Speech Translation and the *End-to-End* Promise: Taking Stock of Where We Are

Matthias Sperber
Apple
sperber@apple.com

Matthias Paulik
Apple
mpaulik@apple.com

Abstract

Over its three decade history, speech translation has experienced several shifts in its primary research themes; moving from loosely coupled cascades of speech recognition and machine translation, to exploring questions of tight coupling, and finally to end-to-end models that have recently attracted much attention. This paper provides a brief survey of these developments, along with a discussion of the main challenges of traditional approaches which stem from committing to intermediate representations from the speech recognizer, and from training cascaded models separately towards different objectives.

Recent end-to-end modeling techniques promise a principled way of overcoming these issues by allowing joint training of all model components and removing the need for explicit intermediate representations. However, a closer look reveals that many end-to-end models fall short of solving these issues, due to compromises made to address data scarcity. This paper provides a unifying categorization and nomenclature that covers both traditional and recent approaches and that may help researchers by highlighting both trade-offs and open research questions.

1 Introduction

Speech translation (ST), the task of translating acoustic speech signals into text in a foreign language, is a complex and multi-faceted task that builds upon work in automatic speech recognition (ASR) and machine translation (MT). ST applications are diverse and include travel assistants (Takezawa et al., 1998), simultaneous lecture translation (Fügen, 2008), movie dubbing/subtitling (Saboo and Baumann, 2019; Matusov et al., 2019), language documentation and crisis response (Bansal et al., 2017), and developmental efforts (Black et al., 2002).

Until recently, the only feasible approach has been the cascaded approach that applies an ASR to the speech inputs, and then passes the results on to an MT system. Progress in ST has come from two fronts: general improvements in ASR and MT models, and moving from the loosely-coupled cascade in its most basic form toward a tighter coupling. However, despite considerable efforts toward tight coupling, a large share of the progress has arguably been owed simply to general ASR and MT improvements.[1]

Recently, new modeling techniques and in particular end-to-end trainable encoder-decoder models have fueled hope for addressing challenges of ST in a more principled manner. Despite these hopes, the empirical evidence indicates that the success of such efforts has so far been mixed (Weiss et al., 2017; Niehues et al., 2019).

In this paper, we will attempt to uncover potential reasons for this. We start by surveying models proposed throughout the three-decade history of ST. By contrasting the extreme points of loosely coupled cascades vs. purely end-to-end trained direct models, we identify foundational challenges: erroneous early decisions, mismatch between spoken-style ASR outputs and written-style MT inputs, and loss of speech information (e.g. prosody) on the one hand, and data scarcity on the other hand. We then show that to improve data efficiency, most end-to-end models employ techniques that re-introduce issues generally attributed to cascaded ST.

Furthermore, this paper proposes a categorization of ST research into well-defined terms for the particular challenges, requirements, and techniques that are being addressed or used. This multi-dimensional categorization suggests a modeling

[1] For instance, Pham et al. (2019)'s winning system in the IWSLT 2019 shared ST task (Niehues et al., 2019) makes heavy use of recent ASR and MT modeling techniques, but is otherwise a relatively simple cascaded approach.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7409–7421

space with many intermediate points, rather than a dichotomy of cascaded vs. end-to-end models, and reveals a number of trade-offs between different modeling choices. This implies that additional work to more explicitly analyze the interactions between these trade-offs, along with further model explorations, can help to determine more favorable points in the modeling space, and ultimately the most favorable model for a specific ST application.

2 Chronological Survey

This chapter surveys the historical development of ST and introduces key concepts that will be expanded upon later.[2]

2.1 Loosely Coupled Cascades

Early efforts to realize ST (Stentiford and Steer, 1988; Waibel et al., 1991) introduced what we will refer to as the **loosely coupled cascade** in which separately built ASR and MT systems are employed and the best hypothesis of the former is used as input to the latter. The possibility of **speech-to-speech** translation, which extends the cascade by appending a text-to-speech component, was also considered early on (Waibel et al., 1991).

These early systems were especially susceptible to **errors propagated** from the ASR, given the widespread use of interlingua-based MT which relied on parsers unable to handle mal-formed inputs (Woszczyna et al., 1993; Lavie et al., 1996; Liu et al., 2003). Subsequent systems Wang and Waibel (1998); Takezawa et al. (1998); Black et al. (2002); Sumita et al. (2007), relying on data driven, statistical MT, somewhat alleviated the issue, and also in part opened the path towards tighter integration.

2.2 Toward Tight Integration

Researchers soon turned to the question of how to avoid early decisions and the problem of error propagation. While the desirable solution of full integration over transcripts is intractable (Ney, 1999), approximations are possible. Vidal (1997); Bangalore and Riccardi (2001); Casacuberta et al. (2004); Pérez et al. (2007) compute a composition of FST-based ASR and MT models, which approximates the full integration up to search heuristics, but suffers from limited reordering capabilities. A much simpler, though computationally expensive, solution is the n**-best** translation approach which replaces the sum over all possible transcripts by a sum over only the n-best ASR outputs (Woszczyna et al., 1993; Lavie et al., 1996). Follow-up work suggested **lattices** and **confusion nets** (Saleem et al., 2004; Zhang et al., 2005; Bertoldi and Federico, 2005) as more effective and efficient alternatives to n-best lists. Lattices proved flexible enough for integration into various translation models, from word-based translation models to phrase-based ST Matusov et al. (2005, 2008) to neural lattice-to-sequence models (Sperber et al., 2017a, 2019b; Zhang et al., 2019; Beck et al., 2019).

Another promising idea was to limit the detrimental effects of early decisions, rather than attempting to avoid early decisions. One way of achieving this is to train **robust translation** models by introducing synthetic ASR errors into the source side of MT corpora (Peitz et al., 2012; Tsvetkov et al., 2014; Ruiz et al., 2015; Sperber et al., 2017b; Cheng et al., 2018, 2019). A different route is taken by Dixon et al. (2011); He et al. (2011) who directly optimize ASR outputs towards translation quality.

Beyond early decisions, research moved towards tighter coupling by addressing issues arising from ASR and MT models being trained separately and on different types of corpora. **Domain adaptation** techniques were used by Liu et al. (2003); Fügen (2008) to adapt models to the spoken language domain. Matusov et al. (2006); Fügen (2008) propose re-segmenting the ASR output and inserting **punctuation**, so as to provide the translation model with well-formed text inputs. In addition, **disfluency removal** (Fitzgerald et al., 2009) was proposed to avoid translation errors caused by disfluencies that are often found in spoken language.

Aguero et al. (2006); Anumanchipalli et al. (2012); Do et al. (2017); Kano et al. (2018) propose **prosody transfer** for speech-to-speech translation by determining source-side prosody and applying transformed prosody characteristics to the aligned target words.

2.3 Speech Translation Corpora

It is important to realize that all efforts to this point had used separate ASR and MT corpora for training. This often led to a mismatch between ASR trained on data from the spoken domain, and MT trained on data from the written domain. **End-to-**

[2]For a good comparison of empirical results, which are not the focus of this paper, we refer to concurrent work (Sulubacak et al., 2019). Moreover, for conciseness we do not cover the sub-topic of simultaneous translation (Fügen, 2008).

end ST data (translated speech utterances) was only available in small quantities for test purposes.

Paulik (2010) proposes the use of audio recordings of interpreter-mediated communication scenarios, which is not only potentially easier to obtain, but also does not exhibit such domain mismatches. Post et al. (2013) manually translate an ASR corpus to obtain an end-to-end ST corpus, and show that training both ASR and MT on the same corpus considerably improves results compared to using out-of-domain MT data. Unfortunately, high annotation costs prevent scaling of the latter approach, so follow-up work concentrates on compiling ST corpora from available web sources (Godard et al., 2018; Kocabiyikoglu et al., 2018; Sanabria et al., 2018; di Gangi et al., 2019a; Boito et al., 2020; Beilharz et al., 2020; Iranzo-Sánchez et al., 2020; Wang et al., 2020a). Note that despite these efforts, publicly available ST corpora are currently strongly limited in terms of both size and language coverage. For practical purposes, the use of separate ASR and MT corpora is therefore currently unavoidable.

2.4 End-to-End Models

The availability of end-to-end ST corpora, along with the success of end-to-end models for MT and ASR, led researchers to explore ST models trained in an end-to-end fashion. This was fueled by a hope to solve the issues addressed by prior research in a principled and more effective way. Duong et al. (2016); Berard et al. (2016); Bansal et al. (2018) explore **direct ST models** that translate speech without using explicitly generated intermediate ASR output. In contrast, Kano et al. (2017); Anastasopoulos and Chiang (2018); Wang et al. (2020b) explore **end-to-end trainable cascades and triangle models**, i.e. models that do rely on transcripts, but are optimized in part through end-to-end training. **Multi-task training** and **pre-training** were proposed as a way to incorporate additional ASR and MT data and reduce dependency on scarce end-to-end data (Weiss et al., 2017; Bérard et al., 2018; Bansal et al., 2019; Stoian et al., 2020; Wang et al., 2020b). As these techniques were not able to exploit ASR and MT data as effectively as the loosely coupled cascade, other approaches like **subtask training** for end-to-end-trainable cascades (Sperber et al., 2019a), **data augmentation** (Jia et al., 2019a; Pino et al., 2019), knowledge distillation (Liu et al., 2019), and meta-learning (Indurthi et al., 2020) were proposed. Salesky et al. (2019a) propose pre-segmenting speech frames, (Jia et al., 2019b; Tjandra et al., 2019) explore speech-to-speech translation. Sung et al. (2019); di Gangi et al. (2019b); Di Gangi et al. (2020); Bahar et al. (2019); Inaguma et al. (2019); di Gangi et al. (2019c) transfer ideas from MT and ASR fields to ST.

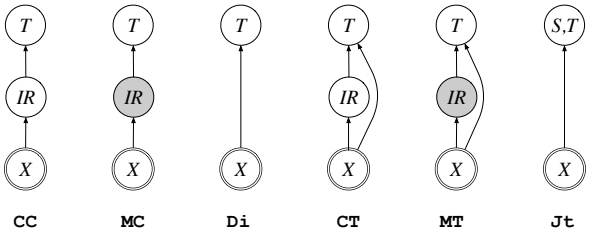

Figure 1: Illustration of inference strategies (§4.2): Committed/marginalizing cascade (CC/MC), direct (Di), committed/marginalizing triangle (CT/MT), joint (Jt). Double lines differentiate the observed variable (speech input *X*) from random variables (intermediate representations *IR* and translations *T*). Shaded circles marginalize over random variables.

3 Central Challenges

Given the abundance of prior work, a clear picture on where we currently stand is needed. For purposes of identifying the key challenges in ST research, this section will contrast the extreme cases of the *loosely coupled cascade* (CC in Fig. 1)[3] against the *vanilla direct model* (Di in Fig. 1).[4] We emphasize that these models are only extreme points in a modeling space with many intermediate points, as we will see in §4. We assume appropriate speech features X as inputs. $T, \hat{T} \in \mathcal{T}$ denote candidate/best translations, respectively, from the MT hypothesis space. $S \in \mathcal{H}$ denotes a graphemic transcript from the ASR hypothesis space.

3.1 Challenges of Loosely Coupled Cascades

The loosely coupled cascade justifies its decomposition into MT model $P_{\text{MT}}(T|S)$ and ASR model $P_{\text{ASR}}(S|X)$ as follows:

[3] ASR and MT models trained separately on different corpora; intermediate representation is ASR 1-best output.

[4] Encoder-decoder model trained on speech utterances paired with translations; no intermediate representations used.

$$\hat{T} = \underset{T \in \mathcal{T}}{\operatorname{argmax}} P(T \mid X) \quad (1)$$

$$= \underset{T \in \mathcal{T}}{\operatorname{argmax}} \sum_{S \in \mathcal{H}} P(T|S, X)\, P(S|X) \quad (2)$$

$$\approx \underset{T \in \mathcal{T}}{\operatorname{argmax}} \sum_{S \in \mathcal{H}} P_{\text{MT}}(T|S)\, P_{\text{ASR}}(S|X) \quad (3)$$

$$\approx \underset{T \in \mathcal{T}}{\operatorname{argmax}} \sum_{S \in \mathcal{H}'} P_{\text{MT}}(T|S)\, P_{\text{ASR}}(S|X) \quad (4)$$

Note that here the set $\mathcal{H}'$ contains only a single entry, the 1-best ASR output. The approximations in these derivations directly result in the following three foundational challenges:

Erroneous early decisions: *Committing to a potentially erroneous S during inference.* This leads to the well-known problem of **error propagation** (Ruiz and Federico, 2014) and is caused by avoiding the intractable full integration over transcripts (Eq. 3) and using only the 1-best ASR output instead (Eq. 4). Typical countermeasures include increasing $\mathcal{H}'$ to cover a larger space using lattices or confusion nets, or improving the robustness of MT models.

Mismatched source-language: *ASR and MT components model the source-language (transcript) priors $P_{\text{MT}}(S)$ and $P_{\text{ASR}}(S)$ differently.*[5] Causes include both modeling assumptions, e.g. ASR modeling only unpunctuated transcripts; and mismatched training data, leading to stylistic and topical divergence. Typical countermeasures are domain adaptation techniques, disfluency removal, text normalization, and segmentation/punctuation insertion.

Information loss: *Assumed conditional independence between inputs and outputs, given the transcript: $(T \perp\!\!\!\perp X) \mid S$.* This can be seen in Eq. 3 and results in any information not represented in S to be lost for the translation step. In particular, the MT model is unaware of **prosody** which structures and disambiguates the utterances, thus playing a role similar to punctuation in written texts; and provides ways to emphasize words or parts of the messages that the speaker think are important. Prosody also conveys information on the speaker's attitude and emotional state (Jouvet, 2019).

[5]Note that our definition does not entail covariance shift and other forms of domain mismatch (Kouw and Loog, 2018) which, though relevant, are not unique to cascaded ST and are widely covered by general ASR and MT literature (Cuong and Sima'an, 2018).

3.2 Challenges of the Vanilla Direct Model

Consider instead the other extreme case: an encoder-decoder model trained to directly produce translations from speech (Eq. 1). Because this model avoids the decomposition in Eq. 2-4, it is not subject to the three issues outlined in §3.1. Unfortunately, this second extreme case is often impractical due to its dependency on scarce end-to-end ST training corpora (§2.3), rendering this model unable to compete with cascaded models that are trained on abundant ASR and MT training data.

Most recent works therefore depart from this purely end-to-end trained direct model, and incorporate ASR and MT back into training, e.g. through weakly supervised training, or by exploring end-to-end trainable cascades or triangle models (`CT`/`MT` in Fig. 1). This departure raises two questions: (1) To what extent does the re-introduction of ASR and MT data cause challenges similar to those found in loosely coupled cascades? (2) Are techniques such as weakly supervised training effective enough to allow competing with the loosely coupled cascade? To address the second question, we propose the notion of data efficiency as a fourth key challenge.

Data efficiency: *The increase in accuracy achievable through the addition of a certain amount of training data.* To assess data efficiency, data ablations that contrast models over at least two data conditions are required. We argue that empirical evidence along these lines will help considerably in making generalizable claims about the relative performance between two ST models. Generalizable findings across data conditions are critical given that ST models are trained on at least three types of corpora (ASR, MT, and end-to-end corpora), whose availability vastly differs across languages.

3.3 Data Efficiency vs. Modeling Power – A Trade-Off?

Consider how the incorporation of MT and ASR data into ST models of any kind may inherently cause the problems as outlined in §3.1: Training on MT data may weaken the model's sensitivity to prosody; the effectiveness of training on ASR+MT data may be impacted by mismatched source-language issues; even some types of end-to-end-trainable models make (non-discrete) early decisions that are potentially erroneous.

This suggests a potential trade-off between data efficiency and modeling power. In order to find

English	Japanese
	kochira wa suekko no lucy desu
this is my niece , lucy	こちら は 姪っ子 の ルーシー です 。
	lucy, kono ko ga watashi no suekko desu
this is my niece , lucy	ルーシー 、 この 子 が 私 の 姪っ子 です 。
	chiizu toka jamu toka, dore ni shimasu ka
will you have /cheese or /jam	チーズ とか ジャム とか、 どれ に します か ？
	chiizu ka jamu, docchi ni shimasu ka
will you have /cheese or \jam	チーズ か ジャム、 どっち に します か ？

Table 1: Motivating examples for prosody-aware translation from English to Japanese. In the first example, prosody disambiguates whether the speaker is talking about *Lucy* as a third person or directly addressing *Lucy*. In the second example, prosody disambiguates whether *cheese or jam* is an open set or a closed set. In both cases, the surface form of the Japanese translation requires considerable changes depending on the prosody.

models that trade off advantages and disadvantages in the most favorable way, it is therefore necessary to thoroughly analyze models across the dimensions of early decisions, mismatched source-language, information loss, and data efficiency.

Analyzing early decisions: Problems due to erroneous early decisions are inference-time phenomena in which upstream ASR errors are responsible for errors in the final translation outputs. It follows that the problem disappears for hypothetical utterances for which the ASR can generate error-free intermediate representations. Thus, models that do not suffer from erroneous early decisions will expectedly exhibit an advantage over other models especially for acoustically challenging inputs, and less so for inputs with clean acoustics. This angle can provide us with strategies for isolating errors related to this particular phenomenon. Prior work in this spirit has demonstrated that lattice-to-sequence translation is in fact beneficial especially for acoustically challenging inputs (Sperber et al., 2017a), and that cascaded models with non-discrete intermediate representations are less sensitive to artificially perturbed intermediate representations than if using discrete transcripts as an intermediate representation (Sperber et al., 2019a).

Analyzing mismatched source-language: End-to-end ST corpora allow for controlled experiments in which one can switch between matched vs. mismatched (out-of-domain) MT corpora. Post et al. (2013) demonstrated that using a matched corpus can strongly improve translation quality for loosely coupled cascades. We are not aware of such analyses in more recent work.

Analyzing information loss: Prior work (Aguero et al., 2006; Anumanchipalli et al., 2012; Do et al., 2017; Kano et al., 2018) has addressed prosody transfer in speech-to-speech translation, but to our knowledge the question of how such information should inform textual translation decisions is still unexplored. Table 1 shows examples that may motivate future work in this direction.

Analyzing data efficiency: While several prior works aim at addressing this problem, often only a single data condition is tested, limiting the generalizability of findings. We are aware of three recent works that do analyze data efficiency across several data conditions (Jia et al., 2019a; Sperber et al., 2019a; Wang et al., 2020b). Findings indicate that both pretraining and data synthesizing outperform multi-task training in terms of data efficiency, and that end-to-end trainable cascades are on par with loosely coupled cascades, while strongly outperforming multi-task training.

4 Modeling Techniques

Let us now break apart modeling techniques from prior literature into four overarching categories, with the aim of exposing the ST modeling space between the extreme points of vanilla direct models and loosely coupled cascades.

4.1 Intermediate Representations

Almost all models use intermediate representations (IRs) in some form: non-direct models to support both training and inference, and direct models to overcome data limitations. IRs are often speech transcripts, but not necessarily so. A number of factors must be considered for choosing an appropriate IR, such as availability of supervised data, inference accuracy, expected impact of erroneous early decisions, and the feasibility of backpropagation through the IR for end-to-end training. We list several possibilities below:

Transcripts: Generally used in the loosely coupled cascade. Being a discrete representation, this option prevents end-to-end training via back-propagation, although future work may experiment with work-arounds such as the straight-through gradient estimator (Bengio et al., 2013). Besides graphemic transcripts, phonetic transcripts are another option (Jiang et al., 2011).

Hidden representations: Kano et al. (2017); Anastasopoulos and Chiang (2018); Sperber et al. (2019a) propose the use of hidden representations that are the by-product of a neural decoder generating an auxiliary IR such as a transcript. Advantages of this representation are differentiability, prevention of information loss, and weakened impact of erroneous early decisions. A downside is that end-to-end ST data is required for training.

Lattices: Lattices compactly represent the space over multiple sequences, and therefore weaken the impact of erroneous early decisions. Future work may explore lattices over continuous, hidden representations, and end-to-end training for ST models with lattices as intermediate representation.

Other: Prior work further suggests pre-segmented speech frames (Salesky et al., 2019a) or unsupervised speech-unit clusters (Tjandra et al., 2019) as intermediate representation. Further possibilities may be explored in future work.

4.2 Inference Strategies

The conditioning graph (Fig. 1) reveals independence assumptions and use of IRs at inference time. Some strategies avoid the problem of early decisions (`MC`, `Di`, `MT`, `Jt`), while others remove the conditional independence assumption between inputs and outputs (`Di`, `CT`, `MT`, `Jt`).

Committed cascade (`CC`): *Compute one IR, rely on it to generate outputs (Eq. 4).* Includes both the loosely coupled cascade, and recent end-to-end trainable cascaded models such as by Kano et al. (2017); Sperber et al. (2019a).

Marginalizing cascade (`MC`): *Compute outputs by relying on IRs, but marginalize over them instead of committing to one (Eq. 3).* As marginalization is intractable, approximations such as n-best translation or lattice translation are generally used.

Direct (`Di`): *Compute outputs without relying on IRs (Eq. 1).* To address data limitations, techniques such as multi-task training or data augmentation can be used, but may reintroduce certain biases.

Committed triangle (`CTr`): *Commit to an IR, then produce outputs by conditioning on both inputs and intermediate representation.* Anastasopoulos and Chiang (2018), who introduce the triangle model, use it in its marginalizing form (see below). Unexplored variations include the use of discrete transcripts as IR, which interestingly could be seen as a strict generalization of the loosely coupled cascade and should therefore never perform worse than it if trained properly.

Marginalizing triangle (`MTr`): *Produce output by conditioning on both input and IR, while marginalizing over the latter (Eq. 2).* Anastasopoulos and Chiang (2018) marginalize by taking an n-best list, with n set to only 4 for computational reasons. This raises the question of whether the more computationally efficient lattices could be employed instead. Similar considerations apply to the end-to-end trainable marginalizing cascade.

Joint (`Jt`): *Changes the problem formulation to* $\hat{S}, \hat{T} = \text{argmax}_{S \in \mathcal{H}, T \in \mathcal{T}} Pr(S, T \mid X)$. This is a useful optimization for many applications which display both transcripts and translations to the user, yet to our knowledge has never been explicitly addressed by researchers.

4.3 Training Strategies

This group of techniques describes the types of supervision signals applied during *training*.

Subtask training: *Training of sub-components by pairing IRs with either the speech inputs or the output translations.* Loosely coupled cascades rely on this training technique while recently proposed cascaded and triangle models often combine subtask training and end-to-end training.

Auxiliary task training: *Training by pairing either model inputs or outputs with data from an arbitrary auxiliary task through multi-task training.*[6] This technique has been used in two ways in literature: (1) To incorporate ASR and MT data into direct models by using auxiliary models that share parts of the parameters with the main model (Weiss et al., 2017). Auxiliary models are introduced for training purposes only, and discarded during inference. This approach has been found

[6]This definition subsumes pretraining, which is simply using a specific multitask training schedule.

inferior at exploiting ASR and MT data when compared to subtask training (Sperber et al., 2019a). (2) To incorporate various types of less closely related training data, such as the use of multitask training to exploit ASR data from an unrelated third language (Bansal et al., 2019; Stoian et al., 2020).

End-to-end: *Supervision signal that directly pairs speech inputs and output translations.* This technique is appealing because it jointly optimizes all involved parameters and may lead to better optima. The main limitation is lack of appropriate data, which can be addressed by combined training with one of the alternative supervision types, or by training on augmented data, as discussed next.

4.4 End-to-End Training Data

Manual: *Speech utterances for training are translated (and possibly transcribed) by humans.* This is the most desirable case, but such data is currently scarce. While we have seen growth in data sources in the past two years (§2.3), collecting more data is an extremely important direction for future work.

Augmented: *Data obtained by either augmenting an ASR corpus with automatic translations, or augmenting an MT corpus with synthesized speech.* This has been shown more data efficient than multitask training in the context of adding large MT and ASR corpora (Jia et al., 2019a). Pino et al. (2019) find that augmented ASR corpora are more effective than augmented MT corpora. This approach allows training direct models and end-to-end models even when no end-to-end data is available. Knowledge distillation can be seen as an extension (Liu et al., 2019). An important problem that needs analysis is to what extent mismatched source-language and information loss degrade the augmented data.

Zero-Shot: *Using no end-to-end data during training.* While augmented data can be used in most situations in which no manual data is available, it suffers from certain biases that may harm the ST model. Similarly to how zero-shot translation enables translating between unseen combinations of source and target languages, it may be worth exploring whether some recent models, such as direct models or cascades with non-discrete IRs, can be trained without resorting to any end-to-end data for the particular language pair of interest.

5 Applications and Requirements

While we previously described the task of ST simply as the task of generating accurate text translations from speech inputs, the reality is in fact much more complicated. Future work may exploit new modeling techniques to explicitly address the aspects drawn out below.

5.1 Mode of Delivery

Batch mode: *A (potentially large) piece of recorded speech is translated as a whole.* Segmentation into utterances may or may not be given. This mode allows access to future context, and imposes no strict computational restrictions. Typical applications include movie subtitling (Matusov et al., 2019) and dubbing (Saboo and Baumann, 2019; Federico et al., 2020).

Consecutive: *Real-time situation where inputs are provided as complete utterances or other translatable units, and outputs must be produced with low latency.* A typical example is a two-way translation system on a mobile device (Hsiao et al., 2006). This is the only mode of delivery that allows interaction between speaker and translator (Ayan et al., 2013).

Simultaneous: *Real-time situation where latency is crucial and outputs are produced incrementally based on incoming audio stream.* Simultaneous translation is faced with an inherent delay vs. accuracy trade-off, such as in a typical lecture translation application (Fügen, 2008). In addition to computational latency, which is relevant also with consecutive translation, simultaneous translation suffers from inherent modeling latency caused by factors including reordering.

5.2 Output Medium

Text: This is a standard setting, but is nevertheless worth discussing in more detail for at least two reasons: (1) as is well-known in the subtitling industry, reading speeds can be slower than speaking and listening speeds (Romero-Fresco, 2009), implying that a recipient may not be able to follow verbatim text translations in case of fast speakers, and that summarization may be warranted. (2) Text display makes repair strategies possible that are quite distinct from spoken outputs: One can alter, highlight, or remove past outputs. One possible way of exploiting this is Niehues et al. (2018)'s strategy of simultaneous translation through re-translation.

ES	también tengo um eh estoy tomando una clase ..
EN	i also have um eh i'm taking a marketing class ..
ES	porque qué va, mja ya te acuerda que ..
EN	because what is, mhm do you recall now that ..

Table 2: Examples for faithful Spanish to English translations, taken from (Salesky et al., 2019b).

Speech: Speech outputs have been used since the early days (Lavie et al., 1996), but whether to apply text-to-speech on top of translated text has often been seen as a question to leave to user interface designers. Here, we argue that ST researchers should examine in what ways speech outputs should differ from text outputs. For example, is disfluency removal (Fitzgerald et al., 2009) beneficial for speech outputs, given that human listeners are naturally able to repair disfluencies (Lickley, 1994)? Further examples that need more exploration are prosody transfer (Aguero et al., 2006) and models that directly translate speech-to-speech (Jia et al., 2019b).

5.3 The Role of Transcripts

Mandatory transcripts: *User interface displays both transcripts and translations to the user.* This scenario has been implemented in many applications (Hsiao et al., 2006; Cho et al., 2013), but has received little attention in the context of end-to-end ST research. It ties together with the *joint* inference model (§4.3). Note that with loosely coupled cascades, there is little need to consider this scenario explicitly because the application can simply display the by-product transcripts to the user. But this is not easily possible with direct models or with models using IRs other than transcripts.

Auxiliary transcripts: *Transcriptions are not needed as user-facing model outputs, but may be exploited as IRs during training and possibly inference.* This is the most typical formal framing of the ST task, assuming that transcribed training data is useful mainly for purposes of improving the final translation.

Transcript-free: *No transcribed training data exists, so the model cannot rely on supervised transcripts as IR.* The main scenario is endangered language preservation for languages without written script, where it is often easier to collect translated speech than transcribed speech (Duong et al., 2016).

5.4 Translation Method

The method of translation is an especially relevant factor in ST, which commonly includes a transfer from the spoken into the written domain. Here, we provide two reference points for the method of translation, while referring to Newmark (1988) for a more nuanced categorization.

Faithful: *Keeps the contextual meaning of the original as precisely as possible within the grammatical constraints of the target language.* With text as output medium, faithful translation may result in poor readability, e.g. due to the translation of disfluencies (Table 2). Arguably the most appropriate output medium for faithful ST would be speech, although user studies are needed to confirm this. Another application are high-stake political meetings in which translations must stay as close to the original sentence as possible. As we move toward more distant language pairs, the practicability of faithful translation of spoken language with disfluencies becomes increasingly questionable.

Communicative: *Renders the contextual meaning of the original such that both content and style are acceptable and comprehensible by the target audience.* An important example for improving communicativeness is disfluency removal (Fitzgerald et al., 2009). Given that human translators and interpreters adapt their translation method depending on factors that include input and output medium (He et al., 2016), more research is needed beyond disfluency removal. Communicative translations are especially relevant in casual contexts where convenience and low cognitive effort are mandative. Arguably the closest neighbor of spoken language style in the text realm is social media, it would be interesting to attempt speech-to-text translation with social-media style outputs.

6 Discussion

Recent works on end-to-end modeling techniques are motivated by the prospect of overcoming the loosely coupled cascade's inherent issues, yet of the issues outlined in §2.1, often only the goal of avoiding early decisions is mentioned motivationally. While early decisions and data efficiency have been recognized as central issues, empirical insights are still limited and further analysis is needed. Mismatched source-language and information loss are often not explicitly analyzed.

We conjecture that the apparent trade-off between data efficiency and modeling power may explain the mixed success in outperforming the loosely coupled cascade. In order to make progress in this regard, the involved issues (early decisions, mismatched source-language, information loss, data efficiency) need to be precisely analyzed (§3), and more model variants (§4) should be explored. As a possible starting point one may aim to extend, rather than alter, traditional models, e.g. applying end-to-end training as a fine-tuning step, employing a direct model for rescoring, or adding a triangle connection to a loosely coupled cascade. We further suggest that more principled solutions to the different application-specific requirements (§5) should be attempted. Perhaps it is possible to get rid of segmentation as a separate step in batch delivery mode, or perhaps text as output medium can be used to visualize repairs more effectively. Several of the application-specific requirements demand user studies and will not be sufficiently solved by relying on automatic metrics only.

7 Conclusion

We started this paper with a chronological survey of three decades of ST research, focusing on carving out the key concepts. We then provided definitions of the central challenges, techniques, and requirements, motivated by the observation that recent work does not sufficiently analyze these challenges. We exposed a significant space of both modeling ideas and application-specific requirements left to be addressed in future research.

Our hope is to encourage meaningful and generalizable comparisons on our quest toward overcoming the long-standing issues found in ST models.

References

P. D. Aguero, Jordi Adell, and Antonio Bonafonte. 2006. Prosody Generation for Speech-to-Speech Translation. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France.

Antonios Anastasopoulos and David Chiang. 2018. Tied Multitask Learning for Neural Speech Translation. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, New Orleans, USA.

Gopala Krishna Anumanchipalli, Luis C. Oliveira, and Alan W. Black. 2012. Intent transfer in speech-to-speech machine translation. In *Workshop on Spoken Language Technology (SLT)*, pages 153–158, Miami, USA.

Necip Fazil Ayan, Arindam Mandal, Michael Frandsen, Jing Zheng, Peter Blasco, Andreas Kathol, Frederic Bechet, Benoit Favre, Alex Marin, Tom Kwiatkowski, Mari Ostendorf, Luke Zettlemoyer, Philipp Salletmayr, Julia Hirschberg, and Svetlana Stoyanchev. 2013. 'Can you give me another word for hyperbaric?': Improving speech translation using targeted clarification questions. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8391–8395, Vancouver, Canada.

Parnia Bahar, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2019. On Using SpecAugment for End-to-End Speech Translation. In *International Workshop on Spoken Language Translation (IWSLT)*, Hongkong.

Srinivas Bangalore and Giuseppe Riccardi. 2001. A Finite-State Approach to Machine Translation. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, Pittsburgh, USA.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2018. Low-Resource Speech-to-Text Translation. In *Annual Conference of the International Speech Communication Association (InterSpeech)*, Hyderabad, India.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2019. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, Minneapolis, USA.

Sameer Bansal, Herman Kamper, Adam Lopez, and Sharon Goldwater. 2017. Towards speech-to-text translation without speech recognition. In *European Chapter of the Association for Computational Linguistics (EACL)*.

Daniel Beck, Trevor Cohn, and Gholamreza Haffari. 2019. Neural Speech Translation using Lattice Transformations and Graph Networks. In *Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 26–31, Hongkong.

Benjamin Beilharz, Xin Sun, Sariya Karimova, and Stefan Riezler. 2020. LibriVoxDeEn: A Corpus for German-to-English Speech Translation and Speech Recognition. In *Language Resources and Evaluation (LREC)*, Marseille, France.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. *arXiv:1308.3432*.

Alexandre Bérard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. 2018. End-to-End Automatic Speech Translation of Audiobooks.

In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada.

Alexandre Berard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. 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*, Barcelona, Spain.

Nicola Bertoldi and Marcello Federico. 2005. A New Decoder for Spoken Language Translation Based on Confusion Networks. In *Automatic Speech Recognition & Understanding (ASRU)*, pages 86–91, San Juan, Puerto Rico.

Alan W. Black, Ralf D. Brown, Robert Frederking, Kevin Lenzo, John Moody, Alexander Rudnicky, Rita Singh, and Eric Steinbrecher. 2002. Rapid Development of Speech-to-Speech Translation Systems. In *International Conference on Spoken Language Processing (ICSLP)*, Denver, USA.

Marcely Zanon Boito, William N. Havard, Mahault Garnerin, Éric Le Ferrand, and Laurent Besacier. 2020. MaSS: A Large and Clean Multilingual Corpus of Sentence-aligned Spoken Utterances Extracted from the Bible. In *Language Resources and Evaluation (LREC)*, Marseille, France.

Francisco Casacuberta, Hermann Ney, Franz Josef Och, Enrique Vidal, J. M. Vilar, S. Barrachina, I. García-Varea, D. Llorens, C. Martínez, S. Molau, F. Nevado, M. Pastor, D. Picó, A. Sanchis, and C. Tillmann. 2004. Some approaches to statistical and finite-state speech-to-speech translation. *Computer Speech and Language*, 18(1):25–47.

Qiao Cheng, Meiyuan Fang, Yaqian Han, Jin Huang, and Yitao Duan. 2019. Breaking the Data Barrier: Towards Robust Speech Translation via Adversarial Stability Training. In *International Workshop on Spoken Language Translation (IWSLT)*, Hongkong.

Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. 2018. Towards Robust Neural Machine Translation. In *Association for Computational Linguistic (ACL)*, Melbourne, Australia.

Eunah Cho, Christian Fügen, Teresa Hermann, Kevin Kilgour, Mohammed Mediani, Christian Mohr, Jan Niehues, Kay Rottmann, Christian Saam, Sebastian Stüker, and Alex Waibel. 2013. A real-world system for simultaneous translation of German lectures. In *Annual Conference of the International Speech Communication Association (InterSpeech)*, pages 3473–3477, Lyon, France.

Hoang Cuong and Khalil Sima'an. 2018. A survey of domain adaptation for statistical machine translation. In *Conference on Computational Linguistics (COLING)*, Santa Fe, USA.

Mattia Antonino Di Gangi, Viet-Nhat Nguyen, Matteo Negri, and Marco Turchi. 2020. Instance-Based Model Adaptation For Direct Speech Translation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain.

Paul R Dixon, Andrew Finch, H Chiori, and H Kashioka. 2011. Investigation on the Effects of ASR Tuning on Speech Translation Performance. In *International Workshop on Spoken Language Translation (IWSLT)*, pages 167–174, San Francisco, USA.

Quoc Truong Do, Tomoki Toda, Graham Neubig, Sakriani Sakti, and Satoshi Nakamura. 2017. Preserving Word-level Emphasis in Speech-to-speech Translation. *IEEE Transactions on Audio, Speech and Language Processing*, 25(3):544–556.

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 2016. An Attentional Model for Speech Translation Without Transcription. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 949–959, San Diego, USA.

Marcello Federico, Robert Enyedi, and Roberto Barrachicote. 2020. From Speech-to-Speech Translation to Automatic Dubbing. *arXiv:2001.06785v3*.

Erin Fitzgerald, Keith Hall, and Frederick Jelinek. 2009. Reconstructing false start errors in spontaneous speech text. In *European Chapter of the Association for Computational Linguistics (EACL)*, pages 255–263, Athens, Greece.

Christian Fügen. 2008. *A System for Simultaneous Translation of Lectures and Speeches*. Ph.D. thesis, University of Karlsruhe.

Antonino Mattia di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019a. MuST-C : a Multilingual Speech Translation Corpus. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, Minneapolis, USA.

Mattia A. di Gangi, Matteo Negri, Roldano Cattoni, Roberto Dessi, and Marco Turchi. 2019b. Enhancing Transformer for End-to-end Speech-to-Text Translation. In *Machine Translation Summit XVII Volume 1: Research Track*, pages 21–31, Dublin, Ireland.

Mattia Antonino di Gangi, Matteo Negri, and Marco Turchi. 2019c. One-to-Many Multilingual End-to-End Speech Translation. In *Automatic Speech Recognition & Understanding (ASRU)*, December.

Pierre Godard, Gilles Adda, Martine Adda-Decker, J. Benjumea, L. Besacier, J. Cooper-Leavitt, G-N. Kouarata, L. Lamel, H. Maynard, M. Mueller, A. Rialland, S. Stueker, F. Yvon, and M. Zanon-Boito. 2018. A Very Low Resource Language Speech Corpus for Computational Language Documentation Experiments. In *Language Resources and Evaluation (LREC)*, Miyazaki, Japan.

He He, Jordan Boyd-Graber, and Hal. 2016. Interpretese vs. Translationese : The Uniqueness of Human Strategies in Simultaneous Interpretation. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 971–976, San Diego, USA.

Xiaodong He, Li Deng, and Alex Acero. 2011. Why word error rate is not a good metric for speech recognizer training for the speech translation task? In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5632–5635, Prague, Czech Republic.

Roger Hsiao, Ashish Venugopal, Thilo Köhler, Ting Zhang, Paisarn Charoenpornsawat, Andreas Zollmann, Stephan Vogel, Alan W. Black, Tanja Schultz, and Alex Waibel. 2006. Optimizing components for handheld two-way speech translation for an English-Iraqi Arabic system. In *Annual Conference of the International Speech Communication Association (InterSpeech)*, pages 765–768, Pittsburgh, USA.

Hirofumi Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. 2019. Multilingual End-to-End Speech Translation. In *Automatic Speech Recognition & Understanding (ASRU)*, Singapore.

Sathish Indurthi, Houjeung Han, Nikhil Kumar Lakumarapu, Beomseok Lee, Insoo Chung, Sangha Kim, and Chanwoo Kim. 2020. Data Efficient Direct Speech-to-Text Translation with Modality Agnostic Meta-Learning. In *Conference on Artificial Intelligence (AAAI)*, New York City, USA.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-ST: A Multilingual Corpus For Speech Translation Of Parliamentary Debates. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain.

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J Weiss, Yuan Cao, Chung-cheng Chiu Naveen, Ari Stella, and Laurenzo Yonghui. 2019a. Leveraging Weakly Supervised Data to Improve End-to-End Speech-to-Text Translation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom.

Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. 2019b. Direct speech-to-speech translation with a sequence-to-sequence model. In *Annual Conference of the International Speech Communication Association (InterSpeech)*, Graz, Austria.

Jie Jiang, Zeeshan Ahmed, Julie Carson-Berndsen, Peter Cahill, and Andy Way. 2011. Phonetic Representation-Based Speech Translation. In *Machine Translation Summit (MTS)*, pages 81–88, Xiamen, China.

Denis Jouvet. 2019. Speech Processing and Prosody. In *International Conference of Text, Speech and Dialogue (TSD)*, Ljubljana, Slovenia.

Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura. 2017. Structured-based Curriculum Learning for End-to-end English-Japanese Speech Translation. In *Annual Conference of the International Speech Communication Association (InterSpeech)*, pages 2630–2634.

Takatomo Kano, Shinnosuke Takamichi, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. 2018. End to end Model for Cross-Lingual Transformation of Paralinguistic Information. *Machine Translation*, 32(4):353–368.

Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif. 2018. Augmenting Librispeech with French Translations: A Multimodal Corpus for Direct Speech Translation Evaluation. In *Language Resources and Evaluation (LREC)*, Miyazaki, Japan.

Wouter M. Kouw and Marco Loog. 2018. An introduction to domain adaptation and transfer learning. Technical report, Delft University of Technology, Delft, Netherlands.

Alon Lavie, Donna Gates, Marsal Gavalda, Laura Mayfield, Alex Waibel, and Lori Levin. 1996. Multilingual Translation of Spontaneously Spoken Language in a Limited Domain. In *International Conference on Computational Linguistics*, pages 252–255, Copenhagen, Denmark.

Robin J Lickley. 1994. *Detecting disfluency in spontaneous speech*. Ph.D. thesis, University of Edinburgh.

Fu-hua Liu, Liang Gu, Yuqing Gao, and Michael Picheny. 2003. Use of Statistical N-Gram Models in Natural Language Generation for Machine Translation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 636–639, Hongkong.

Yuchen Liu, Hao Xiong, Zhongjun He, Jiajun Zhang, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-End Speech Translation with Knowledge Distillation. In *Annual Conference of the International Speech Communication Association (InterSpeech)*, Graz, Austria.

Evgeny Matusov, Björn Hoffmeister, and Hermann Ney. 2008. ASR word lattice translation with exhaustive reordering is possible. In *Annual Conference of the International Speech Communication Association (InterSpeech)*, pages 2342–2345, Brisbane, Australia.

Evgeny Matusov, Arne Mauser, and Hermann Ney. 2006. Automatic Sentence Segmentation and Punctuation Prediction for Spoken Language Translation. In *International Workshop on Spoken Language Translation (IWSLT)*, pages 158–165, Kyoto, Japan.

Evgeny Matusov, Hermann Ney, and Ralph Schluter. 2005. Phrase-based translation of speech recognizer word lattices using loglinear model combination. In *Automatic Speech Recognition and Understanding (ASRU)*, pages 110–115, San Juan, Puerto Rico.

Evgeny Matusov, Patrick Wilken, and Yota Georgakopoulou. 2019. Customizing Neural Machine Translation for Subtitling. In *Conference on Machine Translation (WMT)*, pages 82–93, Florence, Italy.

Peter Newmark. 1988. *Approaches to Translation*. Prentice Hall, Hertfordshire.

Hermann Ney. 1999. Speech Translation: Coupling of Recognition and Translation. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 517–520, Phoenix, USA.

Jan Niehues, Roldano Cattoni, Sebastian Stüker, Matteo Negri, Marco Turchi, Thanh-Le Ha, Elizabeth Salesky, Ramon Sanabria, Loic Barrault, Lucia Specia, and Marcello Federico. 2019. The IWSLT 2019 Evaluation Campaign. In *International Workshop on Spoken Language Translation (IWSLT)*, Hongkong.

Jan Niehues, Ngoc-Quan Pham, Thanh-Le Ha, Matthias Sperber, and Alex Waibel. 2018. Low-Latency Neural Speech Translation. In *Annual Conference of the International Speech Communication Association (InterSpeech)*, Hyderabad, India.

Matthias Paulik. 2010. *Learning Speech Translation from Interpretation*. Ph.D. thesis, University of Karlsruhe.

Stephan Peitz, Simon Wiesler, Markus Nussbaum-Thom, and Hermann Ney. 2012. Spoken Language Translation Using Automatically Transcribed Text in Training. In *International Workshop on Spoken Language Translation (IWSLT) 2011*, pages 276–283, Hongkong.

Alicia Pérez, Victor Guijarrubia, Raquel Justo, M. Inés Torres, and Francisco Casacuberta. 2007. A comparison of linguistically and statistically enhanced models for speech-to-speech machine translation. In *International Workshop on Spoken Language Translation (IWSLT)*, Trento, Italy.

Ngoc-quan Pham, Thai-son Nguyen, Thanh-Le Ha, Juan Hussain, Felix Schneider, Jan Niehues, Sebastian Stüker, and Alexander Waibel. 2019. The IWSLT 2019 KIT Speech Translation System. In *International Workshop on Spoken Language Translation (IWSLT)*, Hongkong.

Juan Pino, Liezl Puzon, Jiatao Gu, Xutai Ma, Arya D. McCarthy, and Deepak Gopinath. 2019. Harnessing Indirect Training Data for End-to-End Automatic Speech Translation: Tricks of the Trade. In *International Workshop on Spoken Language Translation (IWSLT)*, Hongkong.

Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. 2013. Improved Speech-to-Text Translation with the Fisher and Callhome Spanish–English Speech Translation Corpus. In *International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany.

Pablo Romero-Fresco. 2009. More haste less speed: Edited versus verbatim respoken subtitles. *Vigo International Journal of Applied Linguistics*, 6:109–134.

Nicholas Ruiz and Marcello Federico. 2014. Assessing the impact of speech recognition errors on machine translation quality. *Conference of the Association for Machine Translation in the Americas (AMTA)*, 1(November):261–274.

Nicholas Ruiz, Qin Gao, William Lewis, and Marcello Federico. 2015. Adapting Machine Translation Models toward Misrecognized Speech with Text-to-Speech Pronunciation Rules and Acoustic Confusability. In *Annual Conference of the International Speech Communication Association (InterSpeech)*, pages 2247–2251, Dresden, Germany.

Ashutosh Saboo and Timo Baumann. 2019. Integration of Dubbing Constraints into Machine Translation. In *Conference on Machine Translation (WMT)*, pages 94–101, Florence, Italy.

Shirin Saleem, Szu-Chen Jou, Stephan Vogel, and Tanja Schultz. 2004. Using Word Lattice Information for a Tighter Coupling in Speech Translation Systems. In *International Conference on Spoken Language Processing (ICSLP)*, pages 41–44, Jeju Island, Korea.

Elizabeth Salesky, Matthias Sperber, and Alan W Black. 2019a. Exploring Phoneme-Level Speech Representations for End-to-End Speech Translation. In *Association for Computational Linguistics (ACL)*, pages 1835–1841, Florence, Italy.

Elizabeth Salesky, Matthias Sperber, and Alex Waibel. 2019b. Fluent Translations from Disfluent Speech in End-to-End Speech Translation. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, Minneapolis, USA.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. In *Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS.

Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. 2017a. Neural Lattice-to-Sequence Models for Uncertain Inputs. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1380–1389, Copenhagen, Denmark.

Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. 2019a. Attention-Passing Models for Robust and Data-Efficient End-to-End Speech Translation. *Transactions of the Association for Computational Linguistics (TACL)*.

Matthias Sperber, Graham Neubig, Ngoc-Quan Pham, and Alex Waibel. 2019b. Self-Attentional Models for Lattice Inputs. In *Association for Computational Linguistic (ACL)*, pages 1185–1197, Florence, Italy.

Matthias Sperber, Jan Niehues, and Alex Waibel. 2017b. Toward Robust Neural Machine Translation for Noisy Input Sequences. In *International Workshop on Spoken Language Translation (IWSLT)*, Tokyo, Japan.

F. W. M. Stentiford and M. G. Steer. 1988. Machine Translation of Speech. *British Telecom technology journal*, 6(2):116–123.

Mihaela C. Stoian, Sameer Bansal, and Sharon Goldwater. 2020. Analyzing ASR pretraining for low-resource speech-to-text translation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain.

Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. 2019. Multimodal Machine Translation through Visuals and Speech. *arXiv:1911.12798*.

Eiichiro Sumita, Tohru Shimizu, and Satoshi Nakamura. 2007. NICT-ATR speech-to-speech translation system. In *Association for Computational Linguistic (ACL)*, pages 25–28, Prague, Czech Republic. Association for Computational Linguistics.

Tzu-Wei Sung, Jun-You Liu, Hung-yi Lee, and Lin-shan Lee. 2019. Towards End-to-End Speech-to-Text Translation with Two-Pass Decoding. In *ICASSP*, pages 7175–7179, Brighton, United Kingdom.

Toshiyuki Takezawa, Tsuyoshi Morimoto, Yoshinori Sagisaka, Nick Campbell, Hitoshi Iida, Fumiaki Sugaya, Akio Yokoo, and Seiichi Yamamoto. 1998. A Japanese-to-English Speech Translation System: ATR-MATRIX. In *International Conference on Spoken Language (ICSLP)*, Sydney, Australia.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2019. Speech-to-speech Translation between Untranscribed Unknown Languages. In *Automatic Speech Recognition & Understanding (ASRU)*, Singapore.

Yulia Tsvetkov, Florian Metze, and Chris Dyer. 2014. Augmenting translation models with simulated acoustic confusions for improved spoken language translation. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 616–625, Gothenburg, Sweden.

Enrique Vidal. 1997. Finite-State Speech-to-Speech Translation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 111–114, Munich, Germany.

Alex Waibel, Ajay N. Jain, Arthur E. McNair, Hiroaki Saito, Alexander G. Hauptmann, and Joe Tebelskis. 1991. JANUS: a speech-to-speech translation system using connectionist and symbolic processing strategies. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 793–796, Toronto, Canada.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020a. CoVoST: A Diverse Multilingual Speech-To-Text Translation Corpus. In *Language Resources and Evaluation (LREC)*, Marseille, France.

Chengyi Wang, Yu Wu, Shujie Liu, Zhenglu Yang, and Ming Zhou. 2020b. Bridging the Gap between Pre-Training and Fine-Tuning for End-to-End Speech Translation. In *Conference on Artificial Intelligence (AAAI)*, New York City, USA.

Ye-Yi Wang and Alex Waibel. 1998. Modeling with structures in statistical machine translation. In *Association for Computational Linguistics and International Conference on Computational Linguistics (COLING-ACL)*, pages 1357–1363, Montréal, Canada.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-Sequence Models Can Directly Transcribe Foreign Speech. In *Annual Conference of the International Speech Communication Association (InterSpeech)*, Stockholm, Sweden.

M. Woszczyna, N. Coccaro, A. Eisele, A. Lavie, A. McNair, T. Polzin, I. Rogina, C. P. Rose, T. Sloboda, M. Tomita, J. Tsutsumi, N. Aoki-Waibel, A. Waibel, and W. Ward. 1993. Recent advances in JANUS: A Speech Translation System. In *Workshop on Human Language Technology (HLT)*, pages 211–216, Plainsboro, USA.

Pei Zhang, Boxing Chen, Niyu Ge, and Kai Fan. 2019. Lattice Transformer for Speech Translation. In *Association for Computational Linguistic (ACL)*, pages 6475–6484, Florence, Italy.

Ruiqiang Zhang, Genichiro Kikui, Hirofumi Yamamoto, and Wai-Kit Lo. 2005. A Decoding Algorithm for Word Lattice Translation in Speech Translation. In *International Workshop on Spoken Language Translation (IWSLT)*, pages 23–29, Pittsburgh, USA.