Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6504–6513
Marseille, 11–16 May 2020

Crowdsourcing Latin American Spanish for Low-Resource Text-to-Speech

Adriana Guevara-Rukoz[1,2,3], Işın Demirşahin[1], Fei He[1], Shan-Hui Cathy Chu[1], Supheakmungkol Sarin[1], Knot Pipatsrisawat[1], Alexander Gutkin[1], Alena Butryna[1], Oddur Kjartansson[1]

[1]Google Research, Japan, Singapore, United States and United Kingdom
[2]Laboratoire de Sciences Cognitives et Psycholinguistique, ENS/EHESS/CNRS/PSL/INRIA, France
[3]Graduate School of Engineering, The University of Tokyo, Japan
{aguekoz,isin,mungkol,thammaknot,agutkin}@google.com

Abstract

In this paper we present a multidialectal corpus approach for building a text-to-speech voice for a new dialect in a language with existing resources, focusing on various Latin American dialects of Spanish. We first present public speech datasets for Argentinian, Chilean, Colombian, Peruvian, Puerto Rican, and Venezuelan Spanish, specifically constructed with text-to-speech applications in mind using crowdsourcing. We then compare monodialectal TTS voices built with minimal data to voices made with a multidialectal model built by pooling all the resources from all dialects. Our results show that the multidialectal model outperforms the monodialectal baseline models. We also experiment with a "zero-resource" dialect scenario where we build a multidialectal voice for a dialect while holding out target dialect recordings from the training data.

Keywords: corpora, text-to-speech, phonology, Spanish dialects, Latin America

1. Introduction

Building a high-quality text-to-speech (TTS) voice is costly and requires significant effort both in terms of collecting nontrivial amounts of recorded speech in each language, as well as building the corresponding linguistic resources. In low-resource language scenarios several approaches have been proposed to address the issue of data and resource scarcity. These techniques range from utilizing general speech data found online that was not necessarily recorded for speech applications in mind (Baljekar, 2018; Cooper, 2019) to pooling the data from multilingual corpora (Li and Zen, 2016). It has been shown that constructing the models from multiple speakers (Gutkin et al., 2016), as well as using the data from related languages (i.e., languages from close families or language areas (Emeneau, 1956)), consistently improves the quality of the voice with minimal resources (Baljekar et al., 2018; Wibawa et al., 2018; Demirsahin et al., 2018). We believe that this approach can also be extended to constructing the models for new dialects of a given language.

In this paper we explore different compositions of datasets for building voices for Argentinian (AR), Chilean (CL), Peruvian (PE) and Venezuelan (VE) Spanish dialects using multi-speaker data crowdsourced[1] in each locale, combining it with the professional studio recordings of Peninsular (ES) and United States (US)[2] Spanish dialects. Past TTS research on Argentinian (Torres et al., 2012; Violante, 2012), Peruvian (Florentino, 2016), Venezuelan (Rodríguez et al., 2006) and other Latin American Spanish dialects, such as Colombian (Correa et al., 2010), focused mostly on single-speaker nonparametric concatenative systems (Hunt and Black, 1996). An arguably less brittle, parametric, approach to TTS (Zen et al., 2013; Zen et al., 2016), mediated by explicit acoustic models that we employ in this work, allows us to better leverage the data from these dialects. We approach the design of the phonemic inventory for each dialect in a principled way which, on the one hand, leverages the heavy overlap between the sound systems and, on the other, emphasizes the prominent dialectal differences established in the literature (Canfield, 1981; Lipski, 1994; Penny and Penny, 2004; Real Academia Española y Asociación de Academias de la Lengua Española, 2011; Resnick, 2012), some of which, like prosodic differences, are notoriously hard to model (Ortiz-Lira, 1999; Colantoni and Gurlekian, 2004; Feldhausen et al., 2011; O'Rourke, 2012).

The main contributions of this paper are as follows:

- We introduce new open-source speech corpora for six dialects of Latin American Spanish: Argentinian, Chilean, Colombian, Peruvian, Puerto Rican and Venezuelan. To the best of our knowledge these are the first high-quality *free* (unencumbered by a restrictive license) *multi-speaker* datasets available for these dialects. In addition to the approach described in this paper, these datasets have many more potential uses that include cross-lingual or cross-dialectal transfer learning in text-to-speech (Chen et al., 2019) and multi-dialectal acoustic modeling in automatic speech recognition (Li et al., 2018). We hope these datasets become a welcome addition to the growing body of Latin American Spanish speech resources, such as the single-speaker corpus of Argentinian Spanish recently announced by Torres et al. (2019).
- We show that a joint multidialectal model constructed by combining some of the dialect-specific datasets described above with the large in-house corpora for Peninsular (ES) and United States (US) Spanish outperforms the low-resource dialect-specific baselines.
- Furthermore, we demonstrate that, given a linguistic front-end (i.e., component for converting input text

[1]As opposed to data recorded by professional voice actors.
[2]This refers to the dialect used in US media in Spanish (e.g., Univision, Telemundo).

Dialect	Code	Locations	ISLRN	Gender	Name	Lines	Words		Duration (hours)	Speakers
							Total	Unique		
Argentinian	AR	Buenos Aires	395-001-133-368-2	F	arf	3,921	35,360	4,107	5.61	31
				M	arm	1,818	16,914	3,343	2.42	13
Chilean	CL	Santiago	048-218-632-043-6	F	clf	1,738	16,591	3,279	2.84	13
				M	clm	2,636	25,168	4,171	4.31	18
Colombian	CO	Bogota	169-985-498-793-0	F	cof	2,369	22,228	4,460	3.74	16
				M	com	2,534	23,957	4,459	3.84	17
Peruvian	PE	Lima	923-742-092-167-6	F	pef	2,529	23,806	4,278	4.35	18
				M	pem	2,918	27,547	4,268	4.87	20
Puerto Rican	PR	US	721-732-548-994-0	F	prf	617	6,092	1,738	1.00	5
				M	–	–	–	–	–	–
Venezuelan	VE	US and UK	697-927-390-879-1	F	vef	1,603	15,182	3,419	2.41	11
				M	vem	1,754	16,613	3,612	2.40	12
Total:						24,437	229,458	5,783	37.79	174

Table 1: Latin American Spanish multi-speaker dataset details.

into phonemic representation of its corresponding pronunciation), one can still build a satisfactory model of a particular dialect with the acoustic data for this dialect omitted during the training process.

The rest of this paper is organized as follows: In the next section we introduce the six new corpora for Latin American Spanish dialects. We then present phonological design for the set of dialects selected for the experiments (Section 3). This is followed by a series of experiments that investigate different combinations of corpora for building voices with or without acoustic data for the target dialect during training (Section 4). Finally, in Section 5 we discuss our results and set the roadmap for future experiments.

2. Latin American Spanish Corpora

We built the datasets for six dialects of Latin American Spanish: Argentinian (Google, 2019a), Chilean (Google, 2019b), Colombian (Google, 2019c), Peruvian (Google, 2019d), Puerto Rican (Google, 2019e) and Venezuelan (Google, 2019f). The basic information about the released datasets is given in Table 1, where each of the six datasets is shown along the corresponding BCP-47 region code (Phillips and Davis, 2009), recording locations and the International Standard Language Resource Numbers (ISLRNs) (Mapelli et al., 2016).

The corpora contain the crowdsourced recordings from both male and female speakers, along with accompanying orthographic transcriptions. Each corpus consists of two subsets corresponding to female and male speakers, respectively. For each subset, its symbolic name, the total number of recorded lines, the total and number of unique words, the duration in hours and the number of distinct speakers are shown in Table 1.

All recorded volunteers were native speakers of the corresponding dialects. Argentinian, Chilean, Colombian and Peruvian were recorded in the respective locations where the dialect is used, whereas Puerto Rican and Venezuelan were recorded in New York, San Francisco and London.

2.1. Recording Script Design

The original recording script was designed for a conversational system in Mexican Spanish. To adapt this data for our needs, we selected shorter phrases and removed any Mexican Spanish-specific sentences. To increase the variety of the corpus we created additional sentences based on templates, where we varied proper names while leaving the rest of the sentences intact.

Only a small part of the recording script was localized by native speakers of each dialect, however speakers being recorded were allowed to improvise and read the phrases in a way they felt more natural for their dialect, if necessary. Any mismatches between the original transcriptions and the recorded audio were fixed during a quality control process by matching the transcriptions to the improvisations. The transcriptions therefore reflect the spoken speech.

In addition, the recording script contains around 30 "canonical" sentences that all the speakers of all the dialects were required to read. These sentences were specially selected to include salient phonological contrasts known to exist across the dialects in question. Examples of such sentences are provided in Table 2.

Example	Phonetic variation
*El caballo está amarra**d**o.*	Possible deletion of intervocalic consonants: [amara(ð̞)o]
*Los cora**z**ones de po**ll**o son una deli**c**ia.*	<z> as [θ, s], <ll> as [ʎ, ʝ, ɟ, j, ʃ, ʒ], <ci> as [θi, si]
*El via**j**e fue muy divertido.*	<je> as [xe, he, χe, çe]

Table 2: Some examples of sentences where strong dialectal variation is attested.

The total number of unique words in all the recording scripts was relatively low (below 6,000). To increase word coverage in the crowdsourced varieties, dialect-specific pronunciation lexicons were enriched by adding lexical entries from an existing lexicon for Peninsular (ES) or United States (US) Spanish. When applicable, pronunciations were adapted to the dialectal variants by using automated rules, with exceptions being manually adjusted (e.g., loanwords).

2.2. Audio Recording Details

Speakers recorded themselves in a quiet room, using hardware and custom built software provided by an experi-

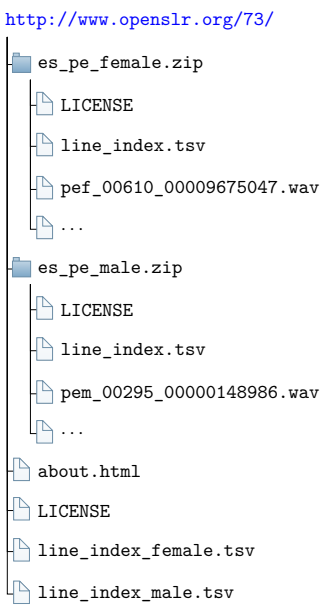

Figure 1: Layout of the Peruvian Spanish corpus.

menter. While the experimenter was initially present to demonstrate how to properly use the materials, speakers were left alone in the room during the recording session, in order to ensure speech naturalness. Each speaker read 150 sentences presented on-screen in a self-paced manner. At each sentence, they were asked to listen and validate the recording. The utterances were re-recorded if noticeable background noise was detected by the recording software, if there were pronunciation problems (e.g., stuttering during reading, laughing), or if speakers found their prosody to be unnatural. Recordings were done using an ASUS Zenbook UX305CA fanless laptop, a Neumann KM 184 microphone and a Blue Icicle XLR-USB A/D converter. The audio was recorded as 48 kHz single-channel and is provided in 16 bit linear PCM RIFF format.

Apart from the Argentinian Spanish data, which was recorded in an office recreational music room, the data for all other crowdsourced dialects were recorded in a portable acoustic vocal booth, which provides an alternative to the commercial studio. Acoustic vocal booths are designed to reduce echo and reverberations and the affordable versions typically reduce the noise levels by about 10–13 dB.

With the exception of Puerto Rican Spanish, over ten speakers were recorded for each dialect. Also, as can be seen from Table 1, no male speakers were recorded for Puerto Rican Spanish. Utterances with wrong pronunciations or recordings artifacts (such as significant background noise, mouth clicks, heavy breathing and so on) were filtered out during the post-recording quality control (QC).

2.3. Distribution and Licensing

The corpora are open-sourced under "Creative Commons Attribution-ShareAlike" (CC BY-SA 4.0) license (Creative Commons, 2019) and hosted on Open Speech and Language Resources (OpenSLR) repository (Povey, 2019).

The corpora structure follows the same lines for each dialect, similar to Figure 1, which shows the structure for Peruvian Spanish. Collections of audio and the corresponding transcriptions are stored in a separate compressed archive for each gender. Transcriptions are stored in a *line index* file, which contains a tab-separated list of pairs consisting of the audio file names and the corresponding transcriptions. The transcriptions have not been text normalized and may contain non-standard word (NSW) tokens (Sproat et al., 2001), such as abbreviations and cardinal numbers. The name of each utterance consists of three parts: symbolic dataset name (e.g., Peruvian male is denoted `pem`), the five-digit speaker ID and the 11-digit hash.

It is important to note that the amount of data we are releasing is rather small (24,437 utterances in total for all the dialects): it may not be enough to build a reasonable modern single-dialect single-gender model using a traditional approach to text-to-speech. On the other hand, the purpose of the data collection was to assemble high quality –rather than high volume– data that can be used in applications that involve combining multiple datasets together (as it is done in this paper) or using the datasets as adaptation data.

	Crowdsourced				Professional	
Dialect	AR	CL	PE	VE	ES	US
UTT	998	1,021	1,044	1,017	1,014	758
WRD	9,633	9,857	10,079	9,738	9,842	9,844
SPK	7	7	7	7	1	1

Table 3: Training set: Number of utterances (UTT), words (WRD), and speakers (SPK) per dialect.

3. Dialect Selection and Phonological Design

3.1. Corpus Selection

For the purpose of experimenting with various dialects of Latin American Spanish, the training set for our models combines the open-source datasets for Argentinian (AR), Chilean (CL), Peruvian (PE) and Venezuelan (VE) Spanish (described in Section 2) with existing proprietary single-speaker corpora for Peninsular (ES) and United States (US) Spanish recorded by the voice actors in professional studios[3]. We restricted the training data to recordings by female speakers only.

In order to ensure an equal representation of all dialects in the multidialectal TTS system, we selected subsets of recorded speakers to be included in the training dataset. For crowdsourced datasets (AR, CL, PE and VE), we selected the seven speakers with the lowest number of utterances discarded during the quality control process (for VE this represented all seven speakers available at the time of the experiments). For professionally recorded data (ES and US), an algorithm randomly selected utterances so that the resulting number of words approximated the average number of words in the recordings from the crowdsourced locales. The details of the resulting training set are shown in Table 3.

[3]Data collection and experiments described here were performed in parallel. As such, the corpora for Colombian (CO) and Puerto Rican (PR) Spanish were not available when the experiments were being designed. Similarly, only a subset of the other crowdsourced locales (AR, CL, PE, VE) had been recorded at the time.

	IPA	ES	US	AR	CL	PE	VE
	m	m	m	m	m	m	m
	n	n	n	n	n	n/ŋ	n/ŋ
	ɲ	ɲ	ɲ	ɲ	ɲ	ɲ	ɲ
	p	p	p	p	p	p	p
	t	t	t	t	t	t	t
	tʃ	tʃ	tʃ	tʃ	tʃ/ttʃ/ts	tʃ	tʃ
c	k	k	k	k	k	k	k
o	b	b	b	b	b	b	b
n	d	d	d	d	d	d	d
s	g	g	g	g	g	g	g
o	ʝ	ʝ/ɟ	ʝ/ɟ	ʝ	ʝ/ɟ	ʝ/ɟ	ʝ
n	f	f	f	f	f	f	f
a	θ	θ	θ	-	-	-	-
n	s	s̪	s	s	s	s	s
t	ʃ	-	ʃ	ʃ	-	ʃ	-
	x	x/χ	x	x	x/ç	x	h
	l	l	l	l	l	l	l
	ʎ	ʎ	-	-	-	-	-
	ɾ	ɾ	ɾ	ɾ	ɾ	ɾ	ɾ
	r	r	r	r	r	r	r
v	a	a	a	a	a	a	a
o	e	e	e	e	e	e	e
w	i	i	i	i	i	i	i
e	o	o	o	o	o	o	o
l	u	u	u	u	u	u	u
semi-	j	j	j	j	j	j	j
vowel	w	w	w	w	w	w	w

Table 4: Multidialectal phonemic inventory.

3.2. Phonemic Inventory

Based on the phonetic and phonological descriptions of the relevant dialects (Lipski, 1994; Penny and Penny, 2004; Real Academia Española y Asociación de Academias de la Lengua Española, 2011), we established a multidialectal phonemic inventory. We used the International Phonetic Alphabet (IPA) (International Phonetic Association, 1999) as the underlying representation. Table 4 shows the list of phonemes in the unified inventory and IPA, and the mapping of the phonetic realisations in each individual dialect in the corresponding columns. In order to keep the salient differences across dialects while making the most of the similarities in a limited resources scenario, this inventory differs from traditional phonological inventories established for Spanish.

In some cases, one phoneme in a dialect was mapped into two salient variants. For instance, the phoneme /ʝ/ (with spellings <ll> and <y>), was mapped to a separate /ʃ/ category in the case of AR, as it is realized as [ʃ ~ʒ] in Rio Platense (Buenos Aires) Spanish (*e.g. muelle, yo, lluvia*). However, the same phoneme in the same dialect is realized as [ʝ] in words like *Youtube*, and in these cases it was transcribed as /ʝ/.

In other cases, a traditionally acknowledged phone in one dialect was kept within its corresponding phoneme in spite of salient cross-dialectic phonetic differences. For example, the VE phoneme /x/ is phonetically produced as [h], but it was not assigned to a separate /h/ phoneme since it does not raise a linguistic contrast with /x/ phonemes within or across dialects.

Finally, the diphthongs are represented as vowel-glide sequences. This allows us to represent all possible vowel-glide combinations by adding only two glide symbols (marked as *semivowel* in the last two rows of Table 4). The validity of the inventory was assessed by cross-referencing with our recordings.

4. Experiments

In this section we describe and evaluate the voices built using different subsets of the multidialectal corpus, as well as investigate choosing different speaker identities (speaker ID) as an input feature to guide the acoustic model during synthesis (i.e., perform dialect selection)[4].

4.1. Phoneme Alignments

In order to build the voices, phoneme boundaries are required for all the corpora used in the experiments. The speech data was downsampled to 16 kHz and then parameterized into HTK-style Mel Frequency Cepstral Coefficients (MFCC) (Ganchev et al., 2005) using a 10 ms frame shift, a 25 ms Hamming window and a first order pre-emphasis filter with a coefficient of 0.97. The dimension of the MFCC parameters is 39 (13 static + Δ + ΔΔ coefficients). To determine the phoneme time boundaries, the acoustic parameter sequences were then force-aligned with the corresponding transcriptions (Young et al., 2006). Each dataset was force-aligned individually.

4.2. Model Architecture Details

We used long short term memory recurrent neural network (LSTM-RNN) acoustic model configuration, as described by Gutkin (2017) and originally proposed by Zen and Sak (2015), to build a set of experimental voices. LSTM-RNNs are designed to model temporal sequences and long-term dependencies within them (Hochreiter and Schmidhuber, 1997).

Two unidirectional LSTM-RNNs for duration and acoustic parameter prediction are used in tandem in a streaming fashion. Given the input features, the goal of the duration LSTM-RNN is to predict the duration (in frames) of the phoneme in question. This prediction, together with the input features, is then provided to the acoustic model which predicts smooth acoustic vocoder parameter trajectories. The smoothing of transitions between consecutive acoustic frames is achieved in the acoustic model by using recurrent units in the output layer.

The input features used by both the duration and the acoustic models consisted of one-hot linguistic features that describe the utterance including the phonemes, stress, syllable counts and distinctive phonological features (such as place and manner of articulation). An additional important feature that we use is a one-hot speaker identity feature. When using a model trained on multiple dialects, this feature is instrumental in forcing the consistent speaker characteristics on the output of the model. In other words, it forces the voice to sound like the requested speaker.

[4]Here speaker ID and dialect are conflated in the speaker ID feature. Therefore, selecting a speaker ID during synthesis determines the dialect used by the voice. Ideally, future models could have an independent dialect feature, allowing to separate speaker-dependent acoustic idiosyncrasies from dialect-specific characteristics.

The original speech data was downsampled to 22.05 kHz. Then mel-cepstral coefficients (Fukada et al., 1992), logarithmic fundamental frequency ($\log F_0$) values (interpolated in the unvoiced regions), voiced/unvoiced decision (boolean value) (Yu and Young, 2011), and 7-band aperiodicities were extracted every 5 ms, similar to previous work (Zen et al., 2016). These values form the output features for the acoustic LSTM-RNN and serve as input vocoder parameters (Agiomyrgiannakis, 2015). The output features for the duration LSTM-RNN are phoneme durations (in seconds). The input features for both the duration and the acoustic LSTM-RNN are linguistic features. The acoustic model supports multi-frame inference (Zen et al., 2016) by predicting four frames at a time, hence the training data for the model is augmented by frame shifting up to four frames. Both the input and output features were normalized to zero mean and unit variance. At synthesis time, the acoustic parameters were synthesized using the Vocaine vocoding algorithm (Agiomyrgiannakis, 2015).

The architecture of the acoustic LSTM-RNN consists of 2×512 ReLU layers (Zeiler et al., 2013) followed by 3×512-cell LSTM with recurrent projection (LSTMP) layers (Sak et al., 2014) with 256 recurrent projection units and a linear recurrent output layer (Zen and Sak, 2015). The architecture of the duration LSTM-RNN consists of 1×512 ReLU layer followed by a single 512-cell LSTMP layer with a feed-forward output layer with linear activation. For both types of models the input and forget gates in each memory cell are coupled since distributions of gate activations for input and forget gates were previously reported as being correlated (Miao et al., 2016). The duration LSTM-RNN was trained using an ε-contaminated Gaussian loss function (Zen et al., 2016), whereas for acoustic LSTM-RNN the L_2 loss function was used as per Gutkin (2017).

4.3. Configurations

Here we provide an overview of the combinations that were experimented on across dialects using the datasets described in Section 3.1.

Monodialectal Baseline For each crowdsourced dialect (i.e., AR, CL, PE and VE), we trained a single monodialectal multi-speaker baseline model using only that dialect's dataset (i.e., 7 speakers per dialect). We then synthesized five sentences using each of the 7 speaker IDs available. For each dialect, a native speaker of the dialect selected the (subjective) best speaker ID for that dialect in terms of both naturalness and nativeness based on these five sentences.

Multidialectal Baseline We trained a single multidialectal multi-speaker model using all the available datasets (i.e., 7 speakers per crowdsourced dialect, 1 speaker per professionally-recorded dialect, with matched amount of words for all dialects, as described in Section 3.1). From it we built the multidialectal baseline voice of each dialect, using the same speaker IDs used for the corresponding monodialectal models.

Multidialectal Holdout Lastly, we built multidialectal voices for the new dialects holding out all or most of the acoustic data for the target dialect during training. The holdout voice for a given target dialect uses all of the training data except for that of the target locale. The speaker ID was then selected from one of the other crowdsourced locales by a native speaker of the target dialect. For example, the training data for the VE holdout voice includes recordings from all locales except VE (i.e., 7 speakers for AR, CL, and PE, plus 1 speaker for ES and US, respectively), with a PE speaker having been selected by a VE native speaker as the most natural and least non-native sounding for the VE holdout voice.

Dialect	Monodialectal	Multidialectal	Holdout
AR	3.43 ± 0.14	3.91 ± 0.11	4.17 ± 0.11
CL	3.52 ± 0.10	3.79 ± 0.09	-
PE	3.73 ± 0.13	3.98 ± 0.09	3.69 ± 0.08
VE	4.13 ± 0.09	3.71 ± 0.08	4.70 ± 0.06

Table 5: Mean-Opinion Scores (MOS). Raters evaluated voice naturalness using a scale from 5 ('Excellent') to 1 ('Bad').

4.4. Evaluation

The overall naturalness of the voices was evaluated by Mean Opinion Scores (MOS) (Streijl et al., 2016), as presented in Table 5. For each dialect, we compiled a set of 100 sentences that were neither in the training data, nor in the five-sentence set used for best speaker ID selection. Each sentence was rated by 3 different native speakers from the locale using a 5-point scale, as follows: 5 = 'Excellent - Completely natural speech', 4 = 'Good - Mostly natural speech', 3 = 'Fair - Equally natural and unnatural speech', 2 = 'Poor - Mostly unnatural speech', 1 = 'Bad - Completely unnatural speech'. This resulted in 300 datapoints per model evaluation; however, each rater did not necessarily evaluate all 100 sentences. In MOS tests, all multidialectal voices were rated fair-to-good[5].

Furthermore, we conducted pairwise comparisons between different voice configurations for the same locale, using the same evaluation set as the MOS tests. The raters listened to a sentence, synthesized with both voices A and B, and were asked to select which voice sounded better as a virtual assistant of their dialect (e.g., "a Chilean virtual assistant") using a 7-point scale. The A/B pair for each sentence in the set was evaluated by 3 different raters, resulting in 300 datapoints per model comparison. As for MOS evaluations, each rater did not necessarily evaluate all 100 sentence comparisons. In the following sections, we present the experimental setups and the results of the pairwise evaluations.

4.5. Multidialectal Bootstrapping

In this section we evaluate the effectiveness of building voices for new dialects with a single multidialectal baseline model, as opposed to training individual monodialectal baseline models for each dialect separately. We investigate

[5]MOS might not be directly comparable across voices (e.g., different raters for a same sentence across models, effects of order of presentation, range-equalization bias); we provide them as a general idea of voice quality (i.e., naturalness only) but base ourselves on more reliable A/B testing for model comparison.

A	B	Winner	Score
Multi(AR)	Mono(AR)	Multi	-1.570 ± 0.146
Multi(CL)	Mono(CL)	Multi	-0.540 ± 0.176
Multi(PE)	Mono(PE)	Multi	-0.487 ± 0.158
Multi(VE)	Mono(VE)	Multi	-0.487 ± 0.146

Table 6: Monodialectal vs. multidialectal baselines. Negative A/B scores denote preference for voice A over voice B as a virtual assistant of the raters' native locales (e.g. 'a Peruvian virtual assistant'). A winner is determined when the preference is statistically significant.

A	B	Winner	Score
Hold(VE)	Mono(VE)	Hold(VE)	-0.600 ± 0.166
Hold(VE)	Multi(VE)	Hold(VE)	-0.633 ± 0.148
Multi(VE)	Multi(PE)	Multi(PE)	0.580 ± 0.128
Hold(VE)	Multi(PE)	Multi(PE)	0.423 ± 0.182

Table 7: A/B test results involving the VE multidialectal holdout voices. Negative scores denote preference for voice A over voice B as a Venezuelan virtual assistant, and vice versa for positive scores. A winner is determined when the preference is statistically significant.

if the voices built with a single multidialectal model outperform their monolingual baseline counterparts. On the one hand, the multidialectal model is trained using more data than the baseline models. On the other hand, the multidialectal data is also more acoustically varied, whether it is at the level of phonemes or prosody, which may counter the data quantity advantage. Indeed, the perceived quality of the resulting voices depends not only on them sounding human-like (i.e., naturalness), but also on them being able to properly display the dialect of interest (i.e., nativeness).

For the A/B tests, every sentence in the evaluation set was read by the monodialectal baseline voice on one side and the multidialectal baseline voice of the target dialect on the other. Both voices were built using the same speaker ID, meaning that they only differed in the data used for training the models. Native listeners of AR, CL, PE, and VE Spanish compared each sentence read by the two voices side by side, in their respective native dialects only.

As seen in Table 6, for all dialects, the multidialectal voices were preferred over the respective monodialectal voices. In this low-resource setting, adding data from closely-related dialects consistently enhanced the perceived quality across all dialects. Based on these findings on low-resource settings, the next step is to transfer the methodology to the zero-resource setting.

4.6. Dialect-Specific Zero-Resource TTS

Collecting data for a target dialect is costly both in terms of time and money, and might be logistically infeasible even through crowdsourcing. Is it possible to bootstrap an extant multidialectal corpus to build a satisfying voice in a dialect that is not present in the corpus?

In this experiment we set our target dialect to VE Spanish[6]. We compared the perceived quality of the VE monolingual baseline voice to what we refer to as a VE multidialectal holdout voice, Hold(VE). Namely, this corresponds to a multidialectal TTS model trained with data from all Spanish varieties except VE Spanish (i.e., AR, CL, PE, ES, US). Since this model does not include data from the chosen VE speaker ID, a speaker ID was selected from the pool of crowdsourced speakers by a native speaker of VE Spanish. The criteria for choosing the voice was an optimisation of both voice naturalness and nativeness (here: which voice subjectively sounds least foreign). As a result, a PE speaker was chosen as the speaker ID for the VE multidialectal holdout voice[7].

As seen in Table 7, A/B tests revealed that native VE raters preferred the VE multidialectal holdout voice (thus, with a PE speaker ID) over the VE monodialectal baseline. As such, a non-native-sounding voice was preferred.

Results from Section 4.5 show that multidialectal baseline models outperformed their monodialectal counterparts, possibly due to having more data for training the acoustic models. So it is possible that the VE multidialectal holdout performs better than the VE monodialectal voice due to differences in training corpus size. In order to investigate this, we compared the VE multidialectal holdout voice to the VE multidialectal baseline, which was previously shown to be preferred to the VE monodialectal baseline. As seen in Table 7, the VE multidialectal holdout voice was once again preferred over the VE multidialectal baseline voice. In fact, only one rater showed the opposite preference (Figure 2). Note that the multidialectal baseline voices are trained with more data than the multidialectal holdout voices, suggesting that the preference for the holdout voice with the PE speaker ID is not simply due to difference in training corpus size.

Furthermore, we confirmed this tendency to prefer a PE voice by our VE raters when comparing the PE multidialectal baseline to the VE multidialectal baseline. As seen in Table 7, the former, non-native voice was preferred to the latter on average.

As a result, we find that it is possible to build a voice for a target dialect that is not present in a multidialectal corpus[8]. In addition, we found that the PE-sounding holdout voice was rated to be more appropriate as a Venezuelan virtual assistant than the mono- and multidialectal VE voices by Venezuelan raters. Therefore, if a suitable speaker ID is available (e.g., from a dialect with relatively neutral prosody), it may be possible to skip costly data collection and build a non-native voice that outperforms native voices. However, depending on the differences across dialects and locales, a suitable speaker might not be available, or a region-neutral voice may not be preferred. In these cases, is it possible to enhance a holdout voice and make

[6]Due to VE being the corpus with the least amount of data, at the time of the experiments.

[7]Coincidentally, the same speaker ID as all non-holdout PE voices.

[8]Since speaker ID and dialect are conflated in our models, we do not claim to have been able to successfully build a voice *in* the target dialect, but a voice *deemed acceptable by speakers of the target dialect*.

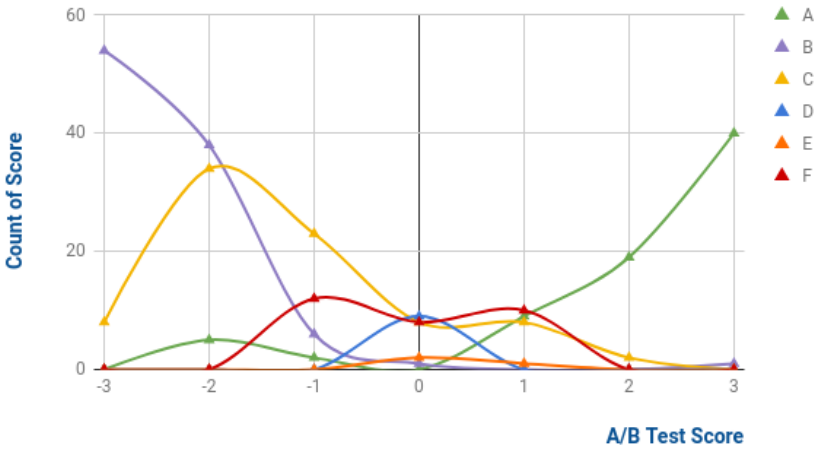

Figure 2: Distribution of scores in the A/B task comparing multidialectal holdout (PE speaker ID) and baseline (VE speaker ID) voices for VE, as rated by native VE listeners A to F (one line per rater). Negative and positive scores show a preference for the holdout and baseline voices, respectively. The further away from zero, the stronger the preference.

it sound more native with minimal additions to the multidialectal training corpus?

4.7. Low-Resource Dialect Tuning

As pointed out in Section 3.2, in addition to the differences at the level of prosody, lexicons, grammar, or syntax, dialects often differ in their phonemic inventories and/or in the phonetic realizations of common phonemes. Some of these phonetic and phonemic peculiarities may be perceived as quintessential characteristics of a target dialect. For instance, the pronunciation of graphemes <ll> and <y> as [ʃ∼ʒ] by (Rio Platense) Argentinian speakers is a distinct characteristic of this dialect (Lipski, 1994; Real Academia Española y Asociación de Academias de la Lengua Española, 2011). We examined the possibility of enhancing an AR multidialectal holdout voice by maximizing the amount of [ʃ] tokens in the training corpus. We first built an AR multidialectal holdout voice, containing data from CL, PE, VE, ES, and US Spanish varieties. This training corpus (i.e., the corpus described in Section 4.3.2, but without AR data) contained randomly selected US utterances, resulting in the presence of only 16 [ʃ] tokens (8 from US, 8 from other locales). By actively biasing the US utterance selection as to maximize the ratio of utterances containing the phoneme, we obtained a corpus with 724 [ʃ] tokens (the total number of words was kept constant). We call the resulting voice the AR multidialectal phoneme-selection holdout voice, differing from the AR multidialectal holdout in the percentage of [ʃ] tokens. Argentinian raters compared these two voice in an A/B test with 12 sentences containing the target phone [ʃ]. Results show that sentences from the phoneme-selection holdout were significantly preferred over those by the non-biased holdout (mean = -1.389). Note that despite this phoneme-level enhancement, the phoneme-selection holdout voice was not rated significantly better than the non-biased holdout on the A/B testing with 100 sentences (A/B score: -0.080 ± 0.135).

Following these results, we investigated the possibility of adding minimal target dialect data to an extant training corpus as an alternative to needing to compile an entire target corpus. This was done by adding only the data for the VE speaker used as speaker ID to the corpus used to train the multidialectal holdout model for VE. The resulting model, referred to as the VE multidialectal semi-holdout, was compared to the VE monodialectal baseline. The multidialectal semi-holdout voice was preferred over the monodialectal baseline voice (0.330 ± 0.122), but the VE multidialectal baseline still outperforms the VE multidialectal semi-holdout (0.360 ± 0.148). These results show that if a multidialectal corpus already exists, it might be preferable to record one speaker of the new target dialect and bootstrap from the rest of the existing data than compiling a full target corpus. On the other hand, if resources allow, the more target data available, the better.

5. Discussion and Conclusions

5.1. Summary

In this work we presented a crowdsourced multidialectal corpus of various dialects of Latin American Spanish. In total, we recorded 44 Argentinian, 31 Chilean, 33 Colombian, 38 Peruvian, 5 Puerto Rican, and 23 Venezuelan native speakers. In total, almost 40 hours of speech were recorded. We believe that this corpus can prove to be a valuable resource for developing speech technologies such as TTS and automatic speech recognition (ASR) in these dialects. We also hope that this corpus will assist the practitioners in the field of Latin Americal Spanish dialectology.

In parallel to data collection, we explored using a multidialectal corpus (subset of all the data collected) in order to build TTS voices for various low-resource dialects, as well as dialects not present in the training corpus. We found that voices built from a multidialectal model were preferred as the voice of a native virtual assistant over the corresponding voices built from monodialectal corpora. The multidialectal model appears as a more parsimonious option, as only one shared model is trained instead of one per dialect.

We also showed that it is possible to use a multidialectal model to build a satisfactory voice for a dialect not present in the training corpus, given that there is a suitable replace-

ment dialect available in the training corpus. In cases such as these, it might be worth evaluating the need to build a new dialect-specific voice before committing to costly data collection. However, if nativeness of the resulting voice is of paramount importance, our results show that it is possible to obtain a native voice from the multidialectal corpus with minimal data collection (here, by merely adding 150 utterances from one native speaker to the training corpus). It is also possible to selectively enhance a specific phonetic realization (as seen with the AR phoneme /ʃ/) by manipulating the amount of exemplars available in the training dataset, but the overall impact may be limited.

5.2. Future Work

A logical segue from our last result is that the multidialectal model presented here may benefit from becoming a multidialectal-multilingual model. Namely, adding data from related languages (e.g., Catalan, Italian, Portuguese) in order to increase the number of instances for each phoneme. It might even be interesting to add seemingly unrelated languages to the mix, if the phonetics are similar. For instance, adding Japanese to the Latin American multidialectal model would allow [h], the VE phonetic realisation of /x/, to be separated from the ES Spanish phoneme /x/, and similarly for CL phones [ts] and [ç] (instead of /tʃ/, and /x/ in certain environments, respectively). Going even further, the ultimate goal would be to establish a universal phonemic inventory in order to bootstrap existing data to build voices in various languages, even low- or zero-resource ones.

It was unexpected for the VE multidialectal holdout voice (i.e. with a PE speaker ID) to be rated higher than not only the VE monodialectal baseline, but also the VE multidialectal voices, as it means that VE raters preferred a voice following a dialect with different prosody. Investigating the reason for this preference was out of the scope of these experiments. Some possible variables that could be controlled for in future studies include recording quality (VE data were collected in multiple locations, while PE data were collected in a single venue), the perceived neutrality of the accent (raters might be biased towards a more region-neutral accent as a virtual assistant voice because of market expectations, media exposure, or other sociolinguistic constraints) and the subjective pleasantness of the speaker ID. Ideally, future models should allow the decorrelation of dialect and speaker ID.

Concerning TTS evaluation methods, we observed some discrepancy between MOS and A/B test results. Specifically, the baseline monolingual VE voice was rated numerically higher than the multidialectal VE voice, which had outperformed the monolingual voice in A/B tests. It should be noted that the MOS task asked raters to evaluate voices' naturalness, while the A/B task required them to identify the best voice for a virtual assistant in their locale, a task where not only naturalness, but also nativeness was to be considered in the comparison. Additionally, even though the raters come from the same rater pool, due to the blind rating nature of the tests we could not investigate more controlled scenarios. For example, we could not ensure that the same rater evaluated the same sentences with different voices across MOS tests of different voices, or the MOS and the A/B tests. Different evaluation setups can provide more insight into the relation between A/B tests and MOS tests (e.g. effect of doing separate, explicit evaluations of voice naturalness and nativeness; effect of various types of MOS calibration on MOS-A/B score correlations; effect of varying the number of raters). Given the limited number of available raters in the context of low-resource languages and dialects, such as some presented in this work, it becomes critical to do a more thorough examination of the popular, yet subjective, MOS evaluation method in the near future.

6. Acknowledgements

The first author acknowledges support by the Japan Society for the Promotion of Science (Postdoctoral Fellowship). The authors would like to thank Ana Sofía Rosas, Álvaro Muñoz Brandon and Héctor Fernández Alcalde for their help with data collection and transcription, and Martin Jansche, Clara Rivera, and Linne Ha for continuous feedback. Lastly, the authors thank the crowdsourcing speakers for their recordings, the anonymous raters for their evaluations, and the anonymous reviewers for many helpful comments and suggestions.

7. Bibliographical References

Agiomyrgiannakis, Y. (2015). VOCAINE the vocoder and applications in speech synthesis. In *Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, pages 4230–4234, Brisbane, Australia, April. IEEE.

Baljekar, P., Rallabandi, S., and Black, A. W. (2018). An Investigation of Convolution Attention Based Models for Multilingual Speech Synthesis of Indian Languages. In *Proc. Interspeech 2018*, pages 2474–2478, Hyderabad, India.

Baljekar, P. (2018). *Speech Synthesis from Found Data*. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA.

Canfield, D. L. (1981). *Spanish Pronunciation in the Americas*. Language and Linguistics: Language Studies. The University of Chicago Press.

Chen, Y.-J., Tu, T., Yeh, C.-c., and Lee, H.-y. (2019). End-to-End Text-to-Speech for Low-Resource Languages by Cross-Lingual Transfer Learning. In *Proc. of Interspeech 2019*, pages 2075–2079, Graz, Austria, September.

Colantoni, L. and Gurlekian, J. (2004). Convergence and intonation: historical evidence from Buenos Aires Spanish. *Bilingualism: Language and cognition*, 7(2):107–119.

Cooper, E. L. (2019). *Text-to-Speech Synthesis Using Found Data for Low-Resource Languages*. Ph.D. thesis, Columbia University, New York.

Correa, P., Rueda, H., and Arguello, H. (2010). Síntesis de voz por concatenación de difonemas para el español de Colombia. *Revista Iberoamericana en Sistemas, Cibernética e Informática*, 7(1):19–24.

Creative Commons. (2019). Attribution-ShareAlike 4.0 International (CC BY-SA 4.0). `http://creativecommons.org/licenses/by-sa/4.0/deed.en`.

Demirsahin, I., Jansche, M., and Gutkin, A. (2018). A Unified Phonological Representation of South Asian Languages for Multilingual Text-to-Speech. In *Proc. of SLTU-2018*, pages 80–84, Gurugram, India, August.

Emeneau, M. (1956). India as a Linguistic Area. *Language*, 32(1):3–16.

Feldhausen, I., Pesková, A., Kireva, E., and Gabriel, C. (2011). Categorical Perception of Porteño Nuclear Accents. In *Proc. 17th International Congress of Phonetic Sciences (ICPhS)*, pages 116–119, Hong Kong, China.

Florentino, Y. I. A. (2016). Base de datos de difonos para la síntesis del habla del español peruano. Master's thesis, Universidad Nacional del Santa, Nuevo Chimbote, Peru.

Fukada, T., Tokuda, K., Kobayashi, T., and Imai, S. (1992). An adaptive algorithm for mel-cepstral analysis of speech. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 137–140. IEEE.

Ganchev, T., Fakotakis, N., and Kokkinakis, G. (2005). Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task. In *Proc. SPECOM*, volume 1, pages 191–194.

Gutkin, A., Ha, L., Jansche, M., Kjartansson, O., Pipatsrisawat, K., and Sproat, R. (2016). Building Statistical Parametric Multi-Speaker Synthesis for Bangladeshi Bangla. In *5th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU '16)*, pages 194–200.

Gutkin, A. (2017). Uniform Multilingual Multi-Speaker Acoustic Model for Statistical Parametric Speech Synthesis of Low-Resourced Languages. In *Proc. of Interspeech 2017*, pages 2183–2187, Sweden, August.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hunt, A. J. and Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings (ICASSP)*, volume 1, pages 373–376.

International Phonetic Association. (1999). *Handbook of the International Phonetic Association*. Cambridge University Press.

Li, B. and Zen, H. (2016). Multi-Language Multi-Speaker Acoustic Modeling for LSTM-RNN based Statistical Parametric Speech Synthesis. In *Proc. of Interspeech*, pages 2468–2472, San Francisco, September. ISCA.

Li, B., Sainath, T., Sim, K. C., Bacchiani, M., Weinstein, E., Nguyen, P., Chen, Z., Wu, Y., and Rao, K. (2018). Multi-Dialect Speech Recognition with a Single Sequence-to-Sequence Model. In *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4749–4753. IEEE.

Lipski, J. M. (1994). *Latin American Spanish*. Longman London.

Mapelli, V., Popescu, V., Liu, L., and Choukri, K. (2016). Language Resource Citation: the ISLRN Dissemination and Further Developments. In *Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1610–1613, Portorož, Slovenia, May. ELRA.

Miao, Y., Li, J., Wang, Y., Zhang, S.-X., and Gong, Y. (2016). Simplifying long short-term memory acoustic models for fast training and decoding. In *Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, pages 2284–2288, Shanghai, China, March. IEEE.

O'Rourke, E. (2012). Intonation in Spanish. In José Ignacio Hualde, et al., editors, *The Handbook of Hispanic Linguistics*, pages 173–191. Blackwell Publishing.

Ortiz-Lira, H. (1999). La aplicación de ToBI a un corpus del español de Chile. *Onomázein*, (4):429–442.

Penny, R. and Penny, R. J., (2004). *Variation and Change in Spanish*, chapter 5, pages 136–173. Cambridge University Press.

Phillips, A. and Davis, M. (2009). BCP 47 – Tags for Identifying Languages. *IETF Trust*.

Povey, D. (2019). Open SLR. `http://www.openslr.org/resources.php`. Accessed: 2019-03-30.

Real Academia Española y Asociación de Academias de la Lengua Española. (2011). *Nueva gramática de la lengua española: Fonética y fonología*. Madrid: Espasa.

Resnick, M. C. (2012). *Phonological Variants and Dialect Identification in Latin American Spanish*, volume 201 of *Janua Linguarum. Series Practica*. de Gruyter Mouton, 2nd edition.

Rodríguez, M., Mora, E., and Cavé, C. (2006). Síntesis de voz en el dialecto venezolano por medio de la concatenación de difonos. *Ciencia e Ingeniería*, 27(1):17–24.

Sak, H., Senior, A. W., and Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Proc. of Interspeech*, pages 338–342, Singapore, September. ISCA.

Sproat, R., Black, A. W., Chen, S., Kumar, S., Ostendorf, M., and Richards, C. (2001). Normalization of non-standard words. *Computer Speech & Language*, 15(3):287–333.

Streijl, R. C., Winkler, S., and Hands, D. S. (2016). Mean Opinion Score (MOS) revisited: Methods and applications, limitations and alternatives. *Multimedia Systems*, 22(2):213–227.

Torres, H. M., Gurlekian, J. A., and Mercado, C. (2012). Aromo: Argentine Spanish TTS System. In *Proc. VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop, IberSPEECH*, pages 416–421.

Torres, H. M., Gurlekian, J. A., Evin, D. A., and Mercado, C. G. C. (2019). Emilia: a speech corpus for Argentine Spanish text to speech synthesis. *Language Resources and Evaluation*, 53(3):419–447.

Violante, L. (2012). *Construcción y evaluación del backend de un sistema de síntesis de habla en español argentino*. Ph.D. thesis, Tesis de Licenciatura, Universidad de Buenos Aires, Buenos Aires, Argentina, August.

Wibawa, J. A. E., Sarin, S., Li, C., Pipatsrisawat, K., Sodimana, K., Kjartansson, O., Gutkin, A., Jansche, M., and Ha, L. (2018). Building Open Javanese and Sundanese Corpora for Multilingual Text-to-Speech. In *Proc. of the 11th International Conference on Language Resources*

and Evaluation (LREC), pages 1610–1614, Miyazaki, Japan, May.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2006). *The HTK Book*. Cambridge University Engineering Department.

Yu, K. and Young, S. (2011). Continuous F0 modeling for HMM based statistical parametric speech synthesis. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(5):1071–1079.

Zeiler, M. D., Ranzato, M., Monga, R., Mao, M., Yang, K., Le, Q. V., Nguyen, P., Senior, A., Vanhoucke, V., Dean, J., and Hinton, G. (2013). On rectified linear units for speech processing. In *Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, pages 3517–3521, Vancouver, Canada, May. IEEE.

Zen, H. and Sak, H. (2015). Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In *Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, pages 4470–4474, Brisbane, Australia, April. IEEE.

Zen, H., Senior, A., and Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. In *Proc. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7962–7966, Vancouver, Canada, May.

Zen, H., Agiomyrgiannakis, Y., Egberts, N., Henderson, F., and Szczepaniak, P. (2016). Fast, Compact, and High Quality LSTM-RNN Based Statistical Parametric Speech Synthesizers for Mobile Devices. In *Proc. Interspeech 2016*, pages 2273–2277, San Francisco, September.

8. Language Resource References

Google. (2019a). *Crowd-sourced high-quality Argentinian Spanish speech data set by Google*. Google, distributed by Open Speech and Language Resources (OpenSLR), `http://www.openslr.org/61/`, Google crowd-sourced resources, 1.0, ISLRN 395-001-133-368-2.

Google. (2019b). *Crowd-sourced high-quality Chilean Spanish speech data set by Google*. Google, distributed by Open Speech and Language Resources (OpenSLR), `http://www.openslr.org/71/`, Google crowd-sourced resources, 1.0, ISLRN 048-218-632-043-6.

Google. (2019c). *Crowd-sourced high-quality Colombian Spanish speech data set by Google*. Google, distributed by Open Speech and Language Resources (OpenSLR), `http://www.openslr.org/72/`, Google crowd-sourced resources, 1.0, ISLRN 169-985-498-793-0.

Google. (2019d). *Crowd-sourced high-quality Peruvian Spanish speech data set by Google*. Google, distributed by Open Speech and Language Resources (OpenSLR), `http://www.openslr.org/73/`, Google crowd-sourced resources, 1.0, ISLRN 923-742-092-167-6.

Google. (2019e). *Crowd-sourced high-quality Puerto Rican Spanish speech data set by Google*. Google, distributed by Open Speech and Language Resources (OpenSLR), `http://www.openslr.org/74/`, Google crowd-sourced resources, 1.0, ISLRN 721-732-548-994-0.

Google. (2019f). *Crowd-sourced high-quality Venezuelan Spanish speech data set by Google*. Google, distributed by Open Speech and Language Resources (OpenSLR), `http://www.openslr.org/75/`, Google crowd-sourced resources, 1.0, ISLRN 697-927-390-879-1.