End-to-End Speech Translation for Code Switched Speech

**Orion Weller[1]*, Matthias Sperber[2], Telmo Pires[2],
Hendra Setiawan[2], Christian Gollan[2], Dominic Telaar[2], Matthias Paulik[2]**
[1]Johns Hopkins University
[2]Apple
oweller@cs.jhu.edu,sperber@apple.com

Abstract

Code switching (CS) refers to the phenomenon of interchangeably using words and phrases from different languages. CS can pose significant accuracy challenges to NLP, due to the often monolingual nature of the underlying systems. In this work, we focus on CS in the context of English/Spanish conversations for the task of speech translation (ST), generating and evaluating both transcript and translation. To evaluate model performance on this task, we create a novel ST corpus derived from existing public data sets.[1] We explore various ST architectures across two dimensions: cascaded (transcribe then translate) vs end-to-end (jointly transcribe and translate) and unidirectional (source → target) vs bidirectional (source ↔ target). We show that our ST architectures, and especially our bidirectional end-to-end architecture, perform well on CS speech, even when no CS training data is used.

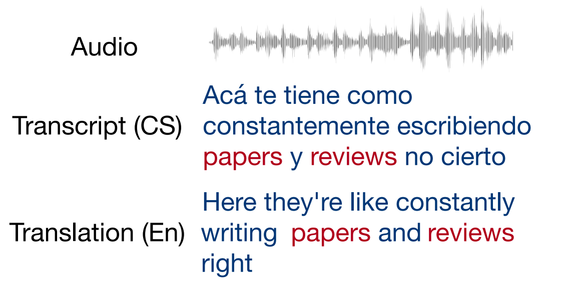

Figure 1: An example instance of the joint speech recognition and translation task for code-switching (CS). Red indicates English words in the transcript and their corresponding words in the translation, whereas blue indicates Spanish words in the transcript and their corresponding translation.

1 Introduction

Over half of the world's population is estimated to be bilingual. [2] Those that know multiple languages are prone to code switch, i.e., to interchangeably use words and phrases from two (or more) languages in situations such as casual dialog, while traveling abroad, or simply to use a word they find more fitting (Myers-Scotton and Ury, 1977; Heredia and Altarriba, 2001). In CS, the base language is referred to as the *matrix* language while the contributing language is called the *embedded* language (Myers-Scotton, 1995), where speakers often use the matrix language the majority of the time.

Code switched language is challenging to both automatic speech recognition (ASR) and machine translation (MT) - and therefore also to the composite task of speech translation (ST). While a rich amount of prior works exist on CS in the context of ASR (Lyu et al., 2006; Ahmed and Tan, 2012; Vu et al., 2012; Johnson et al., 2017; Yue et al., 2019) and MT (Sinha and Thakur, 2005; Winata et al., 2021; Zhang et al., 2021; Yang et al., 2020), there is little prior work in the context of ST.

The aforementioned challenges to ASR, MT and ST arise largely due to the lack of CS data as well as the often monolingual nature of ASR systems, and of encoders of MT and ST systems. The lack of CS data is often addressed via synthetic data, e.g. as seen in Xu and Yvon (2021); Nakayama et al. (2019). Instead, in this work we derive two novel natural CS datasets from existing public corpora. CS is also difficult for modeling due to its mixed multilingual nature. In order to support multiple languages on the utterance level, automatic language identification (LID) is often performed before applying monolingual systems on a per utterance basis. However, this does not address within-utterance CS, where embedded foreign words and phrases result in recognition errors for monolingual ASR systems, making multilingual models an attractive alternative. Furthermore, CS increases speech recognition errors, significantly increasing the problem of error propagation (Ruiz and Fed-

[1]We make instructions and extra data needed to construct our CS data set available at https://github.com/apple/ml-code-switched-speech-translation

[2]BBC: https://bbc.in/3jgwzZ2

* Work done as an intern.

Findings of the Association for Computational Linguistics: ACL 2022, pages 1435 - 1448
May 22-27, 2022

erico, 2014) in cascaded ST systems, where MT is then performed on the erroneous ASR output. Thus, multilingual end-to-end (E2E) ST systems may be especially appropriate to tackle CS speech.

As both the transcript and translation are important in many CS ST use cases, we focus on the joint transcription and translation ST setting (Anastasopoulos and Chiang, 2018; Weller et al., 2021), extending it to CS data. We follow the methodology of these previous works and focus on the triangle E2E ST model to jointly generate both a transcript of the CS utterance and a translation of that utterance into text containing only one language (c.f. Figure 1 for an illustration). We perform a comparison along two axes: (1) comparing this E2E model to the standard cascaded ST systems, and (2) exploring the difference between bilingual systems and primarily monolingual systems gated by utterance-level LID. Following recent work that has shown the effectiveness of pre-trained models for ST (Li et al., 2020; Gállego et al., 2021), we use Wav2Vec 2.0 (Baevski et al., 2020) as our encoder model and the multilingual mBART 50-50 (Tang et al., 2020) as our decoder model.

We also make several modeling contributions in order to use these pre-trained models for joint transcription and translation. For the E2E ST model, we extend Li et al. (2020) to adapt the mBART decoder to jointly produce both transcription and translation. Furthermore, we introduce a triangle E2E ST model with a shared bilingual decoder and show that this improves transcription and translation accuracy. Our model analysis shows a surprising amount of robustness to CS speech, with the amount (or proportion) of CS words in a sentence not affecting model accuracy. Overall, we observe strong accuracy scores (WER, BLEU) on the CS task, both without CS training data and in the low-resource setting. We believe this opens the door to new and exciting progress in this area.

2 Related Work

Code-switching in NLP has seen a rise of interest in recent years, including a dedicated workshop starting in 2014 (Diab et al., 2014) and still ongoing (Solorio et al., 2021). CS in machine translation also has a long history (Le Féal, 1990; Climent et al., 2003; Sinha and Thakur, 2005; Johnson et al., 2017; Elmadany et al., 2021; Xu and Yvon, 2021), but has seen a rise of interest with the advent of large multilingual models such as mBART (Liu et al., 2020) or mT5 (Xue et al., 2020; Gautam et al., 2021; Jawahar et al., 2021). Due to the lack of available CS data and the ease of single-word translation, most of these recent related MT works have synthetically created CS data for either training or testing by translating one or more of the words in a sentence (Song et al., 2019; Nakayama et al., 2019; Xu and Yvon, 2021; Yang et al., 2020). We differ from those works by using naturally occurring CS data (Section 3) which models the real-world CS distribution rather than arbitrary language mixing.

For spoken input, as present in ASR and ST, synthetically creating realistic CS data is more challenging than it is for MT. However, dedicated ASR corpora that contain natural CS exist, including the Bangor Miami (Deuchar et al., 2014), SEAME (Zeng et al., 2018), and the recent large-scale ASRU 2019 task (Shi et al., 2020). These corpora generally do not contain translations of the ASR annotations, since they were designed for the ASR task only. However, there exist two exceptions, which we leverage to derive our ST CS data set, described in Section 3.

There also exists a wide range of prior modeling work on CS in ASR models, for a variety of strategies (Lyu et al., 2006; Ahmed and Tan, 2012; Seki et al., 2018; Luo et al., 2018; Lu et al., 2020; Du et al., 2021; Zhang et al., 2021). However, the recently introduced large multilingual models for speech, such as Wav2Vec, Wav2Vec 2.0, Schneider et al. (2019); Baevski et al. (2020) and HuBERT (Hsu et al., 2021), are still underexplored with regards to their CS performance.

Handling mixed languages also requires understanding what languages are being spoken. Systems that support mixed language input therefore require some form of automatic LID – either as an explicit component on the utterance (Mabokela et al., 2014; Xu and Yvon, 2021) or word-level (Lyu and Lyu, 2008a; Nakayama et al., 2019), or implicitly learned by the underlying model(s) via a multi-task learning setup (Lyu and Lyu, 2008b; Watanabe et al., 2017; Hou et al., 2020). In our work, we leverage both, exploring utterance-level LID components as well as implicit learning of utterance and word level LID.

In both MT and ASR, prior publications have also included the study of intra-word mixing of languages (Yılmaz et al., 2018; Mager et al., 2019), a phenomenon we do not explore in our work.

Finally, our work builds off of advances made by Gállego et al. (2021); Li et al. (2020) that show that combining large multilingual speech and text models provide consistent improvements. We differ however, by exploring ST in the novel CS setting.

3 Task Description & Data Used

3.1 Task Description

We investigate systems suitable for bilingual English/Spanish conversational scenarios where some of the English and Spanish utterances may include some amount of words and phrases of the respective other language. That is, we are focusing on ST systems that can automatically and seamlessly handle utterances that are either purely English, purely Spanish, English with some Spanish words/phrases embedded or Spanish with some English words/phrases embedded. For transcription, we aim for models to generate the exact mixed-language transcript with each word written in its original spoken language. For translation, we aim to generate purely monolingual translations. See Figure 1 for an example. The experiments and results presented in this paper focus on translating into monolingual English only due to data availability, although we expect similar results for Spanish translations, due the bidirectional model training on standard ST data (Appendix D). We will leave it to future work to more closely examine translation into Spanish – or even a third language not present in the original utterance.

It must be noted that word-level language categorization is sometimes ambiguous. A word in one language may also be considered part of a different language. That is for example true for *loan words* (Baugh, 1935), e.g., *e-mail* in many non-English languages such as German. This issue can be further complicated by attempting to categorize what language named entities fall under: is a Spanish speaker saying *Joe Biden* or *New York* code-switching? Although we acknowledge the complexity of separating words between languages, our work, following previous work (Modipa et al., 2013; Nakayama et al., 2018), uses data annotated by crowd-sourced workers, counting any sentence annotated as having a least one foreign word as being CS. This approach also makes intuitive sense for speech, as the CS words (classified as foreign) will have phonemes that will align more with the embedded language, while the non-CS phonemes will align more with the matrix language.

3.2 Code-Switched Speech Datasets

We use the Fisher (Cieri et al., 2004) and Bangor Miami[3] (Deuchar et al., 2014) corpora for CS data, as they are the only publicly available corpora we are aware of that contains both annotated CS ASR transcripts, as well as translations of those transcripts (Table 1). Although these corpora contain the translations, to our knowledge they have not been used to study CS translation before.

The Miami corpus was collected for linguistic code-switching analysis and gathered from recorded conversations between bilingual English/Spanish speakers in casual settings, primarily in Miami, Florida. These conversations include a high proportion of naturally occurring CS speech. However, in order to collect these naturally occurring conversations, the participants were recorded throughout their day using a small digital recorder worn on belts and lapels. Due to this, the Miami audio contains lower audio quality and much noiser background conditions than standard ASR datasets.

The Fisher dataset was collected for ASR and was gathered by pairing sets of Spanish speakers, located in the U.S. and Canada, to each other through phone calls. Although the Fisher dataset is not a CS focused dataset, we found that it contains a large amount of (annotated) CS utterances, due to the speakers being situated in English-speaking contexts. The recording method (phone recordings in 2004) makes this a noisy ASR dataset, although significantly less so than Miami.

To prepare the data for the joint ST CS task, we separate the data with CS utterances (utterances that contain at least one word annotated as CS) from those with none, creating a CS set and a monolingual set for each dataset. We note that for the Miami dataset the monolingual split contains both English-only and Spanish-only monolingual audio. As the Miami corpus was also annotated with both ambiguous and unambiguous code-switching, we only include utterances in the CS set if the annotations were tagged as unambiguously code-switched (i.e. excluding words such as *ok*, *aha*, and named entities). The Fisher CS dataset consists of majority (matrix[4]) Spanish 77% of the time, English-majority 17%, and 6% evenly

[3]Online audio files can be found at `https://biling.talkbank.org/access/Bangor/Miami.html`

[4]For simplicity, we use the terms majority/matrix language and minority/embedded language interchangeably.

Dataset	Raw Transcript	Clean Transcript
Fisher	un <foreign lang=”English”> show <\foreign>, a mi me gusta ver mucho estos <foreign lang=”English”> shows <\foreign> de la medicina forense	un show, a mi me gusta ver mucho estos shows de la medicina forense
Miami	hay una [/] una que dice (.) it’s@s:eng five@s:eng o’clock@s:eng somewhere@s:eng	hay una una que dice it’s five o’clock somewhere

Table 1: Examples of the raw and clean data for Miami and Fisher. Text in red indicates English text while blue text indicates Spanish. The Miami dataset uses the CHAT annotation format (MacWhinney and Snow, 1990).

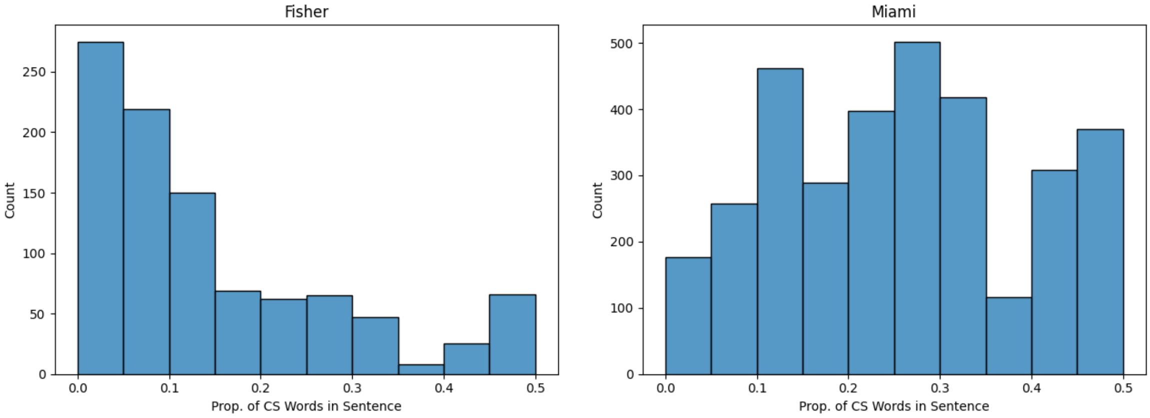

Figure 2: Histogram of the proportions of code-switched words in a sentence for the CS test sets (Fisher on the left, Miami on the right). For example, 0.2 means that 20% of the words in the sentence are CS.

Dataset	Split	Type	Hours	Instances
Miami	Train	Mono	3.60	6,489
	Test	CS	2.82	3,296
		Mono	3.61	6,490
Fisher	Train	CS	13.28	7,398
		Mono	157.3	130,600
	Dev	CS	1.45	821
	Test	CS	1.63	986
		Mono	12.15	10,595

Table 2: Dataset Statistics. CS stands for Code-Switched and Mono for Monolingual.

split between English/Spanish. For the Miami CS dataset the languages are more evenly distributed, with 51% majority-Spanish, 35% majority-English, and 9% evenly split.[5]

The Fisher data consists of three evaluation sets (Dev/Dev2/Test) that together contain approximately a thousand instances of CS with corresponding translations in monolingual English. We combine them into a Fisher CS *Test* set. The Fisher dataset also contains a large amount of CS utterances in the training set (appx. 8k or 15 hrs) which we use as fine-tuning (90%) and validation data (10%). As the Miami dataset contains no splits, we use all CS data for the test set and split the monolingual data into even train/test sets. We include basic summary statistics in Table 2. Note that when compared to standard ST datasets, these CS ST datasets would be considered low-resource settings.

In Figure 2, we see the proportion of CS words in a sentence for the CS test sets. We note that there are no sentences with more than 50% of the words CS since the minority language cannot be more than 50% by definition. For instances that are exactly 50% code switched their language identification was chosen by randomly selecting either English or Spanish. We see that for the Fisher dataset there are more sentences with less than 15% CS with a small uptick around 50%. For Miami it is more uniform, with a large amount of sentences being approximately 25% CS.

To prepare our models for Spanish-English CS, we use the CoVoST (Wang et al., 2020a,b) and MuST-C (Cattoni et al., 2019) datasets for standard

[5]To make these CS datasets reproducible for the broader ST community, we provide a file with instructions for gathering the data (as Fisher is part of the LDC library) as well as files containing a mapping between the original dataset indices to the CS data splits.

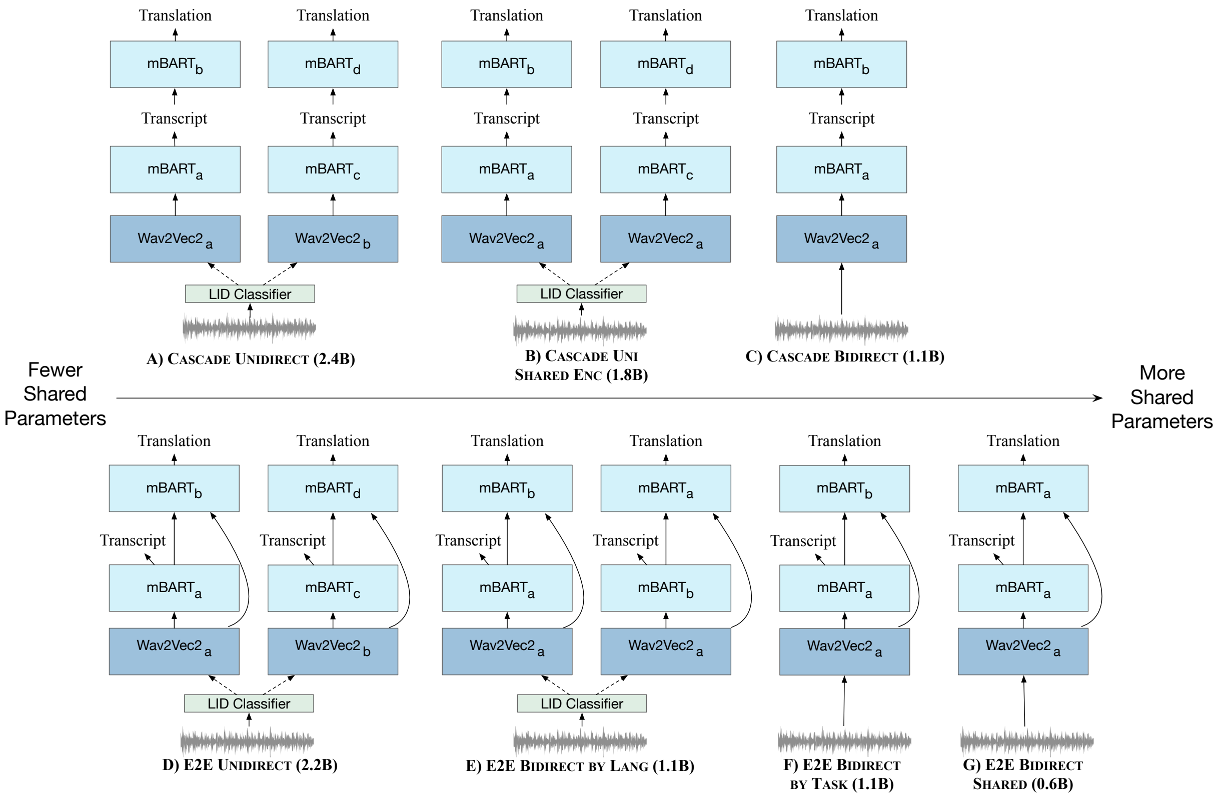

Figure 3: Illustration of model architectures, with cascaded architectures on the top and E2E architectures on the bottom. Left to right shows the progression of models with the least and the most amount of shared parameters respectively. Subscripts are present to indicate shared modules within each model. Dotted lines indicate a decision where only one path is chosen using the LID. Note that there is no cascade equivalent to the BIDIRECTIONAL E2E SHARED model, as the cascaded model by definition generates transcript then translation separately. The numbers in parentheses stands for the number of model parameters in billions.

ST training, as CoVoST contains only Es→En and MuST-C contains only En→Es. Although high scores on these datasets are not our primary target, we note that our scores come close to or improve the state of the art (SoTA) on these tasks (see Appendix A, Table 9) albeit with different data used in training, showing that our base ST models are representative of current SoTA techniques.

4 Experimental Settings

4.1 Models

Joint Transcript/Translation Models Many different types of E2E models exist for joint transcript/translation ST (Sperber and Paulik, 2020). Here, we focus on the *triangle* E2E architecture due to its strong performance in previous work (Anastasopoulos and Chiang, 2018; Sperber et al., 2020). Following recent work (Gállego et al., 2021; Li et al., 2020) we use pre-trained modules as a starting place for our ST model, using a Wav2Vec 2.0 (Baevski et al., 2020) encoder and a mBART 50-50 (Liu et al., 2020; Tang et al., 2020) decoder.

Because our task involves joint ASR and ST, we need to adapt the pre-trained decoder to work with the E2E triangle architecture. Specifically, the triangle model's second decoder computes cross attention separately over both the first decoder and the encoder states. We place an additional cross-attention layer after each encoder-attention layer in mBARTs decoder blocks, initializing them with the pre-trained encoder-attention weights. To make sure these weights converge properly, we freeze the entire model for approximately the first epoch while training only the bridge and additional cross attention layers (c.f. Appendix A).

As described in Section 3, our task involves modeling intra-sentence CS. This means that any model used for this task must either explicitly or implicitly learn to model the language of each word in the sentence. Furthermore, as more than one language is being modeled, each sub-component of the model can either be unidirectional or bidirectional. We can thus categorize potential models by how much information is shared within the parameters: the least shared models would be unidirectional and joined together by explicit LID, whereas the most shared would be bidirectional models that learn the LID implicitly. Models and their categorization along this scale are shown in Figure 3.

For cascade models, the most basic would be separate unidirectional cascaded models joined by an LID model. The LID model will explicitly decide what the matrix language is and send the utterance to the model that is best equipped to handle that language (Figure 3A). Note that this approach may suffer from error propagation issues due to incorrect LID. A more parameter-shared version of this model is to make the cascaded model encoder shared between both unidirectional models (Figure 3B). Finally, we can examine a bidirectional cascade model that shares each component across both languages. This architecture implicitly learns to model the language of the input, removing the need for an explicit LID model (Figure 3C).

We also examine similar analogues for the E2E triangle model: unidirectional models joined by LID (Figure 3D) and a bidirectional model with LID and a shared encoder (Figure 3E). We can also use the standard triangle model (see Anastasopoulos and Chiang (2018) for implementation details) that includes one encoder and two decoders (one for each sub-task) (Figure 3F). Furthermore, we propose to alter the standard triangle model and share both decoder parameters for both languages with a joint bidirectional decoder (Figure 3G, note that the cascade model cannot do this due to the definition of the cascade). By doing so, we hope to provide an inductive bias for the model to more easily handle code-switched data, as the weights of that decoder will already be used to handling multiple languages for both tasks (compared to the bidirectional cascade model, which only shares multilingual parameters for each task of transcript and translation).

Language Identification Model We train the language identification (LID) model to identify the matrix language. For consistency with our other models (and similar to concurrent work, e.g. Tjandra et al. (2021)), we use a pre-trained Wav2Vec2 along with a classifier layer to predict whether the utterance is majority Spanish or majority English. We train the model in the same fashion as the joint transcription and translation models (Section 4.1 and Appendix A) but train on the LID data instead.

The data for the LID model was gathered by taking the CS data[6] from the training set of the Fisher corpus and combining it with randomly sampled data from several different datasets in order to help the model learn despite the domain of the audio. We use MuST-C English audio, CoVoST English audio, CoVoST Spanish audio, and the monolingual Spanish audio from the training sets of Fisher and Miami. We found that upsampling the CS training set by 2 and using the same amount of data (2x the number of the CS set) for CoVoST and MuST-C provided the best results: 98%+ accuracy on CoVoST and MuST-C, 89% on the Fisher CS validation and test sets, and 72% on the Miami CS test set (due to the noisy data). As a large proportion of the CS data is close to 50% code-switched (see Figure 2), it becomes more difficult for the model to predict the matrix language correctly.

4.2 Training Process and Evaluation

For all dataset evaluations, we use word error rate (WER) and character error rate (CER) for the transcript and Charcut (CCT) (Lardilleux and Lepage, 2017) and sacreBLEU (Post, 2018) for the translation. However, we found that there was no difference in conclusions between each of the two metrics (WER vs CER and BLEU vs Charcut) and thus we only report BLEU/WER in the main text (see Appendix A for implementation details). For tables showing all metrics, see Appendix E.

We evaluate our models on the Fisher and Miami test sets (with both CS-only and monolingual-only test sets) in two different settings: (1) without fine-tuning them on CS data (No-FT) and (2) after fine-tuning the already trained ST models on the Fisher CS Training set (FT). For models consisting of two monolingual sub-models we fine-tune both on the CS data. During fine-tuning we employ the same hyperparameters as in the original experiment, but perform early stopping on the Fisher CS Dev set. We use significance tests to verify the reliability of our results (Koehn, 2004). We run bootstrap resampling tests against the best performing model, using $\alpha = 0.05$. More training parameters such as learning rates, etc. can be found in Appendix A.

5 Results

5.1 Scores on Test Sets

In this section, we explore the results of doing ST for CS data along the two axes of unidirectional vs

[6]For the No-FT case (Section 4.2), we exclude the CS data when training the LID model.

Models	Not Fine-Tuned				Fine-Tuned			
	CS		Mono.		CS		Mono.	
	↓ WER	↑ BLEU	↓ WER	↑ BLEU	↓ WER	↑ BLEU	↓ WER	↑ BLEU
CASCADE UNIDIRECT	37.1 (-0.8)	22.5 (-0.4)	26.6 (-3.1)	24.7 (+0.9)	33.5 (-0.4)	24.6 (0.0)	24.8 (-1.0)	25.5 (+0.2)
CASCADE UNI SHARED ENC	36.0 (0.0)	21.6 (+0.6)	25.6 (0.0)	24.3 (+0.5)	31.2 (+0.1)	25.4 (+0.2)	25.6 (-0.3)	24.8 (+0.1)
E2E UNIDIRECT	36.6 (-0.9)	22.3 (-0.1)	26.7 (-3.5)	25.0 (+1.0)	33.4 (-0.2)	24.4 (+0.1)	25.3 (-1.4)	25.5 (+0.4)
E2E BIDIRECT BY LANG	37.0 (-0.9)	23.4 (-0.1)	27.2 (-1.9)	25.0 (+0.5)	36.7 (-0.8)	22.8 (+0.2)	27.3 (-2.0)	25.0 (+0.4)

Table 3: Comparison of Oracle vs Predicted LID results on the Fisher dataset. Numbers in parenthesis are the difference to the corresponding model with oracle LID. Note that the **Oracle LID improves upon the Predicted LID** in most cases. Conclusions are similar for the Miami corpus (see Appendix B Table 7)

	Model	Not Fine-Tuned				Fine-Tuned			
		CS		Mono.		CS		Mono.	
		↓ WER	↑ BLEU	↓ WER	↑ BLEU	↓ WER	↑ BLEU	↓ WER	↑ BLEU
Fisher	CASCADE UNIDIRECT	37.1	22.5	26.6	24.7	33.5	24.6	24.8	25.5
	CASCADE UNI SHARED ENC	36.0	21.6	25.6	24.3	31.2	*25.4	25.6	24.8
	CASCADE BIDIRECT	37.2	21.8	26.5	24.1	33.2	23.2	28.1	23.2
	E2E UNIDIRECT	36.6	22.3	26.7	25.0	33.4	24.4	25.3	25.5
	E2E BIDIRECT BY LANG	37.0	**23.4**	27.2	25.0	36.7	22.8	27.3	25.0
	E2E BIDIRECT BY TASK	*34.1	*23.0	23.6	26.0	*30.1	**25.6**	*24.3	25.6
	E2E BIDIRECT SHARED	**33.8**	*23.3	**23.2**	**26.2**	**30.0**	*25.4	**24.1**	**26.1**
Miami	CASCADE UNIDIRECT	65.2	8.8	52.3	16.8	64.8	10.8	**51.5**	16.8
	CASCADE UNI SHARED ENC	60.2	9.7	53.8	15.7	55.0	**14.7**	55.6	15.3
	CASCADE BIDIRECT	61.4	9.3	54.0	14.8	57.4	10.6	58.2	14.0
	E2E UNIDIRECT	65.6	10.1	53.0	17.2	65.1	11.7	*51.4	**17.6**
	E2E BIDIRECT BY LANG	69.5	**12.4**	55.2	16.5	69.3	11.5	54.5	16.6
	E2E BIDIRECT BY TASK	59.9	11.0	*50.0	*18.1	*53.6	*13.8	52.6	*17.5
	E2E BIDIRECT SHARED	**58.9**	*11.8	**49.9**	**18.3**	**53.0**	*14.1	52.1	*17.4

Table 4: Test set scores, with results from the Fisher corpus on the top half and the Miami corpus on the bottom half. Bold scores indicate the best score in the column, while asterisks indicate results that are statistically similar to the best score in the column group using a bootstrap resampling test with $\alpha = 0.05$.

bidirectional and end-to-end vs cascade.

We see results for models using explicit LID prediction in Table 3, showing that models that use the **predicted LID perform worse than those that use Oracle LID** (e.g. 36.6 vs 35.7 WER for the E2E UNIDIRECT). This provides a slight advantage for the bidirectional models that learn LID implicitly. However, the predicted LID case is the realistic setting, and thus we use it for the remainder of our experiments.

When we examine the models along the scale of unidirectional to bidirectional, we see that **higher amounts of shared parameters are correlated with higher scores**, e.g. bidirectional is better. We see that on all datasets and evaluation settings (Table 4) that the E2E BIDIRECT SHARED model is either statistically similar or outperforms all other models, except for the Miami Monolingual FT case, where it comes in 3rd. Thus, the inductive bias of

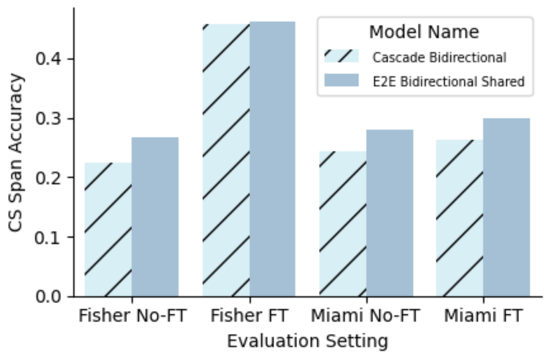

Figure 4: Accuracy of the models in generating the CS spans. Note that this excludes all non-exact matches and is a lower bound on performance.

sharing the multilingual task parameters provides a gain of approximately 3.5 WER points (33.8 vs 37.3) and 1.5 BLEU points (23.3 vs 21.9) for the E2E BIDIRECT SHARED model over the E2E UNIDIRECT model on the Fisher dataset, with similar

Model	Transcript	Translation
Reference	si entonces volví aquí a la casa si el fall break	yes so I returned here to the house yes the fall break
Cascade	si entonces volví aquí a la casa si es folvereak	yes then I returned here at home yes its folvereak
E2E	si entonces volví aquí a la casa si es fallbreak	yes so I came back to the house yes its fallbreak

Table 5: Example generated output from the CASCADE BIDIRECT and E2E BIDIRECT SHARED models. Note the error propagation in the cascade model.

performance on the Miami dataset.

We can also examine Table 4 to see how the cascade models compare to the E2E models. The results show that the **cascaded models perform the same or worse than the E2E models** they compare to w.r.t. parameter sharing, with the best overall model being the E2E BIDIRECT SHARED, beating the CASCADE BIDIRECT (e.g. 33.8 vs 37.2 WER or 23.3 vs 21.8 BLEU on Fisher No-FT).

Table 4 also illustrates that **fine-tuning models on CS data improves scores** on CS test sets (33.8 vs 30.0 WER for the E2E BIDIRECT SHARED on Fisher, 58.9 vs 53.0 for Miami). These gains are consistent for the Fisher dataset, which is the domain of the CS training set, however there are still gains for the out-of-domain Miami CS data. These results suggest that additional pre-training on natural or synthetic data (in both audio/text modalities) would likely be fruitful future work. When we examine how fine-tuning on CS data changes the model's monolingual scores, we find that they generally improve the monolingual results for the unidirectional models, but tend to make bidirectional models slightly worse, perhaps due to interference between the languages and tasks in the same weights. However, overall we find that fine-tuning provides large gains for CS with only minor decreases in monolingual performance.

5.2 Model Analysis

We also provide further analysis of the CS output of the best model and its cascaded counterpart (BIDIRECT CASCADE and E2E BIDIRECT SHARED). We perform three analyses: (1) comparing utterance level scores vs the proportion of CS words in the utterance, (2) computing the exact match accuracy of the CS spans in the model's output, and (3) qualitatively examining model output.

We check the correlation between the proportion of CS words in a sentence and the model's score, using a linear model to find the R^2 values. We found that surprisingly, there was **no correlation between the proportion of CS words and the models score** for any of the different models or metrics ($R^2 < 0.025$ for all models and metrics). A graphical depiction of the model's scores over CS proportions is in the Appendix, Figure 5. We note that this finding was the same for comparing the *number* of CS words instead of the *proportion*. This finding implies that the models are surprisingly robust to the amount of CS in a sentence.

Although BLEU and WER scores show how well the models do on the CS data, we can further isolate the performance of these models on only the code-switched parts of the utterances. To do so, we isolate all CS spans in the sentences and check to see if the model's output contains the exact-match of those spans. We note that this metric does not take into account synonyms or different tenses of the same word, making it a stricter metric serving as a lower bound of absolute performance. We see in Figure 4 that the E2E model still outperforms the cascade on CS spans, with Fisher No-FT scores around 20-30% and Fisher FT scores around 45%.

Finally, we can also examine the model's outputs. We inspected 200 output sentences for the monolingual subsets and found that both models generated the correct language in every case, indicating that they correctly learned the implicit LID. However, we can see that the cascade model does struggle with error propagation (especially so in the CS setting, Table 5), likely causing part of the difference between the E2E and cascade models.

Although the CS WER and BLEU scores are not as high as they are on cleaner monolingual datasets such as CoVoST (Appendix A), their performance is competitive with their respective monolingual performance on Miami and Fisher, even in the No-FT setting. We believe that with additional data and improvements ST models will be well-equipped to handle CS in practical situations and that **overall, models show strong CS performance**.

6 Conclusion

In this work, we expand the ST literature to explore code-switching, contributing a new task framework for ST that extends the joint transcription and translation setup. To further progress, we built and opensourced a new ST corpus for CS from existing public datasets. We evaluated a range of models, showing that using bilingual joint decoders provides gains over using separate task decoders. We also showed that E2E systems provide better performance than their cascading counterparts on the CS task. Overall, our work shows that ST models can perform well on CS applications with both no fine-tuning and in low-resource settings, opening the door to new and exciting areas of future work.

References

Basem HA Ahmed and Tien-Ping Tan. 2012. Automatic speech recognition of code switching speech using 1-best rescoring. In *2012 International Conference on Asian Language Processing*, pages 137–140. IEEE.

Antonios Anastasopoulos and David Chiang. 2018. Tied multitask learning for neural speech translation. *arXiv preprint arXiv:1802.06655*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Albert C Baugh. 1935. The chronology of french loan-words in english. *Modern Language Notes*, 50(2):90–93.

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. Must-c: A multilingual corpus for end-to-end speech translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: A resource for the next generations of speech-to-text. In *LREC*, volume 4, pages 69–71.

Salvador Climent, Joaquim Moré, Antoni Oliver, Míriam Salvatierra, Imma Sànchez, Mariona Taulé, and Lluïsa Vallmanya. 2003. Bilingual newsgroups in catalonia: A challenge for machine translation. *Journal of Computer-Mediated Communication*, 9(1):JCMC919.

Margaret Deuchar, Peredur Davies, Jon Russell Herring, M Carmen Parafita Couto, and Diana Carter. 2014. 5. building bilingual corpora. In *Advances in the Study of Bilingualism*, pages 93–110. Multilingual Matters.

Mona Diab, Julia Hirschberg, Pascale Fung, and Thamar Solorio. 2014. Proceedings of the first workshop on computational approaches to code switching. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*.

Chenpeng Du, Hao Li, Yizhou Lu, Lan Wang, and Yanmin Qian. 2021. Data augmentation for end-to-end code-switching speech recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 194–200. IEEE.

AbdelRahim Elmadany, Muhammad Abdul-Mageed, et al. 2021. Investigating code-mixed modern standard arabic-egyptian to english machine translation. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 56–64.

Gerard I Gállego, Ioannis Tsiamas, Carlos Escolano, José AR Fonollosa, and Marta R Costa-jussà. 2021. Upc's speech translation system for iwslt 2021. *arXiv preprint arXiv:2105.04512*.

Devansh Gautam, Prashant Kodali, Kshitij Gupta, Anmol Goel, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021. Comet: Towards code-mixed translation using parallel monolingual sentences. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 47–55.

Roberto R Heredia and Jeanette Altarriba. 2001. Bilingual language mixing: Why do bilinguals code-switch? *Current Directions in Psychological Science*, 10(5):164–168.

Wenxin Hou, Yue Dong, Bairong Zhuang, Longfei Yang, Jiatong Shi, and Takahiro Shinozaki. 2020. Large-scale end-to-end multilingual speech recognition and language identification with multi-task learning. In *INTERSPEECH*, pages 1037–1041.

Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: How much can a bad teacher benefit asr pre-training? In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6533–6537. IEEE.

Ganesh Jawahar, El Moatez Billah Nagoudi, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. 2021. Exploring text-to-text transformers for english to hinglish machine translation with synthetic code-mixing. *arXiv preprint arXiv:2105.08807*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395.

Adrien Lardilleux and Yves Lepage. 2017. Charcut: Human-targeted character-based mt evaluation with loose differences. In *Proceedings of IWSLT 2017*.

Karla Déjean Le Féal. 1990. A different approach to machine translation. *Meta: journal des traducteurs*, 35(4):710–719.

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2020. Multilingual speech translation with efficient finetuning of pretrained models. *arXiv preprint arXiv:2010.12829*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yizhou Lu, Mingkun Huang, Hao Li, Jiaqi Guo, and Yanmin Qian. 2020. Bi-encoder transformer network for mandarin-english code-switching speech recognition using mixture of experts. In *INTERSPEECH*, pages 4766–4770.

Ne Luo, Dongwei Jiang, Shuaijiang Zhao, Caixia Gong, Wei Zou, and Xiangang Li. 2018. Towards end-to-end code-switching speech recognition. *arXiv preprint arXiv:1810.13091*.

D. Lyu and Ren-Yuan Lyu. 2008a. Language identification on code-switching utterances using multiple cues. In *INTERSPEECH*.

Dau-Cheng Lyu and Ren-Yuan Lyu. 2008b. Language identification on code-switching utterances using multiple cues. In *Ninth Annual Conference of the International Speech Communication Association*.

Dau-Cheng Lyu, Ren-Yuan Lyu, Yuang-chin Chiang, and Chun-Nan Hsu. 2006. Speech recognition on code-switching among the chinese dialects. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE.

Koena Ronny Mabokela, Madimetja Jonas Manamela, and Mabu Manaileng. 2014. Modeling code-switching speech on under-resourced languages for language identification. In *Spoken Language Technologies for Under-Resourced Languages*.

Brian MacWhinney and Catherine Snow. 1990. The child language data exchange system: An update. *Journal of child language*, 17(2):457–472.

Manuel Mager, Özlem Çetinoğlu, and Katharina Kann. 2019. Subword-level language identification for intra-word code-switching. In *arXiv preprint arXiv:1904.01989*.

Thipe Modipa, Marelie Hattingh Davel, and F. D. Wet. 2013. Implications of sepedi/english code switching for asr systems. In *researchgate*.

Carol Myers-Scotton. 1995. *Social motivations for codeswitching: Evidence from Africa*. Oxford University Press.

Carol Myers-Scotton and William Ury. 1977. Bilingual strategies: The social functions of code-switching. *International journal of the sociology of language*, pages 5–20.

Sahoko Nakayama, Takatomo Kano, Andros Tjandra, S. Sakti, and Satoshi Nakamura. 2019. Recognition and translation of code-switching speech utterances. *2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6.

Sahoko Nakayama, Andros Tjandra, S. Sakti, and Satoshi Nakamura. 2018. Speech chain for semi-supervised learning of japanese-english code-switching asr and tts. *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 182–189.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Nicholas Ruiz and Marcello Federico. 2014. Assessing the impact of speech recognition errors on machine translation quality. In *11th Conference of the Association for Machine Translation in the Americas (AMTA), Vancouver, BC, Canada*.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Hiroshi Seki, Shinji Watanabe, Takaaki Hori, Jonathan Le Roux, and J. Hershey. 2018. An end-to-end language-tracking speech recognizer for mixed-language speech. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4919–4923.

Xian Shi, Qiangze Feng, and Lei Xie. 2020. The asru 2019 mandarin-english code-switching speech recognition challenge: Open datasets, tracks, methods and results. *arXiv preprint arXiv:2007.05916*.

R Mahesh K Sinha and Anil Thakur. 2005. Machine translation of bi-lingual hindi-english (hinglish) text. *10th Machine Translation summit (MT Summit X), Phuket, Thailand*, pages 149–156.

Thamar Solorio, Shuguang Chen, Alan W. Black, Mona Diab, Sunayana Sitaram, Victor Soto, Emre Yilmaz, and Anirudh Srinivasan, editors. 2021. *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*. Association for Computational Linguistics, Online.

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and M. Zhang. 2019. Code-switching for enhancing nmt with pre-specified translation. In *NAACL*.

Matthias Sperber and Matthias Paulik. 2020. Speech Translation and the End-to-End Promise: Taking Stock of Where We Are. In *Association for Computational Linguistic (ACL)*, Seattle, USA.

Matthias Sperber, Hendra Setiawan, Christian Gollan, Udhyakumar Nallasamy, and Matthias Paulik. 2020. Consistent Transcription and Translation of Speech. *Transactions of the Association for Computational Linguistics (TACL)*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Andros Tjandra, Diptanu Gon Choudhury, Frank Zhang, Kritika Singh, Alexei Baevski, Assaf Sela, Yatharth Saraf, and Michael Auli. 2021. Improved language identification through cross-lingual self-supervised learning. *arXiv preprint arXiv:2107.04082*.

Ngoc Thang Vu, Dau-Cheng Lyu, Jochen Weiner, Dominic Telaar, Tim Schlippe, Fabian Blaicher, Eng-Siong Chng, Tanja Schultz, and Haizhou Li. 2012. A first speech recognition system for mandarin-english code-switch conversational speech. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4889–4892. IEEE.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020a. Covost: A diverse multilingual speech-to-text translation corpus. *arXiv preprint arXiv:2002.01320*.

Changhan Wang, Anne Wu, and Juan Pino. 2020b. Covost 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310*.

Shinji Watanabe, Takaaki Hori, and John R Hershey. 2017. Language independent end-to-end architecture for joint language identification and speech recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 265–271. IEEE.

Orion Weller, Matthias Sperber, Christian Gollan, and Joris Kluivers. 2021. Streaming models for joint speech recognition and translation. *arXiv preprint arXiv:2101.09149*.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. Are multilingual models effective in code-switching? *ArXiv*, abs/2103.13309.

Jitao Xu and François Yvon. 2021. Can you traducir this? machine translation for code-switched input. *arXiv preprint arXiv:2105.04846*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang, and Qi Ju. 2020. Csp: Code-switching pre-training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2624–2636.

Emre Yılmaz, Astik Biswas, Ewald van der Westhuizen, Febe de Wet, and Thomas Niesler. 2018. Building a unified code-switching asr system for south african languages. *arXiv preprint arXiv:1807.10949*.

Xianghu Yue, Grandee Lee, Emre Yılmaz, Fang Deng, and Haizhou Li. 2019. End-to-end code-switching asr for low-resourced language pairs. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 972–979. IEEE.

Zhiping Zeng, Yerbolat Khassanov, Van Tung Pham, Haihua Xu, Eng Siong Chng, and Haizhou Li. 2018. On the end-to-end solution to mandarin-english code-switching speech recognition. *arXiv preprint arXiv:1811.00241*.

Shuai Zhang, Jiangyan Yi, Zhengkun Tian, Jianhua Tao, and Ye Bai. 2021. Rnn-transducer with language bias for end-to-end mandarin-english code-switching speech recognition. In *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5. IEEE.

Models	CS		Mono.	
	↓ WER	↑ BLEU	↓ WER	↑ BLEU
Random Init	69.6	11.0	59.6	13.2
Pre-trained	33.8	23.3	23.2	26.2

Table 6: Comparison of the E2E bidirectional shared model with pre-training vs random initialization on the Fisher code-switched test sets.

A Training and Evaluation Details

We follow Gállego et al. (2021); Li et al. (2020) and use a triangular learning rate, adapting the step count to depend on the batch size (as not all models could fit the same batch size) with (64 / batch size) * 500 warm up steps, (64 / batch size) * 500 hold steps, (64 / batch size) * 3000 decay steps, a beta of 0.9, and a beta2 of 0.98. The learning rate was selected from running a search over {0.01, 0.005, 0.001, 0.0005, 0.0001, 0.0005}. We found that 0.0005 was best for all models, so we examined learning rates again between 0.0001 to 0.001 (by 0.0001) and found that they all performed similarly, thus we use 0.0005 in our experiments. For efficiency in batch size while training, we removed all instances whose audio length was longer than 20 seconds. We freeze the attention layers for the first 500 * (64 / batch size) steps, which is approximately the first epoch of training.

We initially trained the models on only CoVoST and MuST-C and found there was a large domain shift between these datasets and the comparatively noisier Fisher and Miami datasets. As domain shift was not the focus of this paper, we further trained the models on the Fisher and Miami monolingual training sets to reduce the effect of domain shift.

As a sanity check of the effectiveness of our training, we also include scores in Table 9 for the test sets of CoVoST and MuST-C. We note that our scores are close to the SoTA scores of Li et al. (2020) on CoVoST (and they use the large Wav2Vec2 model while we use the base version) and our MuST-C scores are higher than that of Gállego et al. (2021).

We evaluate using word error rate, character error rate, charcut, and BLEU. As the models learn different punctuation techniques from a variety of sources, including MuST-C, CoVoST, Miami, and Fisher, we remove all punctuation from the output before evaluating on the CS/Mono test sets, in order to only measure scores on the content. For BLEU, we use SacreBLEU with parameters case.lc+numrefs.1+smooth.4.0+tok.13a.

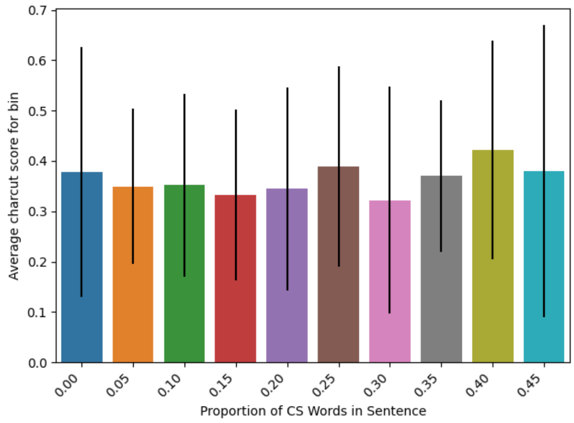

Figure 5: Charcut performance of the E2E BIDIRECT SHARED model on sentences with various levels of CS proportions. Note that there is no clear correlation, as described in Section 5.2. Black lines indicate error bars of 2 standard deviations while the bar represents the average.

B More LID Comparisons

We show results for all models that use LID on both datasets in Table 7. Note the conclusions remain the same as Table 3.

C Random Initialization Results

We also perform an ablation of these pre-trained scores (Table 6) for the E2E BIDIRECT SHARED model, as it is the best performing model overall. We tried many different setups for training it from from scratch rather than loading the pre-trained weights. We found that it was very difficult for this model to converge, and when it did, the results were sub-par.

D Training Results

We include the scores of evaluating our models on the test sets of the ST training data (MuST-C and CoVoST) in Table 9. We also include the results of fine-tuning performance on the CS dev set in Table 8, which roughly mirrors the main results.

E Expanded Results

For brevity, we do not include the CER and Charcut metrics in the main text. In this section we included tables with all metrics for all results (Table 10 for Miami and Table 11 for Fisher). We note however, that the WER and BLEU scores align with the CER and Charcut scores, and thus our conclusions remain the same.

	Model	No Fine-Tuning				Fine-Tuned			
		CS		Mono.		CS		Mono.	
		↓ WER	↑ BLEU	↓ WER	↑ BLEU	↓ WER	↑ BLEU	↓ WER	↑ BLEU
Fisher	CASCADE UNIDIRECT	37.1	22.5	26.6	24.7	33.5	24.6	24.8	25.5
		(36.3)	(22.1)	(23.5)	(25.6)	(33.1)	(24.6)	(23.8)	(25.7)
	E2E UNIDIRECT	36.6	22.3	26.7	25.0	33.4	24.4	25.3	25.5
		(35.7)	(22.2)	(23.2)	(26.0)	(33.2)	(24.5)	(23.9)	(25.9)
	CASCADE UNI SHARED ENC	36.0	21.6	25.6	24.3	31.2	25.4	25.6	24.8
		(36.0)	(22.2)	(25.6)	(24.8)	(31.3)	(25.6)	(25.3)	(24.9)
	E2E BIDIRECT BY LANG	37.0	23.4	27.2	25.0	36.7	22.8	27.3	25.0
		(36.1)	(23.3)	(25.3)	(25.5)	(35.9)	(23.0)	(25.3)	(25.4)
Miami	CASCADE UNIDIRECT	65.2	8.8	52.3	16.8	64.8	10.8	51.5	16.8
		(61.4)	(8.3)	(50.0)	(17.3)	(64.4)	(10.8)	(50.9)	(16.9)
	E2E UNIDIRECT	65.6	10.1	53.0	17.2	65.1	11.7	51.4	17.6
		(63.1)	(9.4)	(51.2)	(17.7)	(65.6)	(11.7)	(50.7)	(17.7)
	CASCADE UNI SHARED ENC	60.2	9.7	53.8	15.7	55.0	14.7	55.6	15.3
		(60.2)	(8.8)	(53.8)	(16.0)	(56.0)	(14.4)	(55.5)	(15.3)
	E2E BIDIRECT BY LANG	69.5	12.4	55.2	16.5	69.3	11.5	54.5	16.6
		(69.7)	(10.7)	(53.4)	(16.7)	(69.7)	(10.4)	(53.2)	(16.6)

Table 7: Scores on the code-switched test sets for the models using LID, with results from zero CS training on the left and results after fine-tuning on the right.

Models	Fisher CS Dev Set			
	↓ WER	↓ CER	↓ CCT	↑ BLEU
CASCADE UNIDIRECT	34.2	19.3	38.4	26.4
E2E UNIDIRECT	33.0	18.9	37.3	27.8
CASCADE UNI SHARED ENC	32.3	17.9	38.4	24.9
E2E BIDIRECT BY LANG	36.3	23.0	39.3	26.3
E2E BIDIRECT BY TASK	31.1	17.0	35.1	29.0
CASCADE BIDIRECT	35.1	19.2	39.7	23.8
E2E BIDIRECT SHARED	31.7	17.5	35.2	28.3

Table 8: Scores on the Fisher CS Dev set. CCT stands for Charcut. Note that this mirrors the main results in Table 4.

Models	MuST-C Test Set				CoVoST Test Set			
	↓ WER	↓ CER	↓ CCT	↑ BLEU	↓ WER	↓ CER	↓ CCT	↑ BLEU
CASCADE UNIDIRECT	11.2	7.6	36.3	29.4	17.2	5.8	35.6	26.9
E2E UNIDIRECT	13.0	8.9	37.3	27.8	18.6	6.4	36.0	26.2
CASCADE UNI SHARED ENC	12.0	8.1	37.7	26.9	22.9	7.3	36.2	26.0
E2E BIDIRECT BY LANG	11.6	7.8	36.6	28.6	19.7	7.7	37.1	25.4
E2E BIDIRECT BY TASK	11.4	7.6	36.6	28.4	17.9	6.0	35.3	26.8
CASCADE BIDIRECT	13.6	9.5	39.7	24.5	22.9	7.3	38.8	22.8
E2E BIDIRECT SHARED	11.6	7.7	36.6	28.5	18.1	6.2	35.0	27.4

Table 9: Scores on the MustC and CovoST datasets. CCT stands for Charcut.

	Models	Miami							
		CS Test Set				Monolingual Test Set			
		↓ WER	↓ CER	↓ CCT	↑ BLEU	↓ WER	↓ CER	↓ CCT	↑ BLEU
Not Fine-Tuned	CASCADE UNIDIRECT	63.6	43.3	65.2	8.7	52.3	34.1	51.9	17.0
	CASCADE UNIDIRECT ORA.	61.4	41.6	67.4	8.3	50.0	32.4	50.9	17.3
	E2E UNIDIRECT	64.0	43.0	64.0	9.9	53.0	34.6	51.0	17.4
	E2E UNIDIRECT ORA.	63.1	42.2	66.5	9.4	51.2	33.1	50.1	17.7
	CASCADE UNI SHARED ENC	60.2	39.7	63.7	9.3	53.8	34.1	52.7	15.9
	CASCADE UNI SHARED ENC ORA.	60.2	39.7	66.1	8.8	53.8	34.1	52.2	16.0
	E2E BIDIRECT BY LANG	68.8	48.4	61.2	11.5	54.6	37.3	52.0	16.7
	E2E BIDIRECT BY LANG ORA.	69.7	49.4	63.6	10.7	53.4	36.3	51.5	16.7
	E2E BIDIRECT BY TASK	59.9	39.6	59.4	11.0	50.0	32.6	49.7	18.1
	CASCADE BIDIRECT	61.4	39.8	62.2	9.3	54.0	34.1	53.1	14.8
	E2E BIDIRECT SHARED	58.9	39.1	58.5	11.8	49.9	32.2	49.3	18.3
Fine-Tuned	CASCADE UNIDIRECT	64.8	42.0	56.5	10.8	51.5	33.4	51.1	16.8
	CASCADE UNIDIRECT ORA.	64.4	41.8	56.4	10.8	50.9	32.9	50.6	16.9
	E2E UNIDIRECT	65.1	43.0	56.9	11.7	51.4	33.7	50.4	17.6
	E2E UNIDIRECT ORA.	65.6	43.1	57.0	11.7	50.7	33.2	49.9	17.7
	CASCADE UNI SHARED ENC	55.0	35.2	51.4	14.7	55.6	35.9	52.9	15.3
	CASCADE UNI SHARED ENC ORA.	56.0	35.7	51.7	14.4	55.5	35.9	52.7	15.3
	E2E BIDIRECT BY LANG	69.3	48.6	61.3	11.5	54.5	37.2	52.1	16.6
	E2E BIDIRECT BY LANG ORA.	69.7	49.5	63.8	10.4	53.2	36.1	51.5	16.6
	E2E BIDIRECT BY TASK	53.6	35.0	53.3	13.8	52.6	34.4	50.5	17.5
	CASCADE BIDIRECT	57.4	36.3	58.8	10.6	58.2	36.6	55.1	14.0
	E2E BIDIRECT SHARED	53.0	35.0	54.4	14.1	52.1	33.9	50.4	17.4

Table 10: Scores on the Miami dataset. CCT stands for Charcut. Results from zero CS training are on the top half and results after fine-tuning are on the bottom half. Ora stands for Oracle.

	Models	Fisher							
		CS Test Set				Monolingual Test Set			
		↓ WER	↓ CER	↓ CCT	↑ BLEU	↓ WER	↓ CER	↓ CCT	↑ BLEU
Not Fine-Tuned	CASCADE UNIDIRECT	37.3	22.2	45.6	21.9	28.0	15.3	40.0	24.4
	CASCADE UNIDIRECT ORA.	36.3	21.5	45.0	22.1	23.5	12.0	38.0	25.6
	E2E UNIDIRECT	36.9	22.0	45.1	21.8	28.2	15.6	39.7	24.7
	E2E UNIDIRECT ORA.	35.7	21.3	44.4	22.2	23.2	12.0	37.6	26.0
	CASCADE UNI SHARED ENC	36.0	20.5	44.7	21.9	25.6	13.0	39.8	24.3
	CASCADE UNI SHARED ENC ORA.	36.0	20.5	44.2	22.2	25.6	13.0	38.8	24.8
	E2E BIDIRECT BY LANG	36.9	23.6	43.0	23.2	27.2	15.5	39.2	25.1
	E2E BIDIRECT BY LANG ORA.	36.1	22.9	42.6	23.3	25.3	14.0	38.4	25.5
	E2E BIDIRECT BY TASK	34.1	19.4	42.3	23.0	23.6	11.9	37.4	26.0
	CASCADE BIDIRECT	37.2	21.3	43.8	21.8	26.5	13.3	39.5	24.1
	E2E BIDIRECT SHARED	33.8	19.3	41.5	23.3	23.2	11.8	37.1	26.2
Fine-Tuned	CASCADE UNIDIRECT	33.5	18.5	39.6	24.6	24.8	12.9	38.2	25.5
	CASCADE UNIDIRECT ORA.	33.1	18.4	39.4	24.6	23.8	12.1	37.7	25.7
	E2E UNIDIRECT	33.4	19.1	40.0	24.4	25.3	13.3	38.3	25.5
	E2E UNIDIRECT ORA.	33.2	19.0	39.9	24.5	23.9	12.2	37.7	25.9
	CASCADE UNI SHARED ENC	31.2	17.1	38.4	25.4	25.6	13.0	38.7	24.8
	CASCADE UNI SHARED ENC ORA.	31.3	17.1	38.2	25.6	25.3	12.8	38.5	24.9
	E2E BIDIRECT BY LANG	36.7	23.3	42.9	22.8	27.3	15.5	39.3	25.0
	E2E BIDIRECT BY LANG ORA.	35.9	22.7	42.5	23.0	25.3	14.0	38.4	25.4
	E2E BIDIRECT BY TASK	30.1	16.2	38.3	25.6	24.3	12.3	37.8	25.6
	CASCADE BIDIRECT	33.2	18.3	41.0	23.2	28.1	14.3	40.1	23.2
	E2E BIDIRECT SHARED	30.0	16.4	38.0	25.4	24.1	12.2	37.3	26.1

Table 11: Scores on the Fisher dataset. CCT stands for Charcut. Results from zero CS training are on the top half and results after fine-tuning are on the bottom half. Ora stands for Oracle.