Recent Advances in Direct Speech-to-text Translation

Chen Xu[1] , **Rong Ye**[2] , **Qianqian Dong**[2] , **Chengqi Zhao**[2] , **Tom Ko**[2] ,
Mingxuan Wang[2*] , **Tong Xiao**[1,3†] and **Jingbo Zhu**[1,3]
[1]School of Computer Science and Engineering, Northeastern University, Shenyang, China
[2]ByteDance
[3]NiuTrans Research, Shenyang, China
xuchennlp@outlook.com, {xiaotong, zhujingbo}@mail.neu.edu.cn
{yerong, dongqianqian, zhaochengqi.d, tom.ko, wangmingxuan.89}@bytedance.com

Abstract

Recently, speech-to-text translation has attracted more and more attention and many studies have emerged rapidly. In this paper, we present a comprehensive survey on direct speech translation aiming to summarize the current state-of-the-art techniques. First, we categorize the existing research work into three directions based on the main challenges — modeling burden, data scarcity, and application issues. To tackle the problem of modeling burden, two main structures have been proposed, encoder-decoder framework (Transformer and the variants) and multitask frameworks. For the challenge of data scarcity, recent work resorts to many sophisticated techniques, such as data augmentation, pre-training, knowledge distillation, and multilingual modeling. We analyze and summarize the application issues, which include real-time, segmentation, named entity, gender bias, and code-switching. Finally, we discuss some promising directions for future work.

1 Introduction

Speech-to-text translation (ST) is a task that aims to translate speech in one language to text in another language. It has numerous practical applications, including global communication, language learning, and accessibility for non-native speakers.

Early solutions for speech translation are to break down the task into smaller and more manageable sub-tasks, such as automatic speech recognition (ASR) and machine translation (MT). This is the idea of the cascaded system. For example, to fulfill the ST task, we can cascade an ASR system to transcribe speech into text with an MT system to translate text into another language in tandem [Stentiford and Steer, 1988]. Research on cascaded systems mainly aims at solving the problem of error accumulation, such as utilizing multiple recognition results and training robust MT models.

Meanwhile, the end-to-end speech translation (E2E ST) model, which eliminates the need for intermediate steps (*e.g.*, ASR and MT), is designed and also has the potential to eliminate error accumulation. In addition, it also has the advantage of reduced latency, more contextual modeling [Bentivogli *et al.*, 2021], and applicability to unwritten languages [Bérard *et al.*, 2016]. In recent years, research on end-to-end models in speech translation has gained momentum, leading to diversity in model architectures and training methods. However, a comprehensive survey that thoroughly reviews their motivations and practices is currently lacking.

ST corpus usually contains the source speech $\mathbf{s}$, transcription $\mathbf{x}$, and translation $\mathbf{y}$. The basic model framework of E2E ST is mainly based on the encoder-decoder structure. The encoder encodes the speech input into a sequence of hidden states, and the decoder outputs the final translation result condition on the hidden states, which is basically autoregressive. The objective training function of the ST model θ is the negative log-likelihood loss:

$$L_\theta = -\mathbb{E}_{s,y} \log p(\mathbf{y}|\mathbf{s};\theta) = -\mathbb{E}_{s,y} \sum_{t=1}^{T} \log p(y_t|\mathbf{y}_{<\mathbf{t}}, \mathbf{s};\theta)$$

where T is the length of $\mathbf{y}$. In the inference stage, we usually apply beam search to generate target sentences.

However, we find that training an E2E ST model is not easy. Although the study also confirms that the performance of the end-to-end model is approaching the results of the cascaded solution, it is still not the best-performing technology. Existing literature mainly attributes and attempts to address the following challenges:

- **Modeling burden**: Conventional cascaded systems decouple it into ASR and MT models, while it is non-trivial and burdensome for speech translation in an E2E manner. This is because the E2E model requires both cross-modal and cross-lingual mapping at the same time. Training the E2E model often encounters poor convergence and low performance [Bérard *et al.*, 2016; Weiss *et al.*, 2017].
- **Data scarcity**: Annotating speech translation data is demanding, so labeled parallel data for training is scarce. For example, an ASR dataset like Librispeech[1] contains 960 hours of speech and the MT dataset typically has

[*] Corresponding Authors.

[1] https://www.openslr.org/12

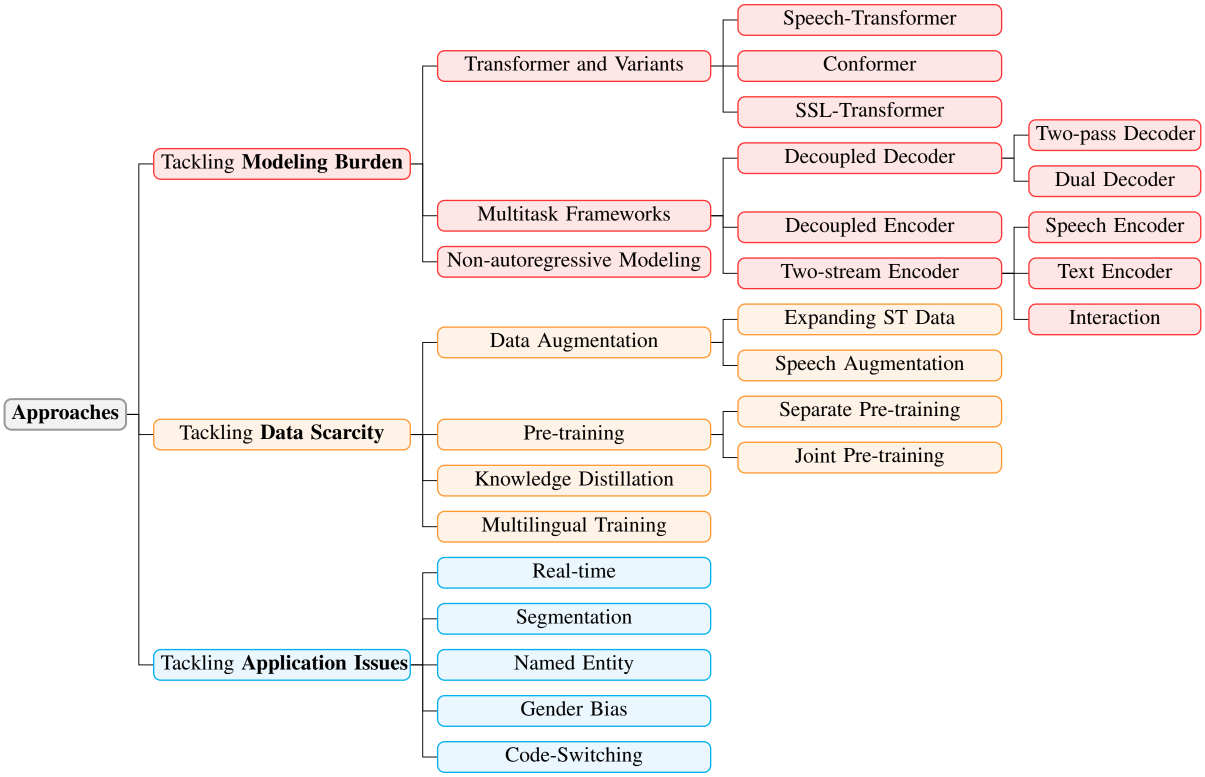

Figure 1: Taxonomy of speech-to-text translation approaches.

millions of bi-texts, while the ST dataset, such as MuST-C[2] contains only about 400 hours of speech with 230k utterances. Data scarcity results in the E2E model being inferior to cascaded systems trained on abundant ASR and MT data [Sperber and Paulik, 2020], which is more severe in the industry.

- **Application issues**: In addition to performance, there are other considerations in the practical implementation, such as real-time, long-form audio segmentation, gender bias, named entity translation, and code-switching speech.

Correspondingly, in this paper, we provide a comprehensive survey of how previous work has tackled the above challenges, and hope to suggest some directions for future research in this community. As shown in the taxonomy in Figure 1, our survey is developed as follows.

- Section 2 describes how to alleviate the modeling burden challenge in the existing literature. Modeling methods can be divided into three categories: Transformer and the variants, multitask frameworks, and non-autoregressive modeling.
- Section 3 summarizes approaches to tackle the problem of data scarcity, including data augmentation, pre-training, knowledge distillation, and multilingual training.
- Section 4 briefly introduces application issues (real-time, segmentation, named entity, gender bias, code-switching) in practice and recent related work.
- Section 5 anticipates some promising directions for future ST research.

2 Tackling Modeling Burden

For the long sequence input, like speech, we require high-capacity E2E models, typically Transformer-based ST models and their variants (Section 2.1). Additionally, to address the issue of modeling burden, the existing literature generally employs a multitask framework to make modifications to the original Transformer-based model. We categorize and introduce them in Section 2.2. Finally, for the consideration of decoding efficiency, there are also non-autoregressive models introduced in Section 2.3.

2.1 Transformer and Variants

The ST task is based on the sequence-to-sequence modeling, which typically adopts the encoder-decoder architecture, as shown in Figure 2(a). Among many well-established networks, Transformer [Vaswani *et al.*, 2017] is chosen for its state-of-the-art performance across almost all sequence generation tasks. Several variants have been proposed to make Transformer more suitable for speech modeling. Here we highlight Speech-Transformer, Conformer, and SSL-Transformer.

Speech-Transformer. Speech-Transformer [Gangi *et al.*, 2019] is built on top of the text-to-text Transformer [Vaswani *et al.*, 2017]. The major difference is that the sequence of

[2] https://ict.fbk.eu/must-c/

audio features (*e.g.* Fbank) is first compressed by the convolutional layers (typically two layers with a stride of 2, compressing the length by a factor of 4) and a normalization layer before the self-attention encoder.

Conformer. Conformer is a convolution-augmented Transformer model [Gulati *et al.*, 2020]. The main feature of the Conformer is the convolution module, which is inserted between the multi-head self-attention module and the feed-forward layer of each encoder block. The convolution module has the attention and convolution modules, sandwiched by two Macaron-net style feed-forward layers and the residual connections. The combination of CNN and Transformer helps to model both local and global information, which is suitable for encoding long-sequence speech.

SSL-Transformer. With the success of self-supervised learned (SSL) speech representations, such as Wav2vec-family [Schneider *et al.*, 2019] and HuBERT [Hsu *et al.*, 2021] on ASR, they have also been utilized in the encoder of ST models, which we collectively refer to as *SSL-Transformer*. The original audio waveform is fed into the SSL model, which subsequently processes the waveform through several convolutional layers and Transformer encoder layers to extract speech features. In the SSL-Transformer model, the SSL model can be incorporated into the decoder either as a standalone encoder [Wu *et al.*, 2020; Wang *et al.*, 2021] or as a speech feature extractor, which is then connected to the whole Transformer [Han *et al.*, 2021; Ye *et al.*, 2021].

2.2 Multitask Frameworks

The idea of the multitask framework is to utilize related auxiliary tasks to enhance the target task. For ST, the auxiliary tasks are often ASR and MT. As for the model structure, some parameters of the target and auxiliary task modules can be shared, while there are parts of the modules that remain independent. We summarize the multitasking frameworks in the extant literature (in chronological order), which can be broadly classified into the following three types, namely **decoupled decoder** (Figure 2(b)), **decoupled encoder** (Figure 2(c)), and **two-stream encoder** (Figure 2(d)).

Decoupled Decoder

To facilitate the burden of direct cross-modal and cross-lingual modeling, the additional decoder is introduced to guide the learning of transcript, while the model is still trained in an E2E manner, as shown in Figure 2(b). The naive approach only decodes the transcription as the auxiliary task [Weiss *et al.*, 2017], and then the researchers further explore how to better prompt translation by the generated transcription (*two-pass decoder*) [Anastasopoulos and Chiang, 2018] or generate transcription and translation synchronously (*dual decoder*) [Liu *et al.*, 2020a].

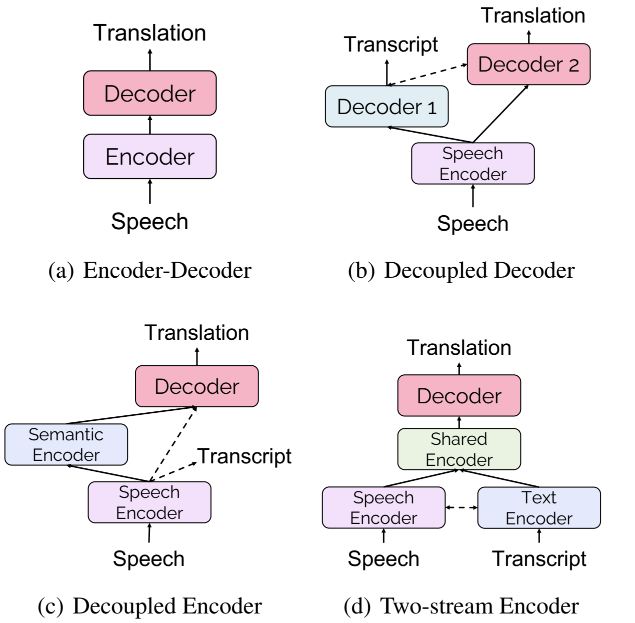

Figure 2: Schematic diagram of the model architectures. The dashed lines indicate possible interactions between the modules.

- **Two-pass decoder:** Taking the acoustic feature as input, the model generates the transcription by the first decoder, then combines the representation of the encoder and the first decoder to generate the translation [Sperber *et al.*, 2019]. However, two-pass decoding loses the inherent advantages of low latency due to sequential generation. To do this, Inaguma *et al.* [2021a] utilize the non-autoregressive method for the first-stage decoding.
- **Dual decoder:** Considering that generation processes of transcription and translation can help each other, interactive decoding [Liu *et al.*, 2020a] is proposed to generate transcription and translation using two decoders synchronously. Also, an additional cross-attention module is used to capture the information from the two decoders to each other. To further improve translation performance, the wait-k policy is introduced, where transcribed tokens are first predicted, in order to provide more useful information for the decoding of the translation tokens.

Decoupled Encoder

Although the E2E model can use the decoupled decoder to partially alleviate the heavy modeling burden, multiple inferences usually lead to complex designs and high latency. Instead of the decoupled decoder, a better choice is to simultaneously recognize and understand the semantics of the original speech input by a decoupled encoder. As shown in Figure 2(c), the decoupled encoder generally has two encoders — the low-level speech encoder first encodes the acoustic information from the speech input, and the semantic encoder further learns the semantic representation needed for translation decoding.

Each stage of encoding can be supervised by the information from the transcription, such as phonemes, text, etc. Such decoupling mimics the cascaded system while transcription provides an alignment to the speech, which helps to ease the encoding burden. For example, studies [Liu *et al.*, 2020b; Dong *et al.*, 2021; Xu *et al.*, 2021] decouple the encoder and add a Connectionist Temporal Classification (CTC) loss term after the speech encoder to predict the transcription. Xu *et al.* [2021] analyze the different behavior of the self-attention in the encoders of ASR, MT and ST models with respect to the attention to local information, and demonstrate that for ST, the CTC module acts more efficiently over the low-level acoustic encoder. Given the length gap between speech and text, Liu *et al.* [2020b] propose a shrinking mechanism based on CTC loss and integrate with multitask learning. In addition, Dong *et al.* [2021] introduce an additional pre-trained BERT model to supervise the output of the high-level acoustic encoder.

Two-stream Encoder

In decoupling methods, it is easy to further boost part of components by utilizing the additional ASR data. However, having the potential to help the ST model improve the ability of semantic understanding, abundant knowledge from MT data is not exploited. As shown in Figure 2(d), the two-stream encoder is proposed to accept either speech or text or both as inputs during training. Speech and text have both their separate encoders (*speech encoder* and *text encoder*, respectively) as the pre-net and a *shared encoder* stacked on top. This structure is usually optimized by multitask training losses, such as NLL losses for both ST and MT [Ye *et al.*, 2021; Han *et al.*, 2021; Tang *et al.*, 2021a; Tang *et al.*, 2021b; Ye *et al.*, 2022]. Its advantage is that better semantic representations can be learned by sharing with the MT encoder to improve translation performance. During inference, we input the speech, and the translated text is generated through the speech encoder, shared encoder, and decoder.

- **Speech encoder:** Speech encoder needs to be more capable to extract acoustic features of the speech input separately. Pre-trained speech models such as Wav2vec2 can be used as the speech encoder better ST performances [Ye *et al.*, 2021; Han *et al.*, 2021; Tang *et al.*, 2021a; Fang *et al.*, 2022; Ye *et al.*, 2022].
- **Text encoder:** Text encoder can be the text embedding layer or a few layers of the textual Transformer encoder. Tang *et al.* [2021b] propose to replace the original transcription with the phoneme of the speech as the text input, which helps to reduce the gap between two input modalities, thus improving performance.
- **Interaction:** Under the overall two-steam encoder framework, there are also multiple variants of interaction between the speech encoder and the text encoder or their output representations. Ye *et al.* [2022] notice the sentence-level representation gap in the vector space between the speech and text, and propose to apply the contrastive learning method to draw close the two representations. Starting from the length gap between the speech and text, Han *et al.* [2021] propose the Chimera model, whose core idea is to align and map audio and text representations to the same length by a fix-size shared semantic memory module, and the decoder cross-attends to the memory module during autoregressive generation. Considering both representations and the length difference, Tang *et al.* [2021a] add a cross-attentive regularization module after the shared encoder. The regularization module first generates two reconstructed sequences from text or speech encoders with the same length via self-attention or cross-attention, and then optimizes the L2 distance between the reconstructed sequences. Indurthi *et al.* [2021] design a Task Characteristics Network (TCN) that produces a task embedding to modulate the parameters of the shared encoder-decoder.

2.3 Non-autoregressive Modeling

E2E modeling reduces latency by almost half compared to cascaded counterpart, which helps applications in real-world scenarios with limited computational resources. However, this advantage only holds in the context of autoregressive decoding, which generates each token depending on the previously predicted tokens. The recently proposed non-autoregressive (NAR) decoding predicts the whole sequence in parallel, eliminating the advantage of the E2E model in inference latency.

To combine E2E modeling and NAR generation, several studies explore NAR speech translation. There are two design concepts. Following the methods in ASR and MT tasks, one can combine multiple existing techniques, like conditional masked language model and re-scoring. Another route explores a more efficient architecture that relies only on CTC for prediction [Chuang *et al.*, 2021], which has the potential to achieve high speed-up. However, the current non-autoregressive model is still inferior to the autoregressive counterpart with a large gap of about $2 \sim 3$ BLEU points. We need more effort to develop a powerful NAR model with comparable performance.

3 Tackling Data Scarcity

Because of the difficulty in accumulating data, the training data for ST is much less, compared to MT or ASR. First, expanding the dataset and data augmentation are the most straightforward ideas (Section 3.1). Second, the majority of the existing research focuses on how to gain more knowledge or information from MT or ASR data or models to improve the performance of the ST model. We divide them into two parts: pre-training (Section 3.2) and knowledge distillation (Section 3.3). Finally, we present some progress on the current multilingual speech translation in Section 3.4.

3.1 Data Augmentation

Data augmentation is the most straightforward idea when training data is scarce.

Expanding ST data. Intuitively, we can expand a large amount of target language translation by using a high-quality off-the-shelf MT system on a large amount of ASR data [Pino *et al.*, 2020; Wang *et al.*, 2021]. This method is often referred to as pseudo-labeling or sequence-level knowledge distillation (SeqKD). Inaguma *et al.* [2021b] propose bidirectional SeqKD that fully leverages knowledge in both source and target language directions (corresponding to forward SeqKD and backward SeqKD) for bilingual E2E-ST models. The source language speech can also be augmented, we can use the text-to-speech (TTS) model to extend the source-side text of MT into speech [Jia *et al.*, 2019].

Speech augmentation. Speech augmentation is also useful to improve performance as well as robustness, because speech is more complex and varied than text, and a single utterance can have different continuous signals depending on the speaker, recorder, environment, and so on. SpecAugment [Park *et al.*, 2019] is commonly applied directly to the filter bank coefficients of speech inputs. The augmentation strategy contains warping features, masking blocks of frequency channels, and time steps. It has been proven to be effective on both ASR and ST tasks [Bahar *et al.*, 2019]. In addition, more diverse ST data can also be constructed by various segmentation methods and recombination to enhance the utility of the original ST data.

3.2 Pre-training

On many tasks in the AI field, pre-training can greatly improve model performance in low-resource situations. Pre-training is generally considered to have the following benefits: (1) Compared to speech-to-translation corpora, the data used for pre-training is usually easy to obtain, such as a large amount of raw data of text sentences or speech. The large-scale data used in pre-training (whether in-domain or out-of-domain) helps to improve the robustness of the model for the downstream tasks. (2) Through basic and various pre-training tasks, such as reconstruction, mask-prediction, and contrastive learning, we can obtain more accurate representations with contextual information. These representations are generally helpful for various downstream tasks.

Pre-training is very effective in improving the performance of end-to-end speech translation, and throughout the state-of-the-art (SOTA) E2E ST models, pre-training is always involved. We outline two pre-training modes, namely *separate pre-training* and *joint pre-training*, based on the percentage of pre-trained modules in the E2E ST model.

Separate pre-training. Separate pre-training refers to the pre-training of a portion of the model parameters or the pre-training of different sub-modules via different tasks. The earlier work explores better pre-training methods to enhance the ability of the encoder in terms of semantic understanding. Wang *et al.* [2020b] pre-train the encoder using a curriculum learning method to improve syntactic and semantic modeling abilities. Chen *et al.* [2020] propose a self-supervised method called Masked Acoustic Modeling (MAM), which randomly masks part of the speech spectrogram and then recovers it on top of the encoder. In addition, as discussed in Section 2.1, the self-supervised model such as Wav2vec [Schneider *et al.*, 2019] can act as a feature extractor [Wu *et al.*, 2020] instead of random parameter initialization, providing effective acoustic features as input.

Joint pre-training. Joint pre-training means that the model (all modules including both encoder and decoder) participates in pre-training as a whole. Joint pre-training usually enjoys a multitask learning framework, which is introduced in Section 2.2). In the multitask pre-training framework, unified models are built to jointly pre-train ASR, MT, masked language modeling, or even speech (re-)synthesis tasks, using speech-text supervised data as well as large amounts of unlabeled text and speech. After pre-training, the models only need to be fine-tuned with the speech-translated parallel corpus to achieve a decent result. For instance, Ao *et al.* [2022] propose SpeechT5, which pre-trains various speech/text-to-speech/text tasks, including ASR, ST, text-to-speech, speech conversion, and speech enhancement. Tang *et al.* [2022] incorporate both self-supervised and supervised pre-training tasks, including (self-)supervised text-to-text, speech SSL learning, speech-to-phoneme, ASR, and ST. In addition, pre-training can also be combined with multilingualism. Bapna *et al.* [2022] propose mSLAM, a large multilingual speech-text Conformer model based on the two-stream encoder, which surprisingly shows some zero-shot learning capability. Cheng *et al.* [2022] further combine mSLAM with multitask learning, propose Mu2SLAM and obtain the SOTA results on CoVoST-2 [Wang *et al.*, 2020a] multilingual ST benchmark.

3.3 Knowledge Distillation

Knowledge distillation (KD) is typically used for model compression, using the output of a larger teacher model that typically performs better to guide the learning of a student model, with the expectation that the student model will achieve the same performance as the teacher model. With limited data, how can we get the ST performance close to that of the MT teacher? The idea of knowledge distillation is then widely used in speech translation. A straightforward approach is to use the ST model and the MT model to predict translation tokens separately, with the prediction probabilities of the MT model serving as the teacher to guide the ST output [Liu *et al.*, 2019; Tang *et al.*, 2021a]. With a two-stream encoder framework, Fang *et al.* [2022] propose to distill the speech-to-text translation module with the translation output from the speech-text manifold mix-up sequence. Experiments show that the mix-up sequence can bridge the representation gap between speech and text, thus making the learned semantic representation of text more readily transferable to speech.

3.4 Multilingual Training

Multilingual speech translation includes one-to-many, many-to-one, and many-to-many scenarios. Like MT, adding language indicators, such as `<2de>`, `<2fr>`, to the decoder is the most straightforward and effective way to evolve from bilingual to multilingual ST [Inaguma *et al.*, 2019]. Wang *et al.* [2020a] also show that with limited data for each translation direction, training a many-to-many multilingual ST model is better than training bilingual ST models individually, because the multilingual model can capture more pronunciation similarity between languages. Current research on multilingual ST mainly focuses on pre-training, such as how to build a unified multilingual speech-text pre-training model [Bapna *et al.*, 2022] and how to design various and effective pre-training tasks [Cheng *et al.*, 2022]. These models can be helpful for translation as well as multilingual ASR. There is also some work focusing on efficient fine-tuning. Li *et al.* [2021] concatenate multilingual pre-trained XLSR speech encoder with mBART decoder and experimentally find that fine-tuning the parameters of layer-norm and attention layers is better than fine-tuning all parameters. Le *et al.* [2021], on the other hand, freeze the pre-trained ASR encoder and the mBART decoder, and complete one-to-many

ST by only tuning the language-specific adapter modules on top of a multilingual system, with only tens of millions of parameters.

4 Tackling Application Issues

Current research is usually conducted under presupposed settings with manual segmentation, noise-free environment, etc. And the demands of practical application are rarely discussed.

Real-time. Simultaneous decoding aims for the quality-latency trade-off for real-time translation on the depend of a decision policy, which determines whether to wait for more audio stream or decode one or more tokens. SimulSpeech [Ren *et al.*, 2020] propose a speech segmenter, based on the CTC criterion, to split the streaming speech in real time. Chang and Lee [2022] adapt Continuous Integrate-and-Fire module to play a role as the adaptive policy, which makes WRITE decisions at each firing step. Liu *et al.* [2021] extend RNN-Transducer into Cross Attention Augmented Transducer, which can jointly optimize the decoding policy and translation quality by considering all possible READ and WRITE action paths.

Segmentation. The direct ST model cannot handle long audios alone, such as a complete speech or a movie, which have to be segmented into shorter utterances using automatic segmentation methods (VAD-based, fixed-length, and hybrid). However, there exists a gap between manual segmentation during training and automatic segmentation at runtime. Tsiamas *et al.* [2022] propose Supervised Hybrid Audio Segmentation (SHAS) method, which uses Wav2vec2 and trains a classifier to predict the split locations supervised by the manual segmentation information.

Named entity. How to handle the translation of named entity (NE) is a critical demand for ST systems in real-world scenarios. Gaido *et al.* [2022a] discover that the nationality of the referred person is the key factor for the failures in person name translation, and propose multilingual models to increase the robustness of varied pronunciations. Gaido *et al.* [2022b] design two methods to jointly perform ST and recognize NE, of which the *inline* method generates NE tags and tokens successively, while the *parallel* method predicts NE tags and tokens in parallel with two linear layers.

Code-switching. Code-switching (CS) speech commonly exists in casual situations, and as blending different languages, translating CS speech is a challenge. Weller *et al.* [2022] create a CS corpus and explore both cascaded and end-to-end architecture to perform CS speech translation. Huber *et al.* [2022] propose a unified Language Agnostic E2E ST model (LAST) by training both ASR and ST tasks, as well as enlarging CS data through concatenation.

Gender bias. Addressing gender bias in translation is a relatively new area of NLP and speech research. For ST task, audio input contains more clues about gender identity. Bentivogli *et al.* [2020] release MuST-SHE benchmark allowing for the fine-grained analysis of gender bias in ST. They also find that the end-to-end approach can directly use audio information and have more potential to better address gender issues. Savoldi *et al.* [2022] later extend MuST-SHE with two additional linguistic information, part-of-speech and agreement chains. Gaido *et al.* [2021] investigate how segmentation methods influence the translation of gender, and propose a combined segmentation method with both subword splitting and character-based splitting.

5 Future

In this paper, we thoroughly present recent advances in direct speech-to-text translation. Specifically, we review and summarize existing approaches in this field for the first time with an original taxonomy. Despite the recent attractive progress of direct ST technology, there remain many unresolved problems to be explored. Finally, we discuss some promising topics for the future.

LLM. Today, large language models (LLMs), such as ChatGPT, have shown powerful dexterity in a variety of NLP applications, such as text generation and even in the zero-shot scenario. First of all, we believe that it is worthwhile to further explore how to integrate the powerful generative capabilities of LLMs into ST tasks and to incorporate speech data into the training of LLMs. As an initial step, for instance, we may optimize speech representation to be comparable to the text representation as a prompt function to interact with LLMs. We conjecture that speech discrete representations as pseudo-language [Hsu *et al.*, 2021; Wu *et al.*, 2022] may be interesting prompts. Furthermore, pre-training large-scale acoustics-aware LLMs is also a promising direction that will greatly promote the entire NLP and speech community. We anticipate that after scaling up further, the models will have the capability for few-shot ST, zero-shot ST, and transfer learning.

Multimodality. Numerous human-computer interaction (HCI) application scenarios have emerged with the recent worldwide surge in AI-generated content (text, images, voice, and video, etc.), which drives the ST field to explore more sophisticated directions, like speech-to-speech translation and video translation. With the explosive growth of multimodal resources, how to perform in-context learning (ICL) on multimodal data is also a promising research topic. Recently, multimodal pre-training has already been proved to be effective in many fields. However, the interactions and interrelated information between multiple modalities (e.g., the speech of characters in videos and their corresponding image frames and prosodic environments) remain underutilized. We believe that a more unified and robust pre-training paradigm, aimed at learning universal cross-lingual cross-modal representations, is important for ST and the more demanding HCI scenarios mentioned above.

Acknowledgements

This work was supported in part by the National Science Foundation of China (No. 62276056), the National Key R&D Program of China, the China HTRD Center Project (No. 2020AAA0107904), the Natural Science Foundation of Liaoning Province of China (2022-KF-16-01), the Yunnan Provincial Major Science and Technology Special Plan

Projects (No. 202103AA080015), the Fundamental Research Funds for the Central Universities (Nos. N2216016, N2216001, and N2216002), and the Program of Introducing Talents of Discipline to Universities, Plan 111 (No. B16009).

Contribution Statement

Chen Xu, **Rong Ye**, and **Qianqian Dong** have the equal contribution to the conception and design of the survey. They performed the literature search, collected relevant papers, and drafted the main sections of the manuscript. They also contributed to the synthesis and analysis of the surveyed work and were responsible for creating the figures.

Chengqi Zhao provided substantial contributions to the refinement of the survey's scope and focus. He helped in the analysis of the surveyed work and provided insights on potential future research directions. In addition, he was responsible for critically reviewing the manuscript at different stages and provided suggestions for improving the clarity and coherence of the paper.

Tom Ko, **Mingxuan Wang**, **Tong Xiao**, and **Jingbo Zhu** provided guidance on the overall direction and structure of the survey paper. They contributed to the refinement of the paper's scope and focus, and provided critical feedback on the manuscript at various stages of development. They also helped in the identification of relevant literature and offered insights on the state-of-the-art and future trends in end-to-end speech translation.

References

[Anastasopoulos and Chiang, 2018] Antonios Anastasopoulos and David Chiang. Tied multitask learning for neural speech translation. In *NAACL*, 2018.

[Ao *et al.*, 2022] Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *ACL*, 2022.

[Bahar *et al.*, 2019] Parnia Bahar, Albert Zeyer, Ralf Schlüter, and Hermann Ney. On using SpecAugment for end-to-end speech translation. In *IWSLT*, 2019.

[Bapna *et al.*, 2022] Ankur Bapna, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, and Alexis Conneau. mslam: Massively multilingual joint pre-training for speech and text. *CoRR*, 2022.

[Bentivogli *et al.*, 2020] Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. Gender in danger? evaluating speech translation technology on the MuST-SHE corpus. In *ACL*, 2020.

[Bentivogli *et al.*, 2021] Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. Cascade versus direct speech translation: Do the differences still make a difference? In *ACL*, 2021.

[Bérard *et al.*, 2016] Alexandre Bérard, Olivier Pietquin, Laurent Besacier, and Christophe Servan. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *NIPS Workshop on end-to-end learning for speech and audio processing*, 2016.

[Chang and Lee, 2022] Chih-Chiang Chang and Hung-yi Lee. Exploring continuous integrate-and-fire for adaptive simultaneous speech translation. *CoRR*, 2022.

[Chen *et al.*, 2020] Junkun Chen, Mingbo Ma, Renjie Zheng, and Liang Huang. MAM: masked acoustic modeling for end-to-end speech-to-text translation. *CoRR*, 2020.

[Cheng *et al.*, 2022] Yong Cheng, Yu Zhang, Melvin Johnson, Wolfgang Macherey, and Ankur Bapna. Mu2slam: Multitask, multilingual speech and language models. *CoRR*, 2022.

[Chuang *et al.*, 2021] Shun-Po Chuang, Yung-Sung Chuang, Chih-Chiang Chang, and Hung-yi Lee. Investigating the reordering capability in CTC-based non-autoregressive end-to-end speech translation. In *Findings of ACL*, 2021.

[Dong *et al.*, 2021] Qianqian Dong, Rong Ye, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. Listen, understand and translate: Triple supervision decouples end-to-end speech-to-text translation. In *AAAI*, 2021.

[Fang *et al.*, 2022] Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. STEMM: Self-learning with speech-text manifold mixup for speech translation. In *ACL*, 2022.

[Gaido *et al.*, 2021] Marco Gaido, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. How to split: the effect of word segmentation on gender bias in speech translation. In *Findings of ACL*, 2021.

[Gaido *et al.*, 2022a] Marco Gaido, Matteo Negri, and Marco Turchi. Who are we talking about? handling person names in speech translation. In *IWSLT*, 2022.

[Gaido *et al.*, 2022b] Marco Gaido, Sara Papi, Matteo Negri, and Marco Turchi. Joint speech translation and named entity recognition. *CoRR*, 2022.

[Gangi *et al.*, 2019] Mattia Antonino Di Gangi, Matteo Negri, and Marco Turchi. Adapting transformer to end-to-end spoken language translation. In *InterSpeech*, 2019.

[Gulati *et al.*, 2020] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In *InterSpeech*, 2020.

[Han *et al.*, 2021] Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. Learning shared semantic space for speech-to-text translation. In *Findings of ACL*, 2021.

[Hsu *et al.*, 2021] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *TASLP*, 2021.

[Huber *et al.*, 2022] Christian Huber, Enes Yavuz Ugan, and Alexander Waibel. Code-switching without switching: Language agnostic end-to-end speech translation. *CoRR*, 2022.

[Inaguma *et al.*, 2019] Hirofumi Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. Multilingual end-to-end speech translation. In *ASRU*, 2019.

[Inaguma *et al.*, 2021a] Hirofumi Inaguma, Siddharth Dalmia, Brian Yan, and Shinji Watanabe. Fast-md: Fast multi-decoder end-to-end speech translation with non-autoregressive hidden intermediates. In *ASRU*, 2021.

[Inaguma *et al.*, 2021b] Hirofumi Inaguma, Tatsuya Kawahara, and Shinji Watanabe. Source and target bidirectional knowledge distillation for end-to-end speech translation. In *NAACL*, 2021.

[Indurthi *et al.*, 2021] Sathish Indurthi, Mohd Abbas Zaidi, Nikhil Kumar Lakumarapu, Beomseok Lee, Hyojung Han, Seokchan Ahn, Sangha Kim, Chanwoo Kim, and Inchul Hwang. Task aware multi-task learning for speech to text tasks. In *ICASSP*, 2021.

[Jia *et al.*, 2019] Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J. Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *ICASSP*, 2019.

[Le *et al.*, 2021] Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. Lightweight adapter tuning for multilingual speech translation. In *ACL*, 2021.

[Li *et al.*, 2021] Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. Multilingual speech translation from efficient finetuning of pretrained models. In *ACL*, 2021.

[Liu *et al.*, 2019] Yuchen Liu, Hao Xiong, Jiajun Zhang, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. End-to-end speech translation with knowledge distillation. In *InterSpeech*, 2019.

[Liu *et al.*, 2020a] Yuchen Liu, Jiajun Zhang, Hao Xiong, Long Zhou, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. Synchronous speech recognition and speech-to-text translation with interactive decoding. In *AAAI*, 2020.

[Liu *et al.*, 2020b] Yuchen Liu, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. Bridging the modality gap for speech-to-text translation. *CoRR*, 2020.

[Liu *et al.*, 2021] Dan Liu, Mengge Du, Xiaoxi Li, Ya Li, and Enhong Chen. Cross attention augmented transducer networks for simultaneous translation. In *EMNLP*, 2021.

[Park *et al.*, 2019] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. In *InterSpeech*, 2019.

[Pino *et al.*, 2020] Juan Miguel Pino, Qiantong Xu, Xutai Ma, Mohammad Javad Dousti, and Yun Tang. Self-training for end-to-end speech translation. In *InterSpeech*, 2020.

[Ren *et al.*, 2020] Yi Ren, Jinglin Liu, Xu Tan, Chen Zhang, Tao Qin, Zhou Zhao, and Tie-Yan Liu. SimulSpeech: End-to-end simultaneous speech to text translation. In *ACL*, 2020.

[Savoldi *et al.*, 2022] Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. Under the morphosyntactic lens: A multifaceted evaluation of gender bias in speech translation. In *ACL*, 2022.

[Schneider *et al.*, 2019] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. In *InterSpeech*, 2019.

[Sperber and Paulik, 2020] Matthias Sperber and Matthias Paulik. Speech translation and the end-to-end promise: Taking stock of where we are. In *ACL*, 2020.

[Sperber *et al.*, 2019] Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. Attention-passing models for robust and data-efficient end-to-end speech translation. *TACL*, 2019.

[Stentiford and Steer, 1988] Fred WM Stentiford and Martin G Steer. Machine translation of speech. *British Telecom technology journal*, (2), 1988.

[Tang *et al.*, 2021a] Yun Tang, Juan Pino, Xian Li, Changhan Wang, and Dmitriy Genzel. Improving speech translation by understanding and learning from the auxiliary text translation task. In *ACL*, 2021.

[Tang *et al.*, 2021b] Yun Tang, Juan Pino, Changhan Wang, Xutai Ma, and Dmitriy Genzel. A general multi-task learning framework to leverage text data for speech to text tasks. In *ICASSP*, 2021.

[Tang *et al.*, 2022] Yun Tang, Hongyu Gong, Ning Dong, Changhan Wang, Wei-Ning Hsu, Jiatao Gu, Alexei Baevski, Xian Li, Abdelrahman Mohamed, Michael Auli, and Juan Pino. Unified speech-text pre-training for speech translation and recognition. In *ACL*, 2022.

[Tsiamas *et al.*, 2022] Ioannis Tsiamas, Gerard I Gállego, José AR Fonollosa, and Marta R Costa-jussà. Shas: Approaching optimal segmentation for end-to-end speech translation. In *InterSpeech*, 2022.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neurips*, 2017.

[Wang *et al.*, 2020a] Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. CoVoST: A diverse multilingual speech-to-text translation corpus. In *IREC*, 2020.

[Wang *et al.*, 2020b] Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang. Curriculum pre-training for end-to-end speech translation. In *ACL*, 2020.

[Wang *et al.*, 2021] Changhan Wang, Anne Wu, Juan Pino, Alexei Baevski, Michael Auli, and Alexis Conneau. Large-scale self- and semi-supervised learning for speech translation. In *InterSpeech*, 2021.

[Weiss *et al.*, 2017] Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. Sequence-to-sequence models can directly translate foreign speech. In *InterSpeech*, 2017.

[Weller *et al.*, 2022] Orion Weller, Matthias Sperber, Telmo Pires, Hendra Setiawan, Christian Gollan, Dominic Telaar, and Matthias Paulik. End-to-end speech translation for code switched speech. In *Findings of ACL*, 2022.

[Wu *et al.*, 2020] Anne Wu, Changhan Wang, Juan Miguel Pino, and Jiatao Gu. Self-supervised representations improve end-to-end speech translation. In *InterSpeech*, 2020.

[Wu *et al.*, 2022] Felix Wu, Kwangyoun Kim, Shinji Watanabe, Kyu Han, Ryan McDonald, Kilian Q Weinberger, and Yoav Artzi. Wav2seq: Pre-training speech-to-text encoder-decoder models using pseudo languages. *CoRR*, 2022.

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

[Ye *et al.*, 2021] Rong Ye, Mingxuan Wang, and Lei Li. End-to-end speech translation via cross-modal progressive training. In *InterSpeech*, 2021.

[Ye *et al.*, 2022] Rong Ye, Mingxuan Wang, and Lei Li. Cross-modal contrastive learning for speech translation. In *NAACL*, 2022.