IberSPEECH 2022
14-16 November 2022, Granada, Spain

Semisupervised training of a fully bilingual ASR system for Basque and Spanish

Mikel Penagarikano, Amparo Varona, Germán Bordel, and Luis J. Rodriguez-Fuentes

Department of Electricity and Electronics
Faculty of Science and Technology, UPV/EHU
Barrio Sarriena, 48940 Leioa, Spain.
{mikel.penagarikano, amparo.varona, german.bordel, luisjavier.rodriguez}@ehu.eus

Abstract

Automatic speech recognition (ASR) of speech signals with code-switching (an abrupt language change common in bilingual communities) typically requires spoken language recognition to get single-language segments. In this paper, we present a fully bilingual ASR system for Basque and Spanish which does not require such segmentation but naturally deals with both languages using a single set of acoustic units and a single (aggregated) language model. We also present the Basque Parliament Database (BPDB) used for the experiments in this work. A semisupervised method is applied, which starts by training baseline acoustic models on small acoustic datasets in Basque and Spanish. These models are then used to perform phone recognition on the BPDB training set, for which only approximate transcriptions are available. A similarity score derived from the alignment of the nominal and recognized phonetic sequences is used to rank a set of training segments. Acoustic models are updated with those BPDB training segments for which the similarity score exceeds a heuristically fixed threshold. Using the updated models, Word Error Rate (WER) reduced from 16.46 to 6.99 on the validation set, and from 15.06 to 5.16 on the test set, meaning 57.5% and 65.74% relative WER reductions over baseline models, respectively.

Index Terms: Automatic Speech Recognition, Multilingual Speech, Semisupervised Learning, Spoken Language Resources

1. Introduction

In bilingual communities, sometimes speakers start with one language and then, at some point, switch to the other language just for one word or phrase or maybe for longer. Then they might switch back again, and repeat this cycle a number of times. This phenomenon is known as code switching [1] and must be handled by Automatic Speech Recognition (ASR) systems so that adequate acoustic and language models are applied [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) should de 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 English-Mandarin, with international evaluations [11] and open datasets [12].

In this work, we deal with a language pair of relatively low interest (Basque-Spanish) but some common features of Basque and the variety of Spanish spoken in the Basque Country allow us to explore a much simpler alternative. A single set of models is used, able to process speech in both languages so that a code switched transcription would be naturally output. Our proposal consists of using 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 language model, which should allow for code switchings at any point.

A positive side effect of this integrated approach is that sharing acoustic models can alleviate the lack of annotated spoken resources for one of the languages, 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). Also, though the language model will account for a large number of switching points, it may not generalize well and have problems to identify code switchings not seen in the training data.

We also address in this work the task of collecting training data from spoken resources with loose or inexact transcriptions. That is the case of the Basque Parliament (BP) minutes, which approximately reflect what was actually said in plenary sessions: false starts, repetitions, filled pauses, syntactic errors and other issues are either ignored or edited, so that the BP minutes can be easily read while respecting the intended meaning. Note that this is not the case of completely untranscribed speech on which most semi-supervised training approaches have been focusing for more than two decades [13–20]. Those approaches start from seed 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 sophisticatd criteria are applied to select the speech materials) to train a second round of acoustic models which replace the seed models. The same procedure is then iteratively applied until some convergence criterion is met.

Here we adopt 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 transcrip-

10.21437/IberSPEECH.2022-8

tions 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 was initialized 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 was also an objective of this work. In [21], the authors also targeted BP plenary sessions, but adopted a different approach to leverage their speech contents (e.g. they created two separate datasets, for Spanish and Basque, on which two monolingual ASR systems were trained).

Finally, as a result of this work, we obtain a speech database specifically targeted at BP sessions, consisting of the original BP plenary sessions' minutes in Spanish and Basque, along with their translations, a large amount (more than 1000 hours) of speech data for training acoustic models and two small datasets (each 2 hours long), extracted from a BP sesion not included in training, that were manually segmented and transcribed and used as development (validation) and evaluation (test) sets.

2. Dealing with bilingual resources

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. Therefore we decided to reduce and simplify the set of acoustic units considered by our ASR system, loosely taking into account their frequencies and their most common realizations. For instance, we collapsed three Basque africates (ʧ, ts' and ts) into a single africate: the one existing in Spanish (ʧ). Similarly, the Basque fricatives s' (as in *zoroa*) and ʃ (as in *kaixo*) were collapsed into the fricative s, existing in both Basque and Spanish. On the other hand, we decided to keep the Spanish fricative θ (as in *pazo* and *cero*), which does not strictly exist in Basque but it is sometimes 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.

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 vino	begia
t	t	tomo	etorri
d	d	dedo	denda
k	k	casa queso kilo	ekarri
g	g	gata	gaia
f	f	fatal	afaria
θ	z	cero pazo	–
s	s	sala	hasi
s'	s	–	zoroa
ʃ	s	–	kaixo
x	j	mujer	ijito
r	R	rosa torre	arrunta
ɾ	r	puro	dirua
l	l	lejos	lana
ʧ	X	mucho	txikia
ts'	X	–	atzo
ts	X	–	mahatsa
c	X	–	ttakun
ʎ	y	caballo	pilaka
ʝ	y	hielo cónyuge	–
j	y	–	joan
ɟ	y	–	onddo

2.2. Lexical models

A bilingual rule- and dictionary-based grapheme-to-phoneme (G2P) converter was developed and applied to get the phonetic baseforms of words in Basque and Spanish, which were integrated into a single lexicon. Numbers and ordinals are transcribed either in Basque or in Spanish depending on the context, using their most common realization, though sometimes it might not match the actual pronunciation. For instance, the phrase '1.5 millones' is transcribed as 'uno coma cinco millones' while the speaker might have actually said 'uno punto cinco millones' or even 'un millón y medio'. By default, acronyms are written in all-caps and assumed to be spelled, with exceptions being listed in an acronym pronunciation dictionary.

2.3. Language model

A language model was built based on BP minutes and their translations, after text normalization, which involves converting numbers and ordinals into their alphabetical counterparts, putting all words (except for acronyms) in lower case, etc. It must be noted that sentences were considered atomic units in BP texts 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 but positive probability that a word in Basque comes after a word in Spanish (and viceversa).

3. Semisupervised data collection

For each BP plenary session, we have an audio file and the corresponding minutes, with an approximate transcription of the audio contents. In fact, for ease of processing, the audio file is manually split into two or three smaller chunks (each about 2 hours 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 get 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 get 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 minimizing the number of errors (deletions, insertions and substitutions), following the same text-to-speech alignment method that has been succesfully applied in our group for the alignment of subtitles [22–24].

Gaps (silences) longer than 0.5 seconds define potential breaking points. Audio chunks between two breaking points will be called *segments*. Data collection is performed by recursively searching for the segment lasting between 3 and 10 seconds with the highest phone recognition rate (PRR). When two or more segments have the same (maximum) PRR, the longest segment is chosen. Selecting a segment also means splitting the audio chunk from which it was extracted into two new chunks, which will be jointly but independently searched in subsequent steps. The process iterates until no valid segment is left. In this way, we end up with an ordered list of segments, ranked according to the alignment error rate, so that at the top of the list we could find segments for which the alignment error is zero, meaning that the phone recognizer output matched the nominal transcription provided by the minutes. The phone recognition rate (%PRR) used as criterion is defined as:

$$\%\text{PRR} = 100 \cdot \frac{m}{m + d + i + s} \tag{1}$$

where m, d, i and s are the number of matching units, deletions, insertions and substitutions yielded by the alignment, respectively.

Once the ranking is obtained, a new set of acoustic models can be trained by using only the top ranking segments, that is, those segments for which the provided transcription best matches the speech contents (%PRR being higher than a given threshold). The resulting models can be then applied again to perform phone recognition, get new alignments and hopefully a better set of segments for training. This process could be repeated until ASR performance on a validation set did not improve.

4. Experimental setup

The acoustic models for the initial phone recognizer were trained on generic speech databases in Basque and Spanish: CommonVoice (cv-corpus-5.1-2020-06-22) [25], OpenSLR (SLR76) [26], Aditu [27] and Albayzin [28] (see Table 2). The development and test sets of Aditu and Albayzin were used to validate and evaluate phone recognition performance, respectively. The training, development and test sets have durations of 332.21, 3.96 and 4.03 hours, respectively. Note, however, that Spanish and Basque are highly imbalanced in the training set (with a 3:1 ratio).

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 [29]. 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 1117 hours) are used.

The ASR system is also based on wav2letter++. The acoustic models obtained in the semisupervised collection procedure are used but in this case lexical and language models are also computed, based on the minutes (and their translations) of all the BP plenary sessions from 2010 to 2021 (except for the session used to extract the dev and eval data), which comprise more than 33 million words and around 279000 different words. For each word in the vocabulary, a single pronunciation baseform was considered, as provided by our in-house G2P converter. A trigram language model was computed using KenLM [30] (without pruning), including 15.77 million trigrams. The development and test datasets, each around 2 hours long, were extracted from a single BP plenary session which took place in April 2013 and were manually audited, segmented and transcribed. The development dataset is used to measure the improvement of acoustic models during the semisupervised collection procedure and to optimize the hyperparameters of the ASR system, whereas the test set is used just to measure the performance of the ASR system at the end of the process (not for tuning).

Table 2: *Databases used to train the acoustic models of the initial phone recognizer (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 (hours)	77.81	254.40

Three wav2letter++ hyperparameters were found to be critically important for ASR performance: (1) *lmweight*: the language model weight which is accumulated with the acoustic model score; (2) *wordscore*: the score (penalty) added when appending a word to the output; and (3) *silscore*: the silence score (penalty) added whenever a silence unit is appended to the output. A random walk search of these hyperparameters was performed around the default values to optimize ASR performance on the development set. Then the optimal hyperparameters were used when processing the test set.

5. Results

The first part of this work involved the training of baseline acoustic models that were integrated in 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 3 shows the amount of speech that would be collected by applying different thresholds to the list of segments sorted according to their %PRR. 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 initial (baseline) acoustic models were also used to run word recognition experiments using the wav2letter++ ASR system described in Section 4. Table 4 shows the Word Error Rate (WER) and Letter Error Rate (LER) performance obtained by the baseline models on the dev and test sets of the BP database, disaggregated per language. Remind that wav2letter++ hyperparameters were optimized on the development set and then applied to the test set. Two important observations can be made about the dev and test sets: (1) there are remarkably more segments (speech) in Spanish than in Basque (with roughly a 2:1 ratio), which is quite common in the BP plenary sessions; and (2) while WER is almost the same for both languages in the test set, it is much worse for Basque in the dev set, due to one of the

Table 3: *Amount of speech (in hours) accumulated by keeping those segments with a %PRR $\geq$ Threshold.*

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

speakers being highly disfluent and introducing a lot of spontaneous speech events (filled pauses, incomplete words, false starts, repetitions, etc.) which makes the task more difficult. We are now auditing more materials to increase the size of the dev and eval sets, which hopefully will reduce this kind of variability. Leaving this latter issue aside, the performance of the baseline models is quite similar for both sets, and the hyperparameter tuning done on the dev set seems to work 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.

Table 4: *WER and LER performance of the baseline acoustic models (trained on generic speech datasets) on the dev and test sets of the BP database.*

Set	Language	#Segments	WER	LER
Dev	All	810	16.46	7.01
	Basque	280	20.25	7.58
	Spanish	530	15.40	6.83
Test	All	860	15.06	6.15
	Basque	267	15.08	5.14
	Spanish	593	15.08	6.43

The next step of the process consisted of using the speech segments of the BP plenary sessions with %PRR $\geq$ 80, which amount to more than 1000 hours, to train a new set of acoustic models. Note that the baseline acoustic models were trained on around 332 hours obtained from different and heterogeneous sources, which had nothing to do with BP sessions. Now we are about to use 3 times more training data extracted from BP sessions, under the same acoustic conditions and probably including some of the speakers of the dev and test sets on which we will evaluate ASR performance. This was, in fact, one of our main objectives: taking advantage of the speech available from BP sessions to improve the performance of our ASR system when dealing with BP speech. Table 5 shows the WER and LER performance obtained by the new acoustic models on the dev and test sets of the BP database, disaggregated per language. As may be expected, the improvement is huge, from 16.46% to 6.99% WER on the dev set and from 15.06% to 5.16% WER on the test set (meaning 57.5% and 65.7% relative error reductions, respectively). The performance is remarkably worse for Basque in terms of WER, even for the test set in this case. These differences can be explained by two causes: (1) training materials for Spanish are about twice the size of training materials for Basque; and (2) the dev and test sets are only 2 hour long each, so small sample effects could be happening (as in the case commented above).

Table 5: *WER and LER performance on the dev and test sets of the BP database, for acoustic models trained on 1000 hours of BP speech obtained after one iteration of semisupervised validation.*

Set	Language	#Segments	WER	LER
Dev	All	810	6.99	3.02
	Basque	280	10.98	4.02
	Spanish	530	5.84	2.66
Test	All	860	5.16	2.33
	Basque	267	7.47	2.59
	Spanish	593	4.68	2.26

In any case, it is remarkable the high ASR performance attained in this task by a fully bilingual ASR system with a single set of acoustic models and a single aggregated language model, which allows to deal in a natural and computationally efficient fashion with code switching events. The semisupervised data validation procedure was applied a second time, and the top ranking segments were used to estimate a second round of acoustic models (with the same threshold of %PRR = 80), but these second-round models did not outperform the ones obtained in the first round. This might indicate that a single round is enough to get reasonable good segments, or that further research is needed to get WER reductions (using different hyperparameters, thresholds, etc.). It is interesting to note the high tolerance of wav2letter++ to inaccuracies in reference transcriptions. When a segment with a low PRR (meaning that the speech in that segment does not match the reference transcription) enters the training set, neural networks learn to map the acoustic observations of that segment to the reference text. This means that, after training, the PRR for that segment will become closer to 1, making us erroneously think that the reference transcription of that segment matches its acoustic contents. On the other hand, this behaviour helps to deal with spontaneous speech events such as filled pauses, repetitions, false starts, etc. which are common in conversational speech but do not appear in reference transcriptions, and thus are never output by the model.

6. Conclusions

In this paper, we have presented a fully bilingual ASR system for Basque and Spanish and a semi-supervised data validation procedure which leverages the speech and the minutes of plenary sessions of the Basque Parliament to train domain-adapted models, which lead to a remarkable improvement in performance (65.7% relative WER reduction on the test set) with regard to the baseline system. Further research is needed to check whether the semi-supervised method can be effective in successive iterations. Also, a study should be carried out to determine which threshold performs the best when selecting the highest ranked segments for training.

7. Acknowledgements

This work was partially funded by the Spanish Ministry of Science and Innovation (OPEN-SPEECH project, PID2019-106424RB-I00) and by the Basque Government (IT-1355-19).

8. References

[1] P. Gardner-Chloros, *Code-switching*. Cambridge University Press, 2009.

[2] E. Yılmaz, M. McLaren, H. van den Heuvel, and D. A. van Leeuwen, "Semi-supervised acoustic model training for speech with code-switching," *Speech Communication*, vol. 105, pp. 12–22, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639318300190

[3] S. Dalmia, Y. Liu, S. Ronanki, and K. Kirchhoff, "Transformer-transducers for code-switched speech recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5859–5863.

[4] A. Biswas, E. Yılmaz, E. van der Westhuizen, F. de Wet, and T. Niesler, "Code-switched automatic speech recognition in five South African languages," *Computer Speech and Language*, vol. 71, p. 101262, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S088523082100067X

[5] A. Alvarez, H. Arzelus, S. Prieto, and A. del Pozo, "Rich Transcription and Automatic Subtitling for Basque and Spanish," in *Proceedings of Iberspeech 2016*, Lisbon, Portugal, November 2016, pp. 197–206.

[6] E. Yılmaz, H. van den Heuvel, and D. van Leeuwen, "Code-switching detection using multilingual DNNs," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 610–616.

[7] E. Yilmaz, M. McLaren, H. van den Heuvel, and D. A. van Leeuwen, "Language diarization for semi-supervised bilingual acoustic model training," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 91–96.

[8] H. Seki, S. Watanabe, T. Hori, J. L. Roux, and J. R. Hershey, "An end-to-end language-tracking speech recognizer for mixed-language speech," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4919–4923.

[9] Z. Zeng, Y. Khassanov, V. T. Pham, H. Xu, E. S. Chng, and H. Li, "On the End-to-End Solution to Mandarin-English Code-Switching Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2165–2169.

[10] Z. Qiu, Y. Li, X. Li, F. Metze, and W. M. Campbell, "Towards Context-Aware End-to-End Code-Switching Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 4776–4780.

[11] X. Shi, Q. Feng, and L. Xie, "The ASRU 2019 Mandarin-English Code-Switching Speech Recognition Challenge: Open Datasets, Tracks, Methods and Results," *CoRR*, vol. abs/2007.05916, 2020. [Online]. Available: https://arxiv.org/abs/2007.05916

[12] C. Li, S. Deng, Y. Wang, G. Wang, Y. Gong, C. Chen, and J. Bai, "TALCS: An Open-Source Mandarin-English Code-Switching Corpus anda Speech Recognition Baseline," *CoRR (accepted at Interspeech 2022)*, vol. abs/2206.13135, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2206.13135

[13] L. Lamel, J. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Comput. Speech Lang.*, vol. 16, no. 1, pp. 115–129, 2002. [Online]. Available: https://doi.org/10.1006/csla.2001.0186

[14] F. Wessel and H. Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 23–31, Jan. 2005.

[15] D. Yu, B. Varadarajan, L. Deng, and A. Acero, "Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion," *Computer Speech and Language*, vol. 24, no. 3, pp. 433–444, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230809000187

[16] H. Liao, E. McDermott, and A. Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription," in *ASRU*, 2013.

[17] K. Veselý, M. Hannemann, and L. Burget, "Semi-supervised training of deep neural networks," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, December 8-12, 2013*. IEEE, 2013, pp. 267–272. [Online]. Available: https://doi.org/10.1109/ASRU.2013.6707741

[18] V. Manohar, H. Hadian, D. Povey, and S. Khudanpur, "Semi-supervised training of acoustic models using lattice-free MMI," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*. IEEE, 2018, pp. 4844–4848. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8462331

[19] Y. Long, Y. Li, S. Wei, Q. Zhang, and C. Yang, "Large-Scale Semi-Supervised Training in Deep Learning Acoustic Model for ASR," *IEEE Access*, vol. 7, pp. 133 615–133 627, 2019.

[20] S. Wotherspoon, W. Hartmann, M. Snover, and O. Kimball, "Improved data selection for domain adaptation in ASR," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7018–7022.

[21] T. Etchegoyhen, H. Arzelus, H. Gete Ugarte, A. Alvarez, A. González-Docasal, and E. Benites Fernandez, "Mintzai-ST: Corpus and Baselines for Basque-Spanish Speech Translation," in *Proc. IberSPEECH 2021*, 2021, pp. 190–194. [Online]. Available: http://dx.doi.org/10.21437/IberSPEECH.2021-41

[22] G. Bordel, S. Nieto, M. Penagarikano, L. J. Rodriguez-Fuentes, and A. Varona, "Automatic Subtitling of the Basque Parliament Plenary Sessions Videos," in *Interspeech 2011*, Florence, Italy, 28-31 August 2011.

[23] G. Bordel, M. Penagarikano, L. J. Rodriguez-Fuentes, and A. Varona, "Aligning very long speech signals to bilingual transcriptions of parliamentary sessions," in *Iberspeech 2012*, Madrid, Spain, November 21-23 2012.

[24] G. Bordel, M. Penagarikano, L. J. Rodriguez-Fuentes, A. Álvarez, and A. Varona, "Probabilistic kernels for improved text-to-speech alignment in long audio tracks," *IEEE Signal Processing Letters*, vol. 23, no. 1, pp. 126–129, january 2016.

[25] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common Voice: A Massively-Multilingual Speech Corpus," *CoRR*, vol. abs/1912.06670, 2019. [Online]. Available: http://arxiv.org/abs/1912.06670

[26] O. Kjartansson, A. Gutkin, A. Butryna, I. Demirsahin, and C. Rivera, "Open-Source High Quality Speech Datasets for Basque, Catalan and Galician," in *Proceedings of the 1st Joint Workshop on SLTU and CCURL*, Marseille, France, May 2020, pp. 21–27.

[27] I. Odriozola, I. Hernaez, M. Torres, L. J. Rodriguez-Fuentes, M. Penagarikano, and E. Navas, "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: European Language Resources Association (ELRA), May 26-31 2014, pp. 2658–2665.

[28] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. B. Marino, and C. Nadeu, "Albayzin speech database: design of the phonetic corpus," in *Proc. 3rd European Conference on Speech Communication and Technology (Eurospeech 1993)*, 1993, pp. 175–178.

[29] R. Collobert, C. Puhrsch, and G. Synnaeve, "Wav2Letter: an End-to-End ConvNet-based Speech Recognition System," *CoRR*, vol. abs/1609.03193, 2016. [Online]. Available: http://arxiv.org/abs/1609.03193

[30] K. Heafield, "KenLM: Faster and smaller language model queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, Jul. 2011, pp. 187–197. [Online]. Available: https://aclanthology.org/W11-2123