





Article

A Bilingual Basque–Spanish Dataset of Parliamentary Sessions for the Development and Evaluation of Speech Technology

Amparo Varona , Mikel Penagarikano , Germán Bordel  and Luis Javier Rodríguez-Fuentes* 

Department of Electricity and Electronics, Faculty of Science and Technology, University of the Basque Country (UPV/EHU), Barrio Sarriena, 48940 Leioa, Spain; amparo.varona@ehu.eus (A.V.); mikel.penagarikano@ehu.eus (M.P.); german.bordel@ehu.eus (G.B.)

* Correspondence: luisjavier.rodriguez@ehu.eus

Abstract: The development of speech technology requires large amounts of data to estimate the underlying models. Even when relying on large multilingual pre-trained models, some amount of task-specific data on the target language is needed to fine-tune those models and obtain competitive performance. In this paper, we present a bilingual Basque–Spanish dataset extracted from parliamentary sessions. The dataset is designed to develop and evaluate automatic speech recognition (ASR) systems but can be easily repurposed for other speech-processing tasks (such as speaker or language recognition). The paper first compares the two target languages, emphasizing their similarities at the acoustic-phonetic level, which sets the basis for sharing data and compensating for the relatively small amount of spoken resources available for Basque. Then, Basque Parliament plenary sessions are characterized in terms of organization, topics, speaker turns and the use of the two languages. The paper continues with the description of the data collection procedure (involving both speech and text), the audio formats and conversions along with the creation and postprocessing of text transcriptions and session minutes. Then, it describes the semi-supervised iterative procedure used to cut, rank and select the training segments and the manual supervision process employed to produce the test set. Finally, ASR experiments are presented using state-of-the-art technology to validate the dataset and to set a reference for future works. The datasets, along with models and recipes to reproduce the experiments reported in the paper, are released through *Hugging Face*.

Keywords: multilingual speech; basque; spanish; spoken language resources; low-resource languages; semisupervised learning; automatic speech recognition



Citation: Varona, A.; Penagarikano, M.; Bordel, G.; Rodríguez-Fuentes, L.J. A Bilingual Basque–Spanish Dataset of Parliamentary Sessions for the Development and Evaluation of Speech Technology. *Appl. Sci.* **2024**, *14*, 1951. <https://doi.org/10.3390/app14051951>

Academic Editor: Douglas O’Shaughnessy

Received: 14 January 2024

Revised: 24 February 2024

Accepted: 25 February 2024

Published: 27 February 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Despite the rapid advancement of speech technologies in the last decade and the easy access to resources (data, software), there is still a large variability regarding the amount of resources available for each language [1]. Some *important* languages (such as English, Chinese, Hindi, Spanish, Arabic, etc.) are suitably covered due to their commercial interest, while languages spoken by few people or lacking the support of governments struggle to be even considered by major technological giants. This issue is not new and has been addressed in two different ways: (1) by fostering the production of language (spoken and text) resources, many of them from parliamentary speeches [2–8]; and (2) by leveraging the resources produced for other languages, e.g., by adjusting (finetuning) models or systems trained on multilingual data [9,10]. In the case of Basque, to compensate for the lack of interest of private companies, efforts have focused on producing data. However, only part of those resources are publicly available [11]. That is the case of the Basque section of Mozilla Common Voice [12], which is being extended thanks to a public data collection campaign (Gaitu, <https://gaitu.eus/>, accessed on 24 February 2024), and the Basque subset of OpenSLR [13]. Other speech datasets in Basque have been created and used by different research groups and companies but are not publicly available [14–18].

The first and most relevant goal of the work reported in this paper is to help fix the lack of speech data for Basque by releasing and making publicly available a speech dataset extracted from Basque Parliament plenary session recordings. The main part of the database, corresponding to the training dataset, consists of a 1445 h long set of transcribed segments of three types: monolingual in Basque, monolingual in Spanish and bilingual in Basque and Spanish. A separate manually supervised 17-h-long set of segments is also provided. This set can be utilized for fine-tuning hyperparameters and/or for evaluating automatic speech recognition (ASR) performance in both languages. Finally, two bilingual n-gram language models are provided too. The first one is a trigram model obtained from Wikipedia pages in Spanish and Basque and aims to support the development of general-purpose ASR systems. The second one is a 4-gram model obtained from Basque Parliament transcriptions in the period 2013–2021 and aims to support the development of ASR systems for the Basque Parliament domain.

A second goal of this work is to leverage Spanish data to improve speech technology for Basque. This is the reason why everything is designed to be bilingual, i.e., the datasets, the acoustic models and the language models. In contrast to more common approaches to low-resource ASR in which multilingual datasets or models are adapted (finetuned) to a new low-resource language based on a relatively small set of transcribed speech segments [19–21], our approach consists of building a fully bilingual ASR system from scratch. The similarity of Basque and Spanish at the acoustic-phonetic level allows us to represent the sounds of both languages with a single (simplified and small) set of units. On the other hand, their large differences at the lexical and syntactic levels allow us to use a single language model (trained on Basque and Spanish texts) that naturally switches from one language to another based only on the acoustic-phonetic clues found by the acoustic models. By the way, this approach makes it easy to deal with code switchings [22] (quite common in Basque Parliament sessions) compared to other recent approaches in the literature [23–29].

The database includes two datasets (training and test) containing speech segments lasting from 3 to 10 s each. Each dataset is accompanied by an index file that offers comprehensive information on the segments, with each line representing a single segment. This information includes the audio filename, language and speaker tags, text-audio similarity score (ranging from 0 to 100), segment length (duration, in seconds) and transcription. Speaker distribution is balanced in terms of gender. On the other hand, language distribution is not balanced, with Spanish and Basque making up approximately 70% and 30% of the datasets, respectively. The resulting datasets, the models and the recipes prepared to estimate the models and to build and evaluate the ASR systems are released through *Hugging Face* (<https://huggingface.co/>, accessed on 24 February 2024). Since language tags are associated with both training and test segments, monolingual ASR experiments could be carried out if desired. The provided datasets could also benefit tasks beyond ASR, such as speaker recognition or spoken language recognition, using the speaker and language tags, respectively.

The paper is organized as follows. Section 2 briefly describes the differences and similarities between the Basque and Spanish languages. Section 3 gives details about Basque Parliament plenary sessions and the operations carried out to extract audio files and transcriptions from raw materials. Section 4 describes the automatic procedure used to create an ASR training set from the most reliable Basque Parliament sessions and the semi-automatic procedure (including manual supervision) employed to create a test set from an independent set of sessions. Section 5 presents a state-of-the-art fully bilingual ASR system that is trained and evaluated on the Basque Parliament datasets. Performance results at the token (grapheme) and word levels are presented to validate the datasets and to establish a baseline for future developments. Finally, Section 6 summarizes the paper and outlines future work.

2. Basque and Spanish: A Brief Comparison

Basque is an isolate non-Indo-European language spoken by around 900 thousand people in a small region of Western Europe around the Biscay Bay, divided between Spain and France [30,31]. A sizeable amount of Basque words have been borrowed from Latin, later from Romance languages (Spanish and French) and in the last decades from English, but most of the vocabulary is genuine and not related to other languages. Unlike other languages (e.g., Spanish, English), which use connectors (conjunctions, prepositions, articles, etc.) to relate words with each other, Basque is an agglutinative language which combines and adds suffixes to stems, conveying information about the case, number, person, etc. Spanish uses a verb conjugation system based on person and number while Basque combines synthetic and periphrastic conjugations: synthetic forms are built around the basic lexeme by adding markers depending on the subject, object and indirect object; on the other hand, periphrastic forms are built based on a few auxiliary verbs that convey most information. In declarative sentences, the most common order in Spanish is subject-verb-object while in Basque the most common structure is subject-object-verb.

However, phonetically and phonologically, Basque and Castilian Spanish share many features, probably due to the fact that Castilian Spanish was strongly influenced by Basque in its origins. The sound systems of both languages are quite similar, featuring the same five vowels and the same syllable-timed phonology. The sets of consonants are almost identical, with only a few additional sounds in Basque: the phonemes /ts/, /ts'/ and /s'/ and some other less frequent ones [32]. Furthermore, in urban settings home to the majority of the Basque Country's population, speakers (often with Basque as their second language) tend to articulate Basque phonemes in a manner akin to Spanish, resulting in a closer approximation to Spanish phonetic realizations.

Therefore, for the development of our database (more specifically, for the acoustic-phonetic alignment on which segment selection is based), we have defined a reduced set of grapheme units (corresponding to the most prevalent sounds/phonemes) by loosely taking into account their frequencies and their most common realizations [14]. For instance, the three Basque affricates (tʃ, ts' and ts) were merged into a single affricate: the one existing in Spanish (tʃ). Similarly, the Basque fricatives s' (as in *zoroa*) and ʃ (as in *kaixo*) were collapsed into the fricative s, which exist in both Basque and Spanish. On the other hand, the Spanish fricative θ (as in *pazo* and *cero*), not strictly present in Basque, was retained due to its common usage in proper names.

We ended up with a reduced set of 23 phonetic units complemented by an extra unit representing silences and other non-linguistic background events (see Table 1).

Table 1. Reduced set of phonetic units for Spanish and Basque with examples. IPA units are shown as well as the simplified ASCII encoding used in this work.

IPA	ASCII	Examples	
		Spanish	Basque
i	i	pico	ipar
u	u	duro	umore
e	e	pero	hemen
o	o	toro	hori
a	a	valle	kale
m	m	madre	ama
n	n	nunca	neska
ɲ	N	año	arraina
p	p	padre	apeza
		bolsa	
b	b	vino	begia
t	t	tomo	etorri
d	d	dedo	denda

Table 1. Cont.

IPA	ASCII	Examples	
		Spanish	Basque
k	k	casa	ekarri
		queso	
g	g	kilo	gaia
		gata	
f	f	fatal	afaria
		cero	
θ	z	pazo	–
s	s	sala	hasi
s′	s	–	zoroa
ʃ	s	–	kaixo
x	j	mujer	ijito
		rosa	
r	R	torre	arrunta
r	r	puro	dirua
l	l	lejos	lana
tʃ	X	mucho	txikia
ts′	X	–	atzo
ts	X	–	mahatsa
c	X	–	ttakun
ʎ	y	caballo	pilaka
		hielo	
j	y	cónyuge	–
j	y	–	joan
J	y	–	onddo

3. Collecting Data from the Basque Parliament

The Basque Parliament's activities consist of plenary sessions, to which all Members of Parliament (MPs) are summoned, and committee meetings, usually comprising around 15 MPs. The minutes of both plenary sessions and committee meetings are publicly available. Additionally, subtitled videos of plenary sessions are released through the Basque Parliament website (<https://www.legebiltzarra.eus/portal/es/web/eusko-legebiltzarra/>, accessed on 24 February 2024). Only plenary sessions were considered to build the datasets described in this paper. The audio recordings were extracted from the videos while the transcriptions were extracted from the draft minutes handed to us by the Basque Parliament. The draft minutes are acoustically close to what speakers say in their turns, similar to the verbatim reports of United Kingdom parliamentary debates (*Hansards*) (<https://hansard.parliament.uk/>, accessed on 24 February 2024) [33]. The draft minutes are not publicly available and were handed to us only to produce the video subtitles. We chose the draft minutes over the official minutes because we wanted segment transcriptions to be as close as possible to the uttered speech. The official minutes are obtained from the draft minutes after a laborious manual process that involves fixing lexical and grammar issues, removing all kinds of spontaneous speech events and adding translations to the other language (from Basque to Spanish and vice versa). In this way, the official minutes will be grammatically correct and will convey the intended meaning but may differ considerably from the words actually spoken.

Plenary sessions of the Basque Parliament are held almost every week of the year, amounting to around 40 plenary sessions per year. Each session takes between 3 and 7 h, sometimes with a break for lunch. Each plenary session has an agenda with a series of topics to be discussed. Topics include legislative initiatives proposed by the government or by a parliamentary group and specific measures adopted by the government that must be approved by the parliament. This means that topics change from session to session involving new technical terms, new names of places and persons, etc. Speakers participating in plenary sessions include not only MPs but also counselors and officers of the

Basque Government and representatives of organizations (unions, chambers of commerce, associations, etc.). Also, during each 4-year parliamentary period (called *legislature*), some MPs may be replaced by others (e.g., so far in the current legislature, there have been 19 MPs replaced). This explains why we could end up with 100 speakers participating in plenary sessions each year. During a plenary session, each speaker may take the floor several times (generally from the speaker’s stand, sometimes from the Parliament seats). Turns are managed by the president of the Parliament, so speakers do not overlap very often. Speakers can choose to speak in Basque or Spanish, and they can switch from one language to another at any time. Also, guest speakers can speak in other languages such as Catalan, English, etc. Finally, though speakers generally stick to clear, prepared speeches, sometimes they go off script and produce spontaneous speech with restarts, hesitations, repetitions, etc. Beyond these traits of spontaneity, the interactions between speakers remain formal and are tightly regulated by the president of the parliament, ensuring that no conversational speech occurs during Basque Parliament sessions.

3.1. Processing Audio Resources

To build the training dataset, we gathered videos from 408 plenary sessions, totaling 2123.86 h. Each session may consist of one or several videos depending on its duration. The sessions were collected from January 2013 to December 2021, spanning three parliamentary periods, as follows: 173 sessions (9–181) from the 10th Legislature (November 2012–October 2016), 160 sessions (1–160) from the 11th Legislature (October 2016–August 2020), and 75 sessions (1–75) from the 12th Legislature (August 2020–. . .). Each legislature includes different representatives and government counselors, meaning that a different (but partially overlapping) set of speakers participated in each period. Over the entire 2013–2021 period, the dataset comprises 190 different speakers. To build the test dataset, we collected videos from five plenary sessions held in February 2022 (sessions 77 to 81 of the 12th Legislature), amounting to 26.38 h and involving 57 speakers, 56 of whom had previously appeared in the training set. In total, the database includes 191 speakers with a nearly even distribution of 99 men and 92 women, providing a balanced representation of speaker gender (see Figure 1). Details about the raw materials used to create the training and test datasets are summarized in Table 2.

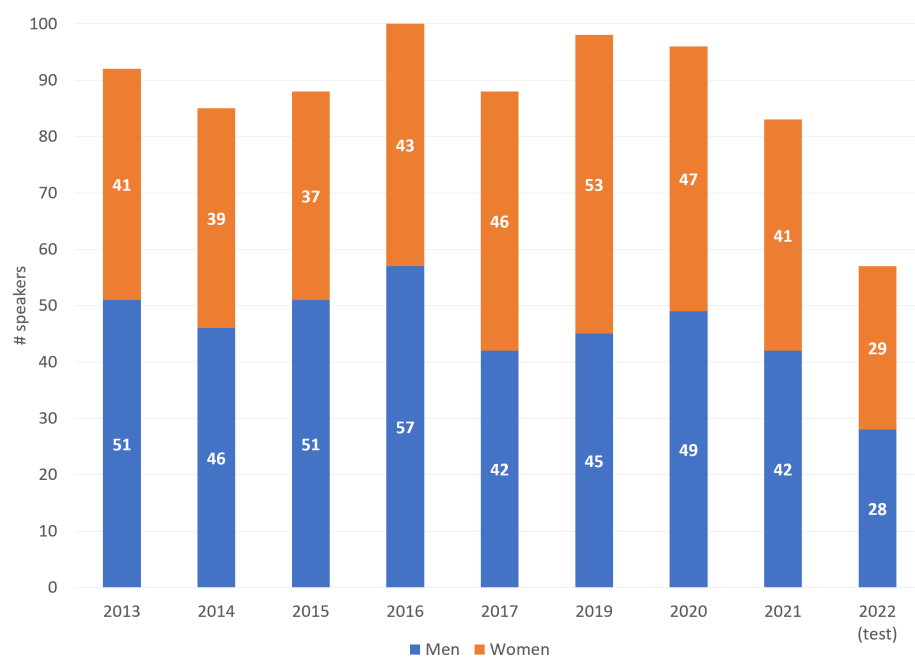


Figure 1. Distribution of male and female speakers by year. Years 2013–2021 include the sessions used to build the training set. Year 2022 includes the five sessions used to build the test set.

Table 2. Raw materials used to build the training and test datasets, disaggregated per year. Durations are given in hours.

Year	#Sessions	#Speakers	Duration
2013	56	92	304.11
2014	47	85	274.84
2015	44	88	220.40
2016	35	100	173.78
2017	52	88	261.16
2018	50	89	255.42
2019	47	98	250.13
2020	27	96	127.07
2021	50	83	256.95
2013–2021 (training)	408	190	2123.86
2022 (test)	5	57	26.38

Note that since almost all speakers in the test set also appear in the training set, the test set is not independent from the training set and ASR performance computed on the test set will be biased, because the models estimated on the training set will be intrinsically adapted to those speakers. However, this speaker dependence is consistent with our primary goal of developing and optimizing an ASR system for the Basque Parliament domain. Since acoustic models will be periodically updated with the latest sessions (including speeches from speakers that will likely appear in the upcoming sessions), our ASR performance scores will closely reflect the performance that can be expected in practice.

Audios were all recorded through stand microphones which could be suitably directed towards the speaker so that the signal-to-noise ratio was always high. As far as we know, the microphones and the recording hardware did not change during the period we made the recordings. However, the audio/video format did change: until the third session of the 11th Legislature (10 November 2016), we received RealMedia videos, the audio channel using a Cook Audio codec with a single channel at 22,050 Hz and 32 bits per sample; starting from the fourth session of the 11th Legislature (23 November 2016), we received MP4 videos, the audio channel using an MPEG AAC audio codec with 2 channels (stereo) at 48,000 Hz and 32 bits per sample. However, since all audios were later converted to WAV format, PCM encoding, single channel at 16,000 Hz and 16 bits per sample (by means of `ffmpeg`, see <https://ffmpeg.org/>, accessed on 24 February 2024), we hope that the switch from RealMedia to MP4 has not introduced any serious bias to our dataset.

Another relevant fact to be noted is that during voting (that may be scheduled in the middle of a plenary session) the debate is interrupted, votes collected and counted and finally, the results are announced by the president. This process may take several minutes, during which no speech is recorded but instead a long quasi-silent interval is included in audio recordings. These long intervals of silence pose a challenge to the alignment procedure that is applied to synchronize the subtitles and to extract the training segments (see Sections 4.4 and 4.5 for details). Thus, specific countermeasures have been integrated into the alignment procedure to avoid including these silent intervals inside a segment. In particular, the kernel used in the dynamic programming algorithm is defined so that insertions are always made between words [34].

3.2. Processing Text Resources

During each session, human live transcribers produce a first draft of the minutes which closely resembles the acoustic content of the recorded audio but contains lexical and grammar errors and format inconsistencies. For instance, numbers may appear either in numeric form or in complete spellings; some words could be repeated and some others left untranscribed; not all sentences may start with an uppercase letter; acronyms could appear in different formats, etc. Note, that this first draft is intrinsically bilingual since it reflects

what the speakers say in the language they speak. In fact, a speaker turn may even include code switchings.

The first draft of the minutes is taken as a starting point to produce video subtitles by aligning them with the audio. Before the proper alignment, an in-house bilingual Grapheme-to-Phoneme (G2P) converter is applied to transform the orthographic sequence of the draft minutes into a phonetic sequence that we call the *reference sequence*. Then, a phone decoder is applied to the audio signal to obtain a second phonetic sequence that we call the *recognized sequence*. The phone decoder is based on bilingual acoustic models trained on speech data in Basque and Spanish, using a common set of phonetic units (the ones represented in Table 1). Finally, the two phonetic sequences are aligned and the timestamps are transferred back from the recognized sequence to the reference sequence, and from the latter to the original draft minutes. The subtitles are generated by applying a heuristic approach to cut the aligned minutes into reasonably short segments so that the subtitles can be easily read. This is how we have been generating the subtitles of the Basque Parliament since 2010 to the present [35–38].

4. Data Extraction

This section summarizes the procedures employed to build the training and test datasets of our bilingual Basque Parliament database. A more in-depth description is provided in [39]. Our goal was to produce two sets of audio excerpts, each audio excerpt lasting between 3 and 10 s, along with the corresponding transcriptions (which are expected to match the audio contents). As a result, two index files for the training and test datasets were created with one line per segment, each line containing the speech segment filename (with information about the session from which it was extracted), language and speaker tags, text-audio similarity (a percentage reflecting how well the transcription reflects segment contents), length (duration, in seconds) and segment transcription (see Figure 2).

filename	language	speaker	similarity	length	transcription
10-065_20140213_03_2548.95_2554.99.mp3	es	290	100.00	6.04	la consejera de educación acata las recomendac...
10-073_20140321_02_4007.08_4011.62.mp3	es	411	89.77	4.54	y en este momento tenemos ochenta y cinco mil ...
10-152_20151203_02_5722.05_5731.28.mp3	eu	205	100.00	9.23	zure egiteak eta zuen esateak ez datoz bat eta...
10-161_20160304_01_488.19_496.81.mp3	es	93	98.28	8.62	ese servicio que la ertzaintza ofrece a la soc...
11-030_20170525_01_11308.36_11318.64.mp3	es	124	100.00	10.28	se reconoce que hay una devaluación y precariz...
11-039_20170629_02_538.24_545.05.mp3	es	282	100.00	6.81	porque llegan antes desde las paradas actuales...
11-028_20170518_01_4067.19_4073.38.mp3	bi	0	84.48	6.19	por no tener no tiene ni un plan amaitzen joan...
12-057_20210923_01_12816.87_12825.70.mp3	es	533	100.00	8.83	y se le dan significados que no son a lo que n...
11-150_20191121_01_8667.66_8675.00.mp3	eu	416	100.00	7.34	erdibideko zuzenketa ez da onartu jarraian eus...
12-030_20210218_01_4839.84_4847.25.mp3	es	505	100.00	7.41	inició hace un año hace más de un año un proce...

Figure 2. A section of an index file, with 10 training segments extracted from the Basque Parliament.

As noted in Section 3.2, we were already producing subtitles for the videos of plenary sessions, where the draft minutes were reasonably aligned with the audio and segmented into readable chunks. However, those subtitles did not meet the requirements of an ASR corpus: first, the texts of the subtitles could be not suitably normalized; second, the aligned texts might not fully match the audio contents; third, the segments presented as subtitles might not meet the duration constraints.

So, we designed an iterative segment extraction procedure from Basque Parliament plenary session videos. Each session could span two or more videos. Audios were extracted from each video in chunks of two or fewer hours to make the alignments computationally feasible; on the other hand, the draft minutes were manually cut in order to match the audio chunks. An alignment was performed between each audio chunk and the corresponding text.

The process starts with text normalization of the draft minutes. Then, the audio and the normalized texts are aligned, which involves applying an ASR system to the audio to

obtain a recognized sequence of tokens/units and transforming the normalized texts into a reference sequence of tokens/units (in the same way as we do to obtain the subtitles in Section 3.2). Finally, starting from those alignments, segments are collected in a recursive fashion: the longest segment meeting the duration constraints and yielding the highest alignment similarity is retrieved and the audio chunks generated at both sides of the retrieved segment are searched until no segments are left. Note, that *not all* of the audio signal is recovered by this procedure but just a part of it; because of the imposed duration constraints many short segments (shorter than 3 s) are left in the way.

To build a training set, we may retain all segments obtained in this way regardless of the quality of the alignments; that is, regardless of how well the transcriptions matched the audio contents. Instead, segments may be ranked by alignment quality and only the top-ranking segments would be kept, that is, those showing a high correspondence between the audios and their transcriptions. No matter how strict the selection of segments is, new acoustic models can be trained on the selected segments and the alignment and selection procedure can be repeated until no further improvement is observed. To measure the goodness of a set of training segments, an ASR system can be trained on those segments and its performance measured on an independent, manually supervised test set, obtained from parliamentary sessions not included in the training set. This iterative procedure stops when ASR performance does not improve or the improvement is too small. In the following subsections, we give details about each step of the procedure.

4.1. Text Normalization

Text normalization involves keeping accented vowels, removing punctuation marks, putting all words in lowercase (including acronyms) and replacing all kind of numbers and ordinals (including Roman numbers) by their orthographic counterparts so that the orthographic text reflects the acoustics as closely as possible. For instance, 'XX mendea' is replaced by 'hogei mendea' ('twentieth century') and '2396' is replaced by 'dos mil trescientos noventa y seis' ('two thousand three hundred ninety six'). Numbers are written in Basque or Spanish depending on the context. The context is given by the surrounding words, initially considering context windows of length 1 and then increasing the window length as needed until a decision can be made. Each surrounding word is assigned the most likely language by means of two dynamic Basque and Spanish dictionaries and the language appearing the most in the context window determine the language to be used. The texts provided with the Basque Parliament database will follow these normalization rules.

4.2. Audio Tokenization

A phone decoder is built using the reduced set of phonetic units described in Table 1 and an off-the-shelf close to state-of-the-art end-to-end neural network-based ASR system: Facebook AI Research wav2letter++ (consolidated into Flashlight), applying the Gated ConvNet recipe presented in [40]. Note, that the phone decoder requires neither lexical models nor a language model and is applied to the audio chunks without any phonological restrictions. In this way, a long sequence of phonetic units along with their corresponding timestamps (the recognized sequence) is obtained from each audio chunk. Initially, the phone decoder is trained on generic speech datasets for Basque and Spanish which amounts to 332.2 h of speech (77.8 h in Basque and 254.4 h in Spanish): Mozilla CommonVoice (cv-corpus-5.1-2020-06-22) [12], OpenSLR (SLR76) [13], Aditu [18] and Albayzin [41]. Then, in successive iterations of the procedure, the phone decoder is re-trained on the set of training segments extracted from Basque Parliament sessions in the previous iteration and then applied to audio chunks to obtain new and hopefully better sequences of phonetic units.

4.3. Text Tokenization

Since audio chunks will be tokenized in terms of phonetic units, alignments must be performed at the phonetic level. Thus, the text to be aligned with each audio chunk is

passed through an in-house bilingual grapheme-to-phoneme (G2P) converter [36,37] to obtain a sequence of phonetic units (the reference sequence). The G2P converter is based on two dynamic dictionaries for Basque and Spanish, initialized on Wikipedia and updated with each new word found in transcriptions. Each word in these dictionaries is mapped to a nominal pronunciation in terms of the reduced set of acoustic units by applying a language-specific set of pronunciation rules [42,43]. When a known word is encountered, the G2P converter outputs the pre-stored pronunciation. Conversely, when an unknown word is detected, pronunciation rules are employed to derive its phonetic baseform, which is then stored in the appropriate dictionary. If a word is either unknown or exists in both lexicons, a choice must be made between the two languages, Basque or Spanish. This decision is based on the context: the language appearing the most in a window around the current word is chosen. This strategy is found to be effective in practice, leading to very few errors. Acronyms are assumed to be spelled; acronyms not following this rule are listed in the dictionaries. Finally, the words added to the dictionaries after processing the transcription corresponding to each audio chunk are supervised and validated by a human expert.

4.4. Alignment

For each audio chunk, the two sequences of phonetic units (the reference sequence derived from the transcription and the recognized sequence derived from the audio) are aligned with one another. This alignment is based on the principle of maximizing the number of matches, which generally equates to minimizing the occurrences of deletions, insertions and substitutions. This approach follows the same text-and-speech alignment methodology that our group has successfully implemented since 2010 for synchronizing subtitles with spoken content in the Basque Parliament [35–38]. As a result, areas with a high concentration of alignment errors likely indicate discrepancies between the draft minutes and the actual audio, suggesting that these segments should be omitted from the training dataset.

The Alignment Similarity (AS) metric is defined as:

$$AS = 100 \cdot \frac{m}{m + d + i + s} \quad (1)$$

where m , d , i and s are the counts of matches, deletions, insertions and substitutions, respectively, obtained from the optimal alignment between the recognized and the reference sequences for a test set.

4.5. Search

For each audio chunk, the recognized sequence of units/tokens sometimes features gaps between two consecutive units, which represent silent pauses. Gaps longer than 0.5 s are defined as potential *breaking points*. A *slice* is defined as an audio chunk between two consecutive breaking points while any audio chunk comprising one or more consecutive slices is called a *segment*. This means that a segment might contain one or more breaking points inside of it. Data collection is performed by searching for the segment lasting between 3 and 10 s with the highest alignment similarity. When the highest similarity is attained by two or more segments, the longest one is chosen. Thus alignment similarity and length are the primary and secondary selection criteria, respectively. Note, that segments may or may not correspond to complete sentences; a segment could comprise just part of a sentence, parts of two sentences or even two or more full sentences.

A single-pass search is conducted (with linear time complexity) to optimize similarity and length across segments that satisfy the duration constraints within an audio chunk U . Due to these constraints, the method only needs to consider a finite number of slices beyond each starting slice, typically one or two. Once the optimal segment s^* within an audio chunk U is identified, the adjacent audio sub-chunks, U_l and U_r , are independently subjected to recursive searches if they are non-empty (see Figure 3). Each recursive call

yields a list of segments resulting in two lists, S_{U_l} and S_{U_r} , which are then combined with the optimal segment s^* to obtain a single unified list S_U . In this way, after processing all audio chunks and merging the resulting segment lists, we obtain the comprehensive list of segments S . Note, that S will not encompass every audio segment considered during the search, as numerous short segments (not satisfying the duration constraints) are excluded from S .

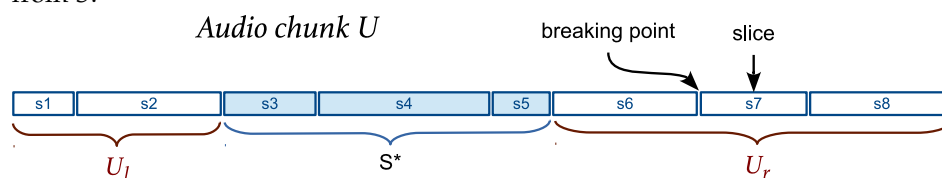


Figure 3. An audio chunk U with 8 slices: the optimal segment s^* is chosen (the longest one among those with the highest similarity); the procedure continues recursively on the left and right chunks, U_l and U_r , until the number of slices is $n \leq 1$.

The time complexity of this recursive procedure is in $O(n \log n)$ on average, with n being the number of slices in an audio chunk U . If K audio chunks are to be processed, the time complexity of the whole procedure will be in $O(\sum_{k=1}^K n_k \log n_k)$, n_k being the number of slices in the k -th audio chunk.

Finally, note that we may keep all of the retrieved segments or only those segments for which the provided transcription best matches the speech contents, either those segments with the highest values of alignment similarity amounting to a given number of hours (e.g., 1000 h) or those segments with alignment similarity higher than a given threshold (e.g., 95%).

4.6. Test Data Extraction

The test set is aimed to evaluate the performance of ASR systems on Basque Parliament data. The test segments are extracted from the audio chunks and transcriptions of five plenary sessions not included in the training set: sessions 77 to 81 of the 12th Legislature, which were held in February 2022.

The draft minutes are normalized and aligned to the audio signals as described above, using the wav2letter++ framework and models. Segments are automatically cut to approximately match the duration constraints (3–10 s) using the silent gaps given by the automatic alignment procedure. Segments with an alignment similarity of 100% are not further processed since, given that the models have not been trained on these audio segments, it is assumed that the transcriptions fully correspond to the audio contents. The remaining segments include explicit indications of the differences between the recognized and reference sequences. These differences are solved by auditing the speech signal. The text obtained after auditing may correspond to the recognized sequence of words, the sequence of words provided in the draft minutes or an entirely different sequence.

Finally, each segment is automatically classified as exclusively Spanish, exclusively Basque or bilingual, the latter likely indicating a code-switching event. Language classification is performed by using word frequencies in Basque and Spanish dictionaries, which allow us to compute two average frequencies for each segment: f_{Basque} and $f_{Spanish}$. A hard decision is automatically made only if the maximum frequency is higher than a strict threshold; otherwise, the decision about the language employed in that segment is left for manual supervision. After manual supervision, the threshold of the automatic classification procedure is tuned on the test set so that it performs the three-way classification task with less than a 1% error. Once tuned on the test set, this automatic language classification method will be applied to automatically assign language tags in the training set. Details of the test dataset, disaggregated per language, are shown in Table 3.

Table 3. Number of segments and duration of the test set, disaggregated per language.

Language	#Segments	Duration
Spanish	6057	11:18:19
Basque	2955	05:27:03
Bilingual	239	00:29:06
Total	9251	17:14:28

The test dataset is used in two different ways in this work. During the extraction of training data, which operates iteratively, ASR experiments are performed to stop iterations. In this case, the test set is used under a cross-validation scheme (see Section 4.7 for details). Then, after the training dataset is defined, the database is characterized through ASR experiments (see Section 5). In this case, the test dataset is split into two fixed subsets: a development set for tuning purposes, comprising segments from sessions 77 and 78 of the 12th Legislature, and an evaluation set for measuring ASR performance, comprising segments from sessions 79, 80 and 81 of the 12th Legislature. These two dev and eval sets comprise the official subsets of the test set included in the Basque Parliament database.

4.7. Training Data Extraction

After the segment extraction procedure (Sections 4.1–4.5) is run on the Basque Parliament sessions from 2014 to 2021, a set of training segments $S_{train}^{(1)}$ is obtained. New acoustic models can be trained on $S_{train}^{(1)}$, models that are expected to perform better than the baseline models when applied to Basque Parliament sessions. Following this logic, successive iterations of the segment extraction procedure can be applied by using models increasingly adapted to the Basque Parliament sessions to obtain successive sets of training segments: $S_{train}^{(2)}$, $S_{train}^{(3)}$, etc.

It is expected that the alignments will improve and a better training set will be obtained, with segments that better match their reference transcriptions. In practice, it is observed that with each iteration of the search procedure, the models increasingly conform to the provided transcripts. So much so, that after several iterations, the similarity score might reach 100% for all segments, making it impossible to distinguish between *truly good* transcripts and *bad* transcripts that the models have adjusted to. In other words, the models might be overfitting to training data so much that their performance would likely degrade when tested on an independent dataset. This unwanted effect forces us to check the performance of our models after each iteration by running ASR experiments on the test set. In this way, the procedure will stop when ASR performance (computed on the test set) does not improve or when the improvement is considered too small.

A fully bilingual ASR system is built using the wav2letter++ framework [40]. Besides the acoustic models, our wav2letter++ system requires a vocabulary and a language model, both extracted from the normalized transcriptions. A single pronunciation baseform is considered for every word in the vocabulary, as generated by our in-house G2P converter. A trigram language model is estimated using KenLM [44]. Since the language model is derived from texts in Basque and Spanish, it inherently makes the ASR system generate a bilingual output which may likely include code-switching events. This is due to the consistent probability that a Basque word may follow a Spanish word, or vice versa. This capability, combined with the employment of a unified set of acoustic units, makes our ASR system thoroughly bilingual and robust to code-switchings.

ASR experiments are carried out on the test set described in Section 4.6 under a cross-validation approach, by considering 20 random partitions and reporting the average ASR performance on them. The Word Error Rate (WER) metric is used to report performance. For each partition, the test set is randomly split into two halves, the first half (tuning set) being used to perform a random walk search of the optimal hyperparameters of the ASR

system (see [39] for details) and the other half (evaluation set) being properly used to evaluate ASR performance.

Three ASR systems are built by using three acoustic models: (1) baseline acoustic models obtained from generic out-of-domain datasets for Basque and Spanish; (2) acoustic models trained on the dataset $S_{train}^{(1)}$ obtained after one iteration of the data extraction procedure using only those segments with alignment similarity $\geq 80\%$ and (3) acoustic models trained on the dataset $S_{train}^{(2)}$ obtained after a second iteration of the data extraction procedure using segments with the highest similarity amounting to the same duration as in (2). The average WER figures obtained on the tuning and evaluation subsets by the three ASR systems in cross-validation experiments are shown in Table 4, disaggregated per language.

Table 4. Average WER performance, disaggregated per language, in cross-validation experiments (using 20 random partitions) on the tuning and evaluation sets for the baseline (out-of-domain) models (iteration 0) and models trained on the segments obtained after the first and second iterations of the segment extraction procedure (in-domain models).

Set	Iteration	Basque	Spanish	Bilingual	All
Tuning	0	16.63	16.19	22.38	16.44
	1	5.43	3.93	4.38	4.29
	2	5.09	3.66	3.95	4.02
Evaluation	0	16.57	16.38	22.44	16.57
	1	5.51	4.04	4.35	4.41
	2	5.13	3.66	3.90	4.02

Significant performance improvements are observed when comparing baseline models to first-iteration models: the average WER goes from 16.44% to 4.29% on the tuning sets, meaning a 73.9% relative reduction in WER, and from 16.57% to 4.41% on the test sets, meaning a 73.4% relative reduction in WER. These improvements are largely due to the increased volume of training data (998 h compared to 332) and especially to the domain-specific nature of the training material, as Basque Parliament data are employed for both training and evaluating the ASR systems. However, second-iteration models perform only slightly better than first-iteration models, suggesting that further iterations of the data extraction pipeline would likely yield even smaller improvements. Therefore, no more iterations are performed.

Due to format issues, the Basque Parliament sessions from 2013 were not included in the experiments reported above. Later, the format issues were fixed and the 2013 sessions were integrated into our datasets. To accomplish that, the models used to obtain the second-iteration set of segments were used to perform a single iteration of the data extraction procedure on 2013 sessions and the resulting set of segments, $S_{2013}^{(2)}$, were added to $S_{train}^{(2)}$ to obtain the final set of training segments: $S_{train} = S_{train}^{(2)} \cup S_{2013}^{(2)}$.

Note, that S_{train} includes *all* of the segments retrieved from the 2013 to 2021 Basque Parliament sessions without applying any similarity threshold, which means that for some segments the transcription will not completely correspond to the audio contents. This could negatively affect the quality of the resulting models. To discard unreliable segments, a more restrictive training set, called *train-clean* ($S_{train-clean}$) has been also defined by keeping only those segments for which $AS \geq 95$.

Note, that not all speakers participating in Basque Parliament sessions are represented in these sets because the segments of some of the speakers were discarded during the data extraction procedure. Out of 191 speakers participating in Basque Parliament sessions, 187 are represented in the training set; on the other hand, the dev and eval subsets of the test set contain 47 and 42 different speakers, respectively, all of them being also represented in the train and train-clean sets. Table 5 shows the duration (in hours) and the number of

segments, disaggregated per language, for all datasets of the Basque Parliament database: train, train-clean, development and evaluation.

Table 5. Duration (in hours) and number of segments in the train, train-clean, dev and eval sets of the Basque Parliament database, disaggregated per language (es: Spanish, eu: Basque, bi: Bilingual).

Set	Duration (h)				#Segments			
	Total	es	eu	bi	Total	es	eu	bi
train	1445.1	1018.6	409.5	17.0	749,945	524,942	216,201	8802
train-clean	1315.5	937.7	363.6	14.2	661,871	469,937	184,950	6984
development	7.6	4.7	2.6	0.3	4095	2567	1397	131
evaluation	9.6	6.4	2.8	0.4	5152	3450	1521	181

5. Validation of the Basque Parliament Database through ASR

In this section, by using the training set to estimate the acoustic models of an state-of-the-art ASR system and by checking its performance on the test set (using part of it for tuning the system and part of it for properly measuring performance), we aim to evaluate the consistency of the Basque Parliament database and, at the same time, to provide a baseline for other authors who may use the database in the future.

The ASR framework consists of an acoustic front end based on a pre-trained Wav2Vec 2.0 speech encoder [45] which produces a sequence of frame-level acoustic representations (speech embeddings), followed by a classification backend consisting of a neural network trained with Connectionist Temporal Classification (CTC) [46] which outputs a vector of grapheme posteriors for each input embedding. Finally, maybe constrained by the phonological and syntactic restrictions introduced by lexical and language models, a search is performed on the sequence of posteriors to output the sequence of graphemes (including blanks) that maximizes the joint acoustic and syntactic likelihood. The framework on which this work is based is available at https://huggingface.co/docs/transformers/model_doc/wav2vec2 (accessed on 24 February 2024).

ASR experiments have been carried out using the train-clean dataset to finetune the classification backend and three different approaches to obtain the output sequence of words from the sequence of frame-level grapheme posteriors. In the first approach, no language model is used so the search for the optimal sequence of words takes into account just the grapheme posteriors. In the second approach, an out-of-domain vocabulary (including 2,159,919 words) and an out-of-domain 3-gram language model (including 33.7 million n-grams) are used, both extracted from Wikipedia pages in Spanish and Basque (downloaded on 1 March 2023), the Spanish part being about 10 times larger than the Basque part. In the third approach, an in-domain vocabulary (including 174,168 words) and an in-domain 4-gram language model (including 22.1 million n-grams) are used, both extracted from transcriptions of the train set. The dev set has been used to tune the hyperparameters of the ASR framework and then the optimal hyperparameter values have been used to obtain the output sequences for both the dev and eval sets.

ASR performance is given in terms of both WER and Character Error Rate (CER). CER is defined the same way as WER but for individual characters (graphemes, including blanks) instead of full words. CER and WER performance figures on the dev and eval sets, disaggregated per language, are shown in Table 6. Figure 4 shows the global CER and WER performance on the dev and eval sets for different language models and Figure 5 shows CER and WER performance, disaggregated per language, on the dev and eval sets when using the Basque Parliament language model, which is the one yielding the best results.

Table 6. CER and WER performance, disaggregated per language, obtained on the dev and eval sets using acoustic models trained on the train-clean dataset and three configurations: not using any language model (no-LM); using a bilingual 3-gram language model estimated on out-of-domain texts: Wikipedia pages in Spanish and Basque (wiki-LM) and using a bilingual 4-gram language model estimated on the Basque Parliament train transcriptions (bp-LM).

Set	LM	CER				WER			
		eu	es	bi	all	eu	es	bi	all
dev	no	1.47	1.26	1.52	1.33	4.90	3.19	3.99	3.66
	wiki	1.52	1.25	1.54	1.34	4.92	2.90	3.99	3.46
	bp	1.50	1.22	1.52	1.31	4.80	2.74	3.56	3.30
eval	no	1.46	1.14	1.50	1.24	4.60	2.84	3.76	3.26
	wiki	1.55	1.14	1.53	1.26	4.80	2.60	3.42	3.11
	bp	1.46	1.10	1.42	1.20	4.35	2.47	3.08	2.90

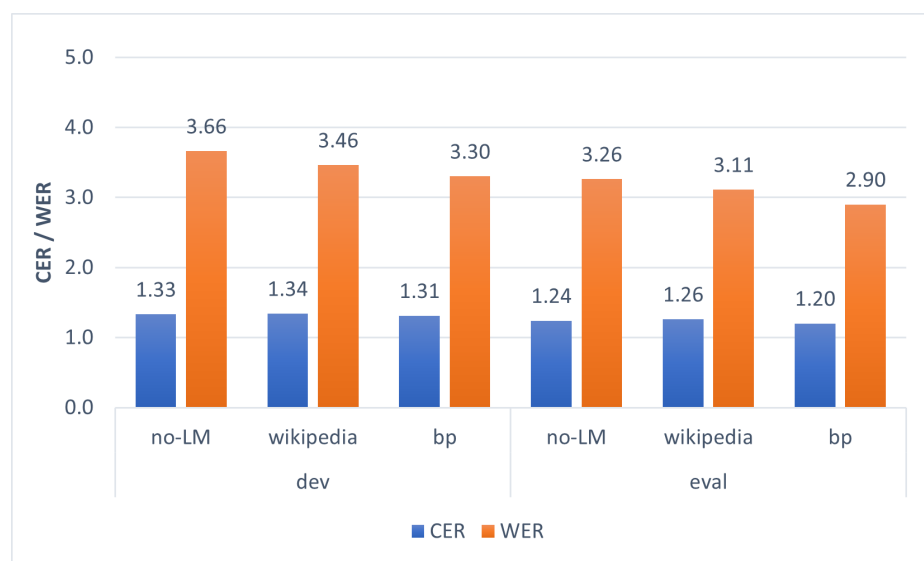


Figure 4. Global CER and WER performance on the dev and eval sets for different language models.

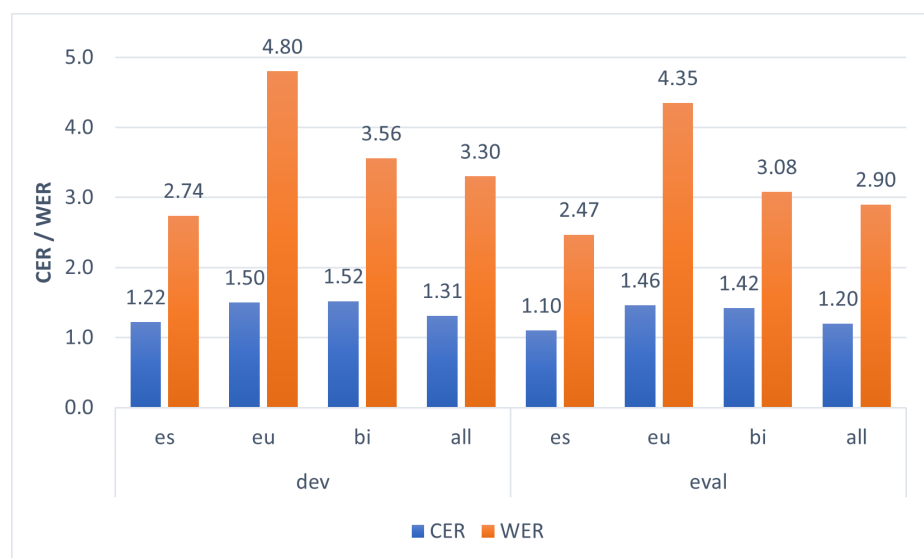


Figure 5. CER and WER performance, disaggregated per language, on the dev and eval sets when using the Basque Parliament language model.

ASR performance is good even when no language model is used (3.26 WER on the eval set), which means that the acoustic models match the background and channel conditions, speakers, etc., of the test set. This is an important result because the main motivation for creating the Basque Parliament database was training in-domain acoustic models and building a competitive ASR system, able to produce highly reliable transcriptions. The integration of lexical and language models in the search for the optimal sequence of graphemes/words does not always improve performance. For instance, the use of generic lexical and language models extracted from Wikipedia (wiki-LM) leads to worse results in terms of CER, although it does generally lead to better performance in terms of WER. On the other hand, the Basque Parliament lexical and language models (bp-LM) do always help, both in terms of CER and WER, which underlines the importance of in-domain lexical and language models. The best overall WER on the eval set (2.90) represents a 28% relative WER reduction with regard to our previous best result (4.02). A language-wise analysis shows that the best performance is found for Spanish (2.47 WER on the eval set) and the worst performance for Basque (4.35 WER on the eval set), meaning 32% and 15% relative WER reductions with regard to our previous best results for Spanish (3.66 WER) and Basque (5.13 WER), respectively. It seems that the imbalance between Spanish and Basque segments in the training set could explain the difference in performance between the two languages. Finally, the good performance attained on bilingual segments (3.08 WER on the eval set) is also remarkable, with some of them featuring code-switching events. In fact, exploring the potential of integrating acoustic and language models into a single bilingual ASR system for other pairs of code-switching languages is an interesting line for future research.

6. Summary and Future Work

In this paper, a new bilingual Basque–Spanish speech database has been presented. The database was extracted from Basque Parliament plenary sessions with the aim of helping the development of speech technology for the Basque language, which is relatively low-resourced. However, the database is well-suited for the development of bilingual ASR systems capable of decoding speech signals in both Basque and Spanish and that would seamlessly transition between languages. Given the similarity between Basque and Spanish at the phonetic/phonological level, the acoustic models can be shared by both languages, which circumvent the lack of training data for Basque. The ASR system becomes fully bilingual by using a single dictionary including words in Basque and Spanish and a single language model trained on texts in both languages.

The database is designed specifically for processing Basque Parliament speeches, which implies that its performance may suffer when dealing with speech from other sources featuring diverse speakers, more variable channels and environments and different or more general domains. Thus, due to the limited number of speakers and the relatively stable conditions of plenary sessions in the Basque Parliament, our database should be supplemented with additional training datasets featuring a broader range of speakers, channels, environments, speech styles (such as conversational speech) and domains to develop ASR systems capable of transcribing speech in Spanish, Basque, or both languages under various conditions.

The database consists of two subsets: the training and test datasets. For each dataset, an index file is provided with each line containing the information corresponding to a single segment: the audio filename, language and speaker tags, text-audio similarity score, segment length and transcription. The training set is aimed at estimating the acoustic models of an ASR system and was extracted from the 2013 to 2021 sessions, amounting to 1445 h of speech (1315 h if only highly reliable segments are considered) with a distribution strongly biased towards Spanish (72% Spanish, 28% Basque). The test set was extracted from five sessions held in February 2022 amounting to 17.2 h of speech and was split into a development set (7.6 h) for tuning purposes and an evaluation set (9.6 h) for measuring the performance of ASR systems. The distribution of languages in the test set is also biased

towards Spanish (68% Spanish, 32% Basque). Both datasets are balanced with regard to gender.

The paper also describes the procedure employed to build the training dataset, which involves: (1) tokenization of the audio signals and the transcriptions at the phonetic level by means of a phonetic decoder and an in-house grapheme-to-phoneme converter, respectively; (2) alignment of the phonetic sequences obtained from the audio signals and transcriptions; (3) an iterative segment extraction procedure which uses the alignment similarity as the main selection criterion. The test dataset was built by applying the same procedure followed by manual supervision of language tags and transcriptions.

Finally, the Basque Parliament database has been characterized and validated by using it to build a state-of-the-art ASR system based on pre-trained Wav2Vec 2.0 models finetuned on the training set and evaluated on the test set. The best overall performance was 2.90 WER and 1.20 CER, obtained when using an in-domain language model. It must be noted that these low error rates are due to an almost perfect matching between the test and training sets in terms of speakers, speech styles, environment and channel conditions and domain. Performance would degrade if our ASR system was tested on other unrelated datasets featuring different speakers, speech styles, channels, environments and/or domains. As noted above, additional training datasets covering the test conditions should be used in that case to supplement our database. The datasets, the acoustic and language models and the Python scripts used to carry out ASR experiments are publicly available at *HuggingFace* (Basque Parliament database: https://huggingface.co/datasets/gttsehu/basque_parliament_1, accessed on 24 February 2024. Acoustic and language models and Python scripts: https://huggingface.co/gttsehu/wav2vec2-xls-r-300m-bp1-es_eu, accessed on 24 February 2024).

The next steps include making this resource available to other researchers in the Spanish and Basque speech technology communities and working to prepare a second and larger version of the Basque Parliament database which will hopefully include data from the years 2010, 2011, 2012, 2022 and 2023. Also, we plan to check how well our acoustic models would perform when tested on other unrelated datasets in Basque and Spanish by using suitable language models and vocabularies. Finally, another promising line of research involves checking the feasibility of using our bilingual ASR paradigm (shared acoustic models and integrated vocabularies and language model) as a generic way to deal with code-switched speech involving any pair of languages.

Author Contributions: All the authors have contributed equally to the work described in this paper, though they have worked on different tasks. Conceptualization, methodology, investigation, resources, data curation, writing—review and editing, all authors; software, experiments, G.B., A.V. and M.P.; writing—original draft preparation, L.J.R.-F.; visualization, M.P., G.B. and L.J.R.-F.; project administration, A.V. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially funded by the Spanish Ministry of Science and Innovation (OPEN-SPEECH project, PID2019-106424RB-I00) and by the Basque Government under the general support program to research groups (IT-1704-22).

Data Availability Statement: The Basque Parliament database, including the train, train-clean, dev and eval datasets, is available at https://huggingface.co/datasets/gttsehu/basque_parliament_1 (accessed on 24 February 2024). Acoustic and language models and Python scripts used to carry out ASR experiments are available at https://huggingface.co/gttsehu/wav2vec2-xls-r-300m-bp1-es_eu (accessed on 24 February 2024).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Rehm, G.; Way, A. *European Language Equality: A Strategic Agenda for Digital Language Equality*; Cognitive Technologies; Springer International Publishing: Berlin/Heidelberg, Germany, 2023. [CrossRef]
2. Geneva, D.; Shopov, G.; Mihov, S. Building an ASR Corpus Based on Bulgarian Parliament Speeches. In Proceedings of the 7th International Conference on Statistical Language and Speech Processing, Ljubljana, Slovenia, 14–16 October 2019; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2019; Volume 11816, pp. 188–197. [CrossRef]
3. Kirkedal, A.; Stepanovic, M.; Plank, B. FT Speech: Danish Parliament Speech Corpus. In Proceedings of the Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25–29 October 2020; pp. 442–446. [CrossRef]
4. Kratochvíl, J.; Polak, P.; Bojar, O. Large Corpus of Czech Parliament Plenary Hearings. In Proceedings of the 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, 11–16 May 2020; pp. 6363–6367. Available online: <https://aclanthology.org/2020.lrec-1.781> (accessed on 24 February 2024).
5. Plüss, M.; Neukom, L.; Vogel, M. Swiss Parliaments Corpus, an Automatically Aligned Swiss German Speech to Standard German Text Corpus. *arXiv* **2020**, arXiv:2010.02810. [CrossRef]
6. Díaz-Munío, G.V.G.; Silvestre-Cerdà, J.A.; Jorge, J.; Giménez-Pastor, A.; Iranzo-Sánchez, J.; Baquero-Arnal, P.; Roselló, N.; de Martos, A.P.G.; Civera, J.; Sanchís, A.; et al. Europarl-ASR: A Large Corpus of Parliamentary Debates for Streaming ASR Benchmarking and Speech Data Filtering/Verbatimization. In Proceedings of the Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August–3 September 2021; pp. 3695–3699. [CrossRef]
7. Solberg, P.E.; Ortiz, P. The Norwegian Parliamentary Speech Corpus. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20–25 June 2022; pp. 1003–1008. Available online: <https://aclanthology.org/2022.lrec-1.106> (accessed on 24 February 2024).
8. Virkkunen, A.; Rouhe, A.; Phan, N.; Kurimo, M. Finnish parliament ASR corpus. *Lang. Resour. Eval.* **2023**, *57*, 1645–1670. [CrossRef] [PubMed]
9. Mohamed, A.; Lee, H.; Borgholt, L.; Havtorn, J.D.; Edin, J.; Igel, C.; Kirchhoff, K.; Li, S.; Livescu, K.; Maaløe, L.; et al. Self-Supervised Speech Representation Learning: A Review. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1179–1210. [CrossRef]
10. Shi, J.; Berrebbi, D.; Chen, W.; Chung, H.; Hu, E.; Huang, W.; Chang, X.; Li, S.; Mohamed, A.; Lee, H.; et al. ML-SUPERB: Multilingual Speech Universal PERFORMANCE Benchmark. *arXiv* **2023**, arXiv:2305.10615. [CrossRef]
11. Etchegoyhen, T.; Arzelus, H.; Gete Ugarte, H.; Alvarez, A.; González-Docasal, A.; Benites Fernandez, E. Mintzai-ST: Corpus and Baselines for Basque–Spanish Speech Translation. In Proceedings of the IberSPEECH, Valladolid, Spain, 24–25 March 2021; pp. 190–194. [CrossRef]
12. Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F.M.; Weber, G. Common Voice: A Massively-Multilingual Speech Corpus. *arXiv* **2019**, arXiv:1912.06670. [CrossRef]
13. Kjartansson, O.; Gutkin, A.; Butryna, A.; Demirsahin, I.; Rivera, C. Open-Source High Quality Speech Datasets for Basque, Catalan and Galician. In Proceedings of the 1st Joint Workshop on SLTU and CCURL, Marseille, France, 11–12 May 2020; pp. 21–27. Available online: <https://aclanthology.org/2020.sltu-1.3> (accessed on 24 February 2024).
14. Lopez de Ipina, K.; Torres, I.; Onederra, L. Design of a phonetic corpus for a speech database in Basque language. In Proceedings of the 4th European Conference on Speech Communication and Technology (Eurospeech 1995), Madrid, Spain, 18–21 September 1995; pp. 851–854. [CrossRef]
15. del Pozo, A.; Aliprandi, C.; Álvarez, A.; Mendes, C.; Neto, J.P.; Paulo, S.; Piccinini, N.; Raffaelli, M. SAVAS: Collecting, Annotating and Sharing Audiovisual Language Resources for Automatic Subtitling. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, 26–31 May 2014; pp. 432–436. Available online: <https://aclanthology.org/L14-1190/> (accessed on 24 February 2024).
16. Alvarez, A.; Arzelus, H.; Prieto, S.; del Pozo, A. Rich Transcription and Automatic Subtitling for Basque and Spanish. In Proceedings of the Iberspeech 2016, Lisbon, Portugal, 23–25 November 2016; pp. 197–206. Available online: https://simonguiroy.github.io/data/OnlineProceedings_IberSPEECH2016.pdf (accessed on 24 February 2024).
17. Hernández, I.; Luengo, I.; Navas, E.; Zubizarreta, M.L.; Gaminde, I.; Sánchez, J. The Basque Speech_DAT(II) Database: A Description and First Test Recognition Results. In Proceedings of the 8th European Conference on Speech Communication and Technology, Interspeech 2003, Geneva, Switzerland, 1–4 September 2003; pp. 1549–1552. [CrossRef]
18. Odriozola, I.; Hernaez, I.; Torres, M.; Rodriguez-Fuentes, L.J.; Penagarikano, M.; Navas, E. Basque Speecon-like and Basque SpeechDat MDB-600: Speech Databases for the Development of ASR Technology for Basque. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland, 26–31 May 2014; pp. 2658–2665. Available online: <https://aclanthology.org/L14-1583/> (accessed on 24 February 2024).
19. Yi, C.; Wang, J.; Cheng, N.; Zhou, S.; Xu, B. Applying Wav2vec2.0 to Speech Recognition in Various Low-resource Languages. *arXiv* **2020**, arXiv:2012.12121. [CrossRef]
20. Pham, N.Q.; Waibel, A.; Niehues, J. Adaptive multilingual speech recognition with pretrained models. In Proceedings of the Interspeech, Incheon, Republic of Korea, 18–22 September 2022; pp. 3879–3883. [CrossRef]
21. Zhao, J.; Zhang, W.Q. Improving Automatic Speech Recognition Performance for Low-Resource Languages With Self-Supervised Models. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1227–1241. [CrossRef]
22. Gardner-Chloros, P. *Code-Switching*; Cambridge University Press: Cambridge, UK, 2009. [CrossRef]

23. Biswas, A.; Yilmaz, E.; van der Westhuizen, E.; de Wet, F.; Niesler, T. Code-switched automatic speech recognition in five South African languages. *Comput. Speech Lang.* **2022**, *71*, 101262. [[CrossRef](#)]
24. Zhang, C.; Li, B.; Sainath, T.N.; Strohmaier, T.; Mavandadi, S.; Chang, S.; Haghani, P. Streaming End-to-End Multilingual Speech Recognition with Joint Language Identification. In Proceedings of the Interspeech 2022, Incheon, Republic of Korea, 18–22 September 2022; pp. 3223–3227. [[CrossRef](#)]
25. Anidjar, O.H.; Yozevitch, R.; Bigon, N.; Abdalla, N.; Myara, B.; Marbel, R. Crossing language identification: Multilingual ASR framework based on semantic dataset creation and Wav2Vec 2.0. *Mach. Learn. Appl.* **2023**, *13*, 100489. [[CrossRef](#)]
26. Dhawan, K.; Rekesh, K.; Ginsburg, B. Unified Model for Code-Switching Speech Recognition and Language Identification Based on Concatenated Tokenizer. In Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching, Singapore, 7 December 2023; pp. 74–82. [[CrossRef](#)]
27. Yu, H.; Hu, Y.; Qian, Y.; Jin, M.; Liu, L.; Liu, S.; Shi, Y.; Qian, Y.; Lin, E.; Zeng, M. Code-Switching Text Generation and Injection in Mandarin-English ASR. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2023, Rhodes Island, Greece, 4–10 June 2023; pp. 1–5. [[CrossRef](#)]
28. Nga, C.H.; Vu, D.; Luong, H.H.; Huang, C.; Wang, J. Cyclic Transfer Learning for Mandarin-English Code-Switching Speech Recognition. *IEEE Signal Process. Lett.* **2023**, *30*, 1387–1391. [[CrossRef](#)]
29. van Vüren, J.J.; Niesler, T. Improving Under-Resourced Code-Switched Speech Recognition: Large Pre-trained Models or Architectural Interventions. In Proceedings of the Interspeech, Dublin, Ireland 20–24 August 2023; pp. 1439–1443. [[CrossRef](#)]
30. King, A.R. *The Basque Language: A Practical Introduction*; University of Nevada Press: Reno, NV, USA, 2012. Available online: <https://archive.org/details/the-basque-language-a-practical-introduction-by-alan-r-king> (accessed on 24 February 2024).
31. Igartua, I.; Olaizola, M.L.O. Basque: The language and its speakers. In *Linguistic Minorities in Europe Online*; Grenoble, L., Lane, P., Røyneland, U., Eds.; De Gruyter Mouton: Berlin, Germany; Boston, MA, USA, 2019. [[CrossRef](#)]
32. Olaizola, M.L.O.; Igartua, I. Basque Sound Segments. In *Linguistic Minorities in Europe Online*; Grenoble, L., Lane, P., Røyneland, U., Eds.; De Gruyter Mouton: Berlin, Germany; Boston, MA, USA, 2019. Available online: <https://www.degruyter.com/database/LME/entry/lme.10310122/html> (accessed on 24 February 2024).
33. Slembrouck, S. The parliamentary Hansard ‘verbatim’ report: The written construction of spoken discourse. *Lang. Lit.* **1992**, *1*, 101–119. [[CrossRef](#)]
34. Bordel, G.; Rodriguez-Fuentes, L.J.; Penagarikano, M.; Varona, A. GTTS Systems for the Albayzin 2022 Speech and Text Alignment Challenge. In Proceedings of the Iberspeech 2022, Granada, Spain, 14–16 November 2022. [[CrossRef](#)]
35. Bordel, G.; Nieto, S.; Penagarikano, M.; Rodriguez-Fuentes, L.J.; Varona, A. Automatic Subtitling of the Basque Parliament Plenary Sessions Videos. In Proceedings of the Interspeech 2011, Florence, Italy, 28–31 August 2011. [[CrossRef](#)]
36. Bordel, G.; Nieto, S.; Penagarikano, M.; Rodriguez-Fuentes, L.J.; Varona, A. A simple and efficient method to align very long speech signals to acoustically imperfect transcriptions. In Proceedings of the Interspeech 2012, Portland, OR, USA, 9–13 September 2012. [[CrossRef](#)]
37. Bordel, G.; Penagarikano, M.; Rodriguez-Fuentes, L.J.; Varona, A. Aligning Very Long Speech Signals to Bilingual Transcriptions of Parliamentary Sessions. In *Advances in Speech and Language Technologies for Iberian Languages*; Communications in Computer and Information Science; Springer: Berlin/Heidelberg, Germany, 2012; Volume 328, pp. 69–78. [[CrossRef](#)]
38. Bordel, G.; Penagarikano, M.; Rodriguez-Fuentes, L.J.; Álvarez, A.; Varona, A. Probabilistic Kernels for Improved Text-to-Speech Alignment in Long Audio Tracks. *IEEE Signal Process. Lett.* **2016**, *23*, 126–129. [[CrossRef](#)]
39. Penagarikano, M.; Varona, A.; Bordel, G.; Rodriguez-Fuentes, L.J. Semisupervised Speech Data Extraction from Basque Parliament Sessions and Validation on Fully Bilingual Basque–Spanish ASR. *Appl. Sci.* **2023**, *13*, 8492. [[CrossRef](#)]
40. Collobert, R.; Puhrsch, C.; Synnaeve, G. Wav2Letter: An End-to-End ConvNet-based Speech Recognition System. *arXiv* **2016**, arXiv:1609.03193. [[CrossRef](#)]
41. Moreno, A.; Poch, D.; Bonafonte, A.; Lleida, E.; Llisterri, J.; Marino, J.B.; Nadeu, C. Albayzin Speech Database: Design of the Phonetic Corpus. In Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech 1993), Berlin, Germany, 21–23 September 1993; pp. 175–178. [[CrossRef](#)]
42. Quilis Morales, A. *Tratado de Fonología y Fonética Españolas*; Gredos: Madrid, Spain, 2019.
43. Hualde, J. *Basque Phonology*; Taylor & Francis: Abingdon, UK, 2004. [[CrossRef](#)]
44. Heafield, K. KenLM: Faster and Smaller Language Model Queries. In Proceedings of the Sixth Workshop on Statistical Machine Translation, Edinburgh, UK, 30–31 July 2011; pp. 187–197. Available online: <https://aclanthology.org/W11-2123> (accessed on 24 February 2024).
45. Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12449–12460. [[CrossRef](#)]
46. Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 369–376. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.