





## Article

# Semisupervised Speech Data Extraction from Basque Parliament Sessions and Validation on Fully Bilingual Basque–Spanish ASR

Mikel Penagarikano , Amparo Varona , 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;  
mikel.penagarikano@ehu.eus (M.P.); amparo.varona@ehu.eus (A.V.); german.bordel@ehu.eus (G.B.)  
\* Correspondence: luisjavier.rodriguez@ehu.eus

**Abstract:** In this paper, a semisupervised speech data extraction method is presented and applied to create a new dataset designed for the development of fully bilingual Automatic Speech Recognition (ASR) systems for Basque and Spanish. The dataset is drawn from an extensive collection of Basque Parliament plenary sessions containing frequent code switchings. Since session minutes are not exact, only the most reliable speech segments are kept for training. To that end, we use phonetic similarity scores between nominal and recognized phone sequences. The process starts with baseline acoustic models trained on generic out-of-domain data, then iteratively updates the models with the extracted data and applies the updated models to refine the training dataset until the observed improvement between two iterations becomes small enough. A development dataset, involving five plenary sessions not used for training, has been manually audited for tuning and evaluation purposes. Cross-validation experiments (with 20 random partitions) have been carried out on the development dataset, using the baseline and the iteratively updated models. On average, Word Error Rate (WER) reduces from 16.57% (baseline) to 4.41% (first iteration) and further to 4.02% (second iteration), which corresponds to relative WER reductions of 73.4% and 8.8%, respectively. When considering only Basque segments, WER reduces on average from 16.57% (baseline) to 5.51% (first iteration) and further to 5.13% (second iteration), which corresponds to relative WER reductions of 66.7% and 6.9%, respectively. As a result of this work, a new bilingual Basque–Spanish resource has been produced based on Basque Parliament sessions, including 998 h of training data (audio segments + transcriptions), a development set (17 h long) designed for tuning and evaluation under a cross-validation scheme and a fully bilingual trigram language model.

**Keywords:** automatic speech recognition; multilingual speech; low-resource languages; code switching; semisupervised learning; spoken language resources



**Citation:** Penagarikano, M.; Varona, A.; Bordel, G.; Rodríguez-Fuentes, L.J. Semisupervised Speech Data Extraction from Basque Parliament Sessions and Validation on Fully Bilingual Basque–Spanish ASR. *Appl. Sci.* **2023**, *13*, 8492. <https://doi.org/10.3390/app13148492>

Academic Editors: Francesc Alías, Zoraida Callejas Carrión, António Joaquim da Silva Teixeira and José Luis Pérez Córdoba

Received: 22 June 2023  
Revised: 15 July 2023  
Accepted: 20 July 2023  
Published: 23 July 2023



**Copyright:** © 2023 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

In bilingual communities, speakers sometimes mix languages and jump spontaneously from one language to another, sometimes just for one word or phrase, sometimes for longer, and go back and forth several times [1]. This phenomenon, known as code switching, appears even in formal settings such as parliamentary sessions and raises some interesting problems from the point of view of automatic speech recognition (ASR) systems [2–4]. Commonly, each language requires a specific ASR system with its own phonetic, phonological, lexical and syntactic constraints. This means that language detection and segmentation (that is, language diarization) must be performed on code-switched speech before applying an ASR system [5–7]. This language identification and segmentation process adds complexity and computational cost, and may introduce unrecoverable ASR errors when language detection fails. Current efforts are being devoted to integrate code switching detection and ASR within end-to-end deep learning approaches [8–10]. In the last years, the interest in

code switching has increased for certain language pairs, especially Mandarin–English, with international evaluations being organized [11] and open datasets being released [12].

In this work, we deal with Basque and Spanish (the two official languages in the Basque Country). Basque is a language of unknown origins, spoken by around 900 thousand speakers in a small region of Spain and France [13,14]. Basque greatly differs from Spanish, especially at the lexical and syntactic levels. Only a relatively small number of words come from the Latin or Romance languages with which Basque has had contact (Spanish, French and, of course, Latin itself). Spanish (like English) builds its structures using individual words with different grammatical functions, while Basque uses a set of cases to mark the grammatical relationships between words in a sentence, each with specific syntactic functions, and words are built in an agglutinative way by adding suffixes to lexemes. Spanish uses a verb conjugation system based on person and number, while Basque includes markers for subject, object and indirect object within the verb itself. In declarative sentences, the most common order in Spanish is subject-verb-object while in Basque it is subject-object-verb. However, on a phonetic level, Basque shares many of its sounds with Spanish (including its five vowels), with only some consonants, such as /ts/, /ts'/, /s'/ and some other less frequent ones (see Table 1) not appearing in Spanish [15].

In fact, Basque and the variety of Spanish spoken in the Basque Country share a great deal of features at the acoustic level, which allows us to use a single set of models, able to process speech in both languages so that a code switched transcription would be naturally output. Our proposed ASR system includes a single set of acoustic models, a single vocabulary (including words in both languages, sometimes with the same transcriptions but different pronunciations, sometimes with different transcriptions but the same pronunciations) and a single (aggregated) language model, which accepts code switchings at any point.

A positive effect of this integrated approach is that sharing acoustic models can alleviate the lack of annotated spoken resources for the low-resource language (Basque, in this case), by taking advantage of the resources available for the other. This will hopefully increase the robustness of the ASR system for the low-resource language, especially if the sets of acoustic units of the two languages are relatively close (as in the case of Basque and Spanish). On the negative side, having a single vocabulary may lead to a higher number of errors, due to words being recognized in the wrong language (those pronounced in the same way or very closely in the two languages). Since the language model has been trained on sentences in both languages, some of them including code switchings, it can naturally accept any sequence of words in any language (the probability of such a sequence will always be nonzero), and this allows it to recognize sentences with code switchings—although the model has not been tuned for this.

Our ASR system is targeted at the plenary sessions of the Basque Parliament (BP), with the final goal of obtaining high-quality automatic subtitles. BP members speak in both languages, Basque and Spanish, and code switchings are relatively abundant, so our fully bilingual approach seems to fit the domain quite well. To achieve the best performance, using in-domain training data is key, so a critical part of our work involves collecting as much BP data (audio + minutes) as possible. An important issue with BP minutes is that they are approximate, not reflecting the audio content of plenary sessions, because false starts, repetitions, filled pauses, syntactic errors and even some words or expressions which are judged too colloquial have been filtered out by human auditors. In this way, the BP minutes would be easily read (being syntactically correct) and fit the intended meaning, but the correspondence with the audio is partially lost.

This forces us to use the BP minutes with caution by applying a semi-supervised method to align the minutes with the audio, extract segments and discard those considered not reliable enough. Note that our method does not match the classical semi-supervised training methods that have been applied for more than two decades [16–23]: while those methods deal with completely untranscribed data, we do have some approximate transcriptions.

**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
b	b	bolsa	begia
t	t	vino	etorri
d	d	tomo	etorri
		dedo	denda
		casa	
k	k	queso	ekarri
		kilo	
g	g	gata	gaia
f	f	fatal	afaria
		cero	-
θ	z	pazo	-
s	s	sala	hasi
s'	s	-	zoroa
ʃ	s	-	kaixo
x	j	mujer	ijito
r	R	rosa	
		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

Classical semi-supervised approaches start from bootstrap acoustic models, typically trained on a relatively low amount of accurately transcribed non-target speech and used to build an initial ASR system, which is applied to transcribe a much larger amount of untranscribed speech, which is the target domain of the ASR system. Typically, the most confident fragments of the transcribed speech are selected (or other more sophisticated criteria are applied to select the speech materials) to train a second round of acoustic models which replace the bootstrap models. The same procedure is then iteratively applied until some convergence criterion is met.

In this work, we follow a similar approach but instead of a full ASR system, we apply a phone recognizer and an in-house bilingual grapheme-to-phoneme converter. Since nominal transcriptions are already available (the parliament minutes), we align the nominal and the recognized transcriptions at the phone level and select those segments that best match. In this way, a large fraction of BP sessions can be leveraged for training acoustic models. Besides increasing the amount of training materials for our ASR system (which is initially trained on generic speech datasets in Basque and Spanish), adding BP segments to the training set will help to improve ASR performance specifically on BP sessions (due to

an implicit adaptation to speakers, acoustic conditions, vocabulary, etc.), which is the main objective of this work. In [24], the authors also targeted BP plenary sessions, but adopted a different approach to leverage their speech contents, by creating two separate datasets for Spanish and Basque, on which two monolingual ASR systems were trained.

As a result of this work, we obtained a speech database specifically targeted at BP sessions. The database includes a large amount (998 h) of speech data for training acoustic models, and a development dataset (comprising more than 17 h of speech) used for tuning and evaluation under a cross-validation scheme. This latter dataset was extracted from a separate set of more recent BP sessions (not included in training) and then manually audited (their transcriptions being edited to match the audio contents). Finally, a bilingual (aggregated) trigram language model, estimated from the original minutes and translations of BP plenary sessions in Spanish and Basque, is also provided.

This paper is an extension of a previous work [25]. The primary purpose of that work was to collect speech data for Basque and Spanish (with particular emphasis on the former) using the Basque Parliament plenary sessions as source. Second, we also aimed to build a fully bilingual ASR system especially targeted at BP sessions, so that its output could be reliably used as a starting point to produce the minutes (which still required human supervision). In this paper, we provide new results and more in-depth analyses. The new contributions of this work with regard to [25] are summarized as follows: (1) the semi-supervised method employed to extract, rank and select training segments from BP sessions is now applied iteratively until the observed improvement is small enough; (2) to increase the statistical significance of performance results, the small (4 h long) development and test datasets used in [25] have been replaced by a larger (17 h long) development set which is used under a cross-validation scheme, with 20 random 50/50 partitions, to perform hyperparameter tuning and then compute ASR performance; (3) the hyperparameter tuning procedure is described in detail and (4) the results section is enriched with figures illustrating the convergence of the training process and the ranking of segment scores.

The rest of the paper is organized as follows. Sections 2 and 3 describe the main components of our bilingual ASR system and the method used to extract, rank and select training segments, respectively. Section 4 provides the details of the experimental framework used to evaluate our fully bilingual ASR system on Basque Parliament data, while Section 5 presents and discusses the results obtained in cross-validation experiments on the new development dataset specifically created in this work. Finally, a summary of the paper, conclusions and further work are outlined in Section 6.

## 2. The Components of a Fully Bilingual ASR System

### 2.1. Acoustic Units

Spanish and Basque phonetic units are not identical but overlap to a great extent, especially if we consider the standard Basque spoken in urban environments where Spanish is dominant. In these urban environments, which gather most of the population in the Basque Country, speakers tend to *soften* their Basque pronunciation, mapping Basque phonemes into something closer to Spanish phonemes. Therefore, the set of acoustic units considered by our bilingual ASR system is reduced and simplified by loosely taking into account the frequency of each unit and its most common realizations [26]. For instance, the three Basque affricates ( $tʃ$ ,  $ts'$  and  $ts$ ) are collapsed into a single affricate: the one existing in Spanish ( $tʃ$ ). Similarly, the Basque fricatives  $s'$  (as in *zoroa*) and  $ʃ$  (as in *kaixo*) are collapsed into the fricative  $s$ , existing in both Basque and Spanish. On the other hand, the Spanish fricative  $\theta$  (as in *pazo* and *cero*), which does not strictly exist in Basque, is kept because it is commonly used for proper names. The reduced set of phonetic units is shown in Table 1, including the original IPA units, their ASCII counterparts (which account for the units actually used in this work) and examples in both languages. We ended up with a reduced set of 23 phonetic units. An additional unit was also defined in our experiments to account for silences and other background (non-linguistic) events.

## 2.2. Lexical Models

An in-house bilingual grapheme-to-phoneme (G2P) converter has been developed and applied to obtain the phonetic baseforms of words in Basque and Spanish [27,28]. All words are then gathered in a single lexicon. The G2P converter is based on two pronunciation dictionaries for Basque and Spanish, each including hundreds of thousands of words, with verb inflections, declined words, numbers, acronyms, etc. These dictionaries, which were initialized from Mozilla CommonVoice (cv-corpus-5.1-2020-06-22) [29], Aditu [30] and Albayzin [31], include nominal pronunciations in terms of the reduced set of acoustic units, obtained by applying a different set of pronunciation rules for each language [32,33]. Dictionaries grow dynamically as new words are found in Basque Parliament sessions: if a known word is found, the G2P converter uses the stored pronunciation; but if an unknown word is found, pronunciation rules are applied to obtain its phonetic baseform, which is stored in the corresponding dictionary. In the case of unknown words or known words existing in both dictionaries, we must decide about the language (Basque or Spanish). This decision is based on the context: the language with more words in a window around the current word is chosen. In fact, a series of window sizes are considered, starting at 1 (one word at each side of the word under analysis) and increasing to 2, 3, etc., (up to the length of the sentence), until a reliable decision can be made (note that some words appear in both dictionaries). This strategy is found to be effective in practice, leading to very few errors. Numbers and ordinals (such as 25, 13.87, 1., etc.)—which are also transcribed either in Basque or in Spanish depending on the context—are assigned their most likely pronunciation, though sometimes it might not match the actual pronunciation. For instance, the Spanish phrase ‘1.5 millones’ (‘1.5 millions’) is transcribed as ‘uno coma cinco millones’ (‘one point five millions’) while the speaker might have actually said ‘un millón y medio’ (‘one million and a half’). Finally, acronyms are expected to be written in all-caps and assumed to be spelled, with exceptions being listed in the pronunciation dictionaries. After processing each Basque Parliament session, the pronunciations of new words added to the dictionaries are supervised and validated by a human expert.

## 2.3. Language Model

A mixed (aggregated) language model has been built based on the Basque Parliament minutes and their translations, including all sessions from 2010 to 2021. In the original BP minutes, Spanish is dominant over Basque, with a 2:1 ratio, but the professionally produced translations included in the minutes have exactly the opposite relation, so that the language model is trained with exactly the same amount of text in both languages. Text normalization has been applied, which involves deleting punctuation marks, converting numbers and ordinals into their alphabetical counterparts (using the most likely realization), putting all words (except for acronyms) in lower case, etc. It must be noted that sentences have been considered atomic units so that they would always feature a single language, except for single words or short phrases (that could be expressed even in a third language, like French or English). In any case, since the language model is estimated from texts in Basque and Spanish, it naturally allows a mix of both languages, including code switching events not seen during training, because there is always a small probability that a word in Basque comes after a word in Spanish (and the other way around). Actually, this feature, along with the use of a common set of acoustic units, is what makes our ASR system truly bilingual and robust to code switchings.

## 3. Iterative Data Collection through Phonetic Decoding and Alignment

For each Basque Parliament plenary session, an audio file and the corresponding minutes are available. In fact, for ease of processing, each audio file is manually split into two or three smaller chunks (each about 2 h long) and the minutes are split accordingly. As a starting point, a phone recognizer, trained on generic datasets for Basque and Spanish (not including BP materials) is applied to the audio files (without any phonological restrictions), to obtain a long sequence of phonetic units with their corresponding timestamps. On

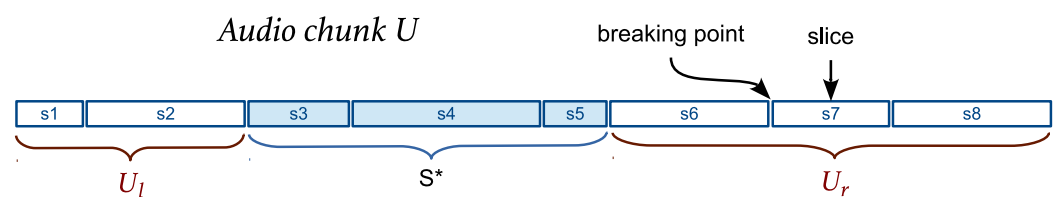
the other hand, the minutes are passed through the above mentioned G2P converter to acquire a reference (nominal) sequence of phonetic units. Finally, the recognized and reference sequences of phonetic units are aligned one with another under the criterion of maximizing the number of matching units (which is approximately the same as minimizing the number of deletions, insertions and substitutions), following the same text-and-speech alignment method that has been successfully applied in our group for the alignment of BP subtitles [28,34,35]. In this way, those regions showing a high density of errors in the alignment would correspond to parts of the minutes which do not match the audio contents.

The recognized phonetic sequence sometimes features gaps between two consecutive units, which represent silent pauses. Gaps longer than 0.5 s are defined as potential *breaking points*. Then, a *slice* is defined as an audio chunk between two consecutive breaking points and a *segment* as an audio chunk comprising one or more consecutive slices. 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 phone recognition rate (PRR), defined as:

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

where  $m$ ,  $d$ ,  $i$  and  $s$  are the number of matching units, deletions, insertions and substitutions yielded by the alignment for a given segment, respectively. When two or more segments attain the same (maximum) PRR, the longest segment is chosen. So PRR and length are the primary and secondary selection criteria, respectively.

A single-pass search is performed (with linear time complexity) to maximize PRR and length over those segments meeting the duration constraints in an audio chunk  $U$ . Note that, because of segment duration constraints, for each starting slice, the method has to consider just a limited number of following slices (usually one or two). Once determined the optimal segment  $s^*$  in an audio chunk  $U$ , the two audio sub-chunks at the left and right sides of  $s^*$ ,  $U_l$  and  $U_r$ , if not empty, are independently searched in two recursive calls (see Figure 1). Each call returns a list of segments, so we acquire two lists  $S_{U_l}$  and  $S_{U_r}$ , which are merged along with the optimal segment  $s^*$  into a single list  $S_U$ . In this way, after searching all the audio files, we end up with a list of segments  $S$  that can be filtered according to PRR.



**Figure 1.** An audio chunk  $U$  with 8 slices: the optimal segment  $s^*$  is chosen (the longest one among those with the highest PRR); 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 recurrence relation defining the time complexity of the search procedure for an audio chunk  $U$  with  $n$  slices would be:

$$T(n) = \begin{cases} 1 & n \leq 1 \\ n + T(i) + T(j) & n > 1 \end{cases} \quad (2)$$

where  $i$  and  $j$  (with  $i + j < n$ ) are the number of slices in the audio chunks  $U_l$  and  $U_r$ , respectively. Note that  $i = 0$  implies that  $U_l$  would be empty and the recursive call would not be carried out; on the other hand, if  $i = 1$ ,  $U_l$  would consist of a single slice and the search would reduce to checking duration constraints. The same stands for  $j$  and  $U_r$ . Recurrence (2) resembles that of the well-known *quicksort* algorithm, which, despite

being  $O(n^2)$  in the worst case, has an average cost of  $O(n \log n)$ . Finally, if  $K$  audio files are to be processed, the time complexity of the segment extraction 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 file.

Once the list of segments  $S$  is obtained for the whole set of BP sessions in the training set, a new set of acoustic models can be trained by using only those segments for which the provided transcription best matches the speech contents, either by requiring PRR to be higher than a given threshold or by using the top ranking segments amounting to a given number of hours (e.g., 998 h). The resulting models can be then applied again to perform phone recognition, obtain new alignments and hopefully a better set of segments for training. Remind that we aim to collect those segments that best match acoustically the provided transcripts. However, after each iteration the models will better adjust to the provided (possibly wrong) transcripts, so that after many iterations we may eventually achieve a PRR of 100% for all segments, with no way to distinguish *truly good* transcripts from *bad* transcripts to which our models have adapted to. This will prevent us from running too many iterations and will force us to carefully set the threshold that separates *good* from *bad* segments after each iteration. We will come back to this issue in Section 5.

#### 4. Experimental Setup

The acoustic models for the initial (bootstrap) phone recognizer have been trained on generic speech databases in Basque and Spanish: CommonVoice (cv-corpus-5.1-2020-06-22) [29], OpenSLR (SLR76) [36], Aditu [30] and Albayzin [31] (see Table 2). The development and test sets of Aditu and Albayzin were used to validate and evaluate phone recognition performance. The training, development and test sets have durations of 332.21, 3.96 and 4.03 h, respectively. Note, however, that Spanish and Basque are highly imbalanced in the training set (with a 3:1 ratio). PRR on the test sets of Aditu (Basque) and Albayzin (Spanish) were of 4.6% and 6.9%, respectively.

**Table 2.** Databases used to train the baseline (bootstrap) acoustic models. Durations are expressed in hours.

Name	Basque	Spanish
CommonVoice	24.75	250.30
Aditu (train)	47.40	-
OpenSLR (SLR76)	5.66	-
Albayzin (train)	-	4.10
Total	77.81	254.40

To build the phone recognizer, an off-the-shelf, close to state-of-the-art end-to-end neural network-based ASR system is used: Facebook AI Research wav2letter++ (consolidated into Flashlight), applying the Gated ConvNet recipe presented in [37]. Note that the phone recognizer requires neither lexical models nor a language model. For the semisupervised data collection step, all the BP plenary sessions from 2014 to 2021 (amounting to more than 1200 h) are used. Despite having access to BP sessions from 2010 to today, the audios prior to 2014 were recorded and stored using different formats and protocols, which prevented us from using those audios in this work.

The ASR system is also based on wav2letter++. In this case, besides the acoustic models, lexical and language models are also estimated, based on the minutes and translations of all BP plenary sessions from 2010 to 2021, which comprise more than 33 million words and around 279 thousand different entries. For each word in the vocabulary, a single pronunciation baseform is considered, as provided by our in-house G2P converter. A trigram language model is computed using KenLM [38] (without pruning), including close to 16 million trigrams.

#### 4.1. Hyperparameter Tuning

Though this work was not oriented towards optimizing the wav2letter++ framework used to build our ASR systems, we realized that three hyperparameters were critical for ASR performance: (1) *lmweight*: the language model weight which is accumulated with the acoustic model score; (2) *silscore*: the silence score (penalty) added whenever a silence unit is appended to the output; and (3) *wordscore*: the score (penalty) added when appending a word to the output. To get the most of the wav2letter++ framework, tuning these parameters really makes a difference. So, a random walk search (see Algorithm 1) is performed to optimize ASR performance on a tuning dataset, and then the optimal hyperparameters are applied when processing a test set. Both the tuning and test datasets are independent from the training set (see Section 4.2 for details).

---

#### Algorithm 1 Random walk optimization

---

```

1: function RWOPT( $D, M, N$ ) ▷  $D$ : tuning data,  $M$ : ASR model,  $N$ : max iterations
2:    $l, s, w \leftarrow 1, -1, 1$ 
3:    $\delta = (\delta_l, \delta_s, \delta_w) \leftarrow (0.3, 0.3, 0.3)$ 
4:    $\delta^{(min)} = (\delta_l^{(min)}, \delta_s^{(min)}, \delta_w^{(min)}) \leftarrow (0.001, 0.001, 0.001)$ 
5:    $E \leftarrow \emptyset$  ▷  $E$ : grid points already evaluated
6:    $min\_wer \leftarrow comp\_wer(D, M, l, s, w)$ 
7:    $i \leftarrow 0$ 
8:   while  $i < N$  and  $\delta \neq \delta^{(min)}$  do
9:      $C \leftarrow \{(x, y, z) \mid x \leftarrow l \pm \delta_l, y \leftarrow s \pm \delta_s, z \leftarrow w \pm \delta_w\}$ 
10:     $C \leftarrow C - C \cap E$  ▷  $C$ : grid points to be evaluated at this iteration
11:    if  $C \neq \emptyset$  then
12:       $i \leftarrow i + 1$ 
13:       $x, y, z \leftarrow pick\_random(C)$ 
14:       $E \leftarrow E \cup \{(x, y, z)\}$ 
15:       $wer \leftarrow comp\_wer(D, M, x, y, z)$ 
16:      if  $wer < min\_wer$  then
17:         $min\_wer \leftarrow wer$ 
18:         $l, s, w \leftarrow x, y, z$ 
19:      end if
20:    else
21:       $\delta_l \leftarrow \max(\delta_l^{(min)}, \delta_l/2)$ 
22:       $\delta_s \leftarrow \max(\delta_s^{(min)}, \delta_s/2)$ 
23:       $\delta_w \leftarrow \max(\delta_w^{(min)}, \delta_w/2)$ 
24:    end if
25:  end while
26:  return  $(l, s, w)$ 
27: end function

```

---

The method sketched in Algorithm 1 includes the initial values of the hyperparameters ( $l$ : *lmweight*,  $s$ : *silscore* and  $w$ : *wordscore*), the initial values of the deltas used to explore the hyperparameter space and the minimum values of those deltas, which mark an exit point when attained. All of them were heuristically adjusted in preliminary experiments. Note that the method also terminates when it reaches a maximum number of iterations  $N$ , which has been set to 500 in this work. Note also that two auxiliary functions are used: (1) *comp\_wer*, which is assumed to perform ASR on a dataset  $D$  using some pretrained models  $M$  and some hyperparameter values, and returns the attained Word Error Rate (WER); and (2) *pick\_random*, which is assumed to return a random element from a given set. Finally, note that the method involves some amount of randomness which might produce convergence issues. To study the impact of randomness, we ran the method a number of times on different datasets and observed that the hyperparameters obtained on a given set may actually differ across runs (due to randomness), but the ASR performance attained



was almost the same in all cases. This means that different hyperparameter values could be equally good and lead to the same (close to optimal) performance.

#### 4.2. Development Dataset and Cross-Validation Procedure

A development dataset was collected and used to carry out cross-validation experiments, first to tune wav2letter++ hyperparameters (using half of the dataset) and then to measure WER performance (using the other half), considering 20 random partitions and reporting the average WER. The development dataset comprises a set of segments extracted from the five BP sessions held in February, 2022 (thus not overlapping with the training set). Segments were extracted in the same way as the training segments, meaning that they did not correspond to complete sentences but to pieces of one or two sentences. These segments were manually audited (the audio listened to and the transcripts fixed) only at sections where the recognized sequence of words did not match the text in the minutes. These sections were located automatically, and involved any number of substitutions, deletions and/or insertions. The transcript resulting after auditing could be either the recognized sequence of words, the text provided in the minutes or a different sequence of words not matching any of them. Finally, each segment was automatically classified as containing only Spanish, only Basque or being bilingual (probably with a code switching event). This allowed us to disaggregate ASR performance by language. Details about this dataset are shown in Table 3.

**Table 3.** Development dataset used to tune and evaluate ASR systems through cross-validation.

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

The 9251 segments of this dataset are organized chronologically, in the same order as they were produced in the original BP sessions. To define each partition, first an index  $k$  is chosen randomly between 0 and 9251, so that the half starting at  $k$  (from  $k$  to  $(k + 4625 - 1) \bmod 9251$ ) is assigned to the tuning set, while the half ending at  $k - 1$  (from  $(k + 4625) \bmod 9251$  to  $(k - 1) \bmod 9251$ ) is assigned to the test set. This guarantees temporal coherence within both subsets, which will possibly contain different speakers and different topics, making the partition more realistic.

For each partition, we compute WER performance on both the tuning and test sets, obtaining one global and three per-language WER figures, for the Basque, Spanish and bilingual subsets of segments. In this way, we end up with eight WER results, by computing averages for the 20 partitions considered in cross-validation experiments. Besides the averages, standard deviations and 95% confidence intervals for the averages (using normal distributions) are also computed.

## 5. Results

### 5.1. Baseline (Out-of-Domain) Models

The first part of this work involved training baseline acoustic models to build a bilingual phone recognizer for Basque and Spanish using wav2letter++ and the datasets in Table 2, as a starting point for the semisupervised data collection procedure. Table 4 shows the amount of speech that would be collected by applying different PRR thresholds to the list of segments (by keeping only those segments with  $PRR > \text{threshold}$ ). By inspecting these numbers, we determined that  $PRR = 80$  was a good compromise between the amount of speech recovered and the quality of reference transcriptions.

The baseline acoustic models were also used to run word-level recognition experiments using the wav2letter++ ASR system described in Section 4. Table 5 shows the average

WER, disaggregated per language, obtained by baseline models on the tuning and test sets in cross-validation experiments. Though WER figures are slightly better (lower) for the tuning set (as may be expected, because hyperparameters are optimized on it), the hyperparameter values seem to be working quite well also for the test set. It is also quite remarkable that sharing the acoustic models and using a single aggregated language model seems to work equally fine for Basque and Spanish. Worst per-language results are obtained on bilingual segments, something that could be expected, due to code switchings. On the other hand, while results are quite similar for Basque and Spanish, higher variabilities (standard deviations and 95% confidence intervals) are found for Basque. This might be due to the particular speakers that are being evaluated. An even larger cross-validation dataset should be used to avoid these variability issues.

**Table 4.** Amount of speech (in hours) obtained from the training set by applying the baseline phone recognizer and keeping segments with  $PRR \geq \text{Threshold}$ .

PRR Threshold	Time (Hours)
100	186
95	490
90	745
85	902
80	1000
75	1054
70	1084
65	1100
60	1108

**Table 5.** WER performance of baseline acoustic models in cross-validation experiments.

Set	Metric	Basque	Spanish	Bilingual	All
Tuning	Avg	16.63	16.19	22.38	16.44
	StdDev	0.79	0.40	0.48	0.38
	95% CI	0.35	0.17	0.21	0.17
Test	Avg	16.57	16.38	22.44	16.57
	StdDev	0.86	0.45	0.51	0.46
	95% CI	0.38	0.20	0.22	0.20

### 5.2. In-Domain Models (Retrained on Basque Parliament Data)

The next step of the process consisted of using the speech segments of the BP plenary sessions with  $PRR \geq 80$ , which amount to around 998 h after discarding segments with a few number of phonetic units, to train a new set of acoustic models. Note that the baseline acoustic models were trained on around 332 h obtained from different and heterogeneous sources, which had nothing to do with BP sessions. At this point, however, we were using three times more training data; moreover, the data were extracted from BP sessions, under the same acoustic conditions and probably including some of the speakers that would appear on the datasets used to tune and evaluate the ASR system. Adapting to the operating conditions was, in fact, one of our main objectives, that is, taking advantage of the speech available from BP sessions to improve the performance of our ASR system when dealing with BP speech.

Average WER figures (disaggregated per language) obtained by the retrained models in cross-validation experiments are presented in Table 6. Large and consistent improvements can be observed with regard to the results shown in Table 5. The global WER goes from 16.44% to 4.29% on the tuning sets (meaning a 73.9% relative WER reduction) and from 16.57% to 4.41% on the test sets (meaning a 73.4% relative WER reduction). Again, as expected, results are slightly better on the tuning set than on the test set. These improvements could be generally due to using a greater amount of training data and these data

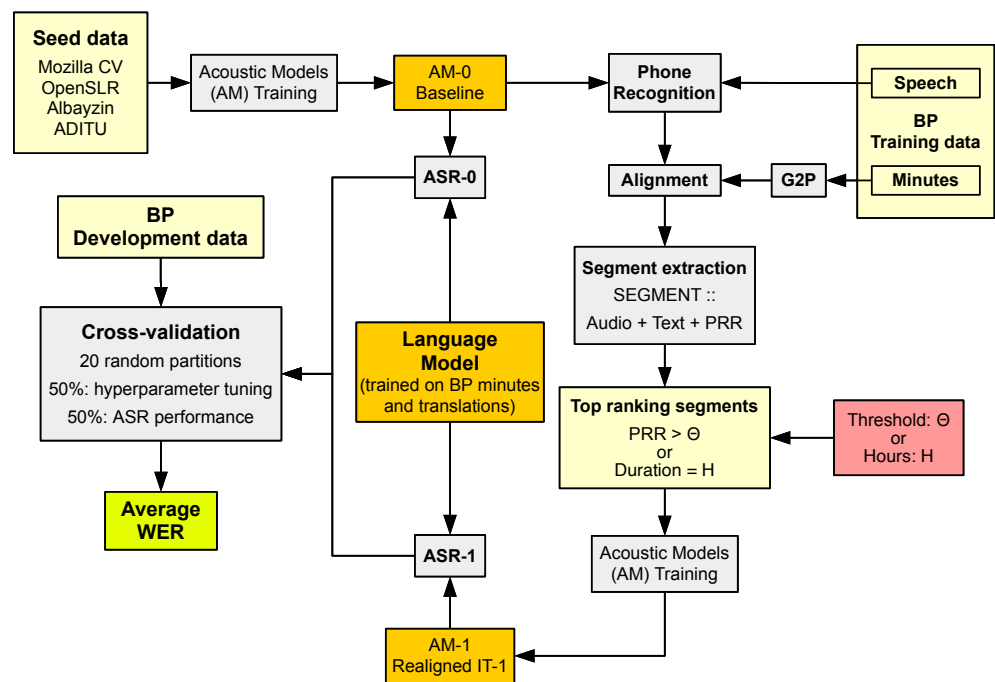
being *in-domain*, that is, the same speakers and environment/channel conditions appear in both training and test datasets. On the other hand, WER figures are better for Spanish than for Basque. This could be explained by different factors: (1) the dominance of Spanish over Basque (with a 2:1 factor) in the training set; (2) a higher variability of accents/dialects in Basque (with not only different pronunciations, but also different vocabularies) compared to Spanish, which features a single accent/dialect in the dataset; and (3) the use of a reduced set of acoustic units might be hindering the discrimination ability of the acoustic models *only for Basque*, because the fused consonants do not exist in Spanish.

**Table 6.** %WER performance of acoustic models obtained after one iteration of the semisupervised data collection method in cross-validation experiments.

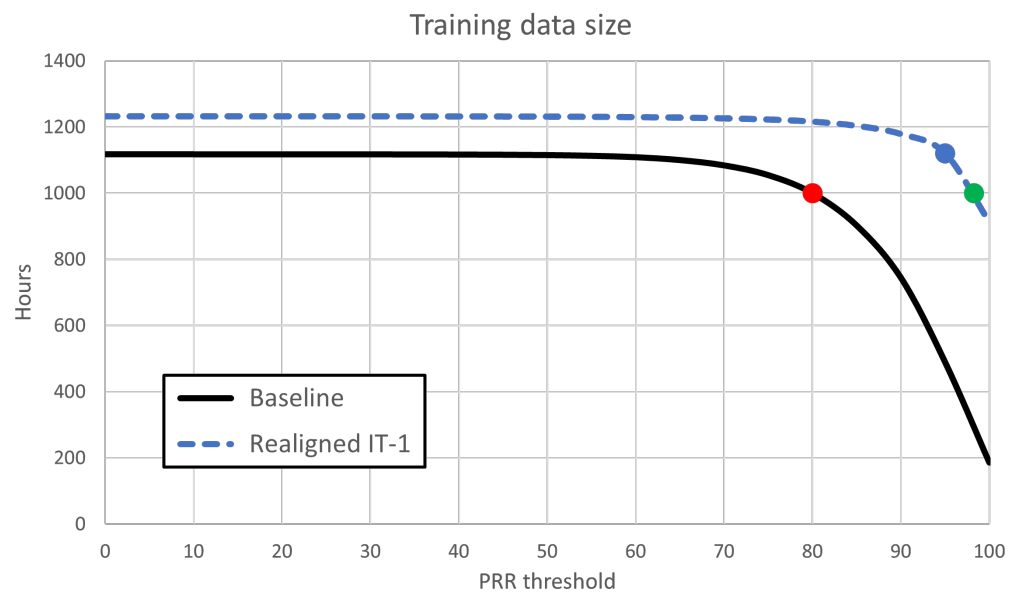
Set	Metric	Basque	Spanish	Bilingual	All
Tuning	Avg	5.43	3.93	4.38	4.29
	StdDev	0.16	0.17	0.55	0.14
	95% CI	0.07	0.07	0.24	0.06
Test	Avg	5.51	4.04	4.35	4.41
	StdDev	0.20	0.17	0.65	0.14
	95% CI	0.09	0.07	0.29	0.06

Figure 2 shows a schematic of what we have done so far: using the baseline models and the models obtained after one iteration of the semisupervised data collection method on the BP training set to run cross-validation experiments, which involve hyperparameter tuning and ASR evaluation on independent BP datasets. Next, we proceed with a second iteration of the process, using the updated models to run the whole data collection pipeline and obtain a new set of hopefully better aligned segments. While running this second iteration, we may decide either to keep the same amount of data as in the first iteration (998 h, which corresponds to using a PRR threshold of 98.28) or instead to apply an alternative threshold ( $PRR > 95$ ) to collect 1118 h (see Figure 3). Counterintuitively, the amount of training data available when the PRR threshold is 0 does not correspond to *all* the training data, because the collection procedure described in Section 3 discards segments smaller than 3 s (*orphan* segments, usually with a low PRR). Since PRR figures are higher for the acoustic models trained after realignment (because they have adapted to Basque Parliament data), the number of discarded low-PRR segments decrease, a sizeable amount of those formerly low-PRR segments are pasted to the surrounding segments and the amount of training data available increases.

We have tried the two options, obtaining practically the same performance in both cases (see Tables 7 and 8). This suggests that those additional 120 h of the second option, besides having questionable transcripts, do not improve the acoustic models and can thus be discarded. It seems that the criterion of choosing an equivalent amount of material (in this case, the top ranking segments amounting to 998 h) is good enough. Figure 4 shows WER results on the test set vs. WER results on the tuning set for 20 random partitions defined of the development set. These results reveal that: (1) hyperparameter tuning is working fine; (2) WER figures are highly and negatively correlated, which makes them somewhat *complementary*: all the partitions considered in the experiments comprise the same set of segments, so the global WER would be the same (around 4%) for all partitions, but segments would be distributed in different ways in the tuning and test sets, so for any given partition, WER figures would be higher on the most difficult set and correspondingly lower on the easiest one; and (3) despite the average WER is statistically identical in both cases, the models trained on the top ranking 998 h yield slightly better results than the models trained on the top ranking 1118 h, maybe because those 120 additional hours do not have reliable transcripts and introduce some noise.



**Figure 2.** Experimental framework, showing the semisupervised data collection pipeline and cross-validation experiments using the baseline acoustic models and the acoustic models obtained after one iteration of the pipeline. Top ranking segments can be chosen using a PRR threshold or a target amount of data.



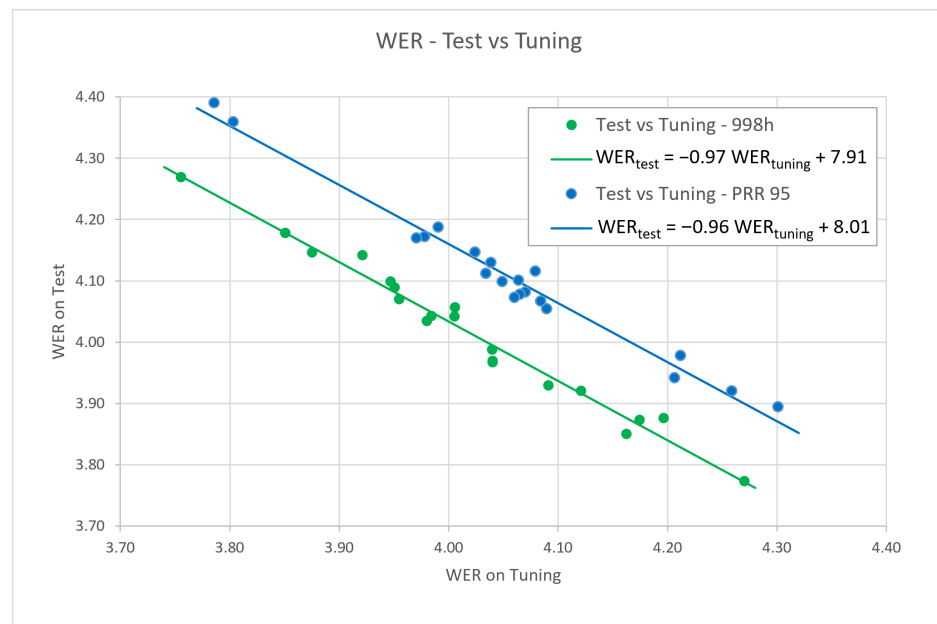
**Figure 3.** Amount of training data collected (in hours) depending on the PRR threshold, for the baseline (black) and first-iteration realigned models (blue). The red point signals the PRR threshold (80%) used to retrain the baseline models (998 h of top ranking segments); the green point signals the PRR threshold (98.28%) corresponding to using the top ranking 998 h in the second retraining stage; and the blue point signals an alternative PRR threshold (95%) for the second retraining stage (1118 h of top ranking segments).

**Table 7.** WER performance of the acoustic models obtained after a second iteration of the semisupervised data collection method (using the top ranking 998 h, PRR threshold = 98.28) in cross-validation experiments.

Set	Metric	Basque	Spanish	Bilingual	All
Tuning	Avg	5.09	3.66	3.95	4.02
	StdDev	0.16	0.12	0.36	0.13
	95% CI	0.07	0.05	0.16	0.06
Test	Avg	5.13	3.66	3.90	4.02
	StdDev	0.16	0.12	0.31	0.12
	95% CI	0.07	0.05	0.14	0.05

**Table 8.** WER performance of the acoustic models obtained after a second iteration of the semisupervised data collection method (using the top ranking 1118 h, PRR threshold = 95) in cross-validation experiments.

Set	Metric	Basque	Spanish	Bilingual	All
Tuning	Avg	5.22	3.67	4.11	4.06
	StdDev	0.14	0.13	0.50	0.13
	95% CI	0.06	0.06	0.22	0.06
Test	Avg	5.29	3.72	4.34	4.10
	StdDev	0.13	0.13	0.51	0.12
	95% CI	0.06	0.06	0.22	0.05

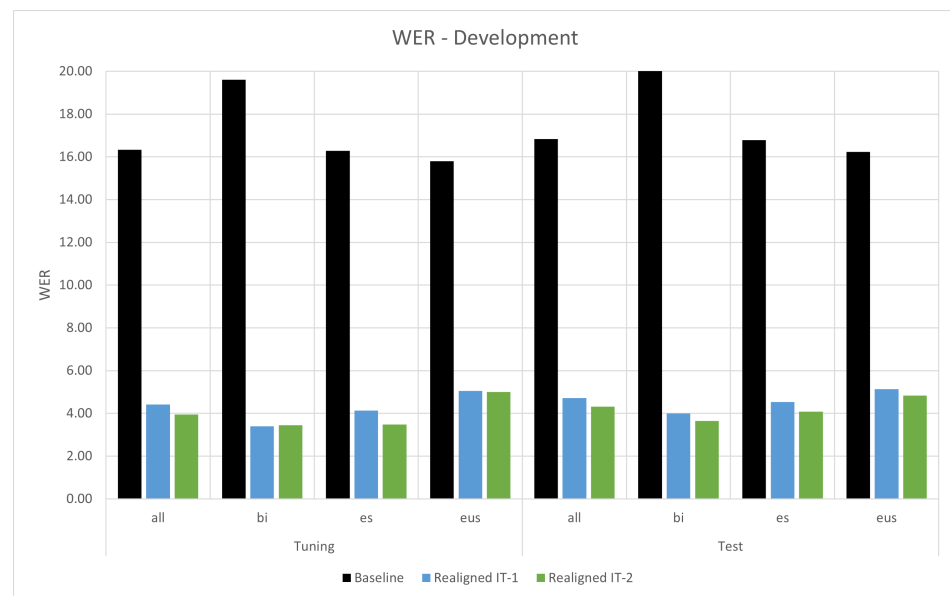


**Figure 4.** WER on test vs WER on tuning for 20 random partitions obtained by the second iteration realigned acoustic models, trained on: (a) the top ranking 998 h of segments (green); and (b) the top ranking segments with PPR > 95 (blue).

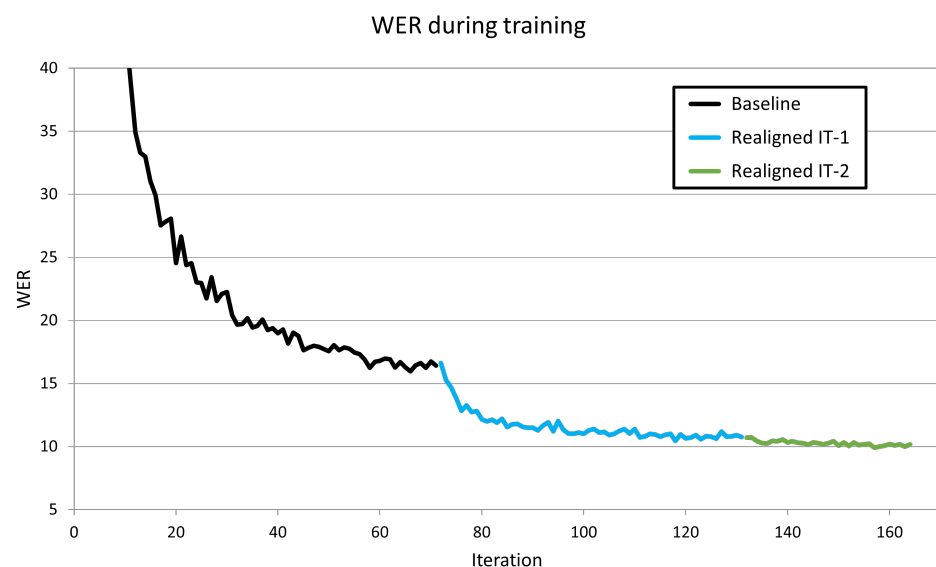
In any case, the performance attained by the second-iteration realigned models is only slightly better than that obtained with the first-iteration realigned models. Nothing compared to the huge improvement observed when going from the baseline to the first-iteration realigned models. This suggests that further iterations of the data collection pipeline are likely to yield even smaller improvements. WER results in cross-validation experiments are summarized in Figure 5.

To support our decision to stop the procedure, Figure 6 shows ASR performance (without language model) on a 4-h validation dataset extracted from a BP session not included in

training, for successive training iterations of the the baseline (black), first-iteration realigned (blue) and second-iteration realigned (green) acoustic models. Apparently, performance improvements get smaller as the training process progresses, and after the second-iteration realigned models converge, there seems to be little margin for improvement. Hopefully, at this point we have already recovered all the *good* segments (those for which the available transcript reasonably reflects the acoustic sequence). Furthermore, we discovered by manual inspection that, after repeatedly training on segments with inaccurate transcripts, our acoustic models did actually learn to recognize those transcripts, making PRR figures get increasingly close to 100%, so that those segments would be mistaken as *good*. In other words, performing more iterations of the data collection pipeline would make our models overfit and eventually all the segments in the training set (including both *good* and *bad* ones) would be collected.



**Figure 5.** WER performance in cross-validation experiments on the BP development dataset (tuning/test, 20 partitions) using baseline, first-iteration and second-iteration realigned acoustic models, disaggregated per language: Spanish (es), Basque (eus), bilingual (bi) and all segments.



**Figure 6.** WER performance (without language model) obtained during the training process on a 4-h validation dataset extracted from a BP session not included in the training set, for the baseline (black), first-iteration realigned (blue) and second-iteration realigned (green) acoustic models.

## 6. Conclusions

In this paper, we have presented and evaluated a semi-supervised data collection pipeline which leverages the audios and minutes of plenary sessions of the Basque Parliament to train domain-adapted models. We have also presented the main features of a fully bilingual ASR system for Basque and Spanish, based on the integration of acoustic, lexical and language models. Our ASR system is able to deal with code switching events in a natural and computationally efficient fashion, and yields remarkable ASR performance in the domain of the Basque Parliament plenary sessions. Global Word Error Rates (WER) reduce (on average) from 16.57% (baseline) to 4.41% (realigned models, first iteration) and 4.02% (realigned models, second iteration), meaning 73.4% and 8.8% relative WER reductions, respectively. Focusing on Basque (the low-resource language this work is targeted at), the average WER goes from 16.57% (baseline) to 5.51% (realigned models, first iteration) and 5.13% (realigned models, second iteration), meaning 66.7% and 6.9% relative WER reductions, respectively. Finally, as a byproduct of the process, a new bilingual database for Basque and Spanish is obtained, consisting of a training set (998 h long), a development dataset (17 h long) designed for cross-validation experiments and a fully bilingual language model (adapted to the Basque Parliament domain) featuring close to 16 million trigrams. Future work involves extending the development dataset for the Basque Parliament domain and trying to export the data collection pipeline to other domains.

**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 the authors; software, experiments, G.B., A.V. and M.P.; writing—original draft preparation, L.J.R.-F.; visualization, M.P. 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 datasets described in this paper will be released through a public repository once the paper is accepted. Software recipes will be also released to reproduce some of the experiments reported in the paper.

**Conflicts of Interest:** The authors declare no conflict 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. Gardner-Chloros, P. *Code-Switching*; Cambridge University Press: Cambridge, UK, 2009.
2. Yilmaz, E.; McLaren, M.; van den Heuvel, H.; van Leeuwen, D.A. Semi-supervised acoustic model training for speech with code-switching. *Speech Commun.* **2018**, *105*, 12–22. [[CrossRef](#)]
3. Dalmia, S.; Liu, Y.; Ronanki, S.; Kirchhoff, K. Transformer-Transducers for Code-Switched Speech Recognition. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, ON, Canada, 6–11 June 2021; pp. 5859–5863. [[CrossRef](#)]
4. 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](#)]
5. 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.
6. Yilmaz, E.; van den Heuvel, H.; van Leeuwen, D. Code-switching detection using multilingual DNNs. In Proceedings of the 2016 IEEE Spoken Language Technology Workshop (SLT), San Diego, CA, USA, 13–16 December 2016; pp. 610–616. [[CrossRef](#)]
7. Yilmaz, E.; McLaren, M.; van den Heuvel, H.; van Leeuwen, D.A. Language diarization for semi-supervised bilingual acoustic model training. In Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 16–20 December 2017; pp. 91–96. [[CrossRef](#)]
8. Seki, H.; Watanabe, S.; Hori, T.; Roux, J.L.; Hershey, J.R. An End-to-End Language-Tracking Speech Recognizer for Mixed-Language Speech. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4919–4923. [[CrossRef](#)]

9. Zeng, Z.; Khassanov, Y.; Pham, V.T.; Xu, H.; Chng, E.S.; Li, H. On the End-to-End Solution to Mandarin-English Code-Switching Speech Recognition. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 2165–2169. [[CrossRef](#)]
10. Qiu, Z.; Li, Y.; Li, X.; Metzger, F.; Campbell, W.M. Towards Context-Aware End-to-End Code-Switching Speech Recognition. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; pp. 4776–4780. [[CrossRef](#)]
11. Shi, X.; Feng, Q.; Xie, L. The ASRU 2019 Mandarin-English Code-Switching Speech Recognition Challenge: Open Datasets, Tracks, Methods and Results. *arXiv* **2020**, arXiv:2007.05916.
12. Li, C.; Deng, S.; Wang, Y.; Wang, G.; Gong, Y.; Chen, C.; Bai, J. TALCS: An Open-Source Mandarin-English Code-Switching Corpus and Speech Recognition Baseline. *arXiv* **2022**. [[CrossRef](#)]
13. King, A. *The Basque Language: A Practical Introduction*; University of Nevada Press: Reno, NV, USA, 2012.
14. Igartua, I.; Onederra-Olaizola, M.L. 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.
15. Igartua, I.; Onederra-Olaizola, M.L. 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.
16. Lamel, L.; Gauvain, J.; Adda, G. Lightly supervised and unsupervised acoustic model training. *Comput. Speech Lang.* **2002**, *16*, 115–129. [[CrossRef](#)]
17. Wessel, F.; Ney, H. Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition. *IEEE Trans. Speech Audio Process.* **2005**, *13*, 23–31. [[CrossRef](#)]
18. Yu, D.; Varadarajan, B.; Deng, L.; Acero, A. Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion. *Comput. Speech Lang.* **2010**, *24*, 433–444. [[CrossRef](#)]
19. Liao, H.; McDermott, E.; Senior, A. Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription. In Proceedings of the ASRU, Olomouc, Czech Republic, 8–12 December 2013.
20. Veselý, K.; Hannemann, M.; Burget, L. Semi-supervised training of Deep Neural Networks. In Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, 8–12 December 2013; pp. 267–272. [[CrossRef](#)]
21. Manohar, V.; Hadian, H.; Povey, D.; Khudanpur, S. Semi-Supervised Training of Acoustic Models Using Lattice-Free MMI. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP 2018), Calgary, AB, Canada, 15–20 April 2018; pp. 4844–4848. [[CrossRef](#)]
22. Long, Y.; Li, Y.; Wei, S.; Zhang, Q.; Yang, C. Large-Scale Semi-Supervised Training in Deep Learning Acoustic Model for ASR. *IEEE Access* **2019**, *7*, 133615–133627. [[CrossRef](#)]
23. Wotherspoon, S.; Hartmann, W.; Snover, M.; Kimball, O. Improved Data Selection for Domain Adaptation in ASR. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, ON, Canada, 6–11 June 2021; pp. 7018–7022. [[CrossRef](#)]
24. 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 2021, Valladolid, Spain, 24–25 March 2021; pp. 190–194. [[CrossRef](#)]
25. Penagarikano, M.; Varona, A.; Bordel, G.; Rodriguez-Fuentes, L.J. Semisupervised training of a fully bilingual ASR system for Basque and Spanish. In Proceedings of the IberSPEECH 2022, Granada, Spain, 14–16 November 2022; pp. 36–40. [[CrossRef](#)]
26. Lopez de Ipinza, 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](#)]
27. 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.
28. Bordel, G.; Penagarikano, M.; Rodriguez-Fuentes, L.J.; Varona, A. Aligning very long speech signals to bilingual transcriptions of parliamentary sessions. In Proceedings of the Iberspeech 2012, Madrid, Spain, 21–23 November 2012.
29. 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.
30. 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.
31. Moreno, A.; Poch, D.; Bonafonte, A.; Lleida, E.; Llisterra, 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, 22–25 September 1993; pp. 175–178. [[CrossRef](#)]
32. Quilis Morales, A. *Tratado de Fonología y Fonética Españolas*; Gredos: Madrid, Spain, 2019.
33. Hualde, J. *Basque Phonology*; Taylor & Francis: Abingdon, UK, 2004.
34. 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.
35. 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](#)]



36. 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.
37. Collobert, R.; Puhersch, C.; Synnaeve, G. Wav2Letter: An End-to-End ConvNet-based Speech Recognition System. *arXiv* **2016**, arXiv:1609.03193.
38. 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.

**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.