CMU's IWSLT 2022 Dialect Speech Translation System

Brian Yan[1] **Patrick Fernandes**[1,2] **Siddharth Dalmia**[1] **Jiatong Shi**[1]
Yifan Peng[3] **Dan Berrebbi**[1] **Xinyi Wang**[1] **Graham Neubig**[1] **Shinji Watanabe**[1,4]
[1]Language Technologies Institute, Carnegie Mellon University, USA
[2]Instituto Superior Técnico & LUMLIS (Lisbon ELLIS Unit), Portugal
[3]Electrical and Computer Engineering, Carnegie Mellon University, USA
[4]Human Language Technology Center of Excellence, Johns Hopkins University, USA
{byan, pfernand, sdalmia, jiatongs}@cs.cmu.edu

Abstract

This paper describes CMU's submissions to the IWSLT 2022 dialect speech translation (ST) shared task for translating Tunisian-Arabic speech to English text. We use additional paired Modern Standard Arabic data (MSA) to directly improve the speech recognition (ASR) and machine translation (MT) components of our cascaded systems. We also augment the paired ASR data with pseudo translations via sequence-level knowledge distillation from an MT model and use these artificial triplet ST data to improve our end-to-end (E2E) systems. Our E2E models are based on the Multi-Decoder architecture with searchable hidden intermediates. We extend the Multi-Decoder by orienting the speech encoder towards the target language by applying ST supervision as hierarchical connectionist temporal classification (CTC) multi-task. During inference, we apply joint decoding of the ST CTC and ST autoregressive decoder branches of our modified Multi-Decoder. Finally, we apply ROVER voting, posterior combination, and minimum bayes-risk decoding with combined N-best lists to ensemble our various cascaded and E2E systems. Our best systems reached 20.8 and 19.5 BLEU on test2 (blind) and test1 respectively. Without any additional MSA data, we reached 20.4 and 19.2 on the same test sets.

1 Introduction

In this paper, we present CMU's Tunisian-Arabic to English ST systems submitted to the IWSLT 2022 dialectal ST track (Anastasopoulos et al., 2022). One of our goals is to investigate dialectal transfer from large MSA ASR and MT corpora to improve Tunisian-Arabic ST performance. We also view this task as setting for extending the sequence-level knowledge distillation (SeqKD) (Kim and Rush, 2016), E2E Multi-Decoder architecture (Dalmia et al., 2021), and system combination methods in our IWSLT 2021 offline ST systems (Inaguma et al., 2021b).

In particular, our contributions are the following:

1. Dialectal transfer from large paired MSA corpora to improve ASR and MT systems (§3.1)
2. MT SeqKD on MSA ASR data for artificial ST triplets to improve E2E ST systems (§3.2.2)
3. Multi-Decoder with hierarchical CTC training for target-oriented speech encodings (§3.2.3)
4. Multi-Decoder with CTC beam search hypothesis re-scoring during ST inference (§3.2.4)
5. Multi-Decoder with surface and posterior-level guidance from external models (§3.3.1)
6. Joint minimum bayes-risk decoding as an ensembling method (§3.3.2)

Results on the blind test set, test2, and ablations on the provided test set, test1, demonstrate the overall efficacy of our systems and the relative contributions of the aforementioned techniques (§5).

2 Task Description and Data Preparation

The Arabic language is not a monolith. Of its estimated 400 million native speakers, many speak in colloquial dialects such as, Tunisian-Arabic, that have relatively less standard orthographic rules and smaller ASR and MT corpora compared to formal MSA (Hussein et al., 2022). Both of these realities present challenges to building effective ST systems, and as such the dialectal speech translation shared task is an important venue for tackling these research problems.

Table 1 shows the corpora relevant to the shared task. The IWSLT22-Dialect corpus consists of ST triplets where 160 hours of 8kHz conversational Tunisian-Arabic speech are annotated with transcriptions and also translated into English. The MGB2 corpus (Ali et al., 2016) consists of 1100 hours of 16kHz broadcast MSA speech and the corresponding transcriptions. The OPUS corpus

Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022), pages 298 - 307

	#Hours	#Sentence	
	of Speech	Arabic	English
IWSLT22-Dialect	160	0.2M	0.2M
MGB2	1100	1.1M	-
OPUS	-	42M	42M

Table 1: Statistics for the three corpora included in the IWSLT 2022 dialect ST shared task. IWSLT22-Dialect has triplets of speech, source Arabic transcription, and target English translation. MGB2 and OPUS have only pairs for ASR and MT respectively.

(Tiedemann et al., 2020) consists of 42M MSA-English translation pairs across several domains. Any systems that use MGB2 or OPUS data for pre-training, fine-tuning, or any other purpose are designated as *dialect transfer* systems.[1]

Following the shared task guidelines, punctuation is removed and English text is lowercased. Buckwalter one-to-one transliteration of Arabic text (Habash et al., 2007) was applied to help non-Arabic speakers with ASR output interpretation. English sentences were tokenized with the `tokenizer.perl` script in the Moses toolkit (Koehn et al., 2007) for training and detokenized for scoring. Language-specific sentencepiece vocabularies were created using the byte pair encoding (BPE) algorithm (Sennrich et al., 2016) with the `sentencepiece` toolkit.[2] Speech data was up-sampled by a factor of 3 using 0.9 and 1.1 speed perturbation ratios (Ko et al., 2015). The IWSLT22-Dialect data was upsampled to 16kHz for consistency using the `sox` toolkit[3].

3 Proposed Methods

In this section, we describe our cascaded (§3.1) and E2E systems (§3.2). Then we describe methods for integrating both approaches §3.3.

3.1 Cascaded ASR→MT Systems

3.1.1 ASR

To train ASR models for our cascaded system, we use the ESPnet (Watanabe et al., 2018) framework. Our ASR architecture is based on hybrid CTC/attention approach (Watanabe et al., 2017) with a Conformer encoder (Gulati et al., 2020). The Conformer, which employs convolutions to model local patterns and self-attention to model long-range context, has shown to be effective on both ASR and E2E ST tasks (Guo et al., 2020; Inaguma et al., 2021b). We also use a bidirectional LSTM (Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005) language model (LM) to re-score beam search hypotheses during inference. We ensemble multiple ASR systems with varying hyper-parameters using Recognizer Output Voting Error Reduction (ROVER) with minimal word-level edit-distance alignment (Fiscus, 1997).

3.1.2 MT

To train MT models for our cascaded system, we use the Fairseq (Ott et al., 2019) framework to train transformers encoder-decoder models (Vaswani et al., 2). To mitigate the exposure bias of training with ground-truth data and using ASR outputs at test time, we introduce *ASR mixing*, where during training, for each sample in the training set, the model maximizes the log-likehood of translation from both the *ground-truth source* and the *ASR source* from an ASR system. This is possible because we have triplet data for training set as well. We use the same system used in the cascaded system to generate ASR outputs for the training set. We ensemble multiple MT systems with varying random seeds using posterior combination of hypotheses during beam search.

We also train an MT model using the ESPnet toolkit (Watanabe et al., 2018) as an auxiliary model used for posterior combinations with our E2E ST systems as described in §3.3.1. These models use BPE vocabulary sizes that are optimal for E2E ST, which we found empirically to be smaller than for MT.

3.1.3 Direct Dialectal Transfer

To leverage MSA annotated speech data to improve our ASR system, we select a subset of the MGB2 data as an augmentation set to be added to the IWSLT22-Dialect data. We first use an ASR model trained on IWSLT22-Dialect data only to compute the cross-entropy of the utterances in the MGB2 data. We then select a percentage of the MGB2 utterances with the lowest cross-entropy. Similar cross-entropy based data selection has shown to effectively reduce noise resulting from domain mismatches in language modeling (Moore and Lewis, 2010) and MT (Junczys-Dowmunt, 2018). After pre-training on the mixture

[1] We do not use self-supervised representations, morphological analyzers, or any other resources reliant on data other than the three aforementioned corpora.

[2] https://github.com/google/sentencepiece

[3] http://sox.sourceforge.net

of MGB2 and IWSLT22-Dialect data, we then fine-tune on IWSLT22-Dialect data only.

To leverage the MSA translation data to improve our MT system, we use the OPUS corpus, cleaning sentences longer than 200 subwords. This results in about 30M sentence pairs of training data for MSA-English. We then train a larger transformer for 20 epochs on this training data. We the use fine-tune this model on the IWSLT22-Dialect data.

3.2 E2E ST Systems

3.2.1 Multi-Decoder Architecture

Multi-decoder model (Dalmia et al., 2021) is an end-to-end sequence model that exploits decomposition of a complex task into simpler tasks in it's model design. For speech translation it decomposes the task into ASR and MT sub-nets while maintaining the end-to-end differentiability. To train Multi-Decoder models, we modified the ESPnet framework (Watanabe et al., 2018).

As shown in figure 1.a, the speech signal, $X = \{\mathbf{x}_t \in \mathbb{R}^D | t = 1, ..., T\}$, is mapped to encoder representations by the *Speech Encoder* which are then in turn mapped autoregressively to decoder representations corresponding to the source language transcription, $Y^{\text{ASR}} = \{y_l^{\text{ASR}} \in \mathcal{V} | l = 1, ..., L\}$, by the *ASR Decoder*. These *ASR Decoder* representations, referred to as searchable hidden intermediates, are passed to the downstream *ST Encoder-Decoder*. In order to avoid error-propagation, the *ST Decoder* performs cross-attention over both the *Speech Encoder* and *ST Encoder* representations. The network is optimized with multi-tasking on cross-entropy losses for both the source and target languages, $\mathcal{L}_{\text{CE}}^{\text{ASR}}$ and $\mathcal{L}_{\text{CE}}^{\text{ST}}$ respectively, along with a CTC (Graves, 2012) loss $\mathcal{L}_{\text{CTC}}^{\text{ASR}}$:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{CE}}^{\text{ASR}} + \lambda_2 \mathcal{L}_{\text{CTC}}^{\text{ASR}} + \lambda_3 \mathcal{L}_{\text{CE}}^{\text{ST}} \quad (1)$$

where λ's are used for interpolation. During inference, the CTC branch of the *Speech Encoder* is also used to re-score beam search hypotheses produced by the *ASR Decoder*, following the Hybrid CTC/Attention method (Watanabe et al., 2017).

Inaguma et al. (2021a) showed that sampling CTC output instead of always using ground truth previous token helps the Multi-Decoder model. With a CTC sampling rate of 0.2, which means that with a probability of 0.2 we would use the CTC output instead of the ground truth during training. This simulates the inference condition where there would be ASR errors. We found this technique to be particularly helpful for this dataset.

3.2.2 SeqKD Dialectal Transfer

Our Multi-Decoder training objective, equation 1, assumes that each speech signal is annotated with both a source language transcription and target language translation. In order to include additional paired MSA data into this training regime, we first generate artificial speech, transcript, and translation triplets. To do so, we first build a MSA MT model using the OPUS data. We then generate pseudo-translations for the paired MGB2 data by feeding the MSA transcriptions as inputs to the MT model. This method is based on SeqKD Kim and Rush (2016) and can be considered as a dialectal application of MT to ST knowledge-distillation. We mix a percentage of the pseudo-translated data using the same cross-entropy based methodology as decribed in §3.1.3 with the Tunisian-Arabic data during training. We refer to this data augmentation as *MT SeqKD* in future sections.

3.2.3 Hierarchical Speech Encoder

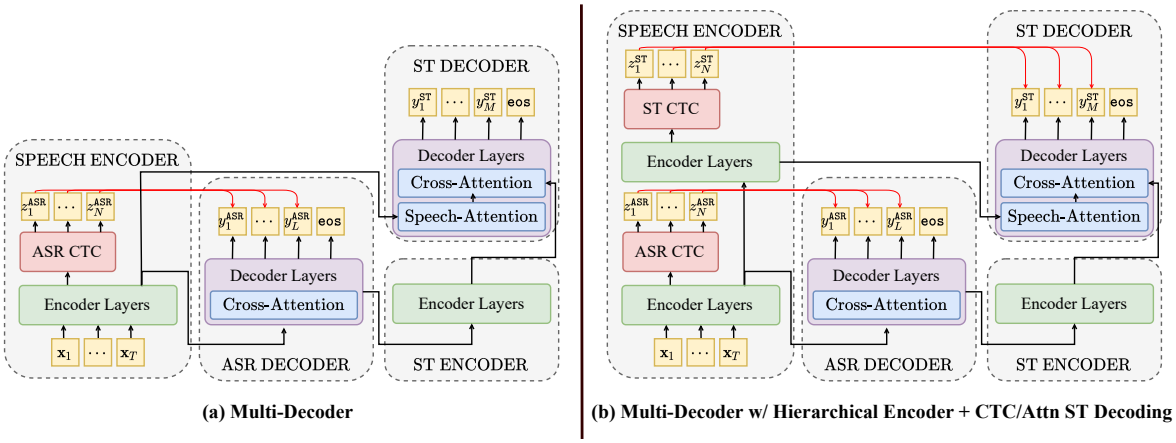
CTC loss is often used as auxiliary loss in attention based encoder decoder models (Watanabe et al., 2017). It helps the attention based decoder by inducing monotonic alignment with the encoder representations (Kim et al., 2017). In this work, we extend this idea by creating a hierarchical encoder that customizes the ordering of the encoder for the individual sub-tasks by using auxiliary CTC loss at each sub-task. Here, we use an auxiliary CTC loss with ASR targets and another CTC loss with ST targets. As shown in figure 1.b, the first 12 layers of the *Speech Encoder* produce ASR CTC alignments, $Z^{\text{ASR}} = \{z_n^{\text{ASR}} \in \mathcal{V} \cup \{\varnothing\} | n = 1, ..., N\}$, while the final 6 layers produce ST CTC alignments, $Z^{\text{ST}} = \{z_n^{\text{ST}} \in \mathcal{V} \cup \{\varnothing\} | n = 1, ..., N\}$, where $\cup\{\varnothing\}$ denotes the blank emission. This creates a hierarchical encoder structure similar to (Sanabria and Metze, 2018; Lee and Watanabe, 2021; Higuchi et al., 2021). The Multi-Decoder with hierarchical encoder is optimized with an additional ST CTC loss, $\mathcal{L}_{\text{CTC}}^{\text{ST}}$:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{CE}}^{\text{ASR}} + \lambda_2 \mathcal{L}_{\text{CTC}}^{\text{ASR}} + \lambda_3 \mathcal{L}_{\text{CE}}^{\text{ST}} + \lambda_4 \mathcal{L}_{\text{CTC}}^{\text{ST}} \quad (2)$$

Note that the *ST Decoder* now performs cross-attention *Speech Encoder* representations that are oriented towards the target language.

3.2.4 Joint CTC/Attention Decoding for ST

The ST CTC branch of the *Speech Encoder* introduced in the previous section allows us to apply



(a) Multi-Decoder

(b) Multi-Decoder w/ Hierarchical Encoder + CTC/Attn ST Decoding

Figure 1: The left side presents the original Multi-Decoder architecture with searchable hidden intermediates produced by the *ASR Decoder*. The red lines indicate joint CTC/Attention decoding of beam search hypotheses produced by an autoregressive decoder. The right side presents a modified Multi-Decoder with both a hierarchical ASR to ST *Speech Encoder* optimized via CTC objectives and joint CTC/Attention ST inference.

joint CTC/Attention decoding using the one-pass beam search algorithm (Watanabe et al., 2017) during ST inference as well. Although previously only applied to ASR decoding, we found that joint CTC/Attention inference for the *ST Decoder* beam search hypotheses were beneficial in this task. Deng et al. (2022) show that joint modeling of CTC/Attention is effective for short contexts of blockwise streaming ST; as far as we know, our work is the first to show the benefit on long context. Our conjecture is that speech to translation transduction with attention mechanisms, as in the original Multi-Decoder, contains irregular alignments between the acoustic information and the target sequence. The hierarchical encoder and joint CTC/Attention decoding methods may alleviate these irregularities by enforcing greater monotonicity. We refer to the Multi-Decoder with hierarchical encoder and joint CTC/Attentionn ST decoding as the *Hybrid Multi-Decoder* in future sections.

3.3 Integrating E2E and Cascaded Systems

3.3.1 Guiding Multi-Decoder Representations

Since the Multi-Decoder (Dalmia et al., 2021) uses hidden representations from the autoregressive *ASR Decoder*, we can perform search and retrieval over this intermediate stage of the model. Dalmia et al. (2021) showed that ST quality improves by using beam search and external models like LMs to improve the representations the ASR sub-task level. We believe this an important property to have when building models for complex sequence tasks like speech translation, as often there is additional data present for the sub-tasks like ASR and MT. In this work, we help guide our Multi-Decoder model to retrieve better decoder representations by using external ASR and MT models.

We experimented with two approaches: 1) posterior level guidance and 2) surface level guidance. The former is similar in concept to posterior combination for model ensembling during inference as described in (Inaguma et al., 2021b), however the Multi-Decoder allows us to incorporate both an external ASR and MT model due to the searchable hidden intermediates whereas a vanilla encoder-decoder ST model would only be compatible with an external MT model. This method requires beam search over both ASR and MT/ST for multiple models. Alternatively, surface level guidance can avoid this expensive search over the ASR intermediates by instead retrieving the hidden representations for an ASR surface sequence produced externally.

We use the ROVER ASR outputs described in §3.1.1 as surface level guides for the Multi-Decoder's ASR intermediates and found this to be more effective than posterior combination with external ASR models. We refer to this method of retrieval as *ROVER intermediates* in future sections. Since ROVER is based on minimal edit-distance alignment, we did not find it compatible with translation sequences. For the *ST Decoder*, we use posterior combination with external ST and MT models and refer to this as *ST/MT Posterior Combination* in future sections.

3.3.2 Minimum Bayes-Risk

Rather than finding the most likely translation, Minimum Bayes-Risk (MBR) decoding aims to find the translation that maximizes the expected *utility* (equivalently, that minimizes *risk*, (Kumar and Byrne, 2002, 2004; Eikema and Aziz, 2020)). Let $\bar{\mathcal{Y}}_{\text{cands}}$, $\bar{\mathcal{Y}}_{\text{samples}}$ be sets containing N *candidate* hypotheses and M *sample* hypothesis. This sets can be obtained from one or multiple model by, for example sampling or taking the top beams in beam search. Let $u(y^*, y)$ be an utility function measuring the similarity between a hypothesis y and a reference y (we only consider BLEU in this work). MBR decoding seeks for

$$\hat{y}_{\text{MBR}} = \arg\max_{y \in \bar{\mathcal{Y}}_{\text{cands}}} \underbrace{\mathbb{E}_{Y \sim p_\theta(y|x)}[u(Y, y)]}_{\approx \frac{1}{M} \sum_{j=1}^{M} u(y^{(j)}, y)} , \quad (3)$$

We experimented with using MBR as a technique for system combination, in two forms:

- *True*: the stronger system (the E2E) is used to generate the N candidates $\bar{\mathcal{Y}}_{\text{cands}}$ and the weaker system (the Cascaded system) is used to generate M samples $\bar{\mathcal{Y}}_{\text{samples}}$. This means that the outputs will guaranteed to generated by the E2E system.
- *Joint*: in this case, both the E2E and the Cascaded generate N hypotheses, with are then concatenated to make both the candidate set and sample set $\bar{\mathcal{Y}}_{\text{samples}} = \bar{\mathcal{Y}}_{\text{cands}}$, with $|\bar{\mathcal{Y}}_{\text{cands}}| = 2N$

We explored using beam search and nucleus sampling (Holtzman et al., 2019) with different p values for both generating candidates and generating samples to compute the expectation over. Overall we found that, for both settings, using beam search to generate hypothesis for the E2E model and nucleus sampling with $p = 0.9$ for the cascaded system yield the best results. We use $N = M = 50$ for both settings.

4 Experimental Setup

ASR: We extracted 80-channel log-mel filterbank coefficients computed with 25-ms window size and shifted every 7-ms with 3-dimensional pitch features.[4] The features were normalized by the mean and the standard deviation calculated on the entire training set. We applied SpecAugment (Park et al., 2019) with mask parameters $(m_T, m_F, T, F) = (5, 2, 27, 0.5)$ and bi-cubic time-warping. We use a BPE vocabulary size of 1000. Our encoder has 2 CNN blocks followed by 12 Conformer blocks following (Guo et al., 2020). Each CNN block consisted of a channel size of 256 and a kernel size of 3 with a stride of 2×2, which resulted in time reduction by a factor of 4. Our decoder has 6 Transformer blocks. In both encoder and decoder blocks, the dimensions of the self-attention layer d_{model} and feed-forward network d_{ff} were set to 256 and 2048, respectively. The number of attention heads H was set to 8. The kernel size of depthwise separable convolution in Conformer blocks was set to 31. We optimized the model with the joint CTC/attention objective with a CTC weight of 0.3. We also used CTC and LM scores during decoding. Models were trained for 60 epochs. We averaged the model parameters of the 10 best epoch checkpoints by validation loss. Our LM is a BLSTM with 4 layers and 2048 unit dimension. Beam search is performed with beam size 20, CTC weight 0.2, and LM weight 0.1.

MT: We use SentencePiece (Kudo and Richardson, 2018) with the Byte-pair Encoding algorithm (Sennrich et al., 2016). We experimented with various vocabularies sizes and found that 4000 vocabulary size to be the best for small models. For the pretrained model, we use a vocabulary size of 16000. The small transformer model used for the non-dialect submissions has 512 embedding dimensions, 1024 feedforward dimensions, 6 layers and 4 heads on each layer on both encoder/decoder. The large transformer model used for dialect transfer has 1024 embedding dimensions, 4096 feedforward dimensions, 6 layers and 16 heads on each layer on both encoder/decoder. Models were trained with early stopping by validation loss. We averaged the model parameters of the last 5 epoch checkpoints. Unless otherwise specified, we use beam search with beam size of 5 and no length penalty in beam search.

Multi-Decoder: We use the same feature extraction as for ASR. We use separate BPE vocabularies for source and target, both of size 1000. The ASR sub-net of the Multi-Decoder is also the same as our ASR configuration, allowing for pre-trained initialization of the ASR encoder, decoder, and CTC. The hierarchical encoder adds 6 additional Trans-

[4] 7-ms shift was found to be helpful due to the presence of many short utterances in the IWSLT22-Dialect data.

ID	Model Type / Name	Dialect Transfer	test1 WER(↓)
A1	ASR Conformer	✗	50.4
A2	+ ROVER Comb.	✗	48.1
A3	ASR Conformer	✓	50.0
A4	+ ROVER Comb.	✓	**47.5**
			MT BLEU(↑)
B1	MT Transformer (Fairseq)	✗	21.8
B2	+ Posterior Comb.	✗	22.8
B3	MT Transformer (Fairseq)	✓	22.4
B4	+ Posterior Comb.	✓	**23.6**
B5	MT Transformer (ESPnet)	✗	21.0

Table 2: Results of the ASR and MT components of our cascaded systems, as measured by % WER and BLEU score on the provided test1 set. ROVER and posterior combinations were applied to ASR and MT respectively.

former layers to the original 12 Conformer layers. The MT sub-net of the Multi-Decoder has a 2 layer Transformer encoder and a 6 layer Transformer decoder. This second encoder has no convolutional subsampling. The MT sub-net has the same d_{model} and d_{ff} as the ASR sub-net. We optimized the model a CTC weight of 0.3 and an ASR weight of 0.3. Models were trained for 40 epochs. We averaged the model parameters of the 10 best epoch checkpoints by validation loss. Beam search over the ASR-subnet uses the same setting as for ASR. Beam search over the MT-subnet uses beam size 5/10 with CTC weight 0.3/0.1 for the basic/dialect conditions. Length penalty 0.1 was used for all cases.

5 Results and Analyses

5.1 Submitted Shared Task Systems

Figure 2 shows the results for ASR and MT systems used as part of the cascaded system as evaluated by WER and BLEU score respectively on the provided test set, test1. Dialectal transfer provides a moderate boosts of 0.4% and 0.6% WER without ROVER and with ROVER respectively. Notably, WER's for all systems are relatively high despite a moderate amount of training data; this is perhaps due to the non-standard orthographic form of the Tunisian-Arabic transcriptions.[5] Another possible cause for the high WER is the conversational nature of the data, which may require normalization similar to the Switchboard dataset (Godfrey et al., 1992). For the MT systems, we see that posterior combination leads to over 1 BLEU point improvements when translating ground-truth source sentences. Interestingly, while there is some benefit from the dialectic transfer, the benefits are relatively small, yielding an additional 0.8 BLEU for the ensembled models. This might be due to the domain mismatch between the Tunisian-Arabic data and MSA data.

Figure 3 shows the results of our cascaded, E2E, and integrated cascaded/E2E systems on both the blind shared task test set, test2, and on the provided test set, test1. The *Hybrid Multi-Decoder* outperforms the *ASR Mixing Cascade* by 1.3 and 0.9 BLEU on test1 without and with dialectal transfer respectively. Both models are boosted by the use of ROVER. The benefit of ROVER for models without dialectal transfer (0.3 BLEU) was larger than for models with dialectal transfer (0.1 BLEU), showing some diminishing returns from isolated improvements of the ASR component of the overall ST task. Posterior combination provided boosts in the range of 0.5-0.8 BLEU across the models. Finally, the *Minimum Bayes Risk Ensembling* yielded additional gains of 0.6-1.3 BLEU. The differences between the final *Minimum Bayes Risk Ensembling* systems and the best single systems without any external model integration are 1.5 and 1.3 BLEU without and without dialectal transfer respectively.

5.2 Ablation Studies

To show the individual contributions of our various methods, we present in this section several ablations. First, we show in figure 4 the impact of dialectal transfer from MGB2 data on ASR (as described in §3.1.3) and on E2E ST (as described in §3.2.2). As subset of MGB2 data selected via the cross-entropy filter outperformed a randomly selected subset, although both were better than when no MGB2 data was included. Since the IWSLT22-Dialect utterances were shorter than the MGB2 utterances on average, one effect of the cross-entropy filter was the removal of long utterances which appeared to benefit the model. We found that using up to 25% of the MGB2 data was best for ASR. For ST, both 25% and 50% of the MGB2 data with *MT SeqKD* yielded 0.5 BLEU gains, which is slightly less than the 0.8 BLEU gains that our cascaded systems obtained from dialectal transfer. This suggests some that there our *MT SeqKD* method may be improved in the future.

Next, in figure 5 we show the results MT and ST

[5] We found that the WER's decreased by about 4% when removing diacritics from the hypothesis and the reference.

ID	Type	Model Name	Child System(s)	Dialect Transfer	test1 BLEU(↑)	test2 BLEU(↑)
C1	Cascade	ASR Mixing Cascade	A1,B1	✗	16.4	-
C2	Cascade	+ ASR Rover Comb.	A2,B1	✗	16.7	-
C3	Cascade	+ MT Posterior Comb.	A2,B2	✗	17.5	18.6
C4	Cascade	ASR Mixing Cascade	A3,B3	✓	17.3	-
C5	Cascade	+ ASR Rover Comb.	A4,B3	✓	17.4	-
C6	Cascade	+ MT Posterior Comb.	A4,B4	✓	**17.9**	**19.4**
D1	E2E ST	Hybrid Multi-Decoder	-	✗	17.7	-
D2	Mix	+ ROVER Intermediates	A2	✗	18.1	19.1
D3	Mix	+ ST/MT Posterior Comb.	A2,B5	✗	18.7	19.7
D4	E2E ST	Hybrid Multi-Decoder	-	✓	18.2	-
D5	Mix	+ ROVER Intermediates	A4	✓	18.3	19.5
D6	Mix	+ ST/MT Posterior Comb.	A4,B5	✓	**18.9**	**19.8**
E1	Mix	Min. Bayes-Risk Ensemble	C3,D3	✗	19.2	20.4
E2	Mix	Min. Bayes-Risk Ensemble	C6,D6	✓	**19.5**	**20.8**

Table 3: Results of our cascaded, E2E, and integrated cascaded/E2E systems as measured by BLEU score on the blind test2 and provided test1 sets. *Dialect Transfer* indicates the use of either MGB2 or OPUS data. Rover, posterior combinations, and minimum bayes-risk ensembling were applied to both cascaded and E2E systems, with *Child System(s)* indicating the inputs to the resultant systems combinations.

Task	MGB2 Training Data	test1 WER(↓)
ASR	none	53.1
ASR	8% w/ random select	52.7
ASR	8% w/ CE filter	**52.4**
ASR	25% w/ CE filter	**52.4**
ASR	50% w/ CE filter	53.0
ASR	75% w/ CE filter	53.5
		BLEU(↑)
ST	none	16.6
ST	25% w/ CE filter + MT SeqKD	**17.1**
ST	50% w/ CE filter + MT SeqKD	**17.1**

Table 4: Ablation study on the effects of additional MGB2 data on ASR and ST performance as measured by WER and BLEU on the test1 set respectively.

Model Name	test1 ST BLEU(↑)	test1 MT BLEU(↑)
MT Transformer	16.2	20.9
+ ASR Mixing Training	**16.7**	**21.8**

Table 5: Ablation study on the effects of ASR mixing on ST and MT as measured by BLEU on the test1 set.

systems trained with and without *ASR mixing* (as described in §3.1.2), both in the cascaded setting and using ground-truth source sentences. Overall we see that *ASR mixing* helps improving the cascaded system. Surprisingly this also improves results for the translating from ground-truth source sentences. We hypothesise that *ASR mixing* acts as a form of regularization for the orthographic inconsistencies in the source transcriptions due to the conversational nature of Tunisian-Arabic.

In table 6, we show the effects of the *ASR CTC Sampling*, *Hierarchical Encoder*, and *Joint CTC/Attention ST Decoding* modifications to the original Multi-Decoder (as described in §3.2). We found that each of these techniques boosts the overall performance and we also found their effects to be additive. Table 6 also shows the performance of a vanilla encoder-decoder for comparison, which performed significantly worse than the Multi-Decoder. Due to time limitations, we did not submit the Multi-Decoder with hierarchical encoder, joint CTC/Attention ST decoding, and ASR CTC sampling for shared task evaluation, but this was our strongest single system as evaluated on the test1 set.

Finally, Figure 7 shows the results for the two

Model Name	test1 BLEU(↑)
Encoder-Decoder	16.0
Multi-Decoder	17.1
+ ASR CTC Sampling	17.6
+ Hierarchical Encoder	17.9
+ Joint CTC/Attn ST Decoding (D4)	18.2
+ ASR CTC Sampling	**18.4**

Table 6: Ablation study on the effects of ASR CTC sampling, hierarchical encoder, and joint CTC/Attn ST decoding as measured by BLEU on the test1 set.

Model Name	MBR Method	Dialect Transfer	test1 BLEU(↑)	test2 BLEU(↑)
MBR Ensemble	True	✗	19.0	20.1
MBR Ensemble (E1)	Joint	✗	**19.2**	**20.4**
MBR Ensemble	True	✓	19.3	20.7
MBR Ensemble (E2)	Joint	✓	**19.5**	**20.8**

Table 7: Comparison of the true vs. joint methods for minimum bayes-risk ensembling as measured by BLEU on the test1 and test2 sets.

different settings for system combination through MBR (as described in §3.3.2). Using the *Joint* setting where the hypothesis from both system are considered as both candidates/samples leads to the best translations compared to the *True* setting. Figure 8 shows that while effective for maximizing BLEU score, MBR did not improve according to human evaluation.[6]

6 Conclusion

In this paper, we have presented CMU's dialect speech translation systems for IWSLT 2022. Our systems encompass various techniques across cascaded and E2E approaches. Of the techniques we presented, the hierarchical encoder and joint CTC/Attention ST decoding modifications to the Multi-Decoder and the minimum bayes-risk ensembling were amongst the most impactful. In future work, we seek to formalize these methods with additional theoretical and experimental backing, including extensions to other corpora and tasks such as pure MT.

[6]Human evaluation methodology is detailed in (Anastasopoulos et al., 2022)

Model Name	test2 BLEU(↑)	test2 DA Ave. / z-score(↑)
Hybrid Multi-Decoder (D6)	19.8	66.5 / 0.119
MBR Ensemble (E2)	**20.8**	66.5 / 0.114

Table 8: Human evaluation results, as measured by DA average and z-score, showing the impact of maximizing BLEU score via minimum bayes-risk ensembling.

Acknowledgements

Brian Yan and Shinji Watanabe are supported by the Human Language Technology Center of Excellence. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) (Towns et al., 2014), which is supported by National Science Foundation grant number ACI-1548562; specifically, the Bridges system (Nystrom et al., 2015), as part of project cis210027p, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center. We'd also like to thank Soumi Maiti, Tomoki Hayashi, and Koshak for their contributions.

References

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2016. The mgb-2 challenge: Arabic multi-dialect broadcast media recognition. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 279–284. IEEE.

Antonios Anastasopoulos, Luisa Bentivogli, Marcely Z. Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Marcello Federico, Christian Federmann, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel M. Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, Juan Pino, Elizabeth Salesky, Jiatong Shi, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alex Waibel, Changhan Wang, and Shinji Watanabe. 2022. FINDINGS OF THE IWSLT 2022 EVALUATION CAMPAIGN. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, Dublin, Ireland. Association for Computational Linguistics.

Siddharth Dalmia, Brian Yan, Vikas Raunak, Florian Metze, and Shinji Watanabe. 2021. Searchable hidden intermediates for end-to-end models of decomposable sequence tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, pages 1882–1896, Online. Association for Computational Linguistics.

Keqi Deng, Shinji Watanabe, Jiatong Shi, and Siddhant Arora. 2022. Blockwise streaming transformer for spoken language understanding and simultaneous speech translation. *arXiv preprint arXiv:2204.08920*.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

J.G. Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 347–354.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society.

Alex Graves. 2012. Connectionist temporal classification. In *Supervised sequence labelling with recurrent neural networks*, pages 61–93. Springer.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610. IJCNN 2005.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for speech recognition. In *Proceedings of Interspeech*, pages 5036–5040.

Pengcheng Guo, Florian Boyer, Xuankai Chang, Tomoki Hayashi, Yosuke Higuchi, Hirofumi Inaguma, Naoyuki Kamo, Chenda Li, Daniel Garcia-Romero, Jiatong Shi, et al. 2020. Recent developments on ESPnet toolkit boosted by Conformer. *arXiv preprint arXiv:2010.13956*.

Nizar Habash, Abdelhadi Soudi, and Timothy Buckwalter. 2007. On arabic transliteration. In *Arabic computational morphology*, pages 15–22. Springer.

Yosuke Higuchi, Keita Karube, Tetsuji Ogawa, and Tetsunori Kobayashi. 2021. Hierarchical conditional end-to-end asr with ctc and multi-granular subword units. *ArXiv*, abs/2110.04109.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *CoRR*, abs/1904.09751.

Amir Hussein, Shinji Watanabe, and Ahmed Ali. 2022. Arabic speech recognition by end-to-end, modular systems and human. *Computer Speech & Language*, 71:101272.

Hirofumi Inaguma, Siddharth Dalmia, Brian Yan, and Shinji Watanabe. 2021a. Fast-md: Fast multi-decoder end-to-end speech translation with non-autoregressive hidden intermediates. *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 922–929.

Hirofumi Inaguma, Brian Yan, Siddharth Dalmia, Pengcheng Gu, Jiatong Shi, Kevin Duh, and Shinji Watanabe. 2021b. Espnet-st iwslt 2021 offline speech translation system. In *IWSLT*.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4835–4839. IEEE.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Proceedings of Interspeech*, pages 3586–3589.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2002. Minimum bayes-risk word alignments of bilingual texts. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, page 140–147, USA. Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Jaesong Lee and Shinji Watanabe. 2021. Intermediate loss regularization for ctc-based speech recognition. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6224–6228.

Robert C Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 conference short papers*, pages 220–224.

Nicholas A Nystrom, Michael J Levine, Ralph Z Roskies, and J Ray Scott. 2015. Bridges: a uniquely flexible hpc resource for new communities and data analytics. In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*, pages 1–8.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Proceedings of Interspeech*, pages 2613–2617.

Ramon Sanabria and Florian Metze. 2018. Hierarchical multitask learning with ctc. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 485–490. IEEE.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Jörg Tiedemann, Santhosh Thottingal, et al. 2020. Opus-mt–building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.

J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkins-Diehr. 2014. Xsede: Accelerating scientific discovery. *Computing in Science & Engineering*, 16(5):62–74.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. ESPnet: End-to-end speech processing toolkit. In *Proceedings of Interspeech*, pages 2207–2211.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tomoki Hayashi. 2017. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.