Consistent Transcription and Translation of Speech

Matthias Sperber, Hendra Setiawan, Christian Gollan, Udhyakumar Nallasamy, Matthias Paulik

Apple
{sperber,hendra,cgollan,udhay,mpaulik}@apple.com

Abstract

The conventional paradigm in speech translation starts with a speech recognition step to generate transcripts, followed by a translation step with the automatic transcripts as input. To address various shortcomings of this paradigm, recent work explores end-to-end trainable direct models that translate without transcribing. However, transcripts can be an indispensable output in practical applications, which often display transcripts alongside the translations to users.

We make this common requirement explicit and explore the task of jointly transcribing and translating speech. Although high accuracy of transcript and translation are crucial, even highly accurate systems can suffer from inconsistencies between both outputs that degrade the user experience. We introduce a methodology to evaluate consistency and compare several modeling approaches, including the traditional cascaded approach and end-to-end models. We find that direct models are poorly suited to the joint transcription/translation task, but that end-to-end models that feature a coupled inference procedure are able to achieve strong consistency. We further introduce simple techniques for directly optimizing for consistency, and analyze the resulting trade-offs between consistency, transcription accuracy, and translation accuracy.[1]

1 Introduction

Speech translation (ST) is the task of translating acoustic speech signals into text in a foreign language. According to the prevalent framing of ST (e.g., Ney, 1999), given some input speech $\mathbf{x}$, ST seeks an optimal translation $\hat{\boldsymbol{t}} \in \mathcal{T}$, while possibly marginalizing over transcripts $\boldsymbol{s} \in \mathcal{S}$:

$$\hat{\boldsymbol{t}} = \operatorname*{argmax}_{\boldsymbol{t} \in \mathcal{T}} \left\{ P\left(\boldsymbol{t} \mid \mathbf{x}\right) \right\} \tag{1}$$
$$\approx \operatorname*{argmax}_{\boldsymbol{t} \in \mathcal{T}} \left\{ \sum_{\boldsymbol{s} \in \mathcal{S}} P_{\text{MT}}\left(\boldsymbol{t} \mid \boldsymbol{s}\right) P_{\text{ASR}}\left(\boldsymbol{s} \mid \mathbf{x}\right) \right\}.$$

According to this formulation, ST models primarily focus on translation quality, while transcription receives less emphasis. In contrast, practical ST user interfaces often display transcripts to the user alongside the translations. A typical example is a two-way conversational ST application that displays the transcript to the speaker for verification, and the translation to the conversation partner (Hsiao et al., 2006). Therefore, there is a mismatch between this practical requirement and the prevalent framing as described above.

While traditional ST models often do commit to a single automatic speech recognition (ASR) transcript that is then passed on to a machine translation (MT) component (Stentiford and Steer, 1988; Waibel et al., 1991), researchers have undertaken much effort to mitigate resulting error propagation issues by developing models that avoid making decisions on transcripts. Recent examples include direct models (Weiss et al., 2017) that bypass transcript generation, and lattice-to-sequence models (Sperber et al., 2017) that translate the ASR search space as a whole. Despite their merits, such models may not be ideal for scenarios that display both a translation and a corresponding transcript to users.

In this paper, we replace Eq. 1 by a joint transcription/translation objective to reflect this requirement:

$$\hat{\boldsymbol{s}}, \hat{\boldsymbol{t}} = \operatorname*{argmax}_{\boldsymbol{s} \in \mathcal{S}, \boldsymbol{t} \in \mathcal{T}} \left\{ P\left(\boldsymbol{s}, \boldsymbol{t} \mid \mathbf{x}\right) \right\}. \tag{2}$$

This change in perspective has significant implications not only on model design but also

[1] We release human annotations of consistency under https://github.com/apple/ml-transcript-translation-consistency-ratings.

Transactions of the Association for Computational Linguistics, vol. 8, pp. 695–709, 2020. https://doi.org/10.1162/tacl_a_00340
Action Editor: David Chiang. Submission batch: 4/2020; Revision batch: 7/2020; Published 11/2020.
© 2020 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00340/1923479/tacl_a_00340.pdf by guest on 25 September 2022

Transcr.	*And I could sort of* ***replay*** *some stuff that I was looking at earlier today.*
Transl.	*Ich konnte irgendwelche Dinge* ***ersetzen****, die ich heute vorhin schaute.* (I could any things replace; that I today earlier viewed)

Figure 1: Example of lexical inconsistencies we encountered when generating transcript and translation independently. Although the transcript correctly contains *replay*, the German translation (mistakenly) chooses *ersetzen* (English: *replace*). The inconsistency is explained by the acoustic similarity between *replay* and *replace*, which is not obvious to a monolingual user.

Transcr.	***Bill Gross*** *has several companies, including one called* ***eSolar*** *that has some great solar* ***thermal technologie****s.*
Transl.1 (gold)	***Bill Gross*** *hat mehrere Firmen, unter anderem eine namens* ***eSolar*** *die großartige Solar****thermaltechnologie****n hat.*
Transl.2 (incons.)	***Bill Gross*** *hat mehrere Firmen, eine nennt sich* ***"Easolare"****, die großartige* ***Solarwärme-Technologie*** *hat.*

Figure 2: Illustration of surface-level consistency between English transcript and German translation. Only translation 1 spells both named entities (*Bill Gross* and *eSolar*) consistently, and the German translation *Solarthermaltechnologie* (translation 1) is preferred over *Solarwärme-Technologie* (translation 2), by itself a correct choice but less similar on the surface level.

on evaluation. First, besides translation accuracy, transcription accuracy becomes relevant and equally important. Second, the issue of *consistency* between transcript and translation becomes essential. For example, let us consider a naive approach of transcribing and translating with two completely independent, potentially erroneous models. These independent models would expectedly produce inconsistencies, including inconsistent lexical choice caused by acoustic or linguistic ambiguity (Figure 1), and inconsistent spelling of named entities (Figure 2). Even if output quality is high on average, such inconsistencies may considerably degrade the user experience.

Our contributions are threefold: First, we introduce the notion of consistency between transcripts and translations and propose methods to assess consistency quantitatively. Second, we survey and extend existing models, and develop novel training and inference schemes, under the hypothesis that both joint model training and a coupled inference procedure are desirable for our goal of accurate and consistent models. Third, we provide a comprehensive analysis, comparing accuracy and consistency for a wide variety of model types across several language pairs to determine the most suitable models for our task and analyze potential trade-offs.

2 Evaluation Beyond Accuracy—The Need for Consistency

To better understand the desiderata of models that perform transcription *and* translation, it is helpful to discuss how one should evaluate such models. A first step is to evaluate transcription accuracy and translation accuracy in isolation. For this purpose, we can use well-established evaluation metrics such as word error rate (WER) for transcripts and BLEU (Papineni et al., 2002) for translations. When considering scenarios in which both transcript and translation are displayed, *consistency* is an essential additional requirement.[2] Let us first clarify what we mean by this term.

Definition: Consistency between transcript and translation is achieved if both are semantically equivalent, with a preference for a faithful translation approach (Newmark, 1988), meaning that stylistic, lexical, and grammatical characteristics should be transferred whenever fluency is not compromised. Importantly, consistency measures are defined over the space of both well-formed and erroneous sentence pairs. In the case of ungrammatical sentence pairs, consistency may be achieved by adhering to a literal or word-for-word translation strategy.

Consistency is only loosely related to accuracy, and can even be in opposition in some cases. For instance, when a translation error cannot be avoided, consistency is improved at the cost of transcription accuracy by placing the back-translated error in the transcript. Because accuracy and error metrics assess transcript or translation quality in isolation, these metrics cannot capture phenomena that involve the interplay between transcript and translation.

[2]Other important ST use cases do not show both transcripts at the same time, such as multilingual movie subtitling. For such cases, consistency may be less critical.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00340/1923479/tacl_a_00340.pdf by guest on 25 September 2022

2.1 Motivational Use Cases

Although ultimately user studies must assess to what extent consistency improves user satisfaction, our intention in this paper is to provide a universally useful notion of consistency that does not depend too much on specific use cases. Nevertheless, our definition may be most convincing when put in the context of specific example use cases.

Lecture Use Case. Here, a person follows a presentation or lecture-like event, presented in a foreign language, by reading transcript and translation on a screen (Fügen, 2008). This person may have partial knowledge of the source language, but knows only the target language sufficiently well. She, therefore, pays attention mainly to the translation outputs, but may occasionally consult the transcription output in cases where the translation seems wrong. In this case, quick orientation can be critical, and inconsistencies would cause distraction and undermine trust and perceived transparency of the transcription/translation service.

Dialog Use Case. Next, consider the scenario of a dialog between two people who speak different languages. One person, the speaker, attempts to convey a message to the recipient, relying on an ST service that displays a transcript and a translation. Here, the transcript is shown to the speaker, who speaks only the source language, for purposes of verification and possibly correction. The translation is shown to the recipient, who only understands the target language, to convey the message (Hsiao et al., 2006). We can expect that if transcript and translation are error-free, then the message is conveyed smoothly. However, when the transcript or translation contains errors, miscommunication occurs. To efficiently recover from such miscommunication, both parties should agree on the nature and details of the mistaken content. In other words, occurring errors are preferred to be consistent between transcript and translation.

3 Estimating Consistency

Having argued for consistency as a desirable property, we now wish to empirically quantify the level of consistency between a particular model's transcripts and translations. To our knowledge, consistency has not been addressed in the context of ST before, perhaps because traditional cascaded models have not been observed to suffer from inconsistencies in the outputs. Therefore, we propose several metrics for estimating transcript/translation consistency in this section. In §7.3, we demonstrate strong agreement of these metrics with human ratings of consistency.

3.1 Lexical Consistency

Our first metric focuses on semantic equivalency in general, and consistent lexical choice in particular, as illustrated in Figure 1. To this end, we use a simple lexical coverage model based on word-level translation probabilities. This approach might also capture some aspects of grammatical consistency by rewarding the use of comparable function words. We sum negative translation log-probabilities for each utterance: $t_{\boldsymbol{t}\rightarrow\boldsymbol{s}} = -\sum_{t_j \in \boldsymbol{t}} \max_{s_i \in \boldsymbol{s}} \log p\left(t_j \mid s_i\right)$. We then normalize across the test corpus $\mathcal{C}$ and average over both translation directions: $\frac{1}{2}\left(\frac{1}{n}\sum_{(\boldsymbol{s},\boldsymbol{t})\in\mathcal{C}} t_{\boldsymbol{t}\rightarrow\boldsymbol{s}} + \frac{1}{m}\sum_{(\boldsymbol{s},\boldsymbol{t})\in\mathcal{C}} t_{\boldsymbol{s}\rightarrow\boldsymbol{t}}\right)$, where n and m denote the number of translated and transcribed words in the corpus, respectively. In practice, we use `fast_align` (Dyer et al., 2013) to estimate probability tables from our training data. When a word has no translation probability assigned, including out-of-vocabulary cases, we use a simple smoothing method by assigning the lowest score found in the lexicon.

Although it may seem tempting to use a more elaborate translation model such as an encoder-decoder model, we deliberately choose this simple lexical approach. The main reason is that we need to estimate consistency for potentially erroneous transcript/translation pairs. In such cases, we found severe robustness issues when computing translation scores using a full-fledged encoder-decoder model.

3.2 Surface Form Consistency

Our consistency definition mentions a preference for a stylistic similarity between transcript and translation. One way of assessing stylistic aspects is to compare transcripts and translations at the *surface level*. This is most sensible when the source and target language are related, and could help capture phenomena such as consistent spelling of named entities, or translations using

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00340/1923479/tacl_a_00340.pdf by guest on 25 September 2022

words with similar surface form as found in the transcript. Figure 2 provides an illustration.

We propose to assess surface form consistency through substring overlap. Our notion of substring overlap follows CharCut, which was proposed as a metric for reference-based MT evaluation (Lardilleux and Lepage, 2017). Following Eq. 2 of that paper, we determine substring insertions, deletions, and shifts in the translation, when compared with the transcript, and compute $1 - \frac{\text{deletions}+\text{insertions}+\text{shifts}}{|\boldsymbol{s}|+|\boldsymbol{t}|}$. Counts are aggregated and normalized at corpus level. To avoid spurious matches, we match only substrings of at least length n (here: 5), compare in case-sensitive fashion, and deactivate CharCut's special treatment of longest common prefixes/suffixes.

We note that surface form consistency is less suited to language pairs that use different alphabets, and leave it to future work to explore alternatives, such as the assessment of cross-lingual phonetic similarity in such cases.

3.3 Correlation of Transcription/Translation Error

This third metric bases consistency on well-established accuracy metrics or error metrics. We posit that a necessary (though not sufficient) condition for consistency is that the accuracy of the transcript should be correlated with the accuracy of the translation, where both are measured against some respective gold standard. We therefore propose to assess consistency through computing statistical correlation between utterance-level error metrics for transcript and translation.

Specifically, for a test corpus of size N, we compute Kendall's τ coefficient across utterance-level error metrics. On the transcript side, we use utterance-level WER as the error metric. Because BLEU is a poor utterance-level metric, we make use of CharCut on the translation side, which has been shown to correlate well with human judgment at utterance level (Lardilleux and Lepage, 2017). Formally, we compute:

$$\text{kendall}_{\,\tau}\left(\text{WER}_{1:N}^{\text{clipped}}, \text{CharCut}_{1:N}\right). \quad (3)$$

Because CharCut is clipped above 1, we also apply clipping to utterance-level WER for stability.

s	t	Result
Good	Good	Immediate success
Bad	Bad	Speaker rephrases
Bad	Good	Speaker rephrases unnecessarily
Good	Bad	Resolve now or in later turn

Figure 3: Dialog use case. Whenever the transcript *or* the translation has errors, additional effort is needed.

3.4 Combined Metric for Dialog Task

The previous metrics estimate consistency in a fashion that is complementary to accuracy, such that it is possible to achieve good consistency despite poor accuracy. This allows trading off accuracy against consistency, depending on specific task requirements. Here, we explore a particular instance of such a task-specific trade-off that arises naturally through the formulation of a communication model. We consider a dialog situation (§2.1), and assume that communication will be successful if and only if both transcript and translation do not contain significant deviations from some reference, as motivated in Figure 3. Conceptually, the main difference to §3.3 is that here we penalize, rather than reward, the *bad/bad* situation (Figure 3). To estimate the probability of some generated transcript and translation allowing successful communication, given reference transcript and translation, we thus require that both the transcript and the translation are sufficiently accurate. For utterance with index k:

$$\begin{aligned} P(\text{succ}_k \mid \text{ref}) &= P\left(\boldsymbol{s}_k \text{ ok} \cap \boldsymbol{t}_k \text{ ok} \mid \text{ref}\right) \\ &= P\left(\boldsymbol{s}_k \text{ ok} \mid \text{ref}\right) \times P\left(\boldsymbol{t}_k \text{ ok} \mid \boldsymbol{s}_k, \text{ref}\right) \quad (4) \\ &\approx P\left(\boldsymbol{s}_k \text{ ok} \mid \text{ref}\right) \times P\left(\boldsymbol{t}_k \text{ ok} \mid \text{ref}\right) \end{aligned}$$

We then use utterance-level accuracy metrics as a proxy, computing accuracy $(\boldsymbol{s}_k) = 1 - \text{WER}_k^{\text{clipped}}$, accuracy $(\boldsymbol{t}_k) = 1 - \text{CharCut}_k$. For a test corpus of size N we compute corpus-level scores as $\frac{1}{N}\sum_{1\le k\le N} P(\text{succ}_k)$.

4 Models for Transcription and Translation

We now turn to discuss model candidates for consistent transcription and translation of speech (Figures 4–5). We hypothesize that there are two desirable model characteristics in our scenario. First, motivated by Eq. 2, models may achieve better consistency by performing joint inference, in the sense that no independence assumption

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00340/1923479/tacl_a_00340.pdf by guest on 25 September 2022

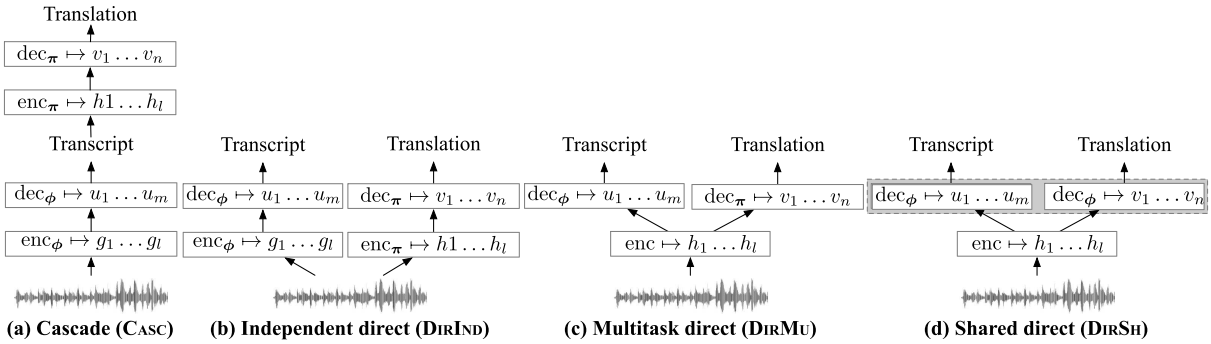

Figure 4: Cascaded and direct model types.

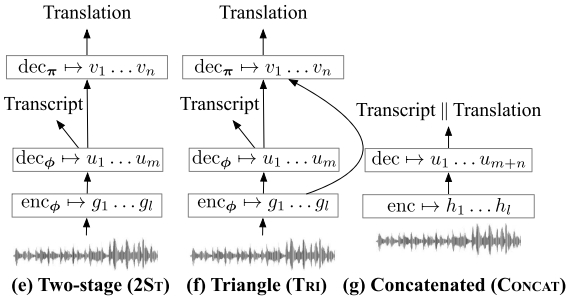

Figure 5: Joint models, featuring both coupled inference and end-to-end training.

between transcript and translation are made. We call this characteristic *coupled inference*. Second, shared representations through *end-to-end* (or joint) training may be of advantage in our scenario. We introduce several model variants, and also discuss whether they match these characteristics.

4.1 Model Basics

For a fair comparison, we keep the underlying architectural details as similar as possible across compared model types. All models are based on the attentional encoder-decoder framework (Bahdanau et al., 2015). For audio encoders, we roughly follow Chiu et al. (2018)'s multilayer bidirectional LSTM model, which encodes log-Mel speech features that are stacked and downsampled by a factor of 3 before being consumed by the encoder. When a model requires a text encoder (§4.2), we utilize residual connections and feed-forward blocks similar to Vaswani et al. (2017), although for simplicity we use LSTMs (Hochreiter and Schmidhuber, 1997) rather than self-attention in all encoder (and decoder) components. Similarly, decoder components use residual blocks of (unidirectional) LSTMs and feed-forward components (Domhan, 2018).

For ease of reference, we use $\texttt{enc}(\cdot)$ to refer to the encoder component that transforms speech inputs (or embedded text inputs) into a hidden encoder representations, $\texttt{dec}(\cdot)$ to refer to the attentional decoder component that produces hidden decoder states auto-regressively, and $\texttt{SoftmaxOut}(\cdot)$ to refer to the output softmax layer that models discrete output token probabilities. We will subscript components with the parameter sets $\boldsymbol{\pi}, \boldsymbol{\phi}$ to indicate cases in which model components are separately parametrized.

4.2 Cascaded Model (CASC)

The cascaded model (Figure 4a) represents ST's traditional approach of using separately trained ASR and MT models (Stentiford and Steer, 1988; Waibel et al., 1991). Here, we use modern sequence-to-sequence ASR and MT components. CASC runs a speech input $\mathbf{x}_{1:l}$ through an ASR model

$$\begin{aligned} \mathbf{g}_{1:l} &= \texttt{enc}_{\boldsymbol{\phi}}(\mathbf{x}_{1:l}) \\ \mathbf{u}_i &= \texttt{dec}_{\boldsymbol{\phi}}(\mathbf{u}_{<i}, \mathbf{g}_{1:l}, s_{i-1}) \\ P(s_i \mid \boldsymbol{s}_{<i}, x_{1:l}) &= \texttt{SoftmaxOut}_{\boldsymbol{\phi}}(\mathbf{u}_i), \end{aligned} \quad (5)$$

decodes the best hypothesis transcript $\hat{\boldsymbol{s}}$, and then applies a separate MT model

$$\begin{aligned} \mathbf{h}_{1:l} &= \texttt{enc}_{\boldsymbol{\pi}}(\hat{\boldsymbol{s}}) \\ \mathbf{v}_i &= \texttt{dec}_{\boldsymbol{\pi}}(\mathbf{v}_{<i}, \mathbf{h}_{1:l}, t_{i-1}) \\ P(t_i \mid \boldsymbol{t}_{<i}, \hat{\boldsymbol{s}}) &= \texttt{SoftmaxOut}_{\boldsymbol{\pi}}(\mathbf{v}_i) \end{aligned} \quad (6)$$

to generate a translation.

With respect to the two desirable characteristics of a consistent model, notice that CASC uses a coupled inference procedure, in the sense that no strong independence assumptions are made between transcript and translation. CASC may therefore be a good candidate for consistent

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00340/1923479/tacl_a_00340.pdf by guest on 25 September 2022

speech transcription/translation. However, it is less straightforward to apply end-to-end training to cascaded models.

4.3 Direct Models

To improve over the cascaded approach, recent work has focused on end-to-end trainable models, with direct ST models being the most prototypical end-to-end model. In the following, we describe straightforward ways of extending direct models in order to apply them to our joint transcription/translation task. Note that these direct models (Figure 4b–d) generate transcripts and translations independently at inference time. In other words, these models do not support coupled inference, which may degrade consistency between transcript and translation.

It is worth discussing how our consistent transcription/translation scenario relates to the issue of error propagation, an important issue in ST in which translations are degraded due to poor transcription decisions. Prior research on direct ST models has often been motivated by the observation that direct ST models elegantly avoid the error propagation problem. However, note that by shifting perspective to the joint transcription/translation goal, error propagation loses much of its relevance. First, error propagation is usually used to describe the negative effect of intermediate decisions, but here transcripts no longer function as intermediates. Second, strategies to mitigate error propagation often seek to make translations *less* influenced by transcription decisions. This is in conflict with our goal of achieving consistency between transcript and translation, which calls for precisely the opposite: Transcription and translation decisions *should* strongly depend on each other.

4.3.1 Independent Direct Model (DirInd)

A simple way of using direct modeling strategies for our purposes is to use two independent direct models, one for transcription, one for translation (Figure 4b). Specifically, we compute

$$\begin{aligned}
\mathbf{g}_{1:l} &= \texttt{enc}_{\boldsymbol{\phi}}(\mathbf{x}_{1:l}) \\
\mathbf{u}_i &= \texttt{dec}_{\boldsymbol{\phi}}(\mathbf{u}_{<i}, \mathbf{g}_{1:l}, s_{i-1}) \\
P(s_i \mid \boldsymbol{s}_{<i}, x_{1:l}) &= \texttt{SoftmaxOut}_{\boldsymbol{\phi}}(\mathbf{u}_i) \\
\mathbf{h}_{1:l} &= \texttt{enc}_{\boldsymbol{\pi}}(\mathbf{x}_{1:l}) \\
\mathbf{v}_i &= \texttt{dec}_{\boldsymbol{\pi}}(\mathbf{v}_{<i}, \mathbf{h}_{1:l}, t_{i-1}) \\
P(t_i \mid \boldsymbol{t}_{<i}, x_{1:l}) &= \texttt{SoftmaxOut}_{\boldsymbol{\pi}}(\mathbf{v}_i).
\end{aligned} \tag{7}$$

We are not aware of prior work using independent models for transcription and translation. We include this model as a contrastive baseline for the subsequent two models.

4.3.2 Multitask Direct Model (DirMu)

A major weakness of DirInd is that transcription and translation models are trained separately. A better solution is to follow Weiss et al. (2017)'s approach and sharing the speech encoder between transcription and translation models while making use of multitask training. Compared with Eq. 7, $\texttt{enc}_{\boldsymbol{\phi}}$ and $\texttt{enc}_{\boldsymbol{\pi}}$ would be collapsed into a shared encoder (Figure 4c). Note that originally, Weiss et al. (2017) and follow-up works use the transcript decoder only to aid training and exploit additional data for ASR as a related task in multitask learning. However, it is straight-forward to utilize the transcript decoder during inference for our purposes.

4.3.3 Shared Direct Model (DirSh)

We can also take the amount of sharing to the extreme by sharing all weights, not just encoder weights. Increasing the number of shared parameters may positively impact transcription/translation consistency. We are not aware of prior work using this model variant for performing speech translation. Compared with Eq. 7, both $\texttt{enc}_{\boldsymbol{\phi}}$/$\texttt{enc}_{\boldsymbol{\pi}}$ and $\texttt{dec}_{\boldsymbol{\phi}}$/$\texttt{dec}_{\boldsymbol{\pi}}$ are collapsed into a shared encoder and a shared decoder (Figure 4d).

4.4 Joint Models

We previously discussed Casc as a model that features coupled inference but does not support end-to-end training. We also discussed several direct models, some of which support end-to-end training, but none of which follow a coupled inference procedure. This section introduces joint models that support both end-to-end training and coupled inference.[3]

4.4.1 Two-Stage Model (2St)

The two-stage model (Kano et al., 2017) is conceptually close to the cascaded approach but is end-to-end trainable because continuous transcript

[3] It is worth noting that the models discussed in §4.4 match our joint optimization goal exactly: $P(\boldsymbol{t}|\boldsymbol{s}, \mathbf{x})P(\boldsymbol{s}|\mathbf{x}) = P(\boldsymbol{t}, \boldsymbol{s}|\mathbf{x})$. This is in contrast to Casc, which assumes conditional independence between translation and input speech, given the transcript. However, we do not expect this to be of major importance for purposes of generating consistent transcripts and translations.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00340/1923479/tacl_a_00340.pdf by guest on 25 September 2022

decoder states are passed on to the translation stage. Following Sperber et al. (2019)'s formulation, we re-use Eq. (5) to model a transcript $\boldsymbol{s}$ and hidden decoder states $\mathbf{u}_1^m$, and then compute

$$\begin{aligned} \mathbf{v}_i &= \texttt{dec}_{\boldsymbol{\pi}}(\mathbf{v}_{<i}, \mathbf{u}_1^m) \\ P(t_i \mid \boldsymbol{t}_{<i}, \mathbf{u}_{1:m}) &= \texttt{SoftmaxOut}_{\boldsymbol{\pi}}(\mathbf{v}_i). \end{aligned} \tag{8}$$

Beam search is applied to decode transcripts, as well as the corresponding hidden decoder states $\mathbf{u}_{1:m}$ that are then translated. Note that in contrast to our paper, Kano et al. (2017) and Sperber et al. (2019) treat transcripts only as intermediate computations and do not report transcription accuracies.

4.4.2 Triangle Model (TRI)

The triangle model (Anastasopoulos and Chiang, 2018) extends 2ST by adding a second attention mechanism to the translation decoder that directly attends to the encoded speech inputs. Eq. 5 is reused for transcription, and translations are computed as

$$\begin{aligned} \mathbf{v}_i &= \texttt{dec}_{\boldsymbol{\pi}}(\mathbf{v}_{<i}, [\mathbf{u}_{1:m}; \mathbf{h}_{1:l}], t_{i-1}) \\ P(t_i|\boldsymbol{t}_{<i}, \mathbf{u}_{1:m}, \mathbf{x}_{1:l}) &= \texttt{SoftmaxOut}_{\boldsymbol{\pi}}(\mathbf{v}_i). \end{aligned} \tag{9}$$

TRI can be seen as combining DIRMU's advantage of featuring a direct connection between speech and translation, and 2ST's advantage of supporting joint inference. Anastasopoulos and Chiang (2018) evaluate both transcription and translation accuracy in a low-resource setting and report consistent improvements for the latter but less reliable gains for the former.

4.4.3 Concatenated Model (CONCAT)

Haghani et al. (2018) propose a sequence-to-sequence model that produces the concatenation of two outputs sequences in the context of spoken language understanding. To our knowledge it has not been utilized in an ST context before, but is a very natural fit for our joint transcription/translation scenario. CONCAT shares both the encoder and the decoder, leading to improved compactness:

$$\begin{aligned} \boldsymbol{r}_{1:m+n} &:= s_1 \dots s_m t_1 \dots t_n \\ \mathbf{g}_{1:l} &= \texttt{enc}(\mathbf{x}_{1:l}) \\ \mathbf{u}_i &= \texttt{dec}(\mathbf{u}_{<i}, \mathbf{g}_{1:l}, r_{i-1}) \\ P(r_i \mid \boldsymbol{r}_{<i}, \mathbf{x}_{1:l}) &= \texttt{SoftmaxOut}(\mathbf{u}_i). \end{aligned} \tag{10}$$

5 Consistency as Training and Inference Objectives

Having surveyed models that are suitable for our task to various degrees, we next explore simple ways to further improve the consistency of the generated outputs through adjusting training or inference objectives.

5.1 Consistency as Training Objective

At training time, we wish to introduce a loss term that penalizes inconsistent outputs. Whereas the consistency measures discussed in §3 are all defined at either the utterance or the corpus level, we define our loss term at the token level for convenient integration with the standard cross entropy loss term. For convenience, we opt to follow the notion of surface-level consistency (§3.2), according to which we may encourage models to assign probability mass to transcript (subword) tokens that appear in the translation, and to translated tokens that appear in the transcript.[4]

Consider the standard cross entropy loss, which is computed against the ground-truth label distribution $q(y_i) = \delta_{y_i, y_i^*}$ for predicted label y_i at target position i, assigning all probability mass to the reference token y_i^*. We modify the ground truth label distribution for transcript and translation outputs, respectively:

$$\begin{aligned} q'_{\text{transl}}(y_i) &= (1-\epsilon)\delta_{y_i, \boldsymbol{t}_i} + \frac{\epsilon}{|\boldsymbol{s}|}\sum_{w \in \boldsymbol{s}} \delta_{y_i, w} \\ q'_{\text{transcr}}(y_i) &= (1-\epsilon)\delta_{y_i, \boldsymbol{s}_i} + \frac{\epsilon}{|\boldsymbol{t}|}\sum_{w \in \boldsymbol{t}} \delta_{y_i, w} \end{aligned} \tag{11}$$

This can be seen as an instance of non-uniform label smoothing with strength ϵ (Szegedy et al., 2016). In practice, we give this loss term a relative weight of 0.1 during training, while at the same time disabling label smoothing. Because this loss requires access to the complete transcript and translation, we do not apply it at inference time.

5.2 Consistency as Inference Objective

We can also modify the inference objective to enforce more consistent outputs. A simple way for accomplishing this is via n-best rescoring. This

[4] Similarly to §3, this strategy targets related languages with shared alphabets, and our results for an English–German speech translation task are encouraging (§7.4). We leave it to future work to explore more elaborate solutions.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00340/1923479/tacl_a_00340.pdf by guest on 25 September 2022

is especially convenient when using consistency measures such as lexical consistency (§3.1), which can be computed without referring to a gold standard. Our approach here follows two simple steps: First, we compute n-best lists using standard beam search. Second, we select the $(\boldsymbol{s}, \boldsymbol{t})$-pair that produces the best lexical consistency score. Expectedly, this rescoring approach will yield improved consistency, while possibly degrading transcript or translation accuracy. Future work may explore ways for more explicitly balancing model and consistency scores.

6 Experimental Setup

6.1 Data

We conduct experiments on the MuST-C corpus (di Gangi et al., 2019), the largest publicly available ST corpus, containing TED[5] talks paired with English transcripts and translations into several languages. We present results for German, Spanish, Dutch, and Russian as the target language, where the data size is 408–504 hours of English speech, corresponding to 234K–270K utterances. In TED, translated subtitles are not displayed simultaneously with the transcribed subtitles, and consistency is therefore not inherently required in this data. In practice, however, the manual translation workflow in TED results in a sufficient level of consistency between transcripts and translations. Specifically, transcripts are generated first, and translators are required to use the transcript as a starting point while also referring to the audio.[6] We use MuST-C *dev* for validation and report results on *tst-COMMON*.

6.2 Model and Training Details

We make use of the 40-dimensional log Mel filterbank speech features provided with the corpus. The only text preprocessing applied to the training data is subword tokenization using `SentencePiece` (Kudo and Richardson, 2018) with the `unigram` setting. Following most recent work on end-to-end ST models, we choose a relatively small vocabulary size of 1024, with transcription/translation vocabularies shared. No additional preprocessing steps are applied for training, but for transcript evaluation we remove punctuation and non-speech event markers such as *(laughter)*, and compute case-insensitive WER. For translations, we remove non-speech markers from the decoded outputs and use `SacreBleu`[7] (Post, 2019) to handle tokenization and scoring.

Model hyperparameters are manually tuned for the highest accuracy with DirMu, our most relevant baseline. Unless otherwise noted, the same hyperparameters are used for all other model types. Weights for the speech encoder are initialized based on a pre-trained attentional ASR task that is identical to the ASR part of the direct multitask model. Other weights are initialized according to Glorot and Bengio (2010). The speech encoder is a 5-layer bidirectional LSTM with 700 dimensions per direction. Attentional decoders consist of 2 Transformer blocks (Vaswani et al., 2017) but use 1024-dimensional unidirectional LSTM instead of self-attention as a sequence model, except for the Concat and DirSh for which we increase to 3 layers. For Casc's MT model, encoder/decoder both contain 6 layers with 1024-dimensional LSTMs. Subword embeddings are of size 1024.

We regularize using LSTM dropout with $p = 0.3$, decoder input word-type dropout (Gal and Ghahramani, 2016), and attention dropout, both $p = 0.1$. We apply label smoothing with strength $\epsilon = 0.1$. We optimize using Adam (Kingma and Ba, 2014) with $\alpha = 0.0005$, $\beta_1 = 0.9$, $\beta_2 = 0.98$, 4000 warm-up steps, and learning rate decay by using the inverse square root of the iteration. We set the batch size dynamically based on the sentence length, such that the average batch size is 128 utterances. The training is stopped when the validation score has not improved over 3 epochs, where the validation score is the product of corpus-level translation BLEU score and corpus-level transcription word accuracy.

For decoding and generating n-best lists, we use beam size 10 and polynomial length normalization with exponent 1.5. Our implementation is based on `PyTorch` (Paszke et al., 2019) and `xnmt` (Neubig et al., 2018), and all trainings are done using single-GPU environments, utilizing Tesla V100 GPUs with 32 GB memory.

6.3 Human Ratings

To obtain a gold standard to compare our proposed automatic consistency metrics against, we collect transcript/translation consistency ratings from

[5] `www.ted.com`.

[6] `www.ted.com/participate/translate`.

[7] hash: case.lc+numrefs.1+smooth.4+tok.13a+version.1.4.3.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00340/1923479/tacl_a_00340.pdf by guest on 25 September 2022

Model	E2E training	$t \mid s$
§4.2 CASC	–	attention
§4.3.1 DIRIND	–	–
§4.3.2 DIRMU	✓	–
§4.3.3 DIRSH	✓	–
§4.4.1 TRI	✓	attention
§4.4.2 2ST	✓	attention
§4.4.3 CONCAT	✓	sequential

Table 1: Overview of models and key properties. All models except CASC/DIRIND are end-to-end (E2E) trained. Models also differ in whether translations are conditioned on transcripts ($\boldsymbol{t}|\boldsymbol{s}$), and whether conditioning is implemented through attention or through sequential decoder states.

human annotators. The annotators are presented a single transcript/translation pair at a time, and are asked to judge the consistency on a 4-point Likert scale. We aimed for a balanced scale which assigned a score of 4 to cases with no or only minor mismatch, a score of 3 to indicate a purely stylistic mismatch, a score of 2 to indicate a partial semantic mismatch, and a score of 1 to a complete semantic mismatch. Instructions given to the annotators include an explanation of the definition given in §2 along with a table of several examples for each of the 4 categories. We displayed transcripts and translations in randomized order, so as to obfuscate the directionality of the translation, and do not provide the source speech utterances. Annotators are recruited from an in-house pool of trusted annotators and required to be proficient English and German speakers.

For each of the 2641 speech utterances in the MuST-C English-German test set, we collect annotations for 8 transcript/translation pairs: 7 system outputs produced by the models in Table 1, and the reference transcript/translation pairs. Each transcript/translation item is rated individually and by at least three different annotators. In total, we used 58 raters to produce 63412 ratings. We fit a linear mixed-effects model on the result using the `lme4` package (Bates et al., 2013), which allows estimating the consistency of the outputs for each system, while accounting for random effects of each annotator and of each input sentence. We refer to Norman (2010) and Gibson et al. (2011) for a discussion of using mixed-effects models in the context of Likert-scale ratings.

7 Results

We start by presenting empirical results across all four language pairs, and will then focus on English–German to discuss details. Table 1 contrasts the different model types that we examine.

7.1 Accuracy Comparison

To validate our implementation and to evaluate the overall model accuracy, Table 2 compares models across four language pairs. The table confirms that, except for DIRIND, our models obtain strong overall accuracies, as compared with prior work on the same data by Di Gangi et al. (2019).[8] Overall, CASC outperforms CONCAT and the 3 direct models in terms of WER and BLEU. 2ST/TRI achieve similar or stronger translation accuracy compared with CASC. Joint model training (used by all models except CASC and DIRIND) seems to hurt transcription accuracy somewhat, although the differences are often not statistically significant. This may be caused by an inherent trade-off between translation and transcription accuracy, as discussed by He et al. (2011). Finally, CONCAT achieves favorable transcription accuracies, and translation accuracies fall between direct models and non-direct models in most cases.

7.2 Lexical Consistency Comparison

Table 2 also shows results for lexical consistency. Without exception, 2ST/TRI achieve the best results, followed by CASC and CONCAT. The direct models perform poorly in all cases. Given that CASC is by design a natural choice for joint transcription/translation, we did not necessarily expect 2ST/TRI to achieve better consistency. This encouraging evidence for the versatility of end-to-end trainable models is also supported by human ratings (§7.3).

To categorize models regarding inference procedure and end-to-end training (Table 1), we observe that coupled inference (all non-direct models) is most decisive for achieving good consistency, with conditioning on generated transcripts through sequential hidden states (CONCAT) being less effective than conditioning through

[8]Concurrent work (Liu et al., 2020) obtains better transcription results, but compiles its own version of the TED corpus, thus it is unclear to what extent differences can be attributed to better data filtering strategies, which are known to be a potential issue in MuST-C.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00340/1923479/tacl_a_00340.pdf by guest on 25 September 2022

	EN→DE			EN→ES			EN→NL			EN→RU		
Model	↓ WER	↑ BLEU	↓ Lex	WER	BLEU	Lex	WER	BLEU	Lex	WER	BLEU	Lex
SOTA *cc*	27.0	18.5	–	26.6	22.5	–	26.6	22.2	–	27.0	11.1	–
SOTA *dir*	–	17.3	–	–	20.8	–	–	18.8	–	–	8.5	–
CASC	**21.6**	19.3	10.4	**20.5**	**25.2**	8.4	**20.6**	**23.5**	10.1	**20.5**	13.4	11.3
DIRIND	**21.6**	11.0	21.1	**20.5**	16.5	17.8	**20.6**	14.9	20.9	**20.5**	3.4	29.0
DIRMU	23.6	18.4	13.9	21.7	24.3	11.6	23.2	22.3	14.3	22.4	13.0	13.9
DIRSH	23.6	19.0	14.7	21.3	24.1	11.5	22.0	22.7	14.2	22.3	13.6	13.6
2ST	*22.2*	**20.1**	*9.9*	*21.4*	24.2	**7.8**	22.6	*23.4*	9.4	21.4	*14.0*	**10.7**
TRI	*22.2*	*19.9*	**9.7**	*21.0*	*24.7*	*7.9*	24.4	22.6	**8.9**	21.2	**14.2**	**10.7**
CONCAT	*21.9*	19.2	12.8	*20.6*	23.7	10.8	21.9	22.8	12.5	21.5	13.3	13.3

Table 2: Comparison of WER, BLEU, lexical consistency (Lex; §3.1) across several language pairs. We compare against state-of-the-art (SOTA) results under same data conditions by Di Gangi et al. (2019), where *cc* denotes a cascaded model, *dir* denotes a direct model. **Bold font** indicates the best score. Results that are not statistically significantly worse than the best score in the same column are in *italics* (pairwise bootstrap resampling (Koehn, 2004), $p<0.05$).

		Transcript	Translation		Consistency				
Model	Params.	↓ WER	↑ BLEU	↓ CharCut	↓ Lex	↑ Sur	↑ Cor	↑ Cmb	↑ Human
CASC	223M	**21.6**	19.2	47.2	10.36	10.65	0.396	0.474	3.119
DIRIND	175M	**21.6**	11.0	60.3	21.13	5.24	0.346	0.374	2.195
DIRMU	124M	23.6	18.4	48.7	13.89	7.07	0.376	0.457	2.715
DIRSH	106M	23.6	19.0	47.9	14.71	8.54	0.371	0.464	2.776
2ST	122M	*22.2*	**20.1**	**46.1**	*9.86*	**12.08**	0.391	**0.484**	3.170
TRI	141M	*22.2*	*19.9*	*46.3*	**9.72**	*11.54*	**0.414**	**0.484**	**3.192**
CONCAT	106M	*21.9*	19.2	47.1	12.79	9.60	0.387	0.477	2.875
Reference	–	0	100	0	12.6	13.3	1	1	3.594

Table 3: Detailed consistency results, including surface form consistency (Sur; §3.2), correlation of error (Cor; §3.3), and the combined task-specific metric (Cmb; §3.4). **Bold font** indicates the best score among automatic outputs. Results that are not statistically significantly worse than the best score in the same column are in *italics*.

attention (other non-direct models). End-to-end training also appears beneficial for consistency (CASC vs. 2ST/TRI and DIRIND vs. DIRMU/DIRSH).

7.3 Analysis of Consistency Metrics

Table 3 presents more details for English–German and includes human ratings as gold standard, along with all four proposed automatic consistency measures. Note that the reported human ratings correspond to the intercepts of the linear mixed-effects model (§6.3). The fitted model estimates the standard deviation of the random effect for annotators at 0.28 and for input sentences at 0.37. All pairwise differences between the systems in the table are statistically significant ($p < 0.01$) according to an ANOVA test.

Encouragingly, lexical and surface form consistencies are aligned, and follow the same trends as the gold standard. The correlation-based measure agrees on the inferior consistency of direct models and the superior consistency of TRI, while producing slightly different orderings among the remaining models. According to our combined dialog-specific measure, TRI/2ST are tied for the best overall model.

One noteworthy observation is that lexical consistency of references is far worse than for 2ST/TRI outputs. This contradicts the gold standard outputs and is possibly caused by both the system outputs and the lexical consistency score being overly literal and biased toward high-frequent outputs. For comparison against references, the surface form consistency therefore appears to be a better choice.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00340/1923479/tacl_a_00340.pdf by guest on 25 September 2022

Model	WER	BLEU	↓ Lex	↑ Sur
TRI	**22.2**	**19.9**	9.72	11.54
training	24.0	**19.9**	9.84	12.09
inference	22.6	19.5	**8.79**	**13.17**
DIRMU	23.6	18.4	13.89	7.07
training	**23.2**	**18.9**	13.94	7.94
inference	24.0	18.7	**12.63**	**9.29**

Table 4: Direct optimization for consistency. We compare training (§5.1) and inference (§5.2) approaches. **Bold font** indicates the best score.

7.4 Directly Optimizing for Consistency

Table 4 considers the English–German translation direction, and examines the effect of using strategies for direct optimization of consistency at training and inference time (§5). All of the examined techniques improve consistency, though often at the cost of degraded accuracy. The training-time techniques appear more detrimental to transcription accuracy, and the inference-time techniques are more detrimental to translation accuracy. Although DIRMU benefits strongly from these techniques, it still falls behind TRI's consistency. For TRI, on the other hand, surface form consistency improves to the point where it almost matches the surface form consistency between reference transcripts and translations (3.594, see Table 3).

7.5 Consistency vs. Accuracy

Tables 2 and 3 tend to assign better consistency scores to models with higher accuracy scores. We wish to verify whether the trend is owed to the model characteristics or whether this indicates that our metrics fail to decouple accuracy and consistency. To this end, we again focus on English–German and introduce two new model variants: First, CINDP performs translation using CASC, but transcribes with an independently trained direct model. Expectedly, such a model shows high accuracy but low consistency, a hypothesis that is confirmed by results in Table 5, contrasted against DIRMU. Second, we train a weaker 2-stage model by using only half the training data. For such a model, we would expect lower accuracy but not lower consistency, which is again confirmed by Table 5, at least to some extent (lexical consistency is worse, but the correlation measure improves). These findings indicate that accuracy and consistency are in fact reasonably well decoupled.

			Consistency			
Model	WER	BLEU	Lex	Sur	Cor	Cmb
DIRMU	23.6	18.4	13.9	7.1	.38	.46
CINDP	21.8	19.2	14.6	8.3	.33	.47
2ST	22.2	20.1	9.9	12.1	.39	.48
2ST/2	30.0	16.6	10.9	11.9	.45	.44

Table 5: Consistency vs. accuracy. CINDP achieves better accuracy than DIRMU, but worse consistency scores. 2ST/2 is trained on less data than 2ST, which hurts its accuracy but not its consistency scores.

DIRMU	s	***Doctor King*** *made that shift* ***in*** *thinking.* (Dr. Keene made this shift)
	t	***Dr. Keene*** *machte diese Verschiebung* (and thinking.) ***und*** *Denkweise.*
TRI	s	***Dr. King*** *made that shift* ***in*** *thinking.* (Dr. King made this shift)
	t	***Dr. King*** *machte diese Verschiebung* (in thinking.) ***im*** *Denken.*
Ref.	s	*See, Dr. Kean made that shift in thinking.*

Figure 6: Example for inconsistently spelled names and an inconsistent function word when generating transcript and translation separately using DIRMU.

7.6 Qualitative Analysis

Manual inspection of the outputs of DIRMU and TRI for the English–German model confirms our intuition and the quantitative findings presented above, namely, that DIRMU suffers from considerable consistency issues due to transcripts and translations being generated separately. Examples in the decoded test data are in fact easy to spot, whereas for TRI we do find any consistency problems. Figures 6–8 show cherry-picked examples.

8 Related Work

To our knowledge there exists no prior work on consistency for joint transcription and translation of speech in particular, or other multitask conditional sequence generation models in general. The closest related prior work is perhaps Ribeiro et al. (2019), who analyze the case of contradictory model outputs in a question answering task in which multiple different but highly related questions are shown to the model. Other prior work examines the trade-off between transcription

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00340/1923479/tacl_a_00340.pdf by guest on 25 September 2022

DIRMU	*s*	*Really advanced civilization based on it, it **dances** in in energy.*
		A really advanced
	t	*Eine wirklich **fortgeschrittene***
		civilization bases on energy.
		Zivilisation basiert auf Energie.
TRI	*s*	*Really advanced civilization is based on **advances** in energy.*
		Really advanced civilization
	t	*Wirklich fortschrittliche Zivilisation*
		bases on advances in energy.
		*basiert auf **Fortschritten** in Energie.*
Ref.	*s*	*Really advanced civilization is based on advances in energy.*

Figure 7: Here, DIRMU makes inconsistent lexical choices for transcript and translation, leading to a correct translation despite an incorrect transcript.

DIRMU	*s*	*So the question is, can you actually get that **desert?***
		The question is thus: can you this
	t	*Die Frage ist also: Können Sie das*
		actually change?
		*tatsächlich **verändern**?*
TRI	*s*	*So the question is: Can you actually get that **to zero?***
		The question is thus: can you this
	t	*Die Frage ist also: Können Sie das*
		actually to zero bring
		*tatsächlich **zu Null** bringen?*
Ref.	*s*	*So the question is: Can you actually get that to zero?*

Figure 8: This is an example where DIRMU produces incorrect outputs on both sides, with seemingly unrelated semantics.

and translation quality in more traditional speech translation models theoretically (He and Deng, 2011) and empirically (He et al., 2011). Findings indicate that optimizing for WER does not necessarily lead to the best translations in a cascaded speech translation model, which is in line with the accuracy trade-offs observed in our experiment. Concurrent work explores synchronous decoding strategies for jointly transcribing and translating speech, but does not discuss the issue of consistency (Liu et al., 2020).

With regard to our consistency evaluation metrics, a closely related line of research is work on quality estimation and cross-lingual similarity metrics (Fonseca et al., 2019). An important difference of transcription/translation consistency is that for purposes of assessing consistency there is no directionality, and both input sequences can be erroneous. It is therefore especially important for metrics to be robust against errors on both sides. Moreover, stylistic differences are often not accounted for in this line of prior work. We note the similarity of our proposed lexical consistency metric to work by Popović et al. (2011), and leave it for future work to explore whether metrics from other related work can and should be employed to measure consistency.

Finally, producing transcripts alongside translations may be framed as producing an explanation (the transcript) alongside the main output (the translation). Research on explainable machine learning systems (Smith-Renner et al., 2020, and references therein) may shed light on desirable properties of these explanation from a usability point of view, as well as questions related to appropriate user interface design.

9 Conclusion

This paper investigates the task of jointly transcribing and translating speech, which is relevant for use cases in which both transcripts and translations are displayed to users. The main theme has been the discussion of consistency between transcripts and translations. To this end, we proposed a notion of consistency and introduced techniques to estimate it. We conducted a thorough comparison across a wide range of models, both traditional and end-to-end trainable, with regards to both accuracy and consistency. As important model ingredients, we found that a coupled inference procedure, where translations are conditioned on transcripts through attention, is particularly helpful. We also found that end-to-end training improves consistency and translations but at the cost of degraded transcripts. We further introduced training and inference techniques that are effective at further improving consistency, which we found to also come with some trade-offs.

Future work should examine how consistency correlates with user experience in practice and establish specific trade-offs for various use cases. Moreover, our techniques are applicable to other multitask use cases that could potentially benefit from consistent outputs.

References

Antonios Anastasopoulos and David Chiang. 2018. Tied Multitask Learning for Neural Speech Translation. In *North American Chapter*

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00340/1923479/tacl_a_00340.pdf by guest on 25 September 2022

of the Association for Computational Linguistics (NAACL). New Orleans, LA, USA. **DOI:** https://doi.org/10.18653/v1/N18-1008

Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Representation Learning (ICLR)*. San Diego, CA, USA.

Douglas Bates, Martin Maechler, Ben Bolker, and Steven Walker. 2013. lme4: Linear mixed-effects models using Eigen and S4. *R package version*, 1(4).

Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Katya Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, and Michiel Bacchiani. 2018. State-of-the-art Speech Recognition With Sequence-to-Sequence Models. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778. Calgary, Canada.

Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. 2019. Adapting Transformer to End-to-end Spoken Language Translation. In *Annual Conference of the International Speech Communication Association (InterSpeech)*, pages 1133–1137. Graz, Austria.

Tobias Domhan. 2018. How Much Attention Do You Need ? A Granular Analysis of Neural Machine Translation Architectures. In *Association for Computational Linguistic (ACL)*, pages 1799–1808. Melbourne, Australia.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 644–648. Atlanta, GA, USA.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 Shared Tasks on Quality Estimation. In *Conference on Machine Translation (WMT)*. Florence, Italy. **DOI:** https://doi.org/10.18653/v1/W19-5401

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

Yarin Gal and Zoubin Ghahramani. 2016. A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. In *Neural Information Processing Systems Conference (NIPS)*, pages 1019–1027. Barcelona, Spain.

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

Edward Gibson, Steve Piantadosi, and Kristina Fedorenko. 2011. Using Mechanical Turk to Obtain and Analyze English Acceptability Judgments. *Language and Linguistics Compass*, 5(8):509–524. **DOI:** https://doi.org/10.1111/j.1749-818X.2011.00295.x

Xavier Glorot and Yoshua Bengio. 2010. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. Sardinia, Italy.

Parisa Haghani, Arun Narayanan, Michiel Bacchiani, Galen Chuang, Neeraj Gaur, Pedro Moreno, Rohit Prabhavalkar, Zhongdi Qu, and Austin Waters. 2018. From Audio to Semantics: Approaches to End-to-End Spoken Language Understanding. In *Spoken Language Technology Workshop (SLT)*. Athens, Greece. **DOI:** https://doi.org/10.1109/SLT.2018.8639043

Xiaodong He and Li Deng. 2011. Speech Recognition, Machine Translation, and Speech Translation – A Unified Discriminative Learning Paradigm. *IEEE Signal Processing Magazine*, 28(5):126–133. **DOI:** https://doi.org/10.1109/MSP.2011.941852

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.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00340/1923479/tacl_a_00340.pdf by guest on 25 September 2022

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780. **DOI:** https://doi.org/10.1162/neco.1997.9.8.1735, **PMID:** 9377276

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, PA, USA.

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. **DOI:** https://doi.org/10.21437/Interspeech.2017-944

Diederik P. Kingma and Jimmy L. Ba. 2014. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*. Banff, Canada.

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

Taku Kudo and John Richardson. 2018. Sentence Piece: A Simple and Language Independent Subword Tokenizer and Detokenizer For Neural Text Processing. *Empirical Methods in Natural Language Processing (EMNLP)*, pages 66–71. **DOI:** https://doi.org/10.18653/v1/D18-2012, **PMID:** 29382465

Adrien Lardilleux and Yves Lepage. 2017. CHARCUT: Human-Targeted Character-Based MT Evaluation with Loose Differences. In *International Workshop on Spoken Language Translation (IWSLT)*. Tokyo, Japan.

Yuchen Liu, Jiajun Zhang, Hao Xiong, Long Zhou, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2020. Synchronous Speech Recognition and Speech-to-Text Translation with Interactive Decoding. In *Conference on Artificial Intelligence (AAAI)*, New York, NY, USA. **DOI:** https://doi.org/10.1609/aaai.v34i05.6360

Graham Neubig, Matthias Sperber, Xinyi Wang, Matthieu Felix, Austin Matthews, Sarguna Padmanabhan, Ye Qi, Devendra Singh Sachan, Philip Arthur, Pierre Godard, John Hewitt, Rachid Riad, and Liming Wang. 2018. XNMT: The eXtensible Neural Machine Translation Toolkit. In *Conference of the Association for Machine Translation in the Americas (AMTA) Open Source Software Showcase*. Boston, MA, USA.

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, AZ, USA.

Geoff Norman. 2010. Likert Scales, Levels of Measurement and the "Laws" of Statistics. *Advances in Health Sciences Education*, 15(5):625–632.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method For Automatic Evaluation of Machine Translation. In *Association for Computational Linguistic (ACL)*, pages 311–318. Philadephia, PA, USA. **DOI:** https://doi.org/10.3115/1073083.1073135

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems (NeuIPS)*. Vancouver, Canada.

Maja Popović, David Vilar, Eleftherios Avramidis, Aljoscha Burchardt, and Maja

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00340/1923479/tacl_a_00340.pdf by guest on 25 September 2022

Popović. 2011. Evaluation Without References: IBM1 Scores as Evaluation Metrics. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 99–103. Edinburgh, Scotland.

Matt Post. 2019. A Call for Clarity in Reporting BLEU Scores. In *Conference on Machine Translation (WMT)*, pages 186–191. Brussels, Belgium. **DOI:** https://doi.org/10.18653/v1/W18-6319

Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. Are Red Roses Red? Evaluating Consistency of Question-Answering Models. In *Association for Computational Linguistic (ACL)*, pages 6174–6184. Florence, Italy. **DOI:** https://doi.org/10.18653/v1/P19-1621

Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-graber, Daniel S. Weld, and Leah Findlater. 2020. No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML. In *Conference on Human factors in computing systems (CHI)*. Honolulu, HI, USA. **DOI:** https://doi.org/10.1145/3313831.3376624

Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. 2017. Neural Lattice-to-Sequence Models for Uncertain Inputs. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1380–1389. Copenhagen, Denmark. **DOI:** https://doi.org/10.18653/v1/D17-1145

Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. 2019. Attention-Passing Models for Robust and Data-Efficient End-to-End Speech Translation. *Transactions of the Association for Computational Linguistics (TACL)*. **DOI:** https://doi.org/10.1162/tacl_a_00270

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

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826. Las Vegas, NV, USA.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Neural Information Processing Systems Conference (NIPS)*, pages 5998–6008. Long Beach, CA, USA.

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. **DOI:** https://doi.org/10.1109/ICASSP.1991.150456

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. **DOI:** https://doi.org/10.21437/Interspeech.2017-503

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00340/1923479/tacl_a_00340.pdf by guest on 25 September 2022