When End-to-End is Overkill: Rethinking Cascaded Speech-to-Text Translation

Anonymous submission

Abstract

Though end-to-end speech-to-text translation has been a great success, we argue that the cascaded speech-to-text translation model still has its place, which is usually criticized for the error propagation between automatic speech recognition (ASR) and machine translation (MT) models. In this paper, we explore the benefits of incorporating multiple candidates from ASR and self-supervised speech features into MT. Our analysis reveals that the primary cause of cascading errors stems from the increased divergence between similar samples in the speech domain when mapped to the text domain. However, by incorporating multiple candidates and self-supervised speech features, we enable the machine translation model to select the correct vocabulary and perform accurate translation by leveraging multiple samples from the speech domain. Through extensive experiments, we demonstrate that cascaded speech-to-text translation not only reduces error propagation but also capitalizes on the advantages of ample ASR and MT datasets, as well as the convenience of pre-trained ASR/MT models while mitigating the issues associated with error propagation.

Keywords: speech-to-text translation,machine translation

1. Introduction

In recent years, the academic community has been intrigued by the rapid advancement of end-to-end speech-to-text translation models. These efficient encoder-decoder architectures provide a direct avenue for translating speech, bypassing the need for complex intermediate symbolic representations. However, the arduous task of assembling and curating end-to-end data poses a significant challenge, entailing considerable costs and extensive efforts. These end-to-end methods need the careful selection of high-quality data, encompassing both speech and translated transcripts, and the scrupulous exclusion of erroneous examples.

Nonetheless, cascaded speech-to-text translation models have encountered substantial criticism due to an intrinsic shortcoming. They are susceptible to the pervasive issue of error propagation, wherein inaccuracies in the automatic speech recognition (ASR) system infiltrate and adversely affect the ensuing machine translation (MT) process.

Recent studies (Bommasani et al., 2021; Han et al., 2021b) have demonstrated the performance improvements achieved by scaling up pre-trained models for downstream natural language processing tasks. However, developing advanced speech translation models is a difficult task, requiring significant computational resources due to the enormous volume of training data and complex model intricacies. Conversely, cascade models offer plentiful latent potential that remains to be exploited. We think the key to fully utilizing the capabilities of cascade models hinges on addressing the issue of error propagation.

In this paper, we present a novel perspective on error propagation and the preservation of essential speech information. We propose the idea of utilizing multiple ASR candidates for machine translation, integrated with self-supervised speech representations to enhance the translation accuracy. Our comprehensive analysis reveals the primary causes of error propagation in cascade systems, which originate from the misalignment between the acoustic and semantic dimensions of speech. Factors such as homophones with different meanings and word elisions contribute to inaccuracies in ASR results, which consequently propagate to the machine translation model.

Furthermore, we explore the use of self-supervised language representations, empowering our model to preserve fine-grained linguistic information in speech. This enables our model to harness finer-grained speech representations and enhance its predictive capabilities. Through extensive experimentation, we demonstrate that a simple augmentation of the machine translation model with additional modules can effectively mitigate error propagation in the cascade S2T model, harnessing the potential of large-scale translation models at low cost. Our method outperforms the end-to-end S2T models, indicating the effectiveness of our model in addressing the limitations of the current paradigm. In addition, we underscore the urgent need to reevaluate the cascade approach for speech-to-text translation, exposing the limitations of end-to-end models, while revealing the vast potential of exploiting multiple candidates and self-supervised representations. By addressing the challenges of error propagation and fine-grained linguistic information loss, our work finds a new way to achieve enhanced performance for speech-to-text translation.

Our model has the following advantages:

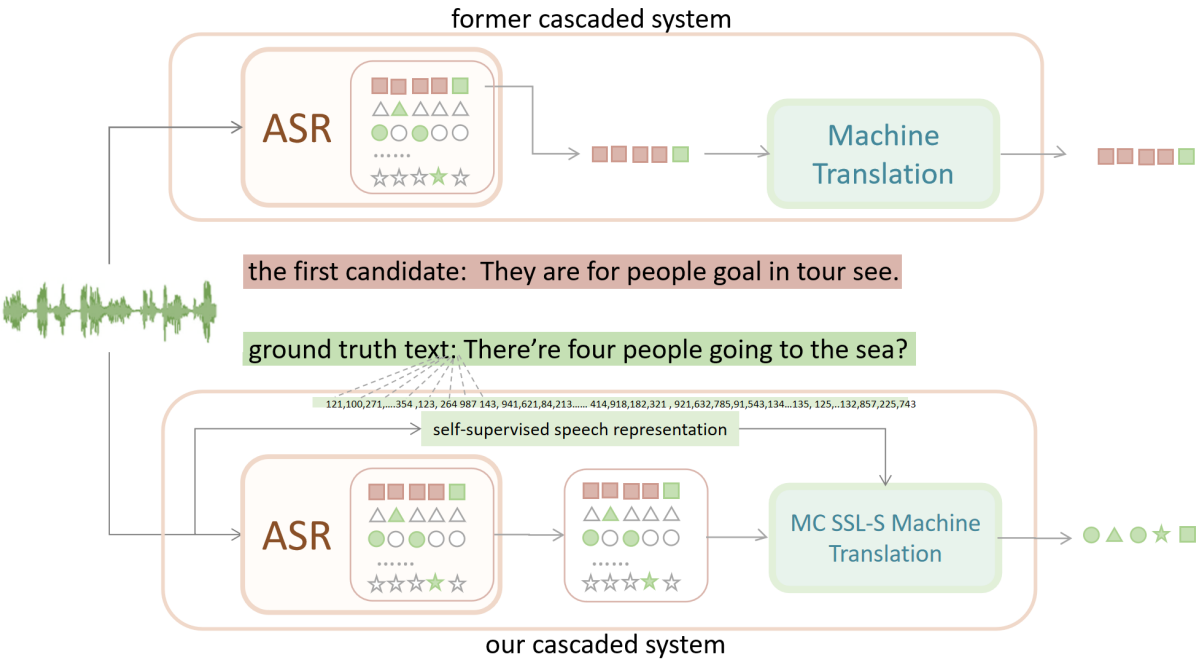

Figure 1: The correct words are scattered among various candidates, while the former cascaded system directly selected the first candidate, resulting in similar pronunciation errors in the ASR output that are further propagated through the translation model, causing cascading losses.

1. Our model achieves the best performance among cascaded models in the current Speech-to-Text (S2T) task.

2. Our model can leverage various known ASR and machine translation (MT) pre-trained models. Specifically, even without modifying the model parameters, our approach utilizes aligning candidates and calculating attention to achieve improved performance, particularly in the MT model. As the effectiveness of the MT pre-trained model improves, our approach has a higher potential to bridge the gap between speech translation and machine translation in terms of performance. This indicates the untapped potential of our method.

3. Unlike end-to-end models, our approach does not require an extensive large amount of costly <speech, transcript, target> paired data.

4. Our model demonstrates rapid training speed and attains exceptional results with minimal data utilization for fine-tuning the Machine Translation (MT) model. Even when not subjected to fine-tuning in the domain of MT, our model still yields noteworthy improvements. Furthermore, our model introduces no additional parameters to the base model, thereby conferring significant adaptability and facilitating lightweight modifications.

2. Related Work

2.1. Cascaded Speech Translation System

Studies on ASR+MT systems have explored various methods to enhance the integration of ASR output lattices into MT models (Matusov et al., 2005). This integration aims to address the challenge of error propagation between the ASR and MT components. To mitigate error propagation, several approaches have been proposed. (Bertoldi and Federico, 2005; Beck et al., 2019; Sperber et al., 2019) suggested feeding the MT system with ASR data structures. (Peitz et al., 2012; Cheng et al., 2019; Di Gangi et al., 2019b) proposed techniques to enhance the robustness of MT to ASR errors, such as training on parallel data that incorporates factual or emulated ASR errors. Several papers (Dalmia et al., 2021; Inaguma et al., 2021a) have explored the integration of ASR output disambiguation into end-to-end (E2E) models using a two-pass decoding strategy. In this approach, the initial pass involves an ASR decoder, followed by a second pass utilizing an S2T decoder. In previous studies (Zhang et al., 2019), approaches have been made to incorporate word lattices into Machine Translation (MT) models. However, these approaches often necessitated the addition of supplementary modules and substantial additional training. In contrast, the method proposed in our research employs an n-best strategy that does not

require additional parameters and can significantly enhance performance with minimal fine-tuning.

2.2. End-to-End Translation System

End-to-end speech translation (ST) (Bérard et al., 2016) directly translates the speech in the source language into sentences in the target language, without outputting the source language transcription (Bérard et al., 2016). With the success of attention-based models for speech and text related tasks, a typical and effective baseline model for speech translation tasks is speech-transformer (Dong et al., 2018), which has much of the same model structure as the commonly used MT model Transformer, except for the pre-processing down sampling module for speech signals. Computational limitations pose a significant challenge to end-to-end speech-to-text (ST) systems. Recent research studies (Li et al., 2020; Le et al., 2021; Yin et al., 2023) have demonstrated the potential of fine-tuning pre-trained models' components, which motivates further exploration of efficient fine-tuning approaches in this work. Another obstacle to end-to-end ST systems is the scarcity of training data. However, recent advancements have addressed this issue through techniques such as multi-task learning (Weiss et al., 2017; Anastasopoulos and Chiang, 2018; Bérard et al., 2018), transfer learning (Liu et al., 2019) and synthetic data generation (Jia et al., 2019; Pino et al., 2020; Bahar et al., 2019). Additionally, self-supervised pre-training with unlabeled speech data (Bérard et al., 2018; Popuri et al., 2022; Stoian et al., 2020; Li et al., 2020; Bahar et al., 2019) and data augmentation (Popuri et al., 2022; Jia et al., 2019; Stoian et al., 2020; Pino et al., 2020; Wang et al., 2021b) have been employed to tackle this challenge. Furthermore, improvements to the loss function have been proposed to effectively utilize the available data (Li and Niehues, 2022).

3. Motivation

3.1. Preliminary experiments

In the cascaded system, we extract the top 20 results based on the scores from the ASR system, and then select the top n candidates. We calculate the lexical overlap between these candidates and the ground truth text of the ASR. In Table 1, "average" refers to the average lexical overlap between each individual candidate and the ground truth, while "cumulative" represents the lexical overlap when considering a combination of n candidates with the ground truth. We observed that as n increases, although the average lexical overlap decreases, the cumulative overlap improves. This indicates that apart from the top-ranked candidate, the ASR system fails to include some vocabulary, which remains in the lower-ranked candidates.

n candidates	average	cumulative
1	92.0%	92.0%
5	90.0%	94.3%
20	89.5%	95.5%

Table 1: Lexical overlap between ASR candidates and GT

Additionally, we used a trained MT model to translate and calculate BLEU scores for the top 5 candidates based on their scores. We then analyzed the index and percentage of the highest BLEU score, and found that only 45.35% of the candidates with the highest BLEU score corresponded to the candidate with the lowest word error rate. We also compared the BLEU score of the translation for the candidate with the lowest word error rate with the average BLEU score of the top 5 translations, and found that the latter was higher by 3.1 BLEU.

n candidates	1	2	3	4	5
percentage(%)	45.35	16.73	14.07	12.12	11.74

Table 2: The proportion of different candidates in the best candidate

	first candidate	best candidate
BLEU	32.9	36.0

Table 3: The translation BLEU score between the best complete sentence candidate and the first candidate (selected in the former cascaded system)

3.2. Source of Cascade Loss

Non-correspondence in pronunciation space and semantic space When similar samples in the speech domain are mapped to the text domain, the differences between them can increase. These dissimilarities have the potential to propagate errors to the machine translation model. For instance, "What did it feel to read that letter" and "What did it fail to read that letter" are two utterances that be pronounced similarly but have significant differences in their intended meaning. The former is the correct ASR result that aligns with human language expression. However, during the ASR's beam search, the latter obtains the highest beam score. In traditional cascaded models, the correct ASR result "What did it feel to read that letter" is discarded as the second-highest probability beam, and subsequently, the machine translation result is poor.

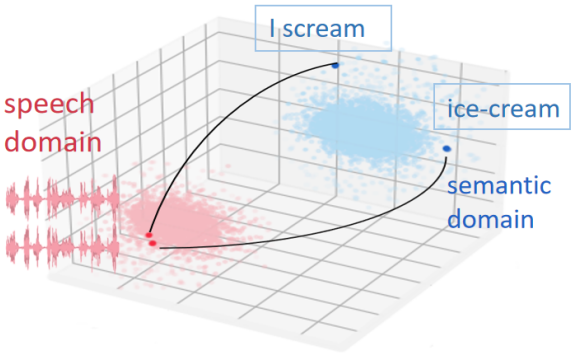

Figure 2: Non-correspondence in pronunciation space and semantic space

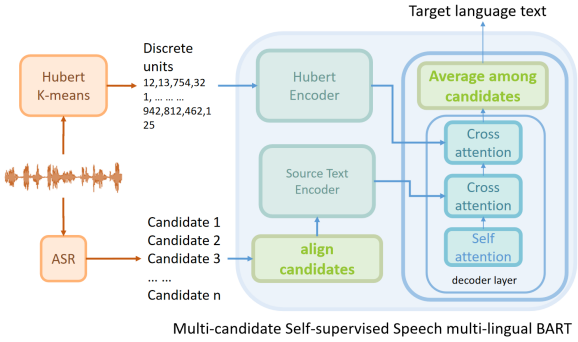

Figure 3: Overview of our proposed MC-SSLS-mBART

ASR models struggle to select the results that best match the language patterns ASR is trained using paired speech and text data, but the language patterns it captures are not as rich as those in translation models. Therefore, when the recognition results are similar, it is difficult for the ASR model to select the candidate that best aligns with human common sense and grammar based on scores alone.

As an example, consider the phrase “has put the race on the top” The highest-scoring candidate in ASR recognizes “race” as “rays” which clearly contradicts common sense. However, subsequent candidates include various results with similar pronunciations, such as “race”,“raised” and “raise”. Another example is “Recording the transaction in an immutable distributed ledger” while the highest-scoring result in ASR is “Recording the transaction in an immutable distributed lecture” which is clearly not a common expression. The subsequent candidates include “legend”,“literature”,“letter” “ledger” and other results.

Spelling errors in professional vocabulary In the ASR part, when encountering domain-specific vocabulary that is not in the lexicon, it tends to be translated as a non-existent word with a similar spelling. This has a significant impact on machine translation models. We have an example sentence: “Where the golgi apparatus sometimes called the golgi body receives them.” The first candidate from ASR misrecognizes the pronunciation of the “golgi apparatus” as two non-existent words, “golgy” and “golji”.

4. Method

4.1. Multi-candidate from ASR

The correspondence between the speech domain and the semantic domain is not always perfect. However, compared to machine translation models, ASR models have limited ability to capture semantic patterns effectively. Consequently, ASR is prone to mapping speech samples that are acoustically similar to outputs that are semantically distant. Moreover, ASR lacks the capability to selectively choose samples that align more closely with human language conventions and patterns. For a given speech input, the correct vocabulary might be dispersed among multiple candidates generated during beam search. Unfortunately, previous cascade systems only consider the top-scoring candidate and pass it to the machine translation model, resulting in error propagation within the cascade system. Our cascade model first uses Wenet (Yao et al., 2021) to perform ASR on GigaST (Ye et al., 2022b), we store the speech recognition texts corresponding to the top 20 cumulative log probability value rankings at the end of beam search.

Aligned by common substrings The top n ranked texts are selected. In the preprocess stage before putting into the model, dynamic programming is used to identify all the longest common subsequences among the multiple texts. Then, the common subsequences are aligned, and the remaining parts are padded with “unk” tokens. Finally, the aligned and padded texts of the same length are used as input to the mBART-based machine translation model.

Average attention among candidates Thus, we propose an innovative approach, incorporating multiple ASR candidates into a single machine translation model(MC-mBART). Remarkably, the multi-candidate model can share identical parameters with the translation model, differing only in its attention averaging technique. This method averages attention at the sentence level, after all decoder layers, for multiple input candidates corresponding to the same translation result, prior to layer normal-

ization.

$$A(Q) = \frac{1}{n_{candidate}} \sum_{i=1}^{n_{candidate}} softmax\left(\frac{Q_i K^T}{\sqrt{d_k}}\right) V$$

During the inference process, the attention mechanism is utilized to compute the average attention score for beams that share the same candidate sequence number. This aggregation of attention scores forms a pool of candidate beams. Subsequently, the translation with the highest score is selected as the final output, presenting a compelling and impactful solution. It's worth noting that this does not add any parameters to the MT model.

4.2. SSL speech feature attention

During human conversation, there are often phenomena such as elision and assimilation, which lead to inaccurate results in automatic speech recognition (ASR). The process of converting audio recognition to text and then to machine translation further results in the loss of a large amount of linguistic information in speech. To enable machine models to more accurately filter out correct results, it is necessary not only to learn language patterns from machine translation, but also to obtain more rich and fine-grained linguistic information. Therefore, we propose a multi-candidate self-supervised learning speech machine translation model. We adopted the approach used in several papers (Lee et al., 2021) and opted to utilize HuBERT (Hsu et al., 2021) for generating the target self-supervised discrete units. This choice was influenced by the superior performance demonstrated by HuBERT in various tasks such as ASR, spoken language modeling, and speech synthesis, as shown in by (Yang et al., 2021; Lakhotia et al., 2021; Polyak et al., 2021). It outperforms other unsupervised representations, including VAE-based representations employed in (Tjandra et al., 2019; Zhang et al., 2021). We use the 11^{th} layer of Hubert as input, which contains richer linguistic information. After being encoded, it interacts with the decoder's cross-attention to obtain deeper language information. The rest of the process is similar to that of the multi-candidate machine translation model.

5. Experiments & Results

5.1. Data

ST Datasets We use GigaST (Ye et al., 2022b), a large-scale pseudo speech translation (ST) corpus containing 750w En-Zh pairs and MuST-C (Di Gangi et al., 2019a), a multilingual speech translation corpus whose size and quality will facilitate the training of end-to-end systems for SLT from English into 8 languages. For each target language, MuST-C comprises at least 385 hours of audio recordings from English TED Talks, which are automatically aligned at the sentence level with their manual transcriptions and translations. It is created by translating the text in GigaSpeech (Chen et al., 2021), an English ASR corpus, into German and Chinese. The training set is translated by a strong machine translation system and the test set is translated by human. We use its training set and use TED dev2010 and tst2015 (Liu et al., 2019) as the validation set. We apply the En-Zh test set contains the translation of all GigaSpeech test utterances. The test sets are produced by human translators looking at the transcriptions.

MT Datasets Our model allows us to use the external MT dataset for better finetuning. We introduce external WMT datasets for En->De/Zh/Es/Fr. The Detailed statistics of all the datasets are shown in Table below.

	train				test	
	GigaST	MuST-C	WMT	Total	tst-C	GigaST
en-zh	7.4M	-	-	7.4M	-	18.7k
en-de	7.4M	0.2M	5.9M	13.5M	2.6k	4.0k
en-fr	-	0.3M	15.2M	15.4M	2.5k	-
en-es	-	0.3M	15.2M	15.4M	2.6k	-

Table 4: Statistics of all datasets

5.2. Model Setup

5.2.1. Cascaded System

Our cascade model first uses Wenet (Yao et al., 2021) to perform ASR on Gigaspeech with 7.4M valid en-zh pairs, and we store the speech recognition texts corresponding to the top 20 cumulative log probability value rankings at the end of beam search. And the top n ranked texts are input as encoder into mBART based machine translation model. The architecture of the ASR model is U2++, a unified two-pass framework with bidirectional attention decoders, which includes the future contextual information by a right-to-left attention decoder. U2++ consists of four parts: a shared encoder, a CTC decoder, a left-to-right attention decoder, and a right-to-left attention decoder. We did not use any self-supervised speech models for initializing the ASR encoder. The encoder is trained from scratch following the WeNet paper. Taking inspiration from the approach proposed by (Lee et al., 2021), We use the multilingual HuBERT (mHuBERT) model and k-means model to encode source speech into a vocabulary of 1000 units. The mHuBERT and the k-means models are learned from the combination of En, Es and French unlabeled speech data from VoxPopuli (Wang et al., 2021a), while we use them to encode En speech only.

Models	External Data				BLEU			
	Speech	Text	ASR	MT	De	Es	Fr	Avg.
End-to-End Model								
MTL (Tang et al., 2021b)	-	-	-	✓	23.9	28.6	33.1	28.5
FAT-ST (Zheng et al., 2021)	✓	✓	✓	✓	25.5	30.8	-	-
JT-S-MT (Tang et al., 2021a)	-	-	-	✓	26.8	31.0	37.4	31.7
Chimera (Han et al., 2021a)	✓	-	-	✓	$27.1^{\dagger}$	30.6	35.6	31.1
XSTNet (Ye et al., 2021)	✓	-	-	✓	27.1	30.8	38.0	32.0
SATE (Xu et al., 2021)	-	-	✓	✓	$28.1^{\dagger}$	-	-	-
STEMM (Fang et al., 2022)	✓	-	-	✓	28.7	31.0	37.4	32.4
TaskAware (Indurthi et al., 2021)	-	-	✓	✓	28.9	-	-	-
STPT (Tang et al., 2022)	✓	✓	✓	✓	-	33.1	39.7	-
ConST (Ye et al., 2022a)	✓	-	-	✓	28.3	32.0	38.3	32.9
Cascade Model								
Espnet (Inaguma et al., 2021b)	-	-	-	-	23.6	-	33.8	-
W2V2-Transformer (Fang et al., 2022)	✓	-	-	✓	26.9	30.0	36.6	31.2
(Ye et al., 2021)	-	-	✓	✓	25.2	-	34.9	-
(Xu et al., 2021)	-	-	✓	✓	28.1	-	-	-
Machine Translation with mBART	-	-	-	✓	31.0	33.8	39.5	34.8
former cascaded system	-	-	-	✓	27.5	28.8	35.7	30.7
MC system(tuned on mBART)	-	-	-	✓	28.5	30.1	36.8	31.8

Table 5: Case-sensitive detokenized BLEU scores on MuST-C tst-COMMON set. "Speech" denotes unlabeled audio data, "Text" denotes unlabeled text data, e.g. Europarl V7 (Koehn, 2005), CC25 (Liu et al., 2020a), † use 40M OpenSubtitles (Lison and Tiedemann, 2016) as external MT data. MC system is the system applying multi-candidate strategy and finetuning on the machine translation model of mBART.

multi-candidate mBART The multi-candidate mBART is built upon the mBART model (Liu et al., 2020b), which is a sequence-to-sequence denoising auto-encoder pre-trained on extensive monolingual corpora from multiple languages using the BART objective (Lewis et al., 2019). The training of mBART involves the application of BART to large-scale monolingual corpora across various languages. In this process, input texts are noised through phrase masking and sentence permuting, and a single Transformer model (Vaswani et al., 2017) is trained to reconstruct the texts. mBART encompasses a fully autoregressive Seq2Seq model, trained once for all languages, providing a parameter set that can be fine-tuned for any language pair in both supervised and unsupervised settings, without requiring task-specific or language-specific modifications or initialization schemes.

In the fine-tuning phase, we experimented with three models: the pre-trained mBART model, a machine translation model fine-tuned on Gigaspeech-aligned text pairs with varying steps. We enhanced the fine-tuning process using the multi-candidate approach, running for n epochs on 8 A100 GPUs. The results show that our method significantly improves mBART model performance in various language pairs and translation tasks.

Our model configuration used several key parameters during training. We built on the pre-trained mBART model, with settings like sharing decoder embeddings, a single adaptor layer, normalization, 0.1 dropout probability, the same dropout for attention and ReLU layers, pretrained decoder loading, masking probabilities, encoder projection, a learning rate of 0.0005, inverse square root learning rate scheduler with warmup, Adam optimizer with specific beta values, gradient clipping, and defined maximum values for updates, tokens, tokens in validation, and source positions. These parameter choices crucially influenced the model's training behavior.

multi-candidate SSL-speech mBART We extract the hubert features at 11th layer for the input speech, and transform them into tokens with a word list of 1000 for the number of speech units using the kmeans model trained on English speech. Due to the high length of the original unit, we reduce the consecutive repetitive units into a single unit.

single-candidate mBART It is a baseline of the cascaded system. It has the same settings with the multi-candidate mBART apart from that it uses the first candidate from ASR results as the input of mBART.

5.2.2. Baseline systems

On the MuST-C dataset, We compare our method with several strong ST systems, including Espnet (Inaguma et al., 2021b), W2V2-Transformer (Fang et al., 2022), MTL (Tang et al., 2021b), FAT-ST (Zheng et al., 2021), JT-SMT (Tang et al., 2021a), Chimera (Han et al., 2021a), XST-Net (Ye et al., 2021), SATE (Xu et al., 2021), STEMM (Fang et al., 2022), TaskAware (Indurthi et al., 2021), STPT (Tang et al., 2022), ConST (Ye et al., 2022a).

On the GigaST dataset, we compare our method with the original cascaded system and two end-to-end ST systems, including SSL-Transformer and Speech-Transformer (Ye et al., 2022b).

5.3. Results

5.3.1. Capability to approximate machine translation without extra training

We applied machine translation models trained at different steps to the multi-candidate cascaded system framework and single-candidate cascaded framework, without modifying any parameters but altering the inference approach and input.

Interestingly, as we increased the training steps used for machine translation, the gap between multi-candidate cascaded system and the machine translation model gradually diminished, while the gap between multi-candidate cascaded system and single-candidate cascaded system grew larger. This discovery sheds light on the potential of harnessing multi-candidate utilization within established commercial machine translation models. By embracing this approach, we can unlock remarkable improvements in S2T accuracy without the necessity of modifying model parameters. It highlights the power of a subtle change in approach, which yields substantial gains and propels us towards superior performance. We can leverage existing commercial machine translation models at minimal additional cost to create high-quality speech-to-text translation models.

5.3.2. Multi-candidate strategy works

We observe that across the three models implemented on mBART with varying degrees of fine-tuning, MC-SSL-S-mBART exhibits the best performance, followed by MC-mBART, and finally single-candidate mBART. This clearly demonstrates the effectiveness of both self-supervised speech representation and the multi-candidate strategy. Moreover, while the transition from high-quality machine translation models to improved multi-candidate translation models requires no parameter changes, in the era of emerging and popular large language models, it is evident that our approach holds greater practical value and prospects than the End-to-End model.

5.3.3. Multi-candidate strategy makes full use of existing models

In Table 6, the term "constrained" refers to training solely on the GigaST's XL dataset, "unconstrained" indicates the use of a commercial machine translation model, while "semi-constrained" represents the results obtained by fine-tuning the pretrained mBART on GigaST. We can observe that in the constrained scenario, the cascaded system significantly lags behind the end-to-end system in terms of BLEU score. However, in the unconstrained scenario, the mature machine translation model achieves an impressive BLEU score of 44.3. In the semi-constrained scenario, our multi-candidate cascaded system approaches the performance of the machine translation model and surpasses the former cascaded system by a significant margin. This clearly demonstrates the success of the multi-candidate approach.

Furthermore, our model achieves a BLEU score of 37.3 after just one epoch of training, requiring only one hour. With a very short duration of fine-tuning, it surpasses the performance of former cascaded models, demonstrating the effectiveness of our approach.

In machine translation, for instance, mBART (Liu et al., 2020b) can be pre-trained by denoising complete text data from various languages. It is capable of transferring knowledge to language pairs without parallel text or those not included in the pre-training corpus. Languages that are not part of the pre-training corpus can benefit from machine translation, which strongly suggests that the initialization process is to some extent language-agnostic, and pre-training can capture common patterns in text. Pre-training is likely to be a promising strategy for future research as end-to-end models face challenges in learning across multiple languages. This further highlights the advantages and potential of cascade models.

5.3.4. Case study

However, by employing the multi-candidate strategy, we can observe that among the top five candidates based on their scores, there exist samples that deviate from conventional human language expressions. Nevertheless, in the attention mechanism of the machine translation process, words that align more closely with human expression and convey correct semantic meaning receive greater attention. As a result, the multi-candidate mBART model effectively filters out erroneous results and selects the correct translation.

Models	en-zh		
	constrained	semi	unconstrained
machine translation	24.9	40.2	44.3
Cascade Model			
former cascaded system	22.3	36.9	39.8
MC system	-	37.8	-
MC-sslS system	-	38.1	-
End-to-End Model			
SpeechTransformer	36.3	-	-
SSL-Transformer	38.0	-	-

Table 6: Giga-ST main results

As for the example of "Where the golgi apparatus sometimes called the golgi body receives them." The first candidate from ASR misrecognizes the pronunciation of the "golgi apparatus" as two non-existent words, "golgy" and "golji". However, the candidate with a BLEU score of 100 is ranked fifth. And the machine translation model that leverages rich text patterns through multi-candidates, allocates more attention to the correct candidate "golgi apparatus" and effortlessly selects the correct translation. Regarding the example of "Recording the transaction in an immutable distributed ledger" while the subsequent candidates include "legend", "literature", "letter", "ledger" and other alternatives. Our model successfully selects the correct phrase "distributed ledger" in terms of vocabulary collocation.

6. Ablation Study

Align or not align? We performed experiments on the same mBART model, both with and without the aligning module. Without the aligning module, the multi-candidate mBART yielded a lower result compared to the former cascaded system, achieving a BLEU score of 36.3. This to some extent indicates that the translation model's sentence-level average attention is related to the selection of vocabulary, thus providing partial validation for our explanation.

Finetune or not finetune? Our n-best strategy offers clear advantages over the lattice method, including resource efficiency, adaptability, and the ability to make lightweight modifications. For both the ASR and translation models, no additional weights are required. The change in the translation model involves adding average attention computation to each network layer (details below). The subsequent computations closely resemble those of a typical machine translation model. The difference between these two models boils down to a single computation step, highlighting the portability of the multi-candidate approach. Furthermore, with minimal fine-tuning, significant improvements in results can be achieved.

The best candidate performance While multiple ASR hypotheses can help mitigate some ASR errors, it's important to note that in certain cases, the overall quality may deteriorate. We conducted experiments using the same model as presented in Table 2 (utilizing 5 candidates), the GigaSpeech test dataset, and the established evaluation methodology. In Table 2, we observed that 45% of sentences with the lowest Word Error Rate (WER) also corresponded to the highest BLEU score. When considering the 20-best ASR candidates for these specific sentences, 45% of them showed a marginal 0.09 decrease in BLEU compared to Machine Translation (MT). This subtle finding suggests that, for other sentences, there might be a more significant improvement, highlighting the role of language patterns within the translation model in selecting the optimal candidate.

7. Conclusion

Our analysis pinpoints factors contributing to error propagation in cascade systems, such as pronunciation disparities, semantic differences, and domain-specific vocabulary errors. Our multi-candidate approach notably enhances speech-to-text (S2T) translation, bridging the S2T-T2T gap without altering model parameters. In summary, our research deepens our understanding of error propagation and linguistic information loss, thereby improving speech translation. With enhanced ASR and MT resources, our method narrows the S2T-T2T divide, providing increased accuracy and efficiency, all without additional parameters or modules.

Limitations

This work enhances speech translation through the utilization of a multi-candidate strategy. However, the current model falls short of achieving industrial-grade implementations. Although the ChatGPT and Whisper models demonstrate superior speech-to-text capabilities compared to our model, we argue that integrating speech and text remains a viable approach in the era of large-scale models. There is a notable limitations in this study that could be addressed in future research. Our model does not address the speed issue in the cascaded model's inference process. To improve speed, it is worth considering the application of the multi-strategy approach to streaming models in both ASR and MT domains.

8. Bibliographical References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

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

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L_1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Parnia Bahar, Tobias Bieschke, and Hermann Ney. 2019. A comparative study on end-to-end speech to text translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 792–799. IEEE.

Daniel Beck, Trevor Cohn, and Gholamreza Haffari. 2019. Neural speech translation using lattice transformations and graph networks. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 26–31.

Alexandre Bérard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. 2018. End-to-end automatic speech translation of audiobooks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6224–6228. IEEE.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *arXiv preprint arXiv:1612.01744*.

Nicola Bertoldi and Marcello Federico. 2005. A new decoder for spoken language translation based on confusion networks. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, pages 86–91. IEEE.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*.

Qiao Cheng, Meiyuan Fang, Yaqian Han, Jin Huang, and Yitao Duan. 2019. Breaking the data barrier: Towards robust speech translation via adversarial stability training. *arXiv preprint arXiv:1909.11430*.

J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.

N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

Siddharth Dalmia, Brian Yan, Vikas Raunak, Florian Metze, and Shinji Watanabe. 2021. Searchable hidden intermediates for end-to-end models of decomposable sequence tasks. *arXiv preprint arXiv:2105.00573*.

Mattia A Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019a. Must-c: a multilingual speech translation corpus. 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. Association for Computational Linguistics.

Mattia Antonino Di Gangi, Robert Enyedi, Alessandra Brusadin, and Marcello Federico. 2019b. Robust neural machine translation for clean and noisy speech transcripts. *arXiv preprint arXiv:1910.10238*.

Linhao Dong, Shuang Xu, and Bo Xu. 2018. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5884–5888. IEEE.

Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.

Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022. Stemm: Self-learning with speech-text manifold mixup for speech translation. *arXiv preprint arXiv:2203.10426*.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. 2021a. Learning shared semantic space for speech-to-text translation. *arXiv preprint arXiv:2105.03095*.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. 2021b. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.

Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

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. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 922–929. IEEE.

Hirofumi Inaguma, Tatsuya Kawahara, and Shinji Watanabe. 2021b. Source and target bidirectional knowledge distillation for end-to-end speech translation. *arXiv preprint arXiv:2104.06457*.

Sathish Indurthi, Mohd Abbas Zaidi, Nikhil Kumar Lakumarapu, Beomseok Lee, Hyojung Han, Seokchan Ahn, Sangha Kim, Chanwoo Kim, and Inchul Hwang. 2021. Task aware multi-task learning for speech to text tasks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7723–7727. IEEE.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. 2019. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7180–7184. IEEE.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Tsz Kin Lam, Shigehiko Schamoni, and Stefan Riezler. 2022. Sample, translate, recombine: Leveraging audio alignments for data augmentation in end-to-end speech translation. *arXiv preprint arXiv:2203.08757*.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2021. Lightweight adapter tuning for multilingual speech translation. *arXiv preprint arXiv:2106.01463*.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, et al. 2021. Direct speech-to-speech translation with discrete units. *arXiv preprint arXiv:2107.05604*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

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*.

Zhaolin Li and Jan Niehues. 2022. Efficient speech translation with pre-trained models. *arXiv preprint arXiv:2211.04939*.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020a. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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

Yuchen Liu, Hao Xiong, Zhongjun He, Jiajun Zhang, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-end speech translation with knowledge distillation. *arXiv preprint arXiv:1904.08075*.

Evgeny Matusov, Stephan Kanthak, and Hermann Ney. 2005. On the integration of speech recognition and statistical machine translation. In *Ninth European Conference on Speech Communication and Technology*.

Stephan Peitz, Simon Wiesler, Markus Nußbaum-Thom, and Hermann Ney. 2012. Spoken language translation using automatically transcribed text in training. In *Proceedings of the 9th International Workshop on Spoken Language Translation: Papers*.

Juan Pino, Qiantong Xu, Xutai Ma, Mohammad Javad Dousti, and Yun Tang. 2020. Self-training for end-to-end speech translation. *arXiv preprint arXiv:2006.02490*.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. Speech resynthesis from discrete disentangled self-supervised representations. *arXiv preprint arXiv:2104.00355*.

Sravya Popuri, Peng-Jen Chen, Changhan Wang, Juan Pino, Yossi Adi, Jiatao Gu, Wei-Ning Hsu, and Ann Lee. 2022. Enhanced direct speech-to-speech translation using self-supervised pre-training and data augmentation. *arXiv preprint arXiv:2204.02967*.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.

Matthias Sperber, Graham Neubig, Ngoc-Quan Pham, and Alex Waibel. 2019. Self-attentional models for lattice inputs. *arXiv preprint arXiv:1906.01617*.

Mihaela C Stoian, Sameer Bansal, and Sharon Goldwater. 2020. Analyzing asr pretraining for low-resource speech-to-text translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7909–7913. IEEE.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.

Yun Tang, Hongyu Gong, Ning Dong, Changhan Wang, Wei-Ning Hsu, Jiatao Gu, Alexei Baevski, Xian Li, Abdelrahman Mohamed, Michael Auli, et al. 2022. Unified speech-text pre-training for speech translation and recognition. *arXiv preprint arXiv:2204.05409*.

Yun Tang, Juan Pino, Xian Li, Changhan Wang, and Dmitriy Genzel. 2021a. Improving speech translation by understanding and learning from the auxiliary text translation task. *arXiv preprint arXiv:2107.05782*.

Yun Tang, Juan Pino, Changhan Wang, Xutai Ma, and Dmitriy Genzel. 2021b. A general multi-task learning framework to leverage text data for speech to text tasks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6209–6213. IEEE.

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, Sakriani Sakti, and Satoshi Nakamura. 2019. Speech-to-speech translation between untranscribed unknown languages. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 593–600. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021a. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*.

Changhan Wang, Anne Wu, Juan Pino, Alexei Baevski, Michael Auli, and Alexis Conneau. 2021b. Large-scale self-and semi-supervised learning for speech translation. *arXiv preprint arXiv:2104.06678*.

Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. *arXiv preprint arXiv:1703.08581*.

Chen Xu, Bojie Hu, Yanyang Li, Yuhao Zhang, Qi Ju, Tong Xiao, Jingbo Zhu, et al. 2021. Stacked acoustic-and-textual encoding: Integrating the pre-trained models into speech translation encoders. *arXiv preprint arXiv:2105.05752*.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. 2021. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*.

Zhuoyuan Yao, Di Wu, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, and Xin Lei. 2021. Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. *arXiv preprint arXiv:2102.01547*.

Rong Ye, Mingxuan Wang, and Lei Li. 2021. End-to-end speech translation via cross-modal progressive training. *arXiv preprint arXiv:2104.10380*.

Rong Ye, Mingxuan Wang, and Lei Li. 2022a. Cross-modal contrastive learning for speech translation. *arXiv preprint arXiv:2205.02444*.

Rong Ye, Chengqi Zhao, Tom Ko, Chutong Meng, Tao Wang, Mingxuan Wang, and Jun Cao. 2022b. Gigast: A 10,000-hour pseudo speech translation corpus. *arXiv preprint arXiv:2204.03939*.

Wenbiao Yin, Zhicheng Liu, Chengqi Zhao, Tao Wang, Jian Tong, and Rong Ye. 2023. Improving speech translation by fusing speech and text. *arXiv preprint arXiv:2305.14042*.

Binbin Zhang, Di Wu, Zhendong Peng, Xingchen Song, Zhuoyuan Yao, Hang Lv, Lei Xie, Chao Yang, Fuping Pan, and Jianwei Niu. 2022. Wenet 2.0: More productive end-to-end speech recognition toolkit. *arXiv preprint arXiv:2203.15455*.

Chen Zhang, Xu Tan, Yi Ren, Tao Qin, Kejun Zhang, and Tie-Yan Liu. 2021. Uwspeech: Speech to speech translation for unwritten languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14319–14327.

Pei Zhang, Boxing Chen, Niyu Ge, and Kai Fan. 2019. Lattice transformer for speech translation. *arXiv preprint arXiv:1906.05551*.

Renjie Zheng, Junkun Chen, Mingbo Ma, and Liang Huang. 2021. Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation. In *International Conference on Machine Learning*, pages 12736–12746. PMLR.