applied sciences

MDPI

Article

Cascade or Direct Speech Translation? A Case Study †

Thierry Etchegoyhen [1,*,‡], Haritz Arzelus [1,‡], Harritxu Gete [1,2,‡], Aitor Alvarez [1,‡], Iván G. Torre [1], Juan Manuel Martín-Doñas [1], Ander González-Docasal [1,3] and Edson Benites Fernandez [4]

[1] Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), 20009 Donostia-San Sebastían, Spain; harzelus@vicomtech.org (H.A.); hgete@vicomtech.org (H.G.); aalvarez@vicomtech.org (A.A.); igonzalez@vicomtech.org (I.G.T.); jmmartin@vicomtech.org (J.M.M.-D.); 814300@unizar.es (A.G.-D.)
[2] Faculty of Informatics, University of the Basque Country, 20018 Donostia-San Sebastián, Spain
[3] School of Engineering and Architecture, University of Zaragoza, 50018 Zaragoza, Spain
[4] Work Done While at Vicomtech, 20009 Donostia-San Sebastían, Spain; benites.ee@gmail.com
* Correspondence: tetchegoyhen@vicomtech.org

† This paper is an extended version of a conference paper published in IberSPEECH2020, Valladolid, Spain, 24–25 March 2021.
‡ These authors contributed equally to this work.

Abstract: Speech translation has been traditionally tackled under a cascade approach, chaining speech recognition and machine translation components to translate from an audio source in a given language into text or speech in a target language. Leveraging on deep learning approaches to natural language processing, recent studies have explored the potential of direct end-to-end neural modelling to perform the speech translation task. Though several benefits may come from end-to-end modelling, such as a reduction in latency and error propagation, the comparative merits of each approach still deserve detailed evaluations and analyses. In this work, we compared state-of-the-art cascade and direct approaches on the under-resourced Basque–Spanish language pair, which features challenging phenomena such as marked differences in morphology and word order. This case study thus complements other studies in the field, which mostly revolve around the English language. We describe and analysed in detail the mintzai-ST corpus, prepared from the sessions of the Basque Parliament, and evaluated the strengths and limitations of cascade and direct speech translation models trained on this corpus, with variants exploiting additional data as well. Our results indicated that, despite significant progress with end-to-end models, which may outperform alternatives in some cases in terms of automated metrics, a cascade approach proved optimal overall in our experiments and manual evaluations.

Keywords: speech translation; Basque; Spanish; corpus; cascade speech translation; direct speech translation

check for updates

Citation: Etchegoyhen, T.; Arzelus, H.; Gete, H.; Alvarez, A.; Torre, I.G.; Martín-Doñas, J.M.; González-Docasal, A.; Fernandez, E.B. Cascade or Direct Speech Translation? A Case Study. *Appl. Sci.* **2022**, *12*, 1097. https://doi.org/10.3390/app12031097

Academic Editors: António Joaquim da Silva Teixeira, Francesc Alías, Valentin Cardeñoso-Payo, David Escudero-Mancebo and César González-Ferreras

Received: 23 December 2021
Accepted: 16 January 2022
Published: 21 January 2022

1. Introduction

Speech translation (ST) systems have been traditionally designed under a cascade approach, where independent automatic speech recognition (ASR) and machine translation (MT) components are chained, feeding the ASR output into the MT component, oftentimes with task-specific bridging to optimise component communication [1–3]. Although this approach has been the dominant paradigm in the field, cascade speech translation is prone to error propagation, requires the assembly of separate components, and cumulates the latency of its two main components.

With the advent of deep learning approaches, significant results have been achieved with end-to-end neural models, notably in the fields of machine translation [4,5] and speech recognition [6,7], building on the ability of neural models to jointly learn different aspects of a task. Following the success of these approaches, end-to-end modelling has also been proposed for the speech translation task [8–10]. In what follows, we use the terms *direct* and *end-to-end* interchangeably to denote the process of modelling speech translation with a neural network trained on speech input in the source language and text output in the target language.

While preliminary results with end-to-end systems had shown promise, the initial cascade systems still obtained better results overall on standard evaluation datasets [11]. One of the main factors for this state of affairs has been the paucity of parallel speech–text corpora, in contrast with the comparatively larger amounts of data available to train separate ASR and MT models, for some language pairs at least. Recent efforts have been made to build parallel corpora suitable for the task, notably the multilingual MuST-C [12] and Europarl-ST [13] corpora. As most available data are built around English, which limits experimental variety, [14] made available the mintzai-ST corpus for Basque–Spanish. This corpus supports further experimentation on languages with a number of marked linguistic properties, such as Basque, notably rich morphology and relatively free word order, which can represent a challenge for natural language processing tasks in general and automated translation in particular [15,16].

With newly available ST corpora supporting research and development in the field, recent variants of direct ST models have closed the quality gap with cascade approaches [17–22], in terms of automated metrics or manual evaluations on standard datasets [23,24]. Nevertheless, additional work is still needed to identify the strengths and weaknesses of end-to-end approaches to the task, and comparative results may fluctuate on standard datasets [25].

In this paper, we extend the work of [14] on Basque–Spanish ST and address in more details the comparative merits of cascade and direct approaches on a relatively difficult language pair and dataset. We first explored the characteristics of the mintzai-ST Basque–Spanish corpus, to provide a complete description of the data and present results for the baseline cascade and direct models trained on this corpus. We then describe additional end-to-end variants, which bridge the gap between the two approaches in terms of automated metrics. Finally, we describe the results of manual evaluations comparing the cascade and direct models.

The remainder of this paper is organised as follows: Section 2 presents related work in the field; in Section 3, we describe the mintzai-ST corpus, including the data acquisition process and data statistics; Section 4 describes the different baseline models that were built for Basque–Spanish speech translation, including cascade and end-to-end models; Section 5 discusses comparative results for the baseline models; in Section 6, we describe several direct ST model variants and their results on automated metrics; Section 7 describes the protocol and results of our manual evaluation of the best cascade and end-to-end models, along with the results of targeted evaluations on specific linguistic phenomena and on the impact of relative input difficulty; finally, Section 8 draws the main conclusions from this work.

2. Related Work

Standard speech-to-text translation systems operate on the basis of separate components for speech recognition and machine translation, feeding the output of the ASR module into the MT component. Chaining these components can be performed by simply selecting the best recognition hypothesis as the input to machine translation, as was often performed with earlier systems via interlingual-based representations [26,27]. To optimise cascade processing, other alternatives have been explored over the years, notably via the exploitation of multiple ASR hypotheses [1,2] or the adoption of statistical methods and finite-state automata, integrating acoustic and translation models within stochastic transducers [28,29]. The issue of error propagation has been also addressed by improving the ASR component [30], the robustness of the MT component with synthetic ASR recognition errors [31], or the use of specific features to improve communication between components [3].

An alternative approach to the speech translation process has focused on performing direct ST via end-to-end artificial neural networks. The first results were obtained with encoder–decoder models coupled with attentions mechanisms [8–10]. Although most studies have focused on speech-to-text, end-to-end architectures have also been explored for speech-to-speech translation [10,32]. Despite promising initial results, which demonstrated the ability of neural networks to model the ST task in an end-to-end fashion, cascade systems tended to outperform end-to-end systems on standard datasets [11]. One of the main reasons for this state of affairs was training data scarcity, i.e., the lack of sufficiently large speech-to-text datasets to train direct ST systems, in contrast with the comparatively larger training data for the ASR and MT components, considered separately. Another relevant factor was the need to improve end-to-end ST architectures or training methods for this type of approach.

The first issue has been partially tackled in recent years, with the preparation and distribution of additional ST datasets. Thus, Reference [33] built and shared a corpus based on 236 h of English speech, from the Librispeech library, aligned with French translations. This corpus was exploited by [34] to train an end-to-end ST model whose performance closed the gap with that of a conventional cascade model. Another important contribution in terms of ST datasets was the release of MuST-C [12], based on translated TED talks from English into eight languages, with audio recordings ranging from 385–504 h. An additional multilingual ST corpus, CoVoST [35,36], provided coverage for translation from 21 languages into English and from English into 15 languages, with audio recording lengths between 1 h and 364 h.

To extend the coverage of language pairs beyond English, Reference [13] released the Europarl-ST corpus, prepared from publicly available videos of the European Parliament, covering six languages (English, French, German, Spanish, Portuguese, and Italian) and thirty translation pairs, with volumes of data ranging from 20–89 h of audio recordings. Supporting multilingual ST beyond English also is the TEDx corpus [37], which covers eight source languages associated with a subset of target languages, with audio recordings ranging from 11–69 h. In [14], a corpus was prepared to address the under-resourced Basque–Spanish language pair, with 180 h for Basque and 468 h for Spanish; a detailed description of this corpus is provided in Section 3. Synthetic data have also been exploited for speech translation, for instance by leveraging high-quality MT models on ASR datasets and speech synthesis models on MT corpora [38].

The second issue for direct ST approaches, namely the improvement of architectural design and training methods, has been addressed via different techniques [19]. Most current frameworks for ST adapt the Transformer architecture [5] to the task, mainly via an adaptation of encoder layers to handle the specificities of the audio input, whose usually large number of frames raises issues given the quadratic complexity of the Self-Attention (SA) mechanism. Thus, Reference [39] proposed a Transformer model, which notably included a 2D attention mechanism to jointly attend to the time and frequency axes of two-dimensional speech spectra input, while [17,40] included convolutional layers alongside 2D SA layers. Recently, an alternative framework was proposed by [22], with decoupled triple supervision signals for acoustic, semantic, and translation processing, improving over the state-of-the-art.

In addition to architectural variant proposals, an important contribution was made by [34], who demonstrated the usefulness of pretraining the encoder and decoder of their end-to-end model on the ASR and MT tasks, respectively, partially closing the gap with corresponding cascade systems. Their work also demonstrated the positive impact for speech-to-text translation of multi-task training, introduced for end-to-end speech-to-speech translation by [10]. Leveraging additional resources has led to significant increases in direct ST quality, with approaches such as the application of knowledge distillation using an MT model as the teacher for direct ST training [18] or the use of meta-learning for knowledge transfer from the ASR and MT tasks to the ST task [41].

Recently, the use of self-supervised models trained on unlabelled speech, such as Wav2Vec [42,43], has demonstrated its potential for the ST task [20,44] and offers an interesting alternative considering the lack of labelled data, which hinders research and development in the field. Combining the use of this type of pretrained speech encoder with an additional pretrained text decoder, via minimalist fine-tuning, has been shown to provide further gains on the multilingual ST task [21]. In Section 6, we describe the impact of pretrained speech models, along with several representative ST variants, on the mintzai-ST datasets.

Recent improvements in ST modelling have closed the gap between direct and cascade approaches on standard datasets. Thus, whereas the latter outperformed the former in the IWSLT 2019 shared task, results from the 2020 edition featured similar performances overall [23]. However, results on the 2021 edition of the shared task have again placed cascade ST as the top-performing approach [25]. Alongside these results, Reference [24] presented an in-depth comparative study of the two main approaches, in three translation directions, via both automatic and manual evaluations based on professional post-editing and annotation. They concluded that, for the language pairs and datasets in their study at least, the gap between the two approaches can be considered closed, as subtle differences between the two are not sufficient for human evaluators to establish a preference. In Section 7, we describe a comparative manual evaluation over several cascade and direct ST model variants, with results that diverge from their conclusions.

3. The mintzai-ST Corpus

The mintzai-ST corpus [14] was created from the proceedings of the Basque Parliament, which provides publicly available video files, along with professional transcriptions and translations, from the parliamentary sessions. The corpus is shared under the Creative Commons CC BY-NC-ND 4.0 license and is available at the following address: https://github.com/vicomtech/mintzai-st (accessed on 11 March 2021). The providers of the original content have granted permission for its use without additional restrictions.) Speakers expressed themselves in either Basque or Spanish, with a majority of interventions in the latter language overall; professional transcriptions and translations were then produced in the other language. In this section, we describe the main processes related to the data acquisition and preparation of the corpus and detail its characteristics.

3.1. Data Acquisition

Raw data were first obtained by crawling the web sites where the official plenary sessions are made available: transcriptions and translations were available at http://www.legebiltzarra.eus, (accessed on 12 November 2019); videos of the sessions at: https://www.irekia.euskadi.eus, (accessed on 12 November 2019).

Texts from the sessions were available as bilingual PDF files, with content in each language provided in a dedicated column: one for the transcription of the session and the other for its translation. Transcriptions are not literal in this corpus, in the sense that repeated or wrongly pronounced words, dialectal variants, and fillers are usually not transcribed. The content was extracted from the PDF files with PDFtoText, which preserved column-based alignments, and boilerplate removal was performed with in-house content-specific scripts. Since the translations were made at the paragraph level for the most part [15], paragraph-level information was kept.

Videos were provided in different formats over the years (.flv, .webm and .mp4), and audio extraction was performed with FFmpeg (https://www.ffmpeg.org/ (accessed on 15 Novemeber 2018).The mapping between videos and reports was performed via inferences from the respective files' metadata whenever possible. For each session, there were between one and seven videos and one or two PDF files. In most cases, the available information was not sufficient to map video and PDF files with absolute confidence, with multiple and sometimes duplicate videos. Therefore, manual revision and mapping were performed throughout this task. The statistics for the collected raw data are shown in Table 1.

Table 1. Collected raw data statistics.

Year	Videos	PDF	Hours	Words
2011	43	21	86.51	132,595
2012	38	21	117.94	173,199
2013	67	38	215.00	306,621
2014	60	30	176.83	252,887
2015	41	27	134.10	195,112
2016	38	21	113.85	170,608
2017	49	33	173.57	250,862
2018	34	26	128.38	207,910
Total	370	217	1146.18	18,625,252

3.2. Alignment and Filtering

As a first step, metadata were filtered from the PDF-extracted text, and source and target files were extracted from the text in the original columns, preserving paragraph-level alignments. Speaker information was usually located at the beginning of a paragraph and was extracted automatically when available. Since sometimes speakers were identified by their function rather than their name, manual speaker identification was also performed (see Section 3.3 for details on speaker distribution in the final datasets).

As a second step, language identification was performed on each paragraph. Although paragraphs were consistently in either one of the two languages overall, in some cases, language switching occurred within a given paragraph or within a sentence. Since any error identifying languages would propagate to subsequent processes, special attention was paid to ensuring correct language identification by employing two separate tools on the content: TextCat and the language identifier of the OpenNER project [45] (available at the following addresses, respectively: https://github.com/Trey314159/TextCat, accessed on 31 January 2017, and https://github.com/opener-project/language-identifier, accessed on 31 July 2014). Paragraphs were discarded if either tool produced different results as their topmost identified language or if neither tool identified either one of the expected languages; in all other cases, we retained the language identified as most probable.

The third step involved forced alignment, where each word in the source transcription was aligned to a section of the corresponding audio file via source and time indications. This step was performed with the Kaldi toolkit [46], using an in-house bilingual model to reduce the impact of remaining language identification uncertainties. Forced alignment was performed with different beam sizes, to capture the largest possible number of alignments in a corpus where audio transcriptions were produced by human experts, and therefore assumed to be complete and correct for the most part. The forced alignment results are shown in Table 2.

Overall, more recent years featured more literal transcriptions, requiring smaller beams to perform alignment. Increasingly larger beams were applied to preserve as much of the content as possible in this initial alignment step, under the assumption that transcriptions of the audio content had been professionally created and were thus correct for the most part.

At this stage, the source and target files were split on the basis of the previous alignment information, with one paragraph per file. Forced alignment was then applied again, this time with a monolingual model and a small beam size of one, along with a retry beam of two, to discard alignment errors and non-literal transcriptions. This process resulted in the perfect and imperfect alignments reported in Table 3, with the former being kept for the final corpus and the latter discarded.

Table 2. Number of aligned audio and transcription files via forced alignment with different beam sizes.

Year	Alignment Beam			
	10	100	1000	10,000
2011	2	51	32	0
2012	1	72	44	1
2013	1	186	7	3
2014	0	136	28	1
2015	0	23	98	1
2016	3	47	55	1
2017	1	151	5	1
2018	0	111	4	0
All	8	776	273	8

Table 3. Forced alignment results at the paragraph level. PARA indicates the number of paragraphs, % speech the percentage of identified speech in the audio, and % perfect the percentage of perfectly aligned content within identified speech.

Year	Duration	Aligned		Discarded		% Speech	% Perfect
	(Hours)	PARA	Hours	PARA	Hours		
2011	86.51	9596	49.81	4572	31.75	94.26	61.07
2012	117.94	13,991	71.49	5547	38.92	93.62	64.75
2013	215.00	24,989	146.60	7788	55.20	93.86	72.65
2014	176.83	19,523	113.44	7093	52.34	93.75	68.43
2015	134.10	13,167	80.15	6004	46.13	94.17	63.47
2016	113.85	11,418	65.78	5372	41.31	94.07	61.42
2017	173.57	19,129	109.44	7474	54.77	94.60	66.65
2018	128.38	16,525	84.69	5711	36.85	94.67	69.68
All	1146.20	128,338	721.40	49,561	357.27	94.11	66.88

Since translation models require specific sentence-based training bitexts, the previously aligned paragraphs were further prepared with sentence splitting, tokenisation, and truecasing. All operations were performed with the scripts from the Moses toolkit [47]. Sentence-level alignments were then computed with the Hunalign toolkit [48], with an alignment probability of 0.4. Table 4 summarises the volumes of data after the alignment and filtering steps on speech data, as described above.

The filtered and aligned data were then randomly split into training, dev, and test subsets of triplets consisting of audio, transcription, and translation. Triplets were removed from the test sets if the transcription–translation pair appeared in the training set as well. This measure was adopted to account for the fact that even minor acoustic differences might make a triple differ from another, even though the transcription–translation pair would be a duplicate for the machine translation component. Stricter removal along the previous lines allowed for a fair comparison between cascade and end-to-end models and made for a more difficult test set, as it mainly discarded acoustic variants of greetings and salutations.

Table 4. Filtered data statistics after each alignment step. Speech indicates the figures for identified speech in the original content; Paragraph indicates the figures after alignment at the paragraph level; Sentence indicates the figures after alignment at the sentence level.

Alignment	SRC	TGT	SRC Hours	SRC Words	TGT Words
Speech	ES	EU	757.31	7,118,590	5,169,324
Paragraph	ES	EU	506.24	4,649,223	3,416,794
Sentence	ES	EU	478.65	4,568,911	3,377,910
Speech	EU	ES	321.36	1,822,122	2,402,048
Paragraph	EU	ES	215.16	1,224,609	1,641,571
Sentence	EU	ES	189.54	1,192,130	1,599,257

Table 5 shows the final statistics of the mintzai-ST corpus. With approximately 191 h for Basque (hereafter, EU) to Spanish (hereafter, ES) and 468 h for Spanish to Basque, the corpus is among the largest available ones for speech translation, featuring a language pair that differs significantly from the ones in publicly available corpora.

Table 5. mintzai-ST: final corpus statistics.

SRC	TGT	Partition	Hours	Sentences	SRC Words	TGT Words
ES	EU	TRAIN	468.16	175,826	4,512,294	3,328,172
EU	ES	TRAIN	180.96	85,409	1,149,803	1,536,695
ES	EU	DEV	2.60	1000	25,359	18,566
EU	ES	DEV	2.23	1000	13,831	18,673
ES	EU	TEST	7.89	2788	74,758	55,283
EU	ES	TEST	6.35	2300	37,706	51,003

3.3. Data Distribution

In this section, we provide additional characteristics of the final corpus in terms of data distribution. As shown in the boxplots of Figure 1, the distribution in terms of audio length was similar across the partitioned datasets, with slightly larger duration on average for the test set. In all three datasets, the data distribution was positively skewed, with audio larger than the median being more represented overall.

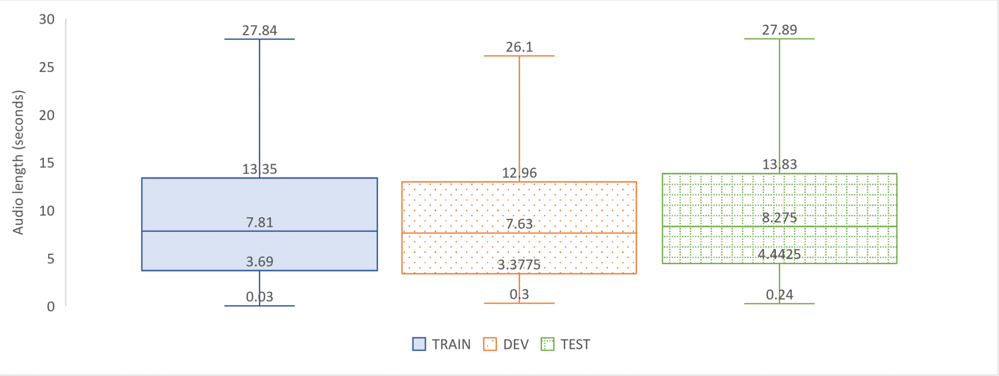

Figure 1. Data distribution in terms of audio length across the training, development, and test partitions for the Basque and Spanish data in the mintzai-ST corpus.

Table 6 lists the number of outliers across partitions, i.e., samples whose duration is above the maximums indicated in Figure 1. Although not all outliers in this sense indicate alignment errors or similar issues, data points above 100 s in the training set may be considered as such and safely discarded at training time, considering the distribution.

Table 6. Outlier distribution per duration range (seconds). MAX indicates the maximum value excluding outliers as computed for each partition.

Range	Train	Dev	Test
300–400	1	0	0
200–250	1	0	0
150–200	1	0	0
100–150	7	0	0
50–100	288	1	5
MAX–50	5058	37	83

Finally, Table 7 indicates the total number of speakers for each dataset and language, along with the median speaker time. Overall, the mintzai-ST corpus features a diverse set of speakers, with a distribution in terms of speakers that is relatively similar across partitions.

Table 7. Number of speakers and median speaker times (minutes).

	EU			ES		
	Train	**Dev**	**Test**	**Train**	**Dev**	**Test**
Number of speakers	151	104	127	175	130	155
Median speaker time	18.62	0.43	1.23	58.96	72.02	1.28

4. Baseline Models

In this section, we describe the baseline models and components built to assess the usefulness of the prepared corpus as a basis for speech translation and to evaluate the two main modelling alternatives, namely: cascade models, based on state-of-the art components for speech recognition and machine translation, and end-to-end neural speech translation models. More advanced variants of the latter are explored in Section 6.

4.1. Cascade Models

The speech-to-text cascade models are based on separate components for speech recognition and machine translation, each trained on their own datasets, either on the in-domain mintzai-ST corpus only or on a combination of the corpus with additional data. We describe each component in turn below.

For the additional dataset, we selected publicly available corpora, close to the mintzai-ST domain, to support a straightforward reproduction of our results. For this language pair, only text-based datasets were available with these characteristics, and we selected the OpenDataEuskadi corpus [16], prepared from public translation memories by the translation services of the Basque public administration (the corpus is available at the following address: http://hltshare.fbk.eu/IWSLT2018/OpendataBasqueSpanish.tgz, accessed on 30 October 2018). This corpus is close enough to the mintzai-ST domain to be meaningfully combined and large enough to contribute significantly to different components of the cascade models. The corpus amounts to 963,391 parallel sentences, with 23,413,116 words in Spanish and 17,802,212 in Basque.

To connect the components, the best hypothesis of the ASR model was fed to the MT model, after generating punctuation as described in Section 4.1.1. Although considering alternative hypotheses in the n-best ASR output might provide additional robustness and accuracy to the overall system, we left an evaluation along these lines for future research.

4.1.1. Speech Recognition

Two speech recognition architectures, based on end-to-end models and Kaldi-based systems, were trained and evaluated on the newly compiled corpus.

The end-to-end speech recognition systems were based on the Deep Speech 2 architecture [49] for both languages and were set up with a sequence of two layers of 2D convolutional neural networks followed by five layers of bidirectional GRU layers and a fully connected final layer. The output corresponds to a softmax function, which computes a probability distribution over characters. The input consisted of Mel-scale-based spectrograms, which were dynamically augmented through the SpecAugment technique [50]. The models were trained using only audio lasting less than 40 s, due to training memory constraints, with a learning rate of 0.0001 annealed by a constant factor of 1.08 for a total of 60 training epochs, and Stochastic Gradient Descent (SGD) was used as the optimiser.

The Kaldi recognition systems were built with the nnet3 DNN setup using the so-called chain acoustic model based on a Factorised Time-Delay Neural Network (TDNN-F) [51], which reduces the number of parameters of the network by factorising the weight matrix of each TDNN layer into the product of two low-rank matrices. Our TDNN-F models consisted of 16 TDNN-F layers with an internal cell dimension of 1536, a bottleneck dimension of 160, and a dropout schedule of "0,0@0.2,0.5@0.5,0". The number of training epochs was set to four, with a learning rate of 0.00015 and a minibatch size of sixty-four. The input vector corresponded to a concatenation of 40-dimensional high-resolution MFCC, augmented through speed (using factors of 0.9, 1.0, and 1.1) [52] and volume (with a random factor between 0.125 and two) [53] perturbation techniques, as well as the appended 100-dimensional iVectors.

Language models were five-gram models with modified Kneser–Ney smoothing, estimated with the KenLM toolkit [54], and were used as either a component of Kaldi-based systems or to rescore the end-to-end models' hypotheses.

The capitalisation models were trained on the mintzai-ST corpus with the recasing tools provided by the Moses open-source toolkit [47], using default parameters that included phrases of length one and a trigram KenLM language model.

Finally, the punctuation module consisted of a Bidirectional Recurrent Neural Network (BRNN) model, which takes advantage of Gated Recurrent Units (GRU) as recurrent layers and an attention mechanism to further increase its capacity to identify relevant parts of the context for punctuation decisions. The models were built using the Punctuator2 toolkit [55] and trained on generic text corpora crawled from digital newspapers and acoustic data from the broadcast news domain [56]. The text corpora were composed of 161.8 million words for Spanish and 139.4 million words for Basque, whilst the acoustic corpus included 200 h of annotated audio per language.

4.1.2. Machine Translation

All machine translation models in the experiments reported below were based on the Transformer architecture [5], built with the MarianNMT toolkit [57].

The models consisted of six-layer encoders and decoders and eight attention heads. The Adam optimiser was used with parameters $\alpha = 0.0003$, $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$. The learning rate was set to increase linearly for the first 16,000 training steps and decrease afterwards proportionally to the inverse square root of the corresponding step. The working memory was set to 8000 MB, and the largest mini-batch was automatically chosen for a given sentence length that fit the specified memory. The validation data were evaluated every 5000 steps for models trained on larger out-of-domain datasets and every epoch otherwise; training ended if there was no improvement in perplexity after five consecutive checkpoints. Embeddings were of dimension 512, tied between source and target, and all datasets were segmented with BPE [58], using 30,000 operations.

Translation metrics were computed in terms of the BLEU [59] and ChrF [60] metrics, obtained with the sacreBLEU toolkit [61]. All statistical significance results were computed via paired bootstrap resampling [62].

4.2. End-to-End Baseline Models

End-to-end ST models were trained on the in-domain speech-text corpus, using the Fairseq-ST toolkit (https://github.com/mattiadg/FBK-Fairseq-S, accessed on 23 April 2020),which supports different types of sequence-to-sequence neural models [40]. The variant selected for the experiments was the Transformer model enhanced for ST described in [17], more specifically the variant the authors referred to as the S-Transformer.

The model follows the standard Transformer architecture with six SA encoder and decoder layers, but adds layers prior to the Transformer encoder to model 2D dependencies. The audio input is provided to the model in the form of sequences of Mel filters, encoded first by two CNNs to model 2D-invariant features, followed by two 2D SA layers to model long-range context. The output of the stacked 2D SA layers underwent a linear transformation, followed by a ReLU non-linearity, and was summed with the positional encoding prior to feeding the Transformer encoder.

We diverged from the implementation described in [17] on one important aspect. Character-based decoding was replaced with subword decoding, using the previously described BPE models, as the former faced consistent issues, resulting in subpar performance; an identical setup with subwords produced significantly better results overall. Further exploration of these differences between translation pairs is left for future research.

5. Comparative Baseline Results

We first performed an evaluation centred on cascade models, where a number of variants were prepared based on different ASR approaches or different types and volumes of training data. The variants included:

- ASR models trained with either an End-to-End neural model (E2E) or the Kaldi toolkit (KAL);
- ASR and MT models trained on either In-Domain data only (IND) or on a combination of in-domain and out-of-domain data (ALL), by integrating the OpenDataEuskadi dataset to train the language and casing models for speech recognition and the translation models for the MT component;
- MT models obtained by fine-tuning a model trained on the out-of-domain dataset with the in-domain data, in addition to the models trained on in-domain data only and all available data.

Table 8 shows the results for the cascade variants on the mintzai-ST test sets, in terms of word error rate (WER) and BLEU [59]. All results in the table were computed with ASR output that included punctuation, generated with the previously mentioned punctuation models. To measure the impact of punctuation on the overall process, differences between BLEU scores obtained with and without punctuation, in that order, are also shown in the table (ΔPUNC).

Overall, cascade models trained on all data performed significantly better than their in-domain counterparts, with improvements of up to five and 1.6 BLEU points for EU–ES and ES–EU, respectively. These results were mostly due to improvements obtained on the MT components, as was expected from adding significantly larger amounts of training data to the small in-domain datasets. For the ASR components, the impact in terms of WER was minor, with around 0.3 gains in either language, mainly due to the use of the same data for acoustic modelling in all cases.

Punctuation had a significant impact on the results, with systematic improvements of up to 2.6 and 1.5 BLEU points in EU–ES and ES–EU, respectively. This trend is not entirely surprising, since the translation models were trained on data that included punctuation marks; the impact of punctuation was amplified for models trained on larger amounts of data.

Regarding the overall translation quality, as measured in terms of the BLEU scores at least, the results were in line with or higher than typical results in similar tasks [11]. One explanation for higher marks is the domain specificity of the corpus, with recurrent topics and typical expressions. Nonetheless, the corpus also features challenging characteristics for automated speech translation, such as the use of Basque dialects or the idiosyncratic properties of the two languages at hand.

Table 8. Evaluation results on cascade variants. The best results, indicated in bold, are statistically significant against all other results, for $p < 0.05$, except where indicated as *; differences between top-performing systems in bold were not statistically significant.

LANG	ASR	MT	WER	BLEU	CHRF	ΔPUNC
EU–ES	E2E IND	IND	14.43	28.4	54.9	+1.0
EU–ES	E2E ALL	IND	14.12	28.4	54.8	+0.8
EU–ES	E2E IND	ALL	14.43	33.3	60.7	+2.4
EU–ES	E2E ALL	ALL	14.12	33.4	60.8	+2.3
EU–ES	E2E IND	FT	14.43	32.6	60.6	+2.4
EU–ES	E2E ALL	FT	14.12	33.0	60.8	+2.6
EU–ES	KAL IND	IND	12.07	29.2	55.7	+0.9
EU–ES	KAL ALL	IND	**11.78**	29.4	55.9	+1.1
EU–ES	KAL IND	ALL	12.07	**34.7**	**61.9** *	+2.6
EU–ES	KAL ALL	ALL	**11.78**	**34.7**	**62.0**	+2.7
EU–ES	KAL IND	FT	12.07	33.7	61.5	+2.5
EU–ES	KAL ALL	FT	**11.78**	33.9	61.7 *	+2.6
ES–EU	E2E IND	IND	8.26	20.6	58.7	+1.3
ES–EU	E2E ALL	IND	8.15	20.6	58.6	+1.3
ES–EU	E2E IND	ALL	8.26	22.0	60.6	+1.0
ES–EU	E2E ALL	ALL	8.15	22.0	60.6	+1.1
ES–EU	E2E IND	FT	8.26	21.5	60.2	+1.3
ES–EU	E2E ALL	FT	8.15	21.5	60.1	+1.2
ES–EU	KAL IND	IND	7.23	20.9	59.0	+1.4
ES–EU	KAL ALL	IND	**7.21**	20.9	58.9	+1.3
ES–EU	KAL IND	ALL	7.23	22.5	**61.0**	+1.2
ES–EU	KAL ALL	ALL	**7.21**	**22.7**	**61.1**	+1.5
ES–EU	KAL IND	FT	7.23	21.9	60.6	+1.2
ES–EU	KAL ALL	FT	**7.21**	22.0	60.7	+1.4

From the previous evaluation, we selected the best cascade variants based on either in-domain or all data and compared with the end-to-end speech translation models, in both translation directions. The comparative results on the mintzai-ST test sets are shown in Table 9, where BP indicates the Brevity Penalty computed within the BLEU metric.

Table 9. Results on cascade and end-to-end baseline models. The best results, indicated in bold, are statistically significant against all other results, for $p < 0.05$.

Lang	Model	ASR	MT	WER	BLEU	BP
EU–ES	CAS	IND	IND	12.07	29.2	0.913
EU–ES	CAS	ALL	ALL	**11.78**	**34.7**	0.978
EU–ES	E2E	-	-	-	17.0	1.000
ES–EU	CAS	IND	IND	7.23	20.9	0.954
ES–EU	CAS	ALL	ALL	**7.21**	**22.7**	0.969
ES–EU	E2E	-	-	-	12.9	1.000

The most notable result from this evaluation was the large difference in terms of the BLEU between the cascade and the end-to-end variants under similar conditions, i.e., using only the in-domain data. Under these conditions, the end-to-end variant was outperformed by 12.2 and eight BLEU points in EU–ES and ES–EU, respectively. Since the conditions were similar, with relatively small amounts of training data, this large gap may be attributed to the relative dependency of the baseline end-to-end models on larger volumes of training data. More robust alternative end-to-end ST approaches are explored in the next section.

Interestingly, the end-to-end model produced translations that were closer in length to the human references, as shown by the results in terms of the brevity penalty. Although further analyses of these aspects will be warranted, these results indicate that the end-to-end systems built for these experiments may be modelling aspects of the speech translation process that are not fully captured by their cascade counterparts.

6. Advanced End-to-End Models

The results in the previous section in terms of end-to-end models were based on a basic S-Transformer architecture, applied directly to the source–target mintzai-ST data. Although this provides useful baseline references, more advanced direct models can be devised to better exploit the available data. In this section, we describe the experimental results with variants based on different architectures, Pretraining (PT) and Knowledge Distillation (KD). The different variants are illustrated in Figure 2 and described below.

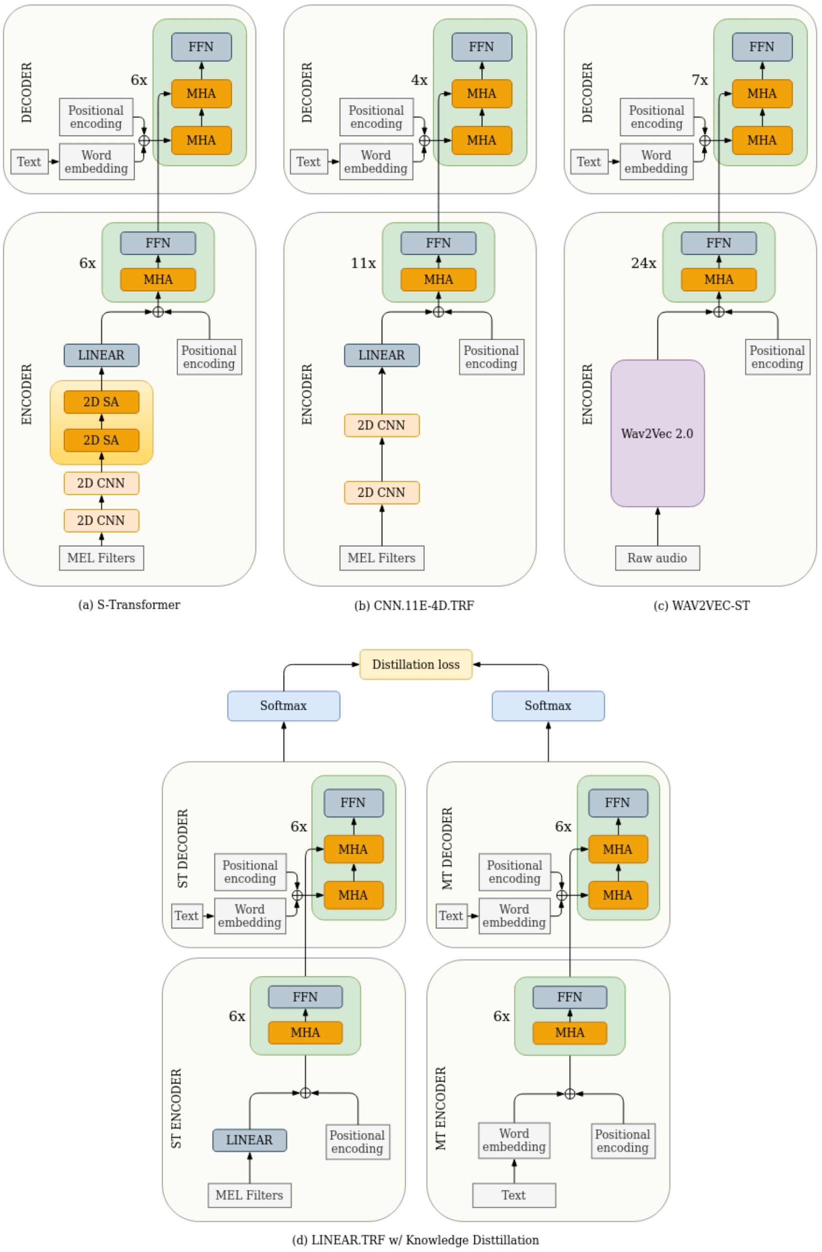

Figure 2. Architectural variants for end-to-end models. FFN denotes the Feed-Forward Networks in a standard Transformer architecture; MHA denotes Multi-Head Attention; CNN denotes convolutional layers.

6.1. Models

As our end-to-end baseline, we used the S-Transformer model [17], denoted as S-TRF in what follows, as described and evaluated in Section 5.

6.1.1. Architectural Variants

The first variant (CNN.11E-4D.TRF) was based on the description in [24] and differs from the S-Transformer architecture by first removing the 2D attention layers and modifying the number of layers, with a stack of eleven SA encoder layers and four SA decoder layers; the model preserves the CNN layers to reduce the dimensionality of the input audio data. This architecture was selected as representative of models that have closed the gap with cascade models in the aforementioned study.

The second variant (LINEAR.TRF) replaces both the CNN and the 2D SA layers of the S-Transformer with a simple linear layer as the first encoder layer, following [18]. This architecture was mainly designed to support knowledge distillation, as it combines a Transformer model for ASR and another one for MT (see Section 6.1.3 for more details). We included it as a variant to provide KD results with the architecture as described in [18], using only the output of the teacher model as it provided the best results in the aforementioned work. We also extended the use of KD to the other ST models described in this section, as it can be straightforwardly applied in all cases.

Finally, we prepared a third variant (WAV2VEC-ST), with WAV2VEC 2.0 [43] set as the encoder in a Transformer-based model [63]. This architecture is of notable interest in its leverage of speech data via self-supervised methods, as one of the main limitations for the development of end-to-end ST models is the lack of parallel audio–text data. This ST model features initial linear and CNN layers, followed by twenty-four SA encoder layers and six SA decoder layers. After preliminary experiments, training on the mintzai-ST corpus was performed by first freezing the encoder layers for the first 20% of the epochs, then allowing updates for all parameters, except the WAV2VEC 2.0 CNN, during the remaining epochs. With this architecture,and contrary to the results obtained with S-Transformer variants, preliminary experiments indicated that character-level decoding outperformed BPE-based decoding. We thus maintained the original configuration and parameters in the FAIRSEQ-ST toolkit used to train the WAV2VEC-ST models, as provided at: https://github.com/pytorch/fairseq/tree/main/examples/speech_to_text (accessed on 18 October 2021).

6.1.2. Pretraining

A key insight to significantly increase the quality of direct models centres on pretraining the model on ASR data [19,34]. We thus pretrained each model variant on the available transcriptions in the mintzai-ST corpus, prior to training on the audio source and target language data in the end-to-end ST scenario. Although pretraining relies on the existence of transcriptions, and might thus be viewed as departing from a strict end-to-end approach to the ST problem, this step is now standardly performed for direct models when transcriptions are indeed available. For the experiments described below, pretraining was applied to all model variants, as it can be performed without changing any of the selected model architectures (models that underwent pretraining are indicated with the suffix .PT in the model name).

6.1.3. Knowledge Distillation

Another relevant technique relies on knowledge distillation, as proposed by [18] for ST. In this approach, a text translation model, based on a standard Transformer architecture, is trained first and taken as the teacher model. In a second step, an ST model, also based on a standard Transformer architecture (with an additional initial linear layer prior to the SA encoder layers in the aforementioned study), is trained on the output of the teacher model. We applied this method to all models that were trained with Transformer encoder–decoder architectures on data segmented with BPE subwords, using the output of the cascade MT system as a reference for the student training step (models that underwent training with knowledge distillation are indicated with the suffix .KD in the model name).

6.2. Comparative Direct Models' Results

We applied each of the previously described variants on the mintzai-ST test sets, with the results shown in Tables 10 and 11, for Spanish to Basque and Basque to Spanish, respectively.

Table 10. Comparative results for end-to-end model variants on Basque to Spanish translation. The best results, indicated in bold, were statistically significant against all other results, for $p < 0.05$.

Lang	Model	BLEU	CHRF
EU–ES	S-TRF	17.0	42.5
EU–ES	S-TRF.PT	26.8	52.6
EU–ES	S-TRF.PT.KD	28.4	53.2
EU–ES	CNN.11E-4D.TRF.PT	25.0	53.5
EU–ES	CNN.11E-4D.TRF.PT.KD	28.1	53.7
EU–ES	LINEAR.TRF.PT	24.6	51.4
EU–ES	LINEAR.TRF.PT.KD	26.8	52.2
EU–ES	WAV2VEC-ST	29.8	57.6
EU–ES	WAV2VEC-ST.PT	**31.4**	**58.5**

Table 11. Comparative results for end-to-end model variants on Spanish to Basque translation. The best results, indicated in bold, were statistically significant against all other results, for $p < 0.05$; differences between top-performing systems in bold were not statistically significant.

Lang	Model	BLEU	CHRF
ES–EU	S-TRF	12.9	48.4
ES–EU	S-TRF.PT	17.8	54.7
ES–EU	S-TRF.PT.KD	20.4	57.3
ES–EU	CNN.11E-4D.TRF.PT	15.6	53.3
ES–EU	CNN.11E-4D.TRF.PT.KD	19.2	56.2
ES–EU	LINEAR.TRF.PT	17.0	53.7
ES–EU	LINEAR.TRF.PT.KD	18.9	55.9
ES–EU	WAV2VEC-ST	**24.5**	**61.9**
ES–EU	WAV2VEC-ST.PT	**24.7**	**62.2**

In both translation directions, the WAV2VEC-ST models performed significantly better than the alternatives, overall. Although these results illustrate the positive impact of WAV2VEC 2.0 cross-lingual speech representations for the ST task as well, it is worth noting that all other models were trained only on the in-domain mintzai-ST data, which feature significantly lower volumes than those used to pretrain WAV2VEC 2.0.

Another notable result is the significant impact of pretraining, with a 9.8 BLEU point increase for the S-Transformer model for Basque to Spanish,and 4.9 points for Spanish to Basque. Knowledge distillation was also impactful across the board, with overall increases of at least two BLEU points over pretrained models.

Turning to architecture variants, in both translation directions, the default S-Transformer model outperformed the variant with eleven SA encoder layers, four decoder layers, and without 2D SA layers, which in turn performed better than the variant with a linear layer replacing both the CNN and SA layers. These comparative metrics' results were consistent for the variants with pretraining only and knowledge distillation in addition to pretraining.

Table 12 summarises the comparative results between cascade and advanced end-to-end models, trained on either in-domain data only or on additional data (as previously indicated, for the WAV2VEC-ST model, additional data refer to the datasets used to pretrain the model). In all cases, the gap between the direct and cascade models was significantly reduced, compared to the baseline results of Section 5. End-to-end models obtained only slightly lower results overall in both directions among systems trained only on in-domain data. In Spanish to Basque, the end-to-end model based on WAV2VEC-ST even outperformed the cascade model trained on all data. For Basque to Spanish though, the cascade model still outperformed the best end-to-end model on the test sets.

Table 12. Summary of the results with cascade and advanced end-to-end models. Best-performing systems are indicated in bold.

Lang	Model	Data	WER	BLEU
EU–ES	CAS	IND	12.07	29.2
EU–ES	S-TRF.PT.KD	IND	-	28.4
EU–ES	CAS	ALL	**11.78**	**34.7**
EU–ES	WAV2VEC-ST.PT	ALL	-	31.4
ES–EU	CAS	IND	7.23	20.9
ES–EU	S-TRF.PT.KD	IND	-	20.4
ES–EU	CAS	ALL	**7.21**	22.7
ES–EU	WAV2VEC-ST.PT	ALL	-	**24.7**

7. Targeted Evaluations of Cascade and Advanced Direct Models

Given the results of the previous sections, we selected the most representative variants to perform a manual evaluation and examine in detail the characteristics of the selected cascade and direct models. Since the results varied significantly between models based strictly on in-domain data and models with access to additional data (where additional data refer to both the datasets described in Section 4.1 and the data used to independently train the WAV2VEC 2.0 models), we selected two different pairs of systems to be compared separately:

- CAS IND vs. S-TRF.PT.KD for systems trained only on in-domain data;
- CAS ALL vs. WAV2VEC-ST.PT for systems trained also on additional data.

In the following sections, we describe the results of the comparative evaluations along different relevant aspects, namely manual ranking, divergences on specific phenomena, and error propagation.

7.1. Manual Ranking Task

This task consisted of a manual evaluation of the translations generated by the different selected systems, previously described, where users were presented the source transcription and two competing translations and had to indicate whether one was preferred or whether they could be considered similar in accuracy and fluency.

As evaluation data for the task, we extracted new data from the 2019 Sessions of the Basque Parliament, not covered by the mintzai-ST corpus, and sampled 100 representative audio inputs. To select the samples, the content of a complete Session was first translated with each of the four selected ST models. The data were then log-normalised and split into quartiles according to the length of the transcriptions, with 25 samples randomly selected from each quartile to provide a representation of different types of input in terms of duration.

Table 13 indicates the BLEU scores obtained by each model on the selected samples, to be taken as a first indication of the relative translation quality on the samples. The results were in line with the previous indications of relative model strength in terms of metrics, with slightly better scores overall obtained by the cascade models.

For the manual three-way ranking task, two separate groups of users were defined to handle, with the following characteristics:

- Group A was tasked with comparing the translations between models trained only on in-domain data, namely CAS IND and S-TRF.PT.KD. For Spanish to Basque, nine users completed the full evaluation and one user only partially completed the task (forty out of one-hundred). For Basque to Spanish, eight users completed the full evaluation and one user only partially completed the task (twenty-three out of one-hundred);
- Group B was tasked with comparing the translations between models trained on additional data, namely CAS ALL and WAV2VEC-ST.PT. For Spanish to Basque, 12 users completed the full evaluation; for Basque to Spanish, 11 users completed the full evaluation.

Table 13. BLEU scores on manual evaluation samples.

Lang	CAS IND	s-trf.pt.kd	CAS All	wav2vec-st.kd
ES–EU	18.4	16.7	22.0	21.7
EU–ES	30.7	28.2	35.5	33.0

The evaluators who volunteered for the task were native speakers of Basque and Spanish, but not professional translators. They were provided guidelines on the task itself and on the use of the evaluation environment, which is based on the Appraise system [64] and provides inter-annotator agreement statistics.

The results of the manual evaluation are shown in Table 14, with inter-annotator agreement measured in terms of Krippendorff's alpha [65], Fleiss's Kappa [66], Bennett' S [67], and Scott's Pi [68] (omitted from the table are the number of cases that were skipped by users, which were none in all cases for Group B and, for Group A, amounted to 2.66% in ES–EU and 0.12% in EU–ES). Overall, translations from the cascade systems were preferred by a large margin for systems trained only on in-domain data, in both translation directions, and in Basque to Spanish for the systems trained on additional data. Only in the latter case, for Spanish to Basque, were the systems considered relatively equal. It also worth noting that between 30% and 40% of the translations, approximately, were considered of equal quality overall.

Table 14. Cascade vs. end-to-end 3-way ranking results. α indicates Krippendorff's alpha; κ: Fleiss's Kappa; s: Bennett's S; pi: Scott's Pi.

Data	Lang	CAS < E2E	Equal	CAS > E2E	α	κ	s	pi
IND	ES–EU	25.74%	31.60%	40.00%	0.31	0.27	0.29	0.25
IND	EU–ES	18.47%	41.56%	39.85%	0.58	0.47	0.47	0.43
All	ES–EU	32.04%	35.18%	32.78%	0.26	0.23	0.16	0.21
All	EU–ES	17.09%	40.36%	42.55%	0.45	0.45	0.48	0.44

Inter-annotator agreement was moderate overall and significantly higher in both cases of Basque to Spanish translation, where the differences were larger between the compared systems. The lowest agreement was obtained in the only case where no clear preference was shown for either system, which may be interpreted as a result of the translations being of similar quality overall, without systematic aspects favouring one or the other.

7.2. Divergence on Specific Phenomena

To assess more detailed differences between the cascade and direct approaches, we selected and evaluated translation subsets addressing three different phenomena, after a preliminary manual evaluation of the data to identify aspects that led to divergent translations.

We thus identified all source transcriptions in the test sets that contained: numbers; a specific subset of punctuation marks, namely question, exclamation, and ellipsis marks; named entities introduced by markers in Basque or Spanish corresponding to *Sir*, *Madam*, or similar, which are almost systematically used to refer to other participants by name in the sessions of the Basque Parliament. The number of identified samples, and BLEU scores for the four selected systems on each subset, are indicated in Table 15.

Table 15. BLEU results on selected test subsets

Subset	Lang	Samples	CAS IND	s-tf.pt.kd	CAS All	wav2vec-st.pt
Number	EU–ES	221	21.8	24.1	28.7	28.2
Punct	EU–ES	348	25.2	27.3	31.9	29.8
Names	EU–ES	252	43.0	41.8	45.4	44.1
Number	ES–EU	332	17.1	19.2	18.3	24.2
Punct	ES–EU	543	16.4	17.0	18.2	21.3
Names	ES–EU	260	24.6	24.7	26.3	27.1

On the numbers subset, the end-to-end models performed better overall, although this was mainly due to the fact that numbers were provided in word rather than numeral form by ASR in the cascade system. Additional processes could perform this conversion in a cascade setup, similar to the use of additional processes to insert punctuation marks. The differences in translations of numbers were thus mainly significant to indicate that some of the gains obtained by direct models were related to this variable way of representing numbers in the final translations.

Translations also differed in terms of punctuation, in this case also with higher marks obtained overall by end-to-end models, except for Basque to Spanish with additional data. One of the main reasons for this divergence comes from the limitations of the specific punctuation model used for the cascade system, which only covered commas and periods. In contrast, the end-to-end models exploited the source–target punctuation data directly and could model the whole spectrum of punctuation marks in the datasets.

Finally, the divergences in terms of names were less marked, with slightly better scores with cascade models overall for Basque to Spanish. For Spanish to Basque, nearly identical results were obtained with models trained only on in-domain data and slightly better scores for WAV2VEC-ST.PT against CAS ALL.

Table 16 provides examples of divergent translations between cascade and direct models. Except the punctuation case, which is only correct in the direct translation example, the other examples illustrated divergences that may all be considered correct. The differences nonetheless impacted the BLEU scores on the single references used in these evaluations, an aspect that needs to be factored in when comparing translation models [16,24].

Table 16. Examples of diverging translations on specific phenomena, with English translations.

Subset	Translation	Example
Number	Reference EN	Lean, si no, el artículo 6. *If not, read article 6.*
	CAS ALL EN	Véase, de lo contrario, el artículo sexto. *See, otherwise, article six.*
	WAV2VEC.PT EN	Si no, se ha visto el artículo 6. *If not, article 6 has been seen.*
Punct	Reference EN	Lan-erreforma indargabetu da? *Is job reform repealed?*
	CAS IND EN	Lan-erreforma indargabetu egin da. *Job reform is repealed.*
	S-TF.PT.KD EN	Indargabetu al da lan-erreforma? *Is job reform repealed?*

Table 16. *Cont.*

Subset	Translation	Example
Name	Reference	Pasando al turno de réplica, tiene usted la palabra, señora Rojo.
	EN	*Turning to the reply, you have the floor, Mrs. Rojo.*
	CAS ALL	Pasando al turno de réplica, señora Rojo, tiene usted la palabra.
	EN	*Turning to the reply, Mrs. Rojo, you have the floor.*
	WAV2VEC-ST.PT	En el turno de réplica, tiene la palabra la señora Rojo.
	EN	*In the turn to reply, Mrs. Rojo has the floor.*

7.3. Error Propagation

As previously noted, one of the expected advantages of direct ST models is avoiding the propagation of ASR errors to the MT component, an aspect that was one of the characteristics of earlier cascade systems. To measure this effect, albeit indirectly, we evaluated the translation results on input classified according to the WER scores obtained with the top-performing ASR component described in Section 4.1.1, with WER scores assumed to reflect the percentage of ASR errors that may impact a cascaded MT model. Clearly, since higher WER scores may be due to audio input that is intrinsically difficult to process for any automated recognition system, cascaded or direct, this evaluation was an approximation explored to determine eventual differences in behaviour between cascade and direct models. For these evaluations, we used the four systems selected in Section 7.

We first measured the correlation between the WER and BLEU scores for each considered model, to determine the linear relationship between the two scores, if any. Additionally, we measured the correlation between WER and the difference of BLEU scores obtained by the compared cascade and end-to-end models, to determine the relation between WER and performance variation for the two types of models. To account for the fact that WER scores are less reliable on short audio, we filtered all audio whose corresponding transcription consisted of 10 words or less; WER scores were also normalised in the 0–100 range. The results in terms of Pearson correlation coefficients are shown in Table 17. It is worth noting that negative coefficients between WER and BLEU indicate a positive correlation, since the metrics are reversed in terms of rank interpretation, with a higher WER indicating worse results, while a higher BLEU indicates better results.

Table 17. Pearson correlation between WER↓ (W) and BLEU↑ (B) scores for the selected Cascade (CAS) and End-to-End (E2E) models. All results were statistically significant ($p < 0.0001$).

Lang	Data	ρ(W, B CAS)	ρ(W, B E2E)	ρ(W, (B CAS) − (B E2E))
ES–EU	IND	−0.14	−0.13	0.40
ES–EU	ALL	−0.15	−0.10	0.35
EU–ES	IND	−0.26	−0.25	0.32
EU–ES	ALL	−0.24	−0.20	0.43

For both cascade and direct models, the correlation between the WER and BLEU scores was small overall, with lower marks for ES–EU than for EU–ES. The correlation was comparatively higher between the WER and BLEU scores' differences, being moderate across the board.

To further evaluate the impact of input difficulty as indicated by the BLEU scores, the test data were further first divided into quartiles according to the WER scores obtained by the ASR component. We then computed the difference in the BLEU scores between paired models on the data in each quartile. The results are shown in Figure 3.

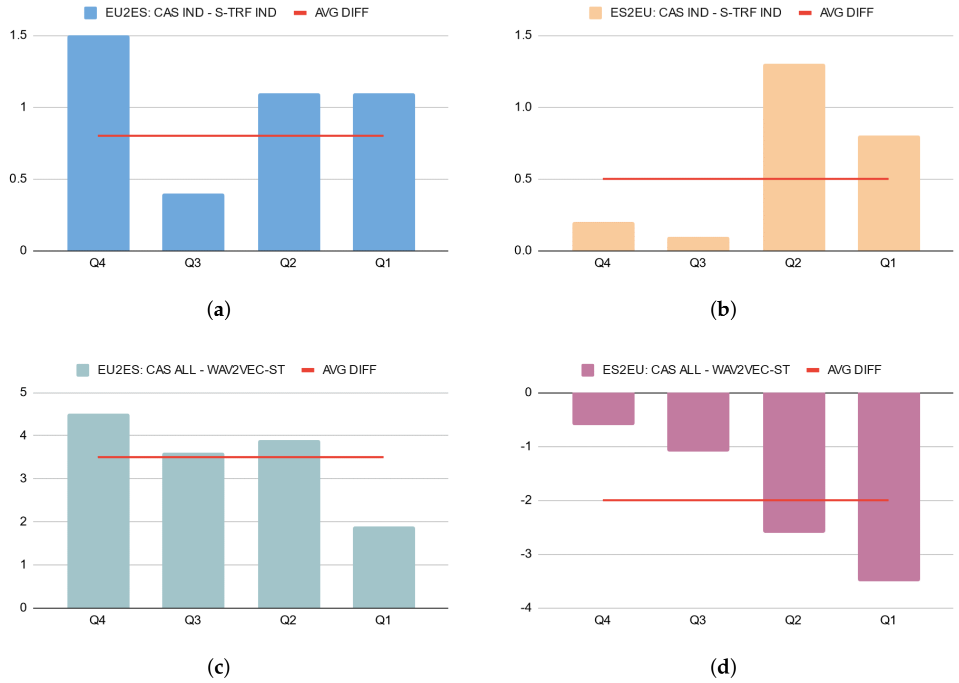

Figure 3. BLEU score differences on WER-based quartile subsets of the mintzai-ST test sets, between: (**a**) CAS IND and S-TRF.PT.KD IND in EU-ES (**b**) CAS IND and S-TRF.PT.KD IND in ES-EU, (**c**) CAS ALL and WAV2VEC-ST.PT in EU2ES, and (**d**) CAS ALL and WAV2VEC-ST.PT in ES2EU. The red line indicates the average difference over joint subsets. Quartiles range from lower WER (Q4) to higher WER (Q1).

Under the assumption that higher WER scores would correlate with a larger number of errors that would impact the MT component of cascade models more than they would an end-to-end model, the expectation would be that the difference in BLEU scores is lower than the average where cascade models outperform direct models and higher than average where direct models score better than their cascade counterpart.

With models trained on in-domain data only, for EU–ES (Figure 3a), only Q3 featured a result under the average difference; in all other quartiles, the difference was larger than average, although the differences on both Q2 and Q1 were lower than with Q4. In ES–EU (Figure 3b), the tendencies were opposite the expectation, with the cascade model performing markedly better than average compared to the direct model on data associated with a higher WER, i.e., Q2 and Q1.

The results for models trained on additional data were more in line with the assumed impact of data associated with higher WER scores. For EU–ES for instance (Figure 3c), the positive difference in favour of the cascade model against the direct model was lesser on Q1 and higher on Q4, with Q1 and Q3 in line with the average difference. For ES–EU (Figure 3d), the direct model scored higher than average compared to the cascade model, with a consistent tendency as the WER scores increased from Q4 to Q1.

The results for this evaluation were thus not uniform, with opposite tendencies for the two pairs of systems depending on their having been trained only on in-domain data or also on additional data. In this case as well, additional references and manual assessments of the audio files in terms of difficulty could help determine the actual tendencies in terms of error propagation. A manual examination of non-literal transcriptions in the test sets would also support a more precise evaluation of WER-related differences between models. These tasks were however beyond the scope of this study and are left for future work.

8. Conclusions

In this work, we presented the results of a case study in Basque–Spanish speech translation, centred on comparative evaluations of cascade and direct approaches to the task.

We first described the mintzai-ST corpus, based on the proceedings of the sessions of the Basque Parliament between 2011 and 2018. We extended the initial description of the corpus in [14] with a detailed examination of the different alignment and filtering steps, along with an analysis of the data distribution in the corpus.

Different ST models were compared, based on cascade and end-to-end approaches, starting with baseline results that included end-to-end models trained strictly on the source–target data, which resulted in cascade systems significantly outperforming their direct counterparts, with or without additional data.

Several variants of advanced end-to-end models were then prepared, exploiting pretraining and knowledge distillation techniques in particular. As was the case in other studies exploiting these techniques [18,34], all advanced variants significantly closed the gap with their cascaded counterparts, in terms of automated metrics, including a variant based on WAV2VEC [63], which outperformed all other models in Spanish to Basque translation. This latter model proved the most efficient in terms of metrics, providing further support to the usefulness of self-supervised training on audio data prior to performing speech translation. The comparison with other models in this study was less direct, however, given the use of large amounts of audio data to pretrain the models. Among other variants, pretraining and knowledge distillation both proved critical to drastically increase the performance of end-to-end models, with the S-Transformer model outperforming similarly trained model variants that featured architectural differences.

To further evaluate the differences between the two main ST approaches, cascade and direct, a manual evaluation was carried out on the translations from the best-performing models in each case. Although between 30% and 40% of the translations overall were considered of similar quality by a panel of native speakers of the two languages, the translations generated by the cascade models were preferred by a significant margin in all but one case, where the preferences were equally distributed. These results complement other comparative manual evaluations such as [24], though reaching differing conclusions, as in our study and specific evaluation protocol, cascade translations were preferred overall.

We also conducted targeted evaluations along two axes. First, we evaluated divergent translations under the cascade and direct approaches, where the latter approach performed better on numbers and punctuation against the specific cascade models in this study, although in both cases, improvements could be straightforwardly achieved in a cascade approach. Additionally, we compared model result differences according to the WER scores, to measure the potential impact of ASR errors on MT results in a cascade approach and the eventual higher robustness of a direct approach in this respect. The results were inconclusive, with better relative scores for direct models on more difficult input in Basque to Spanish translation, but a reverse tendency in Spanish to Basque.

From this study, it appears that the gap between cascade and direct approaches has been reduced significantly with recent approaches to direct ST modelling, in line with similar findings [19,23,24]. Nonetheless, the cascade models still obtained better results overall in terms of reference-based metrics and manual evaluations, in line with the findings reported in [25], where cascade models achieved the best results overall on the IWSLT 2021 shared task datasets. It is worth noting that this was the case even under the controlled conditions of our study, where additional audio and translation data were considered to some extent, but with volumes of data largely under what might be exploited in this language pair [16,69]. The gap between cascade and direct speech translation models is thus likely to be larger under unrestricted conditions, despite the significant progress achieved with direct models in recent years.

In future work, a more detailed examination of the characteristics of the Basque–Spanish language pair, in terms of syntactic and morphological variation in particular, along with further manual examination of the corpus in terms of literality and audio variation, could shed more light on the respective strengths and limitations of cascade and direct approaches.

Author Contributions: Conceptualisation, T.E., A.A., H.A. and H.G.; methodology, T.E., A.A., H.A., H.G., A.G.-D., I.G.T., J.M.M.-D. and E.B.F.; validation, T.E. and A.A.; formal analysis, T.E. and H.G.; investigation, T.E., A.A., H.A., H.G, A.G-D., I.G.T., J.M.M.-D. and E.B.F.; resources, T.E., H.A. and H.G.; data curation, T.E., H.A., H.G. and A.G.-D.; writing—original draft preparation, T.E.; writing—review and editing, T.E., A.A., H.A., H.G., A.G.-D., I.G.T., J.M.M.-D. and E.B.F.; visualisation, T.E. and H.G.; supervision, T.E. and A.A.; project administration, T.E.; funding acquisition, T.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Department of Economic Development and Competitiveness of the Basque Government (Spri) Grant Number KK-2019/00065.

Data Availability Statement: Publicly available datasets were analysed in this study. These data can be found here: https://github.com/vicomtech/mintzai-st, accessed on 11 March 2021.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ney, H. Speech translation: Coupling of recognition and translation. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Phoenix, AZ, USA, 15–19 March 1999; pp. 517–520.
2. Matusov, E.; Kanthak, S.; Ney, H. On the integration of speech recognition and statistical machine translation. In Proceedings of the Ninth European Conference on Speech Communication and Technology, Lisbon, Portugal, 4–8 September 2005; pp. 3176–3179.
3. Kumar, G.; Blackwood, G.; Trmal, J.; Povey, D.; Khudanpur, S. A coarse-grained model for optimal coupling of ASR and SMT systems for speech translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1902–1907.
4. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
5. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 6000–6010.
6. Graves, A.; Mohamed, A.R.; Hinton, G. Speech recognition with deep recurrent neural networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 6645–6649.
7. Chan, W.; Jaitly, N.; Le, Q.; Vinyals, O. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Shanghai, China, 20–25 March 2016; pp. 4960–4964.
8. Duong, L.; Anastasopoulos, A.; Chiang, D.; Bird, S.; Cohn, T. An attentional model for speech translation without transcription. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 949–959.
9. Bérard, A.; Pietquin, O.; Besacier, L.; Servan, C. Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation. In Proceedings of the NIPS Workshop on End-to-End Learning for Speech and Audio Processing, Barcelona, Spain, 10 December 2016.
10. Weiss, R.J.; Chorowski, J.; Jaitly, N.; Wu, Y.; Chen, Z. Sequence-to-Sequence Models Can Directly Translate Foreign Speech. *arXiv* **2017**, arXiv:1703.08581.
11. Niehues, J.; Cattoni, R.; Stüker, S.; Negri, M.; Turchi, M.; Salesky, E.; Sanabria, R.; Barrault, L.; Specia, L.; Federico, M. The IWSLT 2019 Evaluation Campaign. In Proceedings of the 16th International Workshop on Spoken Language Translation, Hong Kong, China, 2–3 November 2019.
12. Di Gangi, M.A.; Cattoni, R.; Bentivogli, L.; Negri, M.; Turchi, M. MuST-C: A Multilingual Speech Translation Corpus. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 3–5 June 2019; pp. 2012–2017.
13. Iranzo-Sánchez, J.; Silvestre-Cerdà, J.A.; Jorge, J.; Roselló, N.; Giménez, A.; Sanchis, A.; Civera, J.; Juan, A. Europarl-ST: A multilingual corpus for speech translation of parliamentary debates. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4–8 May 2020; pp. 8229–8233.
14. Etchegoyhen, T.; Arzelus, H.; Gete, H.; Alvarez, A.; González-Docasal, A.; Fernandez, E.B. mintzai-ST: Corpus and Baselines for Basque–Spanish Speech Translation. In Proceedings of the IberSPEECH 2020, Valladolid, Spain, 24–25 March 2021; pp. 190–194.
15. Pérez, A.; Alcaide, J.M.; Torres, M.I. EuskoParl: A speech and text Spanish-Basque parallel corpus. In Proceedings of the Interspeech, Portland, OR, USA 9–13 September 2012; pp. 2362–2365.
16. Etchegoyhen, T.; Martínez Garcia, E.; Azpeitia, A.; Labaka, G.; Alegria, I.; Cortes Etxabe, I.; Jauregi Carrera, A.; Ellakuria Santos, I.; Martin, M.; Calonge, E. Neural Machine Translation of Basque. In Proceedings of the 21st Annual Conference of the European Association for Machine Translation, Alacant, Spain, 28–30 May 2018; pp. 139–148.
17. Di Gangi, M.A.; Negri, M.; Cattoni, R.; Dessi, R.; Turchi, M. Enhancing Transformer for End-to-end Speech-to-Text Translation. In Proceedings of the Machine Translation Summit XVII Volume 1: Research Track, Dublin, Ireland, 19–23 August 2019; pp. 21–31.

18. Liu, Y.; Xiong, H.; He, Z.; Zhang, J.; Wu, H.; Wang, H.; Zong, C. End-to-End Speech Translation with Knowledge Distillation. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 1128–1132.
19. Sperber, M.; Paulik, M. Speech Translation and the End-to-End Promise: Taking Stock of Where We Are. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 7409–7421.
20. Wu, A.; Wang, C.; Pino, J.M.; Gu, J. Self-Supervised Representations Improve End-to-End Speech Translation. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; pp. 1491–1495.
21. Li, X.; Wang, C.; Tang, Y.; Tran, C.; Tang, Y.; Pino, J.; Baevski, A.; Conneau, A.; Auli, M. Multilingual Speech Translation from Efficient Finetuning of Pretrained Models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Bangkok, Thailand, 1–6 August 2021; Volume 1, pp. 827–838. [CrossRef]
22. Dong, Q.; Ye, R.; Wang, M.; Zhou, H.; Xu, S.; Xu, B.; Li, L. Listen, Understand and Translate: Triple Supervision Decouples End-to-end Speech-to-text Translation. In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, Virtual Event, 2–9 February 2021; pp. 12749–12759.
23. Ansari, E.; Axelrod, A.; Bach, N.; Bojar, O.; Cattoni, R.; Dalvi, F.; Durrani, N.; Federico, M.; Federmann, C.; Gu, J.; et al. Findings of the IWSLT 2020 evaluation campaign. In Proceedings of the 17th International Conference on Spoken Language Translation, Online, 9–10 July 2020; pp. 1–34. [CrossRef]
24. Bentivogli, L.; Cettolo, M.; Gaido, M.; Karakanta, A.; Martinelli, A.; Negri, M.; Turchi, M. Cascade versus Direct Speech Translation: Do the Differences Still Make a Difference? In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Bangkok, Thailand, 1–6 August 2021; Volume 1, pp. 2873–2887. [CrossRef]
25. Anastasopoulos, A.; Bojar, O.; Bremerman, J.; Cattoni, R.; Elbayad, M.; Federico, M.; Ma, X.; Nakamura, S.; Negri, M.; Niehues, J.; et al. Findings of the IWSLT 2021 Evaluation Campaign. In Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021), Virtual Event, 5–6 August 2021; pp. 1–29.
26. Waibel, A.; Jain, A.N.; McNair, A.E.; Saito, H.; Hauptmann, A.G.; Tebelskis, J. JANUS: A speech-to-speech translation system using connectionist and symbolic processing strategies. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Toronto, ON, Canada, 14–17 April 1991; pp. 793–796.
27. Lavie, A.; Gates, D.; Gavalda, M.; Tomokiyo, L.M.; Waibel, A.; Levin, L. Multi-lingual translation of spontaneously spoken language in a limited domain. In Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, Denmark, 5–9 August 1996; pp. 442–447.
28. Vidal, E. Finite-state speech-to-speech translation. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Munich, Germany, 21–24 April 1997; Volume 1, pp. 111–114.
29. Casacuberta, F.; Ney, H.; Och, F.J.; Vidal, E.; Vilar, J.M.; Barrachina, S.; Garcıa-Varea, I.; Llorens, D.; Martınez, C.; Molau, S.; et al. Some approaches to statistical and finite-state speech-to-speech translation. *Comput. Speech Lang.* **2004**, *18*, 25–47. [CrossRef]
30. Dixon, P.; Finch, A.; Hori, C.; Kashioka, H. Investigation of the effects of ASR tuning on speech translation performance. In Proceedings of the 8th International Workshop on Spoken Language Translation: Evaluation Campaign, San Francisco, CA, USA, 8–9 December 2011.
31. Peitz, S.; Wiesler, S.; Nußbaum-Thom, M.; Ney, H. Spoken language translation using automatically transcribed text in training. In Proceedings of the 9th International Workshop on Spoken Language Translation, Hong Kong, China, 6–7 December 2012.
32. Jia, Y.; Weiss, R.J.; Biadsy, F.; Macherey, W.; Johnson, M.; Chen, Z.; Wu, Y. Direct speech-to-speech translation with a sequence-to-sequence model. *arXiv* **2019**, arXiv:1904.06037.
33. Kocabiyikoglu, A.C.; Besacier, L.; Kraif, O. Augmenting Librispeech with French Translations: A Multimodal Corpus for Direct Speech Translation Evaluation. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018; European Language Resources Association (ELRA): Miyazaki, Japan, 2018.
34. Bérard, A.; Besacier, L.; Kocabiyikoglu, A.C.; Pietquin, O. End-to-end automatic speech translation of audiobooks. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 6224–6228.
35. Wang, C.; Wu, A.; Pino, J. CoVoST 2: A Massively Multilingual Speech-to-Text Translation Corpus. *arXiv* **2020**, arXiv:2007.10310.
36. Wang, C.; Pino, J.; Wu, A.; Gu, J. CoVoST: A Diverse Multilingual Speech-To-Text Translation Corpus. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; European Language Resources Association: Marseille, France, 2020; pp. 4197–4203.
37. Salesky, E.; Wiesner, M.; Bremerman, J.; Cattoni, R.; Negri, M.; Turchi, M.; Oard, D.W.; Post, M. The Multilingual TEDx Corpus for Speech Recognition and Translation. In Proceedings of the Interspeech, Brno, Czech Republic, 30 August–3 September 2021; pp. 3655–3659. [CrossRef]
38. Jia, Y.; Johnson, M.; Macherey, W.; Weiss, R.J.; Cao, Y.; Chiu, C.C.; Ari, N.; Laurenzo, S.; Wu, Y. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 7180–7184.
39. Dong, L.; Xu, S.; Xu, B. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5884–5888.

40. Di Gangi, M.A.; Negri, M.; Turchi, M. Adapting Transformer to End-to-End Spoken Language Translation. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 1133–1137.
41. Indurthi, S.; Han, H.; Lakumarapu, N.K.; Lee, B.; Chung, I.; Kim, S.; Kim, C. End-end speech-to-text translation with modality agnostic meta-learning. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7904–7908.
42. Schneider, S.; Baevski, A.; Collobert, R.; Auli, M. wav2vec: Unsupervised Pre-Training for Speech Recognition. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 3465–3469. [CrossRef]
43. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 12449–12460.
44. Nguyen, H.; Bougares, F.; Tomashenko, N.; Estève, Y.; Besacier, L. Investigating self-supervised pre-training for end-to-end speech translation. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020.
45. Agerri, R.; Cuadros, M.; Gaines, S.; Rigau, G. OpeNER: Open polarity enhanced named entity recognition. In *Procesamiento del Lenguaje Natural*; Sociedad Española para el Procesamiento del Lenguaje Natural: Jaén, Spanish, 2013; pp. 215–218.
46. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; et al. The Kaldi speech recognition toolkit. In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, Big Island, HI, USA, 11–15 December 2011.
47. Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; et al. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, 25–27 June 2007; pp. 177–180.
48. Varga, D.; Németh, L.; Halácsy, P.; Kornai, A.; Trón, V.; Nagy, V. Parallel corpora for medium density languages. In Proceedings of the Recent Advances in Natural Language Processing, Borovets, Bulgaria, 21–23 September 2005; pp. 590–596.
49. Amodei, D.; Ananthanarayanan, S.; Anubhai, R.; Bai, J.; Battenberg, E.; Case, C.; Casper, J.; Catanzaro, B.; Cheng, Q.; Chen, G.; et al. Deep speech 2: End-to-end speech recognition in English and Mandarin. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 173–182.
50. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 2613–2617. [CrossRef]
51. Povey, D.; Cheng, G.; Wang, Y.; Li, K.; Xu, H.; Yarmohammadi, M.; Khudanpur, S. Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks. In Proceedings of Interspeech, Hyderabad, India, 2–6 September 2018; pp. 3743–3747.
52. Ko, T.; Peddinti, V.; Povey, D.; Khudanpur, S. Audio augmentation for speech recognition. In Proceedings of the Interspeech, Dresden, Germany, 6–10 September 2015; pp. 3586–3589.
53. Peddinti, V.; Povey, D.; Khudanpur, S. A time delay neural network architecture for efficient modeling of long temporal contexts. In Proceedings of the Interspeech, Dresden, Germany, 6–10 September 2015; pp. 3214–3218.
54. 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.
55. Tilk, O.; Alumäe, T. Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration. In Proceedings of the Interspeech, San Francisco, CA, USA, 8–12 September 2016; pp. 3047–3051.
56. 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, Reykjavik, Iceland, 26–31 May 2014; pp. 432–436.
57. Junczys-Dowmunt, M.; Grundkiewicz, R.; Dwojak, T.; Hoang, H.; Heafield, K.; Neckermann, T.; Seide, F.; Germann, U.; Aji, A.F.; Bogoychev, N.; et al. Marian: Fast Neural Machine Translation in C++. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 116–121.
58. Sennrich, R.; Haddow, B.; Birch, A. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; pp. 1715–1725.
59. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
60. Popović, M. chrF: Character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, Baltimore, ML, USA, 26–27 June 2014; Association for Computational Linguistics: Lisbon, Portugal, 2015; pp. 392–395. [CrossRef]
61. Post, M. A Call for Clarity in Reporting BLEU Scores. In Proceedings of the Third Conference on Machine Translation, Belgium, Brussels, 31 October–1 November 2018; pp. 186–191.
62. Koehn, P. Statistical Significance Tests for Machine Translation Evaluation. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 25–26 July 2004; pp. 388–395.
63. Wang, C.; Wu, A.; Pino, J.; Baevski, A.; Auli, M.; Conneau, A. Large-Scale Self- and Semi-Supervised Learning for Speech Translation. In Proceedings of the Interspeech, Brno, Czech Republic, 30 August–3 September 2021; pp. 2242–2246. [CrossRef]
64. Federmann, C. Appraise evaluation framework for machine translation. In Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, Santa Fe, NM, USA, 20–26 August 2018; pp. 86–88.

65. Krippendorff, K. *Computing Krippendorff's Alpha-Reliability*; Technical Report; University of Pennsylvania: Philadelphia, PA, USA, 2011.
66. Fleiss, J.L. Measuring nominal scale agreement among many raters. *Psychol. Bull.* **1971**, *76*, 378. [CrossRef]
67. Bennett, E.M.; Alpert, R.; Goldstein, A. Communications through limited-response questioning. *Public Opin. Q.* **1954**, *18*, 303–308. [CrossRef]
68. Scott, W.A. Reliability of content analysis: The case of nominal scale coding. *Public Opin. Q.* **1955**, *19*, 321–325. [CrossRef]
69. Alvarez, A.; Arzelus, H.; Torre, I.G.; González-Docasal, A. The Vicomtech Speech Transcription Systems for the Albayzın-RTVE 2020 Speech to Text Transcription Challenge. In Proceedings of the IberSPEECH 2021, Valladolid, Spain, 24–25 March 2021; pp. 104–107.