includes a subset of 50 sessions with 57 interviewees ranging from columnist Ann Landers, driver Mario Andretti, fashion designer Francesco Scavullo and actors Mark Hamill, Annie Potts and Chuck Norris to architect Buckminster Fuller, authors Erica Jong, Chaim Potok, and Isaac Asimov, and politicians Ed Muskie and Joe Biden.

Political speeches: A corpus of political speech consisting of 50 weekly radio addresses given by George W. Bush during 2008 and 127 weekly addresses and prepared statements given by Barak Obama between 2009 and 2011. Audio and transcripts were downloaded from each president's White House website and segmented into turns on silences greater than 200 ms using an existing aligner trained on audiobooks. Presidents Bush and Obama are treated separately in the following analysis.

LibriSpeech: LibriSpeech [14] is a corpus of read English speech consisting of some 5,832 audiobook chapters from the LibriVox[1] project. It comprises 334,345 turns from 2,484 speakers with a total audio duration of 1,600 hours.

Precise figures for number of utterances, speakers, and total duration for each corpus are presented in Table 1.

2.2. Measurements

For each speaker and speech type we computed three measures:

- The proportion of speech segments of duration greater than 600 ms.

[1] https://librivox.org

http://dx.doi.org/10.21437/Interspeech.2016-1355

	Hours	Utterances	Speakers
Fresh Air	8.53	7,148	18
Political	13.93	16,560	2
YouthPoint	14.08	7,984	65
LibriSpeech	1,570.75	334,345	2,484

Table 1: Corpus composition.

- The proportion of silence segments of duration greater than 200 ms.
- F0 range, defined as the difference in semitones between the 90th and 10th percentiles of a given speaker's measured F0 values.

Speech and silence segments were identified from the force-aligned transcripts of each recording with a speech segment defined as a sequence of contiguous non-silence phones in the alignment belonging to a single speaker and bordered by either silence or speech from another speaker. Similarly, silence segments were defined as contiguous sequences of silences in the alignment bordered on both sides by speech from the same speaker. Silences that could not be unequivocally assigned to a single speaker were excluded. Speech and silence segment distributions for LibriSpeech, YouthPoint, Bush, Obama, Terry Gross, and the Fresh Air guests are depicted in Figure 1 and observed proportions of speech and silence segments above threshold in Table 2.

F0 contours were extracted for every recording using an implementation of the Kaldi pitch tracking algorithm [15] and smoothed using a Butterworth low-pass filter as per [16]. To determine F0 range we then excluded all frames with a probability-of-voicing of less than 0.9 and from the surviving frames computed the F0 range, defined as the difference between the 90th and 10th percentiles expressed as semitones. We chose the 90th and 10th percentiles so as to avoid the effects of outliers that may result from pitch tracking errors or from vocal creak or fry (which we plan to measure in other ways).

2.3. Alignments

For each corpus speaker turn segmentations were produced by forced alignment using an aligner trained on all turns from that corpus with the exception of LibriSpeech, where training was restricted to a random 120 hour sample of clean turns[2]. The aligners were trained with the Kaldi ASR toolkit [17] using the CMUdict lexicon with stress markings removed; pronunciations for out-of-vocabulary (OOV) words were generated with the Sequitur G2P toolkit [18] using a model trained on CMUdict. The acoustic frontend 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 phones with the exception of a 5-state silence model and 1-state phone boundary models [19]. 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).

[2]LibriSpeech classified turns as clean if the word-error-rate of an ASR engine was below the median for the corpus.

3. Results

As can be seen in Figures 1 and 2 and Table 2, the measures defined in terms of the distribution of speech and silence segment durations are effective in separating spontaneous from read speaking styles, at least in this collection of material. The two political speakers (who were reading prepared texts), and the modal region of the 2,484 LibriSpeech readers are well separated from the three spontaneous speech samples, represented by Terry Gross, her guests, and the YouthPoint speakers with the spontaneous speech samples exhibiting both shorter speech segments and silence segments than the audiobook and political speech samples. Indeed, the proportion of speech segments $>$ 600 ms jumps from 0.68 for the Fresh Air guests to .82 for the LibriSpeech speaker's mean and to greater than .9 for the Bush speeches and more extreme LibriSpeech readers. Similarly for proportion of silence segments $>$ 200 ms, which is less than .60 for the spontaneous samples, but ranges as high as as .80 for the Bush speeches and even higher for the most extreme LibriSpeech speakers. Unsurprisingly, there is quite a bit of variability among the LibriSpeech readers with the tails of their distribution extending from the edge of the spontaneous speakers' region to far past the political speech region (Figure 2).

We take this as tending to confirm that these dimensions are useful in characterizing prosodic style. But as noted earlier, our goal is not to establish a two-class dichotomy between read and spontaneous speech – rather, we hope to define a space in which linguistically, perceptually, and culturally (and perhaps clinically) relevant aspects of prosodic style can be situated. And this issue is brought out sharply by the next dimension we introduce, namely pitch range, operationalized as the difference in semitones between the 90th and 10th percentile of each speaker's F0 estimates.

Although some studies have found pitch-range differences between read and spontaneous speech, it's intuitively clear that a speaker can use a wider or narrower pitch range in either situation. And this is what we found: as is seen quite clearly in Figure 3, President Obama's range was 11.5 semitones while President Bush's range was only 6.8 semitones; among the Fresh Air guests, Tanehisi Coates used a range of 10.3 semitones while Jill Soloway used a range of only 6.8. This doesn't mean that the pitch range dimension is not relevant or useful in general – it's clearly a significant aspect of prosodic style – but it's not relevant or useful to the specific task of distinguishing spontaneous speech from reading.

Finally, in Figure 4, we've added six speakers from a set of Spanish audiobooks ([20, 21, 22, 23, 24], discussed at greater length in [25]), along with a dotted line representing the 5th through 95th percentiles from the collection of 2,484 LibriSpeech English-language audiobook readers. While the Spanish-language readers clearly fall along the trend line established by the English-language readers, they seem to show a generally higher proportion of longer silence segments. Whether this represents a genuine cultural difference in reading styles, a chance result from a relatively small sample, or an artefact of the way the LibiSpeech material was prepared, is a topic for further investigation.

4. Discussion

As a way of characterizing prosodic style, we have developed and tested several phonetic dimensions which can be calculated automatically and efficiently on large volumes of speech

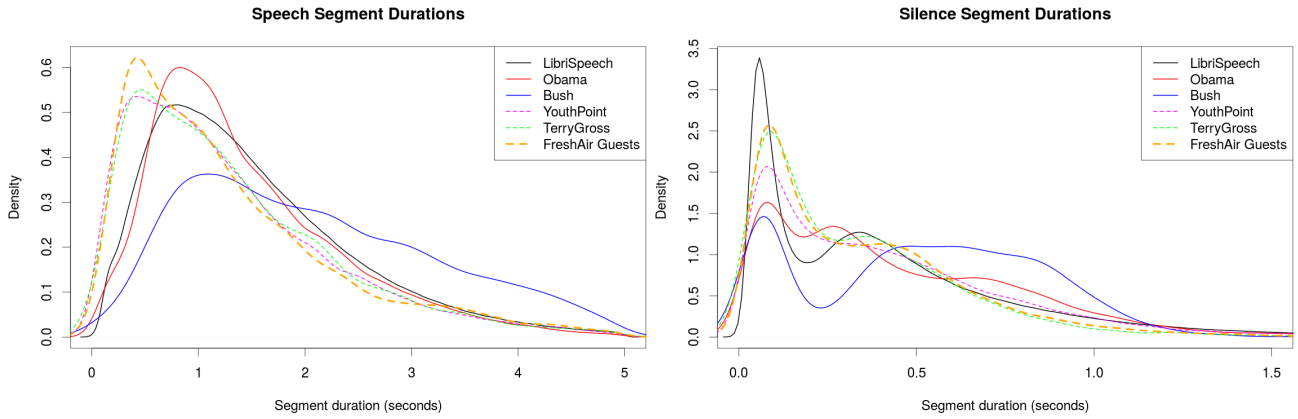

Figure 1: Distributions of speech (left) and silence (right) segment durations across the corpora.

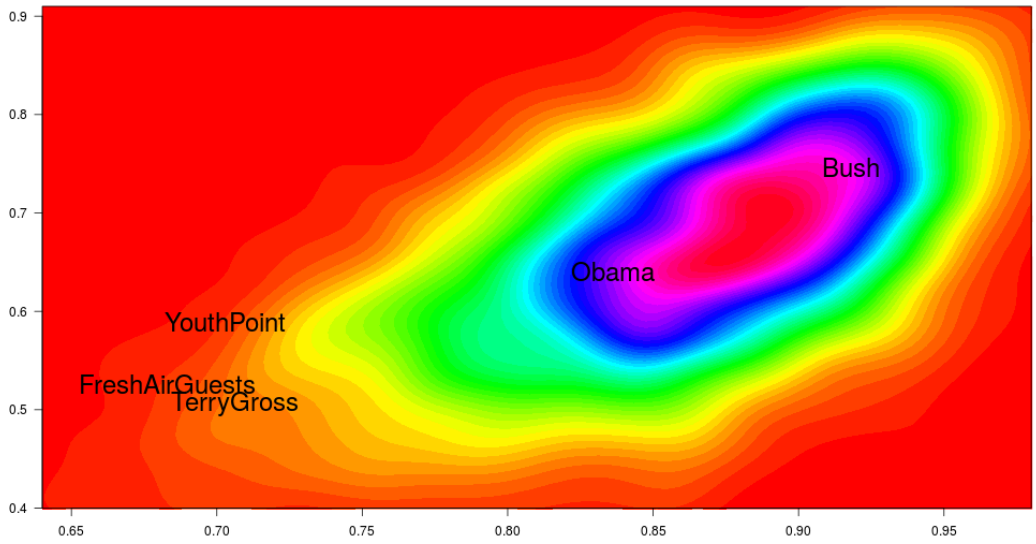

Figure 2: Contour plot for LibriSpeech showing number of speakers as a function of proportion of speech segments longer than 600 ms (x-axis) and proportion of silence segments longer than 200 ms (y-axis) with Bush, Obama, Terry Gross, YouthPoint, and all Fresh Air guests overplotted for reference. For precise proportions of speech segments $>$ 600 ms and silence segments $>$ 200 ms, consult Table 2.

	Speech $>$ 600 ms	Silence $>$ 200 ms
Terry Gross	0.706	0.506
FA Guests	0.683	0.526
YouthPoint	0.703	0.589
Obama	0.836	0.641
Bush	0.918	0.746
LibriSpeech	0.824	0.600

Table 2: Proportion of speech durations $>$ 600 ms and silence durations $>$ 200 ms for each data set.

data. We have tested these and similar dimensions on conversation, news broadcasts and audiobooks in Chinese and Spanish as well as in English and find that they work in a generally similar way across language. We are also using these dimensions of prosodic description, among others, in clinical studies of recorded interviews following the Autism Diagnosis Observation Schedule (ADOS) protocol [26, 27], and picture-description recordings from neurological examinations of elderly patients with various neurodegenerative disorders [28, 29], in studies to be reported on in other upcoming meetings.

It should be obvious that there are many other aspects of prosodic variation that we have not tried to characterize in this study – the nature of syllable-level amplitude contours, the alignment of pitch movements with syllable-level sonority profiles, the degree and distribution of pre-boundary lengthening

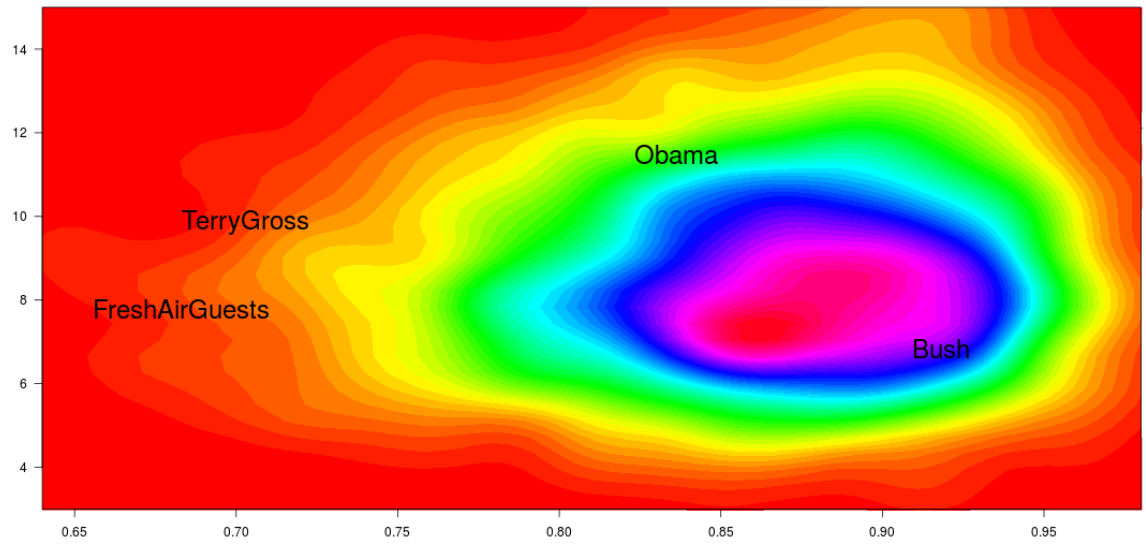

Figure 3: Contour plot for LibriSpeech showing number of speakers as a function of proportion of speech segments longer than 600 ms (x-axis) and F0 range in semitones (y-axis) with Bush, Obama, Terry Gross, YouthPoint, and all Fresh Air guests overplotted for reference.

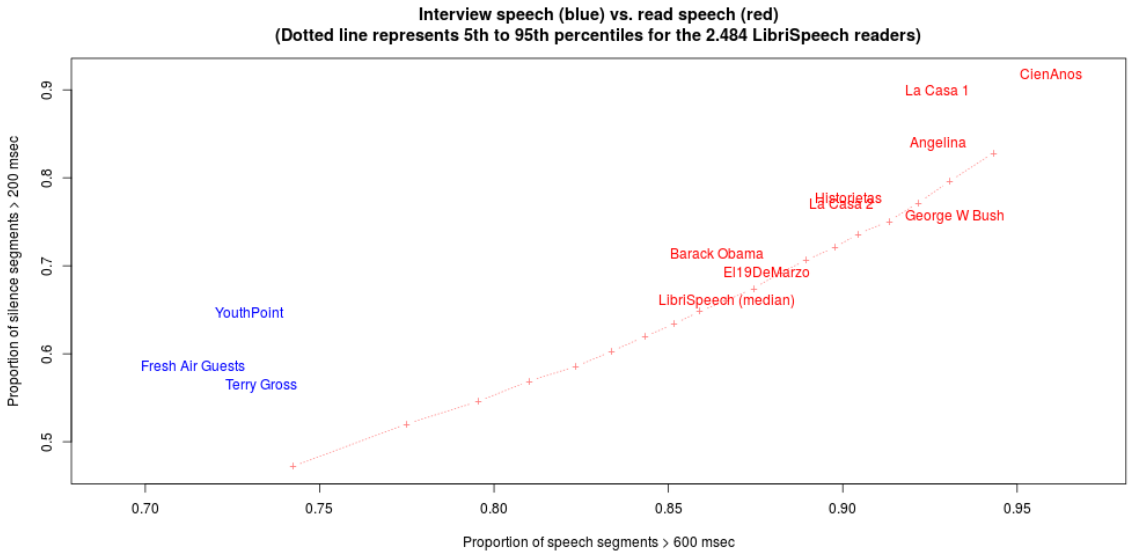

Figure 4: Comparison of interview speech (blue) and read speech (red) in terns of proportion of speech segments longer than 600 ms (x-axis) and proportion of silence segments longer than 200 ms (y-axis). For the English audiobooks we plot the median, while for the Spanish audiobooks – *Angelina*, *Cien años de soledad*, *La Casa de los Esperitus*, *El 19 de Marzo y el 2 de Mayo*, and *Historietas Nacionales* – we plot the speakers individually. The dotted line (–) represents 5th to 95th percentiles of 2,484 LibriSpeech readers.

and emphatic lengthening, the proportions of consonant and vowel segments, the distribution of phrase-final F0 movements, individual differences in voice quality and patterns of voice quality variation, and so on. For all of these we propose to work towards operational measures that can be calculated automatically and efficiently for the increasingly large volumes of increasingly varied speech that are becoming available. This paper frames an approach to these problems that we believe can and will be extended and generalized.

We will be releasing and documenting our code and examples of its applications, and we hope that others will join us in future explorations.

5. References

[1] T. H. Crystal and A. S. House, "Segmental durations in connected speech signals: Preliminary results," *The Journal of the Acoustical Society of America*, vol. 72, no. 3, pp. 705–716, 1982.

[2] A. Batliner, R. Kompe, A. Kießling, E. Nöth, and H. Niemann, "Can you tell apart spontaneous and read speech if you just look at prosody?" in *Speech Recognition and Coding*. Springer, 1995, pp. 321–324.

[3] G. P. Laan, "The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style," *Speech Communication*, vol. 22, no. 1, pp. 43–65, 1997.

[4] V. Dellwo, A. Leemann, and M.-J. Kolly, "The recognition of read and spontaneous speech in local vernacular: The case of Zurich German," *Journal of Phonetics*, vol. 48, pp. 13–28, 2015.

[5] P. Howell and K. Kadi-Hanifi, "Comparison of prosodic properties between read and spontaneous speech material," *Speech Communication*, vol. 10, no. 2, pp. 163–169, 1991.

[6] H. Mixdorff and H. R. Pfitzinger, "Analysing fundamental frequency contours and local speech rate in map task dialogs," *Speech Communication*, vol. 46, no. 3, pp. 310–325, 2005.

[7] M. Swerts, E. Strangert, and M. Heldner, "F0 declination in read-aloud and spontaneous speech," in *ICSLP*, 1996, pp. 1501–1504.

[8] C. Lai, K. Gorman, J. Yuan, and M. Liberman, "Perception of disfluency: language differences and listener bias." in *INTERSPEECH*, 2007, pp. 2345–2348.

[9] R. Carlson and J. B. Hirschberg, "Cross-cultural perception of discourse phenomena," 2009.

[10] Y. Liu, E. Shriberg, and A. Stolcke, "Automatic disfluency identification in conversational speech using multiple knowledge sources." in *INTERSPEECH*, 2003.

[11] H. Medeiros, F. Batista, H. Moniz, I. Trancoso, and L. Nunes, "Comparing different methods for disfluency structure detection," in *OASIcs-OpenAccess Series in Informatics*, vol. 29. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2013.

[12] W. Wang, A. Stolcke, J. Yuan, and M. Liberman, "A cross-language study on automatic speech disfluency detection." in *HLT-NAACL*, 2013, pp. 703–708.

[13] P. Wagner, J. Trouvain, and F. Zimmerer, "In defense of stylistic diversity in speech research," *Journal of Phonetics*, vol. 48, pp. 1–12, 2015.

[14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210.

[15] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *ICASSP*, 2014, pp. 2494–2498.

[16] J. Yuan and M. Liberman, "F0 declination in English and Mandarin broadcast news speech." in *INTERSPEECH*, 2010, pp. 134–137.

[17] 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.

[18] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol. 56, pp. 85–100, 2014.

[19] J. Yuan, N. Ryant, M. Liberman, A. Stolcke, V. Mitra, and W. Wang, "Automatic phonetic segmentation using boundary models." in *INTERSPEECH*, 2013.

[20] R. Delgado, *Angelina*. LibriVox, 1893. [Online]. Available: https://librivox.org/angelina-por-rafael-delgado

[21] G. G. Márquez, *Cien años de soledad*. Audible, 1967. [Online]. Available: http://www.audible.com/pd/Fiction/Cien-anos-de-soledad-Audiobook/B00RDN6DO4

[22] B. Pérez Galdós, *El 19 de Marzo y el 2 de Mayo*. LibriVox, 1881. [Online]. Available: https://librivox.org/el-19-de-marzo-y-el-2-de-mayo-by-benito-perez-galdos

[23] P. A. de Alarcón, *Historietas Nacionales*. LibriVox, 1893. [Online]. Available: https://librivox.org/historietas-nacionales-by-pedro-antonio-de-alarcon-y-ariza

[24] I. Allende, *La Casa de los Esperitus*. Audible, 1982. [Online]. Available: http://www.audible.com/pd/Fiction/La-casa-de-los-espiritus-The-House-of-the-Spirits-Audiobook/B00P025V8E

[25] N. Ryant and M. Liberman, "Large-scale analysis of Spanish /s/-lenition using audiobooks," in *Proceedings of 22nd International Congress on Acoustics*, 2016.

[26] C. Lord, M. Rutter, P. C. DiLavore, S. Risi, K. Gotham, and S. Bishop, *Autism Diagnostic Observation Schedule: ADOS-2*. Western Psychological Services Los Angeles, CA, 2012.

[27] J. Parrish-Morris, C. Cieri, M. Liberman, L. Bateman, E. Ferguson, and R. Schultz, "Building language resources for exploring autism spectrum disorders," in *LREC*, 2016.

[28] S. Ash, E. Evans, J. O'Shea, J. Powers, A. Boller, D. Weinberg, J. Haley, C. McMillan, D. J. Irwin, K. Rascovsky *et al.*, "Differentiating primary progressive aphasias in a brief sample of connected speech," *Neurology*, vol. 81, no. 4, pp. 329–336, 2013.

[29] S. Ash, A. Menaged, C. Olm, C. T. McMillan, A. Boller, D. J. Irwin, L. McCluskey, L. Elman, and M. Grossman, "Narrative discourse deficits in amyotrophic lateral sclerosis," *Neurology*, vol. 83, no. 6, pp. 520–528, 2014.