Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 21–27
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Open-Source High Quality Speech Datasets for Basque, Catalan and Galician

Oddur Kjartansson, Alexander Gutkin, Alena Butryna, Işın Demirşahin, Clara Rivera
Google Research
United Kingdom and United States
{oddur, agutkin, alenab, isin, rivera}@google.com

Abstract

This paper introduces new open speech datasets for three of the languages of Spain: Basque, Catalan and Galician. Catalan is furthermore the official language of the Principality of Andorra. The datasets consist of high-quality multi-speaker recordings of the three languages along with the associated transcriptions. The resulting corpora include over 33 hours of crowd-sourced recordings of 132 male and female native speakers. The recording scripts also include material for elicitation of global and local place names, personal and business names. The datasets are released under a permissive license and are available for free download for commercial, academic and personal use. The high-quality annotated speech datasets described in this paper can be used to, among other things, build text-to-speech systems, serve as adaptation data in automatic speech recognition and provide useful phonetic and phonological insights in corpus linguistics.

Keywords: Speech Corpora, Open Source, Basque, Catalan, Galician

1. Introduction

Castilian Spanish is the official language of entire Spain. In addition, Basque, Catalan and Galician are the official languages of the three respective autonomous communities of the Basque Country, Catalonia and Galicia (Hoffmann, 1996; Lasagabaster, 2011). Catalan is also spoken in Valencia, Balearic Islands, Andorra, French Catalonia, and a small region of Sardinia. Basque is also spoken in Navarre and the French Basque Country. According to Ethnologue (2019), Basque has close to 800,000 native (first language or L1) speakers, Catalan around 4 million native speakers and Galician around 1.5 million native speakers. Out of the three languages, UNESCO considers Basque to be endangered (Moseley, 2010).

Since the 1980s there has been a resurgence of these languages due to democratization of the central government's cultural and language policies towards the regions (Ferrer, 2000; O'Rourke and Ramallo, 2013; Gorter et al., 2014). As part of this process, considerable work has gone into building speech and language technologies for these languages, especially since late 1990s (López de Ipiña et al., 1995; Villarrubia et al., 1998). Despite evident progress, the availability of speech and language technology in these languages is still not on par with Castilian Spanish and the scarcity of linguistic resources available for building competitive systems, especially in Basque and Galician, has often been pointed out by the researchers (Agić et al., 2016; Vania et al., 2019).

Building language resources is expensive. It can be time consuming to set up the recording logistics, collect and analyze the data. In the case of low-resource languages, finding linguistic experts can become an added factor of complexity. When collecting high-quality speech resources for applications such as text-to-speech, further complications arise as one needs to secure a location for the recording, as well as find an adequate voice talent. Our work relies on several methods that were proposed to mitigate some of these issues. In case of the recording script preparation, freely available text resources such as Wikipedia are used. In addition, using templates which are automatically filled out to cover local place names, important holidays and prominent figures help reduce the time required for recording script design (Wibawa et al., 2018). Finally, to mitigate the cost of professional voice talents, multiple volunteer speakers are used instead of relying on one person (Gutkin et al., 2016).

In this paper we present open high-quality speech resources for three languages, Basque (Google, 2019a), Catalan (Google, 2019b) and Galician (Google, 2019c). The corpora are distributed under a "Creative Commons Attribution-ShareAlike" (`CC BY-SA 4.0`) license (Creative Commons, 2019) and are freely available for download from Open Speech and Language Resources (OpenSLR) repository (Povey, 2019). Similar speech resources for these three minority languages of Spain have been developed in the past. These resources, however, vary in either availability (academic-only or unclear licensing terms, non-free distribution) or quality (low quality, e.g. 16kHz, recordings), and sometimes both. The main contribution of the work described in this paper is the corpora that is both free for commercial and academic use, and is of sufficiently high-quality to be used in state-of-the-art speech applications, such as multilingual multi-speaker text-to-speech (Chen et al., 2019). To the best of our knowledge, based on the review of existing databases provided in the next section, our three datasets are among the very first truly free resources available online for public use.

The rest of this paper is organized as follows: Section 2 provides an overview of related corpora. Brief linguistic introduction into the languages in question is given in Section 3. The details of the recording script design, the recording process and corpora details are provided in Section 4. Section 5 concludes the paper.

2. Related Corpora

Considerable effort has gone into developing speech resources for Basque, Catalan and Galician in the past. Among the databases that cover multiple minority languages, Rodriguez-Fuentes et al. (2012) describe a large TV Broadcast database developed for automatic speech recognition (ASR) of Basque, Catalan and Galician in clean and noisy environments. The licensing terms are negotiable with the authors.

Basque Basque is included as part of the open-source CMU Wilderness Multilingual Speech Dataset (Black, 2019) containing Bible translation for over 700 languages. Sainz et al. (2012) introduced a high-quality text-to-speech (TTS) database of Basque containing six hours of speech recorded by 11 speakers, with the availability of the corpus being unclear. The database was used by the authors to successfully build statistical parametric speech synthesis system based on Hidden Markov Models (HMMs) using their prior work (Erro et al., 2010). One of the earliest attempts to develop a parallel corpus of Basque and Spanish was undertaken by Pérez et al. (2006), who developed a weather forecast corpus consisting of 28 months of spoken daily weather forecast reports in Spanish and Basque, which were used in speech-to-speech translation (Pérez et al., 2008), language identification (Guijarrubia and Torres, 2010) and ASR (Guijarrubia et al., 2009). Pérez et al. (2012) later described a more sophisticated parallel corpus of Spanish and Basque that includes both text and speech data and consists of the proceedings of the Basque Parliament. The speech portion of the corpus contains 189 hours of speech from 81 speakers. The licensing of this corpus appears unclear and it cannot be located online. The other, more specialized, corpora developed for Basque include the Emotional Speech Database for corpus-based speech synthesis by Saratxaga et al. (2006) that consists of approximately 20 hours of high-quality recordings, evaluated in detail by Sainz et al. (2008), and a smaller 1.5 hour-long multimodal audiovisual database of emotional speech developed by Navas et al. (2004) for prosody studies. Further Basque speech resources are hopefully going to be developed as part of the BerbaTek project, an joint effort by various academic and commercial organizations in the Basque Autonomous Community to increase the availability of speech and language technologies (Arrieta et al., 2008; Leturia et al., 2018).

Catalan Bonafonte et al. (1997) from Universitat Politècnica de Catalunya (UPC) describe one of the earliest datasets of Catalan developed for bilingual Spanish-Catalan unit selection TTS, detailed in (Bonafonte et al., 1998). Additional small corpus of Catalan consisting of 3,600 short utterances was recorded at UPC for prosodic modeling (Febrer et al., 1998). Around the same time Hernando and Nadeu (1999) from UPC developed SpeechDat – a Catalan speech database that contains recordings of 2,000 speakers (each uttering around 50 sentences) over fixed telephone lines. The database is primarily intended for ASR systems (Mariño et al., 2000; Padrell and Mariño, 2002) and is distributed by ELRA under the restricted license. The lack of emotional speech database for Catalan was first noticed by Iriondo et al. (2004), who built emotional HMM-based TTS piggybacking on the existing corpus of Castilian Spanish. Bonafonte et al. (2008) describe a Catalan text-to-speech database consisting of 10 hours of single male and female speaker recordings. This resource is free for academic and commercial use, but it does not seem to be available online. This database was used by the authors to produce HMM-based TTS voices (Bonafonte et al., 2009). The most recent development is the open ASR database described by Külebi and Öktem (2018) consisting of 240 hours of transcribed Catalan TV broadcasts that is freely available online. A slightly outdated work by Moreno et al. (2006) and Schulz et al. (2008) provides reviews of existing programs for Catalan and a roadmap for constructing speech and language applications.

Galician There has been a reasonably late focus on speech applications in Galician, with one of the earliest efforts undertaken by Dieguez-Tirado et al. (2005), who built a bilingual ASR system for Galician and Spanish based on corpus of TV shows, and by González et al. (2008), who outlined the challenges of the language, such as homograph disambiguation (Mourín et al., 2009), for speech processing. Proprietary database of Galician was used by Microsoft to build HMM TTS system (Braga et al., 2010). This effort was a joint collaboration with University of Vigo that resulted in a speech database consisting of 10,000 utterances (Campillo et al., 2010). Among recent work is the large high-quality corpus of spoken Galician annotated on multiple linguistic levels (García-Mateo et al., 2014) that consists of 98 hours of recordings and the corresponding transcriptions. The corpus is mostly intended for corpus linguistic studies and sociolinguists as it contains recordings of Galician in various styles and dialects over a long period of time starting in 1960s and may not be suitable for building speech applications. Similar to Basque, there is a growing awareness of the need to increase the availability of Galician language resources (Mateo and Rodríguez, 2012).

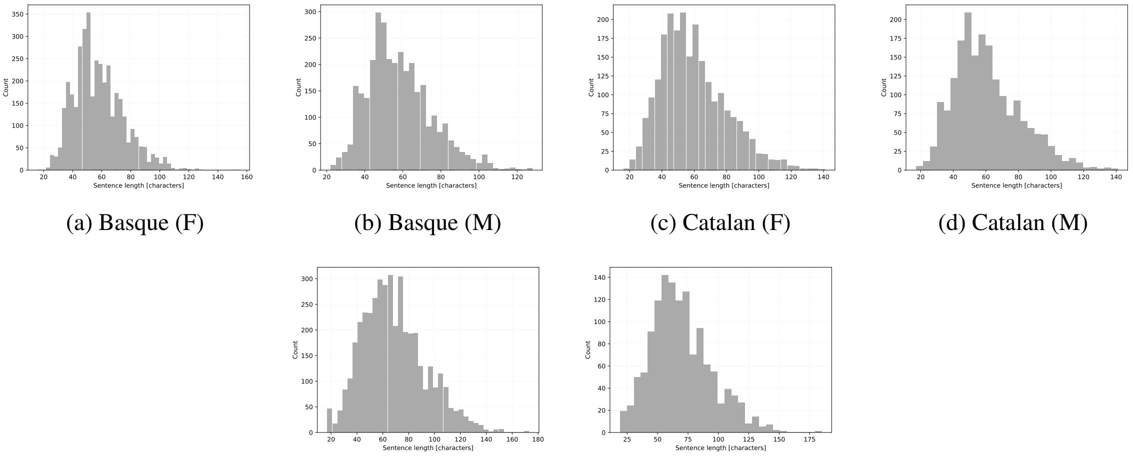
3. Overview of the Languages

There is a general consensus that Basque is a language isolate with no known relatives and uncertain origins (Hualde et al., 1996; Etxeberria, 2008; Trask, 2013). Basque is an inflectional and agglutinative language (Hualde and de Urbina, 2003; King, 2012), its grammatical relations between components within a clause are represented by suffixes, and many words consist of compounded morphemes. Among several uses, the suffixes are used to mark over a dozen cases and four definite determiners (Albizu, 2002; King, 2012). This markedly distinguishes Basque from its neighbouring Romance languages. Although the traditional phonology of Basque is noticeably different from Spanish (Hualde, 1991; Bengtson, 2003), Hualde (2015) mentions the acceleration in the processes of convergence between pronunciations of the two languages, whereby "for many speakers of the younger generations there are not many phonological or phonetic differences between their Basque and their Spanish, if any".

Catalan is an Indo-European language of the Romance family sharing many traits with the neighbouring Romance languages from Ibero- and Gallo-Romance groups (Posner, 1996; Hualde, 2013), yet differing from them in several respects, such as phonology, which places the language roughly between Spanish, French and Italian (Wheeler,

Language	Gender	Lines	Tokens					Chars			Speakers	Audio Duration	
			min	max	avg	Total	Unique	min	max	avg		Total [h:m:s]	Average[s]
Basque	F	3,858	1	20	8.0	30,901	8,583	17	156	58.1	29	7:26:36	6.77
	M	3,278	1	18	8.0	26,383	8,030	23	129	58.3	23	6:36:00	7.25
Catalan	F	2,321	2	24	10.5	24,385	6,568	17	142	59.5	20	5:24:00	8.38
	M	1,919	2	29	10.6	20,261	6,514	28	141	60.8	16	4:01:12	7.53
Galician	F	4,264	3	28	11.6	49,674	6,530	18	174	68.3	34	7:40:12	6.48
	M	1,324	4	28	11.7	15,462	4,336	20	186	69.4	10	2:38:24	7.19
Total	–	**16,963**	–	–	–	167,066	–	–	–	–	132	33:35:19	–

Table 1: Details of the recording script lines and the audio properties of the corpora.



(a) Basque (F)

(b) Basque (M)

(c) Catalan (F)

(d) Catalan (M)

(e) Galician (F)

(f) Galician (M)

Figure 1: Histograms of the utterance length (in characters) by language and gender (x-axis shows the lengths, y-axis the frequency).

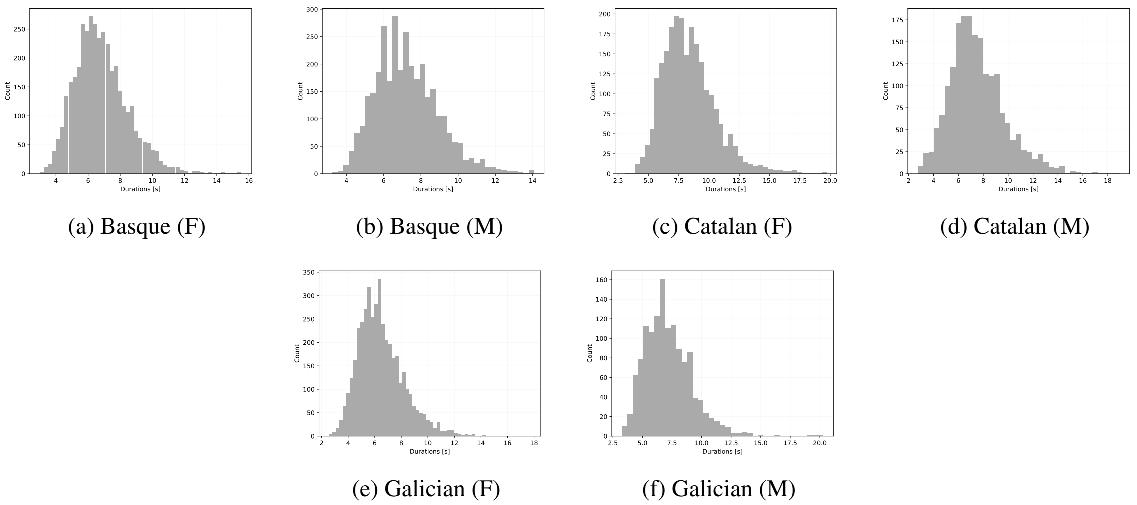
2005). Some of these differences from Spanish are manifest in greater complexity in terms of syllable structure types (Prieto et al., 2012), the existence of vowel reduction (Cabré, 2009) and dissimilar vowel and consonant inventories, such as the existence of three more vowels and voiced fricatives (Pallier et al., 2001).

According to an areal classification Galician belongs to the Ibero-Romance family together with Spanish and Portuguese. Galician shares strong similarities with Portuguese, such as possible use of inflectional endings in infinitive forms, both languages descending from the same medieval ancestor deriving from Vulgar Latin (Holt, 2016; Martìnez-Gil, 2020). However, similar to Vulgar Latin and Italian, Galician distinguishes seven vowels, which differs it from modern Portuguese that has two extra vowels and Spanish with its five vowels (Kabatek and Pusch, 2011; Gibson and Gil, 2019). In other respects, such as phonotactics, the language is more similar to Spanish (Harris, 1983).

4. Corpus Design and Overview

Script Design Recording scripts were generated by native speakers using a mixture of Wikipedia and template sentences. The templates were filled in using local and global entity names including businesses, places and people. Examples of such templates include sentences of the form “*person* traveled from $place_A$ to $place_B$ at *time*” and “*event* is celebrated in *place* on *date*”, where the variable template slots are denoted by italics. Using a template method makes it relatively easy to automatically generate sentences for the script. However, the applicability of this method is language-dependent. In the case of morphologically rich languages, such as Basque, the template wording might need to be changed depending on grammatical context. Using this process we generated up to 5,000 sentences for each of the languages, which were then proofread and hand-tuned (if necessary) by the native speakers. Even though transcriptions mostly contain sequences of natural language words, because they have not been text normalized they also contain non-standard word (NSW) expressions, such as numbers (Sproat et al., 2001). Therefore, here and below we refer to the constituent space-separated elements of transcriptions as “tokens” rather than words. The total number of script lines, the minimum, maximum and average number of tokens and characters (including spaces) per sentence for each language and gender are shown in the first nine columns of Table 1. Please note, some sentences may contain a single token, such as telephone number (e.g., Basque “*zortzi-zazpi-bi-bederatzi-zazpi-bost-zero-zero-zero*”). The corresponding distributions of sentence lengths per language and per gender are represented as histograms in Figure 1. The distribution shapes and the modes for all the datasets are roughly similar, with Galician having the longest orthographic representation if sentences of over 120 characters are considered. According to Table 1, Galician also has the highest average number of characters per sentence.

Recording Process The recordings took place in three different locations in Spain. Catalan was recorded in Barcelona, Basque was recorded in Bilbao and Galician



(a) Basque (F)

(b) Basque (M)

(c) Catalan (F)

(d) Catalan (M)

(e) Galician (F)

(f) Galician (M)

Figure 2: Histograms of the utterance durations (in seconds) by language and gender (x-axis shows duration, y-axis the frequency).

was recorded in Santiago de Compostela. Instead of renting professional recording studios, sound insulated rooms were used in each location. Volunteer amateur native speakers were sourced with the help from local groups in each area and represent a variety of local accents.

The audio was recorded using a Neuman KM184 diaphragm condenser cardioid microphone, a Blue ICICLE XLR to USB analogue to digital (A/D) converter, which also provides power to the microphone. The USB A/D converter was connected to an Asus Zenbook fanless laptop, which the participants used to control the recordings. The microphone was put on a microphone stand, and adjusted for each volunteer. The microphone was kept at a distance of 30 cm from the mouth of the volunteer, slightly off center and pointing either down towards or up towards the volunteer (approximately 5 degrees off the center on both axes).

A proprietary Web-based recording software was used for the recordings. The setup is designed for self-service, so that the speaker both records and controls the recordings. Using a self-serve model eliminates the need for an extra person to control the recordings. Quality of the audio can be monitored from another computer, once the volunteer has saved the recordings. Most speakers were able to record about 150 sentences in the span of an hour, which included a short break about half way through the recordings. A few minutes were needed to familiarize them with the recording software, and the volunteers then took over the process. The volunteers were instructed to keep their voice neutral, and speak clearly.

All the recordings went through a quality control process performed by trained native speakers to ensure that each recording matched the corresponding script, had consistent volume, was noise-free and consisted of fluent speech without unnatural pauses or mispronunciations. Problematic lines that could not be re-recorded were dropped.

Corpora Overview The last three columns of Table 1 show various properties of the resulting corpora that include the total number of speakers, the total duration of each dataset and the average utterance duration for each gender for each language. The corresponding distributions of utterance durations (measured in seconds) for each language and gender are shown in Figure 2. As can be seen from the figure, Catalan has the highest number of long utterances (over 10 seconds long) among the three languages and also has the longest average audio duration (as shown in Table 1).

Each language is distributed in two ZIP archives, one for each gender. The audio is stored in a single channel 48kHz 16-bit signed integer PCM RIFF audio format. No post-processing was performed on the audio files. The file naming scheme of the audio files consists of a three letter code denoting the language and gender (e.g., `caf` represents Catalan female), a 5 digit speaker identifier, followed by an 11 digits number identifying the utterance. All components are separated by underscores (e.g., `caf_00195_00047731813.wav`). The transcriptions for the audio files are stored in a single textual index file (`line_index.tsv`).

5. Conclusion

In this paper, we presented free high quality multi-speaker speech corpora for three official languages of Spain: Basque, Catalan and Galician. The corpora has been designed with speech applications in mind, such as multi-speaker TTS and ASR speaker adaptation. We described the details of the process used to construct the corpora. The data is released with an open-source license with no limitations on academic or commercial use. We hope that this data will contribute to research and development of speech applications for these important languages.

6. Acknowledgments

The authors would like to thank following entities for supporting this work by finding volunteers, recording environment and their help with script generation: Bilboko Udala, Eusko Jaurlaritza and Euskaltzaindia (Basque), Direcció General de Política Lingüística del Departament de Cultura, Generalitat de Catalunya (Catalan) and Secretaría Xeral de Política Lingüística da Consellería de Cultura e Turismo, Xunta de Galicia (Galician).

7. Bibliographical References

Agić, Ž., Johannsen, A., Plank, B., Martínez Alonso, H., Schluter, N., and Søgaard, A. (2016). Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Albizu, P. (2002). Basque Verbal Morphology: Redefining Cases. *Anuario del Seminario de Filología Vasca" Julio de Urquijo"*, pages 1–19.

Arrieta, K., Leturia, I., Iturraspe, U., De Ilarraza, A. D., Sarasola, K., Hernáez, I., and Navas, E. (2008). AnHitz, development and integration of language, speech and visual technologies for Basque. In *2008 Second International Symposium on Universal Communication*, pages 338–343. IEEE.

Bengtson, J. D. (2003). Notes on Basque Comparative Phonology. *Mother Tongue*, 8:23–39.

Black, A. W. (2019). CMU Wilderness Multilingual Speech Dataset. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975. IEEE.

Bonafonte, A., Esquerra Llucià, I., Febrer Godayol, A., and Vallverdú Bayés, S. (1997). A bilingual text-to-speech system in Spanish and Catalan. In *Proc. of EUROSPEECH'97*, pages 2455–2458, Rhodes, Greece.

Bonafonte, A., Esquerra, I., Febrer, A., Fonollosa, J. A., and Vallverdú, F. (1998). The UPC text-to-speech system for Spanish and Catalan. In *Fifth International Conference on Spoken Language Processing*.

Bonafonte, A., Adell, J., Esquerra, I., Gallego, S., Moreno, A., and Pérez, J. (2008). Corpus and Voices for Catalan Speech Synthesis. In *Proc. Language Resources and Evaluation Conference (LREC)*.

Bonafonte, A., Aguilar, L., Esquerra, I., Oller, S., and Moreno, A. (2009). Recent work on the FESTCAT database for speech synthesis. *Proc. SLTECH*, pages 131–132.

Braga, D., Silva, P., Ribeiro, M., Dias, M. S., Campillo, F., and Garcıa-Mateo, C. (2010). Hélia, Heloısa and Helena: new HTS systems in European Portuguese, Brazilian Portuguese and Galician. In *PROPOR: 2010-International Conference on Computational Processing of the Portuguese Language*, pages 27–30.

Cabré, T. (2009). Vowel reduction and vowel harmony in Eastern Catalan loanword phonology. In Sónia Frota Marina Vigário et al., editors, *Phonetics and Phonology: Interactions and interrelations*, pages 267–285.

Campillo, F., Braga, D., Mourín, A. B., García-Mateo, C., Silva, P., Dias, M. S., and Méndez, F. (2010). Building high quality databases for minority languages such as Galician. In *Proc. Language Resources and Evaluation Conference (LREC)*, Valetta, Malta.

Chen, M., Chen, M., Liang, S., Ma, J., Chen, L., Wang, S., and Xiao, J. (2019). Cross-Lingual, Multi-Speaker Text-to-Speech Synthesis Using Neural Speaker Embedding. pages 2105–2109, Graz, Austria.

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

Dieguez-Tirado, J., Garcia-Mateo, C., Docio-Fernandez, L., and Cardenal-Lopez, A. (2005). Adaptation strategies for the acoustic and language models in bilingual speech transcription. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages I–833. IEEE.

Erro, D., Sainz, I., Luengo, I., Odriozola, I., Sánchez, J., Saratxaga, I., Navas, E., and Hernáez, I. (2010). HMM-based speech synthesis in Basque language using HTS. *Proc. FALA*, pages 67–70.

Ethnologue. (2019). Ethnologue. SIL International: `https://www.ethnologue.com`. Accessed: 2020-01-20.

Etxeberria, U. (2008). On Basque Quantification and on How Some Languages Restrict their Quantificational Domain Overtly. In Lisa Matthewson, editor, *Quantification: A Cross-Linguistic Perspective*, volume 64 of *North-Holland Linguistic Series: Linguistic Variation*. Emerald.

Febrer, A., Padrell, J., and Bonafonte, A. (1998). Modeling phone duration: Application to Catalan TTS. In *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*.

Ferrer, F. (2000). Languages, minorities and education in Spain: the case of Catalonia. *Comparative Education*, 36(2):187–197.

García-Mateo, C., López, A. C., Regueira, X. L., Rei, E. F., Martinez, M., Seara, R., Varela, R., and Basanta, N. (2014). CORILGA: a Galician Multilevel Annotated Speech Corpus for Linguistic Analysis. In *Proc. Language Resources and Evaluation Conference (LREC)*, pages 2653–2657.

Gibson, M. and Gil, J. (2019). *Romance Phonetics and Phonology*. Oxford University Press.

González, M. G., Banga, E. R., Díaz, F. C., Pazó, F. M., Liñares, L. R., and Iglesias, G. I. (2008). Specific features of the Galician language and implications for speech technology development. *Speech Communication*, 50(11-12):874–887.

Gorter, D., Zenotz, V., Etxague, X., and Cenoz, J. (2014). Multilingualism and European minority languages: The case of Basque. In *Minority Languages and Multilingual Education*, pages 201–220. Springer.

Guijarrubia, V. G. and Torres, M. I. (2010). Text-and speech-based phonotactic models for spoken language identification of Basque and Spanish. *Pattern Recognition Letters*, 31(6):523–532.

Guijarrubia, V. G., Torres, M. I., and Justo, R. (2009). Morpheme-based automatic speech recognition of Basque. In *Proc. Iberian Conference on Pattern Recognition and Image Analysis*, pages 386–393. Springer.

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 Technolo-*

gies for Under-Resourced Languages (SLTU '16), pages 194–200.

Harris, J. W. (1983). *Syllable structure and stress in Spanish. A nonlinear analysis*. Number 8 in Linguistic Inquiry Monographs. MIT Press, Cambridge, MA.

Hernando, J. and Nadeu, C. (1999). SpeechDat. Catalan Database for the Fixed Telephone Network. *Corpus Design Technical Report, TALP-UPC*.

Hoffmann, C. (1996). Language Planning at the Crossroads: the Case of Contemporary Spain. In Charlotte Hoffmann, editor, *Language, Culture and Communication in Contemporary Europe*, pages 93–110. Multilingual Matters Clevedon.

Holt, D. E. (2016). From Latin to Portuguese: Main Phonological Changes. In W. Leo Wetzels, et al., editors, *The Handbook of Portuguese Linguistics*, pages 457–470. Wiley-Blackwell, UK.

Hualde, J. I. and de Urbina, J. O. (2003). *A Grammar of Basque*. Mouton de Gruyter, Berlin.

Hualde, J. I., Lakarra, J. A., and Trask, R. L. (1996). *Towards a History of the Basque Language*, volume 131. John Benjamins Publishing.

Hualde, J. I. (1991). *Basque Phonology*. Routledge, United Kingdom.

Hualde, J. I. (2013). *Catalan*. Routledge.

Hualde, J. I. (2015). Basque as an Extinct Language. In *Ibon Sarasola, Gorazarre. Homenatge, Homenaje*, pages 319–326. Bilbao: University of the Basque Country.

Iriondo, I., Alías, F., Melenchón, J., and Llorca, M. A. (2004). Modeling and synthesizing emotional speech for Catalan text-to-speech synthesis. In *Tutorial and research workshop on affective dialogue systems*, pages 197–208. Springer.

Kabatek, J. and Pusch, C. D. (2011). The Romance Languages. In Bernd Kortmann et al., editors, *The Languages and Linguistics of Europe: A Comprehensive Guide*, volume 1, pages 69–96. Walter de Gruyter, Berlin.

King, A. R. (2012). *The Basque Language: A Practical Introduction*. University of Nevada Press.

Külebi, B. and Öktem, A. (2018). Building an Open Source Automatic Speech Recognition System for Catalan. In *Proc. IberSPEECH*, pages 25–29.

Lasagabaster, D. (2011). Language policy in Spain: The coexistence of small and big languages. In Catrin Norrby et al., editors, *Uniformity and Diversity in Language Policy. Global Perspectives*, pages 109–125. Multilingual Matters Clevedon.

Leturia, I., Sarasola, K., Arregi, X., de Ilarraza, A. D., Navas, E., Sainz, I., del Pozo, A., Baranda, D., and Iturraspe, U. (2018). The BerbaTek project for Basque: Promoting a less-resourced language via language technology for translation, content management and learning. *Language Technologies for a Multilingual Europe*, 4:181.

López de Ipiña, M., Torres, M., and Oñederra, M. (1995). Design of a phonetic corpus for Automatic Speech Recognition in Basque Language. In *Proc. EUROSPEECH*, volume 95, pages 851–854.

Mariño, J. B., Padrell, J., Moreno Bilbao, M. A., and Nadeu Camprubí, C. (2000). Monolingual and bilingual Spanish-Catalan speech recognizers developed from SpeechDat databases. In *Proc. XLDB-Very Large Telephone Speech Databases*, pages 57–61. C. Draxler.

Martìnez-Gil, F. (2020). Galician. In Christoph Gabriel, et al., editors, *Manual of Romance Phonetics and Phonology*, volume 27 of *Manuals of Romance Linguistics*, pages 1–46. Walter de Gruyter, Berlin.

Mateo, C. G. and Rodríguez, M. A. (2012). Language Technology Support for Galician. In *The Galician Language in the Digital Age*, pages 50–67. Springer.

Moreno, A., Febrer, A., and Márquez, L. (2006). Generation of Language Resources for the Development of Speech Technologies in Catalan. In *Proc. LREC*, Genoa, Italy, May. European Language Resources Association (ELRA).

Moseley, C. (2010). *Atlas of the world's languages in danger*. UNESCO Publishing, Paris, 3 edition. `http://www.unesco.org/culture/en/endangeredlanguages/atlas`.

Mourín, A., Braga, D., Coelho, L., García-Mateo, C., Campillo, F., and Dias, M. (2009). Homograph Disambiguation in Galician TTS Systems. In *IX Congreso Internacional da Asociación Internacional de Estudos Galegos, A Coruña-Santiago de Compostela-Vigo*.

Navas, E., Castelruiz, A., Luengo, I., Sánchez, J., and Hernáez, I. (2004). Designing and Recording an Audiovisual Database of Emotional Speech in Basque. In *Proc. Language Resources and Evaluation Conference (LREC)*.

O'Rourke, B. and Ramallo, F. (2013). Competing ideologies of linguistic authority amongst new speakers in contemporary Galicia. *Language in Society*, 42(3):287–305.

Padrell, J. and Mariño, J. B. (2002). Taking Advantage of Spanish Speech Resources to Improve Catalan Acoustic HMMs. *Co-operating Organisation*, page 67.

Pallier, C., Colomé, A., and Sebastián-Gallés, N. (2001). The influence of native-language phonology on lexical access: Exemplar-based versus abstract lexical entries. *Psychological Science*, 12(6):445–449.

Pérez, A., Torres, I., Casacuberta, F., and Guijarrubia, V. (2006). A Spanish-Basque Weather Forecast Corpus for Probabilistic Speech Translation. In *Proc. 5th SALTMIL Workshop on Minority Languages*, pages 99–101, Genoa, Italy.

Pérez, A., Torres, M. I., and Casacuberta, F. (2008). Joining linguistic and statistical methods for Spanish-to-Basque speech translation. *Speech Communication*, 50(11-12):1021–1033.

Pérez, A., Alcaide, J. M., and Torres, M.-I. (2012). EuskoParl: a speech and text Spanish-Basque parallel corpus. In *Proc. Interspeech*, pages 2362–2365, Portland, Oregon.

Posner, R. (1996). *The Romance Languages*. Cambridge University Press.

Povey, D. (2019). Open Speech and Language Resources (OpenSLR). `http://www.openslr.org/resources.php`. Accessed: 2019-03-30.

Prieto, P., del Mar Vanrell, M., Astruc, L., Payne, E., and Post, B. (2012). Phonotactic and phrasal properties of speech rhythm. Evidence from Catalan, English, and Spanish. *Speech Communication*, 54(6):681–702.

Rodriguez-Fuentes, L. J., Penagarikano, M., Varona, A., Diez, M., and Bordel, G. (2012). KALAKA-2: a TV Broadcast Speech Database for the Recognition of Iberian Languages in Clean and Noisy Environments. In *Proc. Language Resources and Evaluation Conference (LREC)*, pages 99–105.

Sainz, I., Saratxaga, I., Navas, E., Hernáez, I., Sanchez, J., Luengo, I., Odriozola, I., and Madariaga, I. (2008). Subjective Evaluation of an Emotional Speech Database for Basque. In *Proc. Language Resources and Evaluation Conference (LREC)*.

Sainz, I., Erro, D., Navas, E., Hernáez, I., Sanchez, J., Saratxaga, I., and Odriozola, I. (2012). Versatile Speech Databases for High Quality Synthesis for Basque. In *Proc. Language Resources and Evaluation Conference (LREC)*, pages 3308–3312.

Saratxaga, I., Navas, E., Hernáez, I., and Luengo, I. (2006). Designing and Recording an Emotional Speech Database for Corpus Based Synthesis in Basque. In *Proc. Language Resources and Evaluation Conference (LREC)*, pages 2126–2129.

Schulz, H., Costa-Juss, M. R., and Fonollosa, J. A. (2008). TECNOPARLA – Speech technologies for Catalan and its application to Speech-to-speech Translation. *Procesamiento del lenguaje Natural*, 41.

Sproat, R., Black, A. W., Chen, S., Kumar, S., Ostendorf, M., and Richards, C. (2001). Normalization of Non-Standard Words. *Computer Speech and Language*, 15(3):287–333, July.

Trask, R. L. (2013). *The History of Basque*. Routledge.

Vania, C., Kementchedjhieva, Y., Søgaard, A., and Lopez, A. (2019). A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. *arXiv preprint arXiv:1909.02857*.

Villarrubia, L., León, P., Hernández, L., Elvira, J., Nadeu, C., Esquerra, I., Hernando, J., Garcia-Mateo, C., and Docio, L. (1998). VOCATEL and VOGATEL: Two telephone speech databases of Spanish minority languages (Catalan and Galician). In *Proc. of the Workshop on Language Resources for European Minority Languages, LREC*.

Wheeler, M. W. (2005). *The Phonology of Catalan*. Oxford University Press.

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.

8. Language Resource References

Google. (2019a). *Crowd-sourced high-quality Basque speech data set by Google*. Google, distributed by Open Speech and Language Resources (OpenSLR), `http://www.openslr.org/76`, Google crowd-sourced speech and language resources, 1.0, ISLRN 490-901-445-079-2.

Google. (2019b). *Crowd-sourced high-quality Catalan speech data set by Google*. Google, distributed by Open Speech and Language Resources (OpenSLR), `http://www.openslr.org/69`, Google crowd-sourced speech and language resources, 1.0, ISLRN 993-764-975-949-2.

Google. (2019c). *Crowd-sourced high-quality Galician speech data set by Google*. Google, distributed by Open Speech and Language Resources (OpenSLR), `http://www.openslr.org/77`, Google crowd-sourced speech and language resources, 1.0, ISLRN 799-821-375-475-5.