Adaptive Feature Selection for End-to-End Speech Translation

Citation for published version:

Digital Object Identifier (DOI):
10.18653/v1/2020.findings-emnlp.230

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Findings of the Association for Computational Linguistics: EMNLP 2020

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Adaptive Feature Selection for End-to-End Speech Translation

Biao Zhang1  Ivan Titov1,2  Barry Haddow1  Rico Sennrich3,1

1School of Informatics, University of Edinburgh
2ILLC, University of Amsterdam
3Department of Computational Linguistics, University of Zurich

B.Zhang@ed.ac.uk, {ititov, bhaddow}@inf.ed.ac.uk, sennrich@cl.uzh.ch

Abstract

Information in speech signals is not evenly distributed, making it an additional challenge for end-to-end (E2E) speech translation (ST) to learn to focus on informative features. In this paper, we propose adaptive feature selection (AFS) for encoder-decoder based E2E ST. We first pre-train an ASR encoder and apply AFS to dynamically estimate the importance of each encoded speech feature to ASR. A ST encoder, stacked on top of the (frozen) ASR encoder, then receives the filtered features from the ASR encoder. We take $\mathcal{L}_0$Drop (Zhang et al., 2020) as the backbone for AFS, and adapt it to sparsify speech features with respect to both temporal and feature dimensions. Results on LibriSpeech En-Fr and MuST-C benchmarks show that AFS facilitates learning of ST by pruning out $\sim$84% temporal features, yielding an average translation gain of $\sim$1.3–1.6 BLEU and a decoding speedup of $\sim$1.4$x$. In particular, AFS reduces the performance gap compared to the cascade baseline, and outperforms it on LibriSpeech En-Fr with a BLEU score of 18.56 (without data augmentation).

1 Introduction

End-to-end (E2E) speech translation (ST), a paradigm that directly maps audio to a foreign text, has been gaining popularity recently (Duong et al., 2016; Béard et al., 2016; Bansal et al., 2018; Di Gangi et al., 2019; Wang et al., 2019). Based on the attentional encoder-decoder framework (Bahdanau et al., 2015), it optimizes model parameters under direct translation supervision. This end-to-end paradigm avoids the problem of error propagation that is inherent in cascade models where an automatic speech recognition (ASR) model and a machine translation (MT) model are chained together. Nonetheless, previous work still reports that E2E ST delivers inferior performance compared to cascade methods (Niehues et al., 2019).

We study one reason for the difficulty of training E2E ST models, namely the uneven spread of information in the speech signal, as visualized in Figure 1, and the consequent difficulty of extracting informative features. Features corresponding to uninformative signals, such as pauses or noise, increase the input length and bring in unmanageable noise for ST. This increases the difficulty of learning (Zhang et al., 2019b; Na et al., 2019) and reduces translation performance.

In this paper, we propose adaptive feature selection (AFS) for ST to explicitly eliminate uninformative features. Figure 2 shows the overall architecture. We employ a pretrained ASR encoder to induce contextual speech features, followed by an ST encoder bridging the gap between speech and translation modalities. AFS is inserted in-between them to select a subset of features for ST encoding (see red rectangles in Figure 1). To ensure that the selected features are well-aligned to transcriptions, we pretrain AFS on ASR. AFS estimates the informativeness of each feature through a parameterized gate, and encourages the dropping of
We compare against fixed-rate feature selection and The E2E ST task offers new opportunities for performance gains, apart from enhancing efficiency and/or interpretability. 

The acoustic input of speech signals involves two dimensions: temporal and feature, where the latter one describes the spectrum extracted from time frames. Accordingly, we adapt $L_0$DROP to sparsify encoder states along temporal and feature dimensions but using different gating networks. In contrast to (Zhang et al., 2020), who focus on efficiency and report a trade-off between sparsity and quality for MT and summarization, we find that sparsity also improves translation quality for ST.

We conduct extensive experiments with Transformer (Vaswani et al., 2017) on LibriSpeech En-Fr and MuST-C speech translation tasks, covering 8 different language pairs. Results show that AFS only retains about 16% of temporal speech features, revealing heavy redundancy in speech encodings and yielding a decoding speedup of $\sim 1.4 \times$. AFS eases model convergence, and improves the translation quality by $\sim 1.3$–$1.6$ BLEU, surpassing several strong baselines. Specifically, without data augmentation, AFS narrows the performance gap against the cascade approach, and outperforms it on LibriSpeech En-Fr by 0.29 BLEU, reaching 18.56. We compare against fixed-rate feature selection and a simple CNN, confirming that our adaptive feature selection offers better translation quality.

Our work demonstrates that E2E ST suffers from redundant speech features, with sparsification bringing significant performance improvements. The E2E ST task offers new opportunities for follow-up research in sparse models to deliver performance gains, apart from enhancing efficiency and/or interpretability.

2 Background: $L_0$DROP

$L_0$DROP provides a selective mechanism for encoder-decoder models which encourages removing uninformative encoder outputs via a sparsity-inducing objective (Zhang et al., 2020). Given a source sequence $X = \{x_1, x_2, \ldots, x_n\}$, $L_0$DROP assigns each encoded source state $x_i \in \mathbb{R}^d$ with a scalar gate $g_i \in [0, 1]$ as follows:

$$L_0\text{DROP}(x_i) = g_i x_i, \quad (1)$$

with $g_i \sim \text{HardConcrete}(\alpha_i, \beta, \epsilon), \quad (2)$

where $\alpha_i, \beta, \epsilon$ are hyperparameters of the hard concrete distribution (HardConcrete) (Louizos et al., 2018).

Note that the hyperparameter $\alpha_i$ is crucial to HardConcrete as it directly governs its shape. We associate $\alpha_i$ with $x_i$ through a gating network:

$$\log \alpha_i = x_i^T \cdot w, \quad (3)$$

Thus, $L_0$DROP can schedule HardConcrete via $\alpha_i$ to put more probability mass at either 0 (i.e. $g_i \to 0$) or 1 (i.e. $g_i \to 1$). $w \in \mathbb{R}^d$ is a trainable parameter. Intuitively, $L_0$DROP controls the openness of gate $g_i$ via $\alpha_i$ so as to determine whether to remove ($g_i = 0$) or retain ($g_i = 1$) the state $x_i$.

$L_0$DROP enforces sparsity by pushing the probability mass of HardConcrete towards 0, according to the following penalty term:

$$L_0(X) = \sum_{i=1}^{n} 1 - p(g_i = 0|\alpha_i, \beta, \epsilon). \quad (4)$$

By sampling $g_i$ with reparameterization (Kingma and Welling, 2013), $L_0$DROP is fully differentiable and optimized with an upper bound on the objective: $L_{\text{MLE}} + \lambda L_0(X)$, where $\lambda$ is a hyperparameter affecting the degree of sparsity – a larger $\lambda$ enforces more gates near 0 – and $L_{\text{MLE}}$ denotes the maximum likelihood loss. An estimation of the expected value of $g_i$ is used during inference. Zhang et al. (2020) applied $L_0$DROP to prune encoder outputs for MT and summarization tasks; we adapt it to E2E ST. Sparse stochastic gates and $L_0$ relaxations were also by Bastings et al. (2019) to construct interpretable classifiers, i.e. models that can reveal which tokens they rely on when making a prediction.

3 Adaptive Feature Selection

One difficulty with applying encoder-decoder models to E2E ST is deciding how to encode speech features (pushing the gate to 0) that contribute little to ASR. An underlying assumption is that features irrelevant for ASR are also unimportant for ST.

We base AFS on $L_0$DROP (Zhang et al., 2020), a sparsity-inducing method for encoder-decoder models, and extend it to sparsify speech features. The acoustic input of speech signals involves two dimensions: temporal and feature, where the latter one describes the spectrum extracted from time frames. Accordingly, we adapt $L_0$DROP to sparsify encoder states along temporal and feature dimensions but using different gating networks. In contrast to (Zhang et al., 2020), who focus on efficiency and report a trade-off between sparsity and quality for MT and summarization, we find that sparsity also improves translation quality for ST.

We conduct extensive experiments with Transformer (Vaswani et al., 2017) on LibriSpeech En-Fr and MuST-C speech translation tasks, covering 8 different language pairs. Results show that AFS only retains about 16% of temporal speech features, revealing heavy redundancy in speech encodings and yielding a decoding speedup of $\sim 1.4 \times$. AFS eases model convergence, and improves the translation quality by $\sim 1.3$–$1.6$ BLEU, surpassing several strong baselines. Specifically, without data augmentation, AFS narrows the performance gap against the cascade approach, and outperforms it on LibriSpeech En-Fr by 0.29 BLEU, reaching 18.56. We compare against fixed-rate feature selection and a simple CNN, confirming that our adaptive feature selection offers better translation quality.

Our work demonstrates that E2E ST suffers from redundant speech features, with sparsification bringing significant performance improvements. The E2E ST task offers new opportunities for follow-up research in sparse models to deliver performance gains, apart from enhancing efficiency and/or interpretability.
signals. In contrast to text where word boundaries can be easily identified, the spectrum features of speech are continuous, varying remarkably across different speakers for the same transcript. In addition, redundant information, like pauses in-between neighbouring words, can be of arbitrary duration at any position as shown in Figure 1, while contributing little to translation. This increases the burden and occupies the capacity of ST encoder, leading to inferior performance (Duong et al., 2016; Bérard et al., 2016). Rather than developing complex encoder architectures, we resort to feature selection to explicitly clear out those uninformative speech features.

Figure 2 gives an overview of our model. We use a pretrained and frozen ASR encoder to extract contextual speech features, and collect the informative ones from them via AFS before transmission to the ST encoder. AFS drops pauses, noise and other uninformative features and retains features that are relevant for ASR. We speculate that these retained features are also the most relevant for ST, and that the sparser representation simplifies the learning problem for ST, for example the learning of attention strength between encoder states and target language (sub)words. Given a training tuple (audio, source transcription, translation), denoted as $(X,Y,Z)$ respectively, we outline the overall framework below, including three steps:

1. Train ASR model with the following objective and model architecture until convergence:
   \[
   \mathcal{L}_{\text{ASR}} = \eta \mathcal{L}_{\text{MLE}}(Y|X) + \gamma \mathcal{L}_{\text{CTC}}(Y,X), \tag{5}
   \]
   \[
   \mathcal{M}_{\text{ASR}} = D_{\text{ASR}}(Y, E_{\text{ASR}}(X)). \tag{6}
   \]

2. Finetune ASR model with AFS for $m$ steps:
   \[
   \mathcal{L}_{\text{AFS}} = \mathcal{L}_{\text{MLE}}(Y|X) + \lambda \mathcal{L}_0(X), \tag{7}
   \]
   \[
   \mathcal{M}_{\text{AFS}} = D_{\text{ASR}}(Y, F(E_{\text{ASR}}(X))). \tag{8}
   \]

3. Train ST model with pretrained and frozen ASR and AFS submodules until convergence:
   \[
   \mathcal{L}_{\text{ST}} = \mathcal{L}_{\text{MLE}}(Z|X), \tag{9}
   \]
   \[
   \mathcal{M}_{\text{ST}} = D_{\text{ST}}(Z, E_{\text{ST}}(F_{\overline{E}}(E_{\text{ASR}}(X)))). \tag{10}
   \]

We handle both ASR and ST as sequence-to-sequence problem with encoder-decoder models. We use $E^x(\cdot)$ and $D^x(\cdot, \cdot)$ to denote the corresponding encoder and decoder respectively. $F(\cdot)$ denotes the AFS approach, and $F_{\overline{E}}$ means freezing the ASR encoder and the AFS module during training. Note that our framework puts no constraint on the architecture of the encoder and decoder in any task, although we adopt the multi-head dot-product attention network (Vaswani et al., 2017) for our experiments.

**ASR Pretraining** The ASR model $\mathcal{M}_{\text{ASR}}$ (Eq. 6) directly maps an audio input to its transcription. To improve speech encoding, we apply logarithmic penalty on attention to enforce short-range dependency (Di Gangi et al., 2019) and use trainable positional embedding with a maximum length of 2048. Apart from $\mathcal{L}_{\text{MLE}}$, we augment the training objective with the connectionist temporal classification (Graves et al., 2006, CTC) loss $\mathcal{L}_{\text{CTC}}$ as in Eq. 5. Note $\gamma = 1 - \eta$. The CTC loss is applied to the encoder outputs, guiding them to align with their corresponding transcription (sub)words and improving the encoder’s robustness (Karita et al., 2019). Following previous work (Karita et al., 2019; Wang et al., 2020), we set $\gamma$ to 0.3.

**AFS Finetuning** This stage aims at using AFS to dynamically pick out the subset of ASR encoder outputs that are most relevant for ASR performance (see red rectangles in Figure 1). We follow Zhang et al. (2020) and place AFS in-between ASR encoder and decoder during finetuning (see $F(\cdot)$ in $\mathcal{M}_{\text{AFS}}$, Eq. 8). We exclude the CTC loss in the training objective (Eq. 7) to relax the alignment constraint and increase the flexibility of feature adaptation. We use $\mathcal{L}_0$ DROP for AFS in two ways.

1. **AFS’** The direct application of $\mathcal{L}_0$ DROP on ASR encoder results in $\text{AFS’}(\cdot)$, sparsifying encodings along the temporal dimension \{$x_i\}_{i=1}^n$:
   \[
   F^t(x_i) = \text{AFS’}(x_i) = g_i^t x_i, \tag{11}
   \]
   with $\log \alpha_i^t = x_i^T \cdot w^t$,
   \[
   g_i^t \sim \text{HardConcrete}(\alpha_i^t, \beta, \epsilon),
   \]
   where $\alpha_i^t$ is a positive scalar powered by a simple linear gating layer, and $w^t \in \mathbb{R}^d$ is a trainable parameter of dimension $d$. $g^t$ is the temporal gate. The sparsity penalty of $\text{AFS’}$ follows Eq. 4:
   \[
   \mathcal{L}_0^t(X) = \sum_{i=1}^n 1 - p(g_i^t = 0|\alpha_i^t, \beta, \epsilon). \tag{12}
   \]
time frames to form the acoustic input, similar to
the word embedding. As each encoded speech feature
contains temporal information, it is reasonable to extend AFS^t to AFS^{t, f}, including sparsification along the feature dimension \{x_{i,j}^t\}_{j=1}^d:

\[
F^{t, f}(x_i) = \text{AFS}^{t, f}(x_i) = g^t_i x_i \odot g^f_i,
\]
with \( \log \alpha^f = w^f \),

\[
g^f_i \sim \text{HardConcrete}(\alpha^f_i, \beta, \epsilon),
\]

where \( \alpha^f \in \mathbb{R}^d \) estimates the weights of each feature, dominated by an input-independent gating model with trainable parameter \( w^f \in \mathbb{R}^d \). \( g^f_i \) is the feature gate. Note that \( \alpha^f \) is shared for all time steps. \( \odot \) denotes element-wise multiplication. AFS^{t, f} reuses \( g^t_i \)-relevant submodules in Eq. 11, and extends the sparsity penalty \( \mathcal{L}_0 \) in Eq. 12 as follows:

\[
\mathcal{L}_0^{t, f}(X) = \mathcal{L}_0^t + \sum_{j=1}^d 1 - p(g^f_j = 0 | \alpha^f_j, \beta, \epsilon).
\]

We perform the finetuning by replacing \((F, \mathcal{L}_0)\) in Eq. (8-7) with either AFS^t \((F^t, \mathcal{L}_0^t)\) or AFS^{t, f} \((F^{t, f}, \mathcal{L}_0^{t, f})\) for extra \( m \) steps. We compare these two variants in our experiments.

**E2E ST Training**

We treat the pretrained ASR and AFS model as a speech feature extractor, and freeze them during ST training. We gather the speech features emitted by the ASR encoder that correspond to \( g^t_i > 0 \), and pass them similarly as done with word embeddings to the ST encoder. We employ sinusoidal positional encoding to distinguish features at different positions. For the input to the ST encoder, our E2E ST follows the standard encoder-decoder translation model \( M^{ST} \) in Eq. 10 and is optimized with \( \mathcal{L}_{\text{MLE}} \) alone as in Eq. 9. Intuitively, AFS bridges the gap between ASR output and MT input by selecting transcript-aligned speech features.

**4 Experiments**

**Datasets and Preprocessing**

We experiment with two benchmarks: the Augmented LibriSpeech dataset (LibriSpeech En-Fr) (Kocabiyikoglu et al., 2018) and the multilingual MuST-C dataset (MuST-C) (Di Gangi et al., 2019). LibriSpeech En-Fr is collected by aligning e-books in French with English utterances of LibriSpeech, further augmented with French translations offered by Google Translate. We use the 100 hours clean training set for training, including 47K utterances to train ASR models and double the size for ST models after concatenation with the Google translations. We report results on the test set (2048 utterances) using models selected on the dev set (1071 utterances). MuST-C is built from English TED talks, covering 8 translation directions: English to German (De), Spanish (Es), French (Fr), Italian (It), Dutch (Nl), Portuguese (Pt), Romanian (Ro) and Russian (Ru). We train ASR and ST models on the given training set, containing \( \sim 452 \) hours with \( \sim 252 \)K utterances on average for each translation pair. We adopt the given dev set for model selection and report results on the common test set, whose size ranges from 2502 (Es) to 2641 (De) utterances.

For all datasets, we extract 40-dimensional log-Mel filterbanks with a step size of 10ms and window size of 25ms as the acoustic features. We expand these features with their first and second-order derivatives, and stabilize them using mean subtraction and variance normalization. We stack the features corresponding to three consecutive frames without overlapping to the left, resulting in the final 360-dimensional acoustic input. For transcriptions and translations, we tokenize and truecase all the text using Moses scripts (Koehn et al., 2007). We train subword models (Sennrich et al., 2016) on each dataset with a joint vocabulary size of 16K to handle rare words, and share the model for ASR, MT and ST. We train all models without removing punctuation.

**Model Settings and Baselines**

We adopt the Transformer architecture (Vaswani et al., 2017) for all tasks, including \( M^{\text{ASR}} \) (Eq. 6), \( M^{\text{AFS}} \) (Eq. 8) and \( M^{\text{ST}} \) (Eq. 10). The encoder and decoder consist of 6 identical layers, each including a self-attention sublayer, a cross-attention sublayer (decoder alone) and a feedforward sublayer. We employ the base setting for experiments: hidden size \( d = 512 \), attention head 8 and feedforward size 2048. We schedule learning rate via Adam \((\beta_1 = 0.9, \beta_2 = 0.98)\) (Kingma and Ba, 2015), paired with a warmup step of 4K. We apply dropout to attention weights and residual connections with a rate of 0.1 and 0.2 respectively, and also add label smoothing of 0.1 to handle overfitting. We train all models with a maximum step size of 30K and a

---

Footnote: 3 Other candidate gating models, like linear mapping upon mean-pooled encoder outputs, delivered worse performance in our preliminary experiments.
minibatch size of around 25K target subwords. We average the last 5 checkpoints for evaluation. We use beam search for decoding, and set the beam size and length penalty to 4 and 0.6, respectively. We set $\epsilon = -0.1$, and $\beta = 2/3$ for AFS following Louizos et al. (2018), and finetune AFS for an additional $m = 5$K steps. We evaluate translation quality with tokenized case-sensitive BLEU (Papineni et al., 2002), and report WER for ASR performance without punctuation.

We compare our models with four baselines:

**ST:** A vanilla Transformer-based E2E ST model of 6 encoder and decoder layers. Logarithmic attention penalty (Di Gangi et al., 2019) is used to improve the encoder.

**ST + ASR-PT:** We perform the ASR pretraining (ASR-PT) for E2E ST. This is the same model as ours (Figure 2) but without AFS finetuning.

**Cascade:** We first transcribe the speech input using an ASR model, and then pass the results on to an MT model. We also use the logarithmic attention penalty (Di Gangi et al., 2019) for the ASR encoder.

**ST + Fixed Rate:** Instead of dynamically selecting features, we replace AFS with subsampling at a fixed rate: we extract the speech encodings after every $k$ positions.

Besides, we offer another baseline, **ST + CNN**, for comparison on MuST-C En-De: we replace the fixed-rate subsampling with a one-layer 1D depth-separable convolution, where the output dimension is set to 512, the kernel size over temporal dimension is set to 5 and the stride is set to 6. In this way, the ASR encoder features will be compressed to around 1/6 features, a similar ratio to the fixed-rate subsampling.

### 4.1 Results on MuST-C En-De

We perform a thorough study on MuST-C En-De. With AFS, the first question is its feasibility. We start by analyzing the degree of sparsity in speech features (i.e. sparsity rate) yielded by AFS, focusing on the temporal sparsity rate $\#(s_f^{t}=0)/n$ and the feature sparsity rate $\#(s_f^{t}=0)/d$. To obtain different rates, we vary the hyperparameter $\lambda$ in Eq. 7 in a range of $[0.1, 0.8]$ with a step size 0.1.

Results in Figure 3 show that large amounts of encoded speech features ($> 59\%$) can be easily pruned out, revealing heavy inner-speech redundancy. Both AFS$^t$ and AFS$^{t,f}$ drop $\sim 60\%$ temporal features with $\lambda$ of 0.1, and this number increases to $> 85\%$ when $\lambda \geq 0.5$ (Figure 3b), remarkably surpassing the sparsity rate reported by Zhang et al. (2020) on text summarization (71.5%). In contrast to rich temporal sparsification, we get a feature sparsity rate of 0, regardless of $\lambda$’s value, although increasing $\lambda$ decreases $g^t$ (Figure 3a). This suggests that selecting neurons from the feature dimension is harder. Rather than filtering neurons, the feature gate $g^t$ acts more like a weighting mechanism on them. In the rest of the paper, we use **sparsity rate** for the temporal sparsity rate.

We continue to explore the impact of varied sparsity rates on the ASR and ST performance. Figure 4 shows their correlation. We observe that AFS slightly degenerates ASR accuracy (Figure 4a), but still retains $\sim 95\%$ accuracy on average: AFS$^{t,f}$ often performs better than AFS$^t$ with similar sparsity rate. The fact that only 15% speech features successfully support 95% ASR accuracy proves the informativeness of these selected features. These findings echo with (Zhang et al., 2020), where they observe a trade-off between sparsity and quality.

However, when AFS is applied to ST, we find consistent improvements to translation quality by $> 0.8$ BLEU, shown in Figure 4b. Translation quality on the development set peaks at 22.17 BLEU.
We set $\lambda$ with ASR-PT as our real baseline. We observe improvement over the baseline, although CNN offers more flexibility in feature manipulation. By contrast to fixed-rate subsampling, the proposed AFS is data-driven, shifting the decision burden to the data and model themselves. As a result, AFS$^\lambda$ and AFS$^{t,f}$ surpass ASR-PT by 0.9 BLEU and 1.71 BLEU, respectively, substantially narrowing the performance gap compared to the cascade baseline (~0.14 BLEU).

We also observe improved decoding speed: AFS runs ~1.37× faster than ASR-PT. Compared to the fixed-rate subsampling, AFS is slightly slower which we ascribe to the overhead introduced by the gating module. Surprisingly, Table 1 shows that the vanilla ST runs slower than ASR-PT (0.87×) while the cascade model is slightly faster (1.06×). By digging into the beam search algorithm, we discover that ASR pretraining shortens the number of steps in beam-decoding: 94 ASR-PT vs. 112 vanilla ST (on average). The speedup brought by cascading is due to the smaller English vocabulary size compared to the German vocabulary when processing audio inputs.

### 4.2 Why (Adaptive) Feature Selection?

Apart from the benefits in translation quality, we go deeper to study other potential impacts of (adaptive) feature selection. We begin with inspecting training curves. Figure 6 shows that ASR pretraining improves model convergence; feature selection makes training more stable. Compared to other models, the curve of ST with AFS is much smoother, suggesting its better regularization effect.

We then investigate the effect of training data size, and show the results in Figure 7. Overall, we do not observe higher data efficiency by feature selection on low-resource settings. But instead, our results suggest that feature selection delivers larger performance improvement when more training data is available. With respect to data efficiency, ASR pretraining seems to be more important (Figure 7, left) (Bansal et al., 2019; Stoian et al., 2020). Com-
### Position Analysis

Figure 7: BLEU as a function of training data size on MuST-C En-De. We split the original training data into non-overlapped five subsets, and train different models with accumulated subsets. Results are reported on the test set. Note that we perform ASR pretraining on the original dataset. \( \lambda = 0.5, k = 6 \).

Figure 8: Histogram of the cross-attention weights received per ST encoder output on MuST-C En-De test set. For each instance, we collect attention weights averaged over different heads and decoder layers following Zhang et al. (2020). Larger weight indicates stronger impact of the encoder output on translation. Feature selection biases the distribution towards larger weights. \( \lambda = 0.5, k = 6 \).

Figure 9: The number of selected features vs. word duration (left) and position (right) on MuST-C En-De test set. For word duration, we align the audio and its transcription by Montreal Forced Aligner (McAuliffe et al., 2017), and collect each words’ duration and its corresponding retained feature number. For position, we uniformly split each input into 50 pieces, and count the average number of retained features in each piece. \( \lambda = 0.5, k = 6 \).

Figure 10: Illustration of feature gate \( g^f \) with \( \lambda = 0.5 \).

words with longer duration correspond to more speech features. However, when it comes to temporal position, Figure 9b illustrates their difference: fixed-rate subsampling is context-independent, periodically picking up features; while AFS decides feature selection based on context information. The curve of AFS is more smooth, indicating that features kept by AFS are more uniformly distributed across different positions, ensuring the features’ informativeness.

### AFS vs. Fixed Rate

We compare these two approaches by analyzing the number of retained features with respect to word duration and temporal position. Results in Figure 9a show that the underlying pattern behind these two methods is similar: compared to AFS, the fixed-rate subsampling suffers more from small-scale training: it yields worse performance than ASR-PT when data size \( \leq 100K \), highlighting better generalization of AFS.

In addition to model performance, we also look into the ST model itself, and focus on the cross-attention weights. Figure 8 visualize the attention value distribution, where ST models with feature selection noticeably shift the distribution towards larger weights. This suggests that each ST encoder output exerts greater influence on the translation. By removing redundant and noisy speech features, feature selection eases the learning of the ST encoder, and also enhances its connection strength with the ST decoder. This helps bridge the modality gap between speech and text translation. Although fixed-rate subsampling also delivers a distribution shift similar to AFS, its inferior ST performance compared to AFS corroborates the better quality of adaptively selected features.

### AFS vs. AFS\(^t\)/AFS\(^{t,f}\)

Their only difference lies at the feature gate \( g^f \). We visualize this gate in Figure 10. Although this gate induces no sparsification, it offers AFS\(^{t,f}\) the capability of adjusting the weight of each neuron. In other words, AFS\(^{t,f}\) has more freedom in manipulating speech features.

### 4.3 Results on MuST-C and LibriSpeech

Table 2 and Table 3 list the results on MuST-C and LibriSpeech En-Fr, respectively. Over all tasks, AFS\(^{t,f}\) substantially outperforms ASR-PT by 1.34/1.60 average BLEU, pruning out 84.5\% temporal speech features on average and yielding an average decoding speedup of 1.45 \( \times \). Our model narrows the gap against the cascade model to -0.8 average BLEU, where AFS surpasses Cascade on LibriSpeech En-Fr, without using KD (Liu et al.,...
<table>
<thead>
<tr>
<th>Metric</th>
<th>Model</th>
<th>De</th>
<th>Es</th>
<th>Fr</th>
<th>It</th>
<th>Nl</th>
<th>Pt</th>
<th>Ro</th>
<th>Ru</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU↑</td>
<td>Di Gangi et al. (2019)</td>
<td>17.30</td>
<td>20.80</td>
<td>26.90</td>
<td>16.80</td>
<td>18.80</td>
<td>20.10</td>
<td>16.50</td>
<td>10.50</td>
</tr>
<tr>
<td>ST</td>
<td>17.44</td>
<td>23.85</td>
<td>28.43</td>
<td>19.54</td>
<td>21.23</td>
<td>22.55</td>
<td>17.66</td>
<td>12.10</td>
<td></td>
</tr>
<tr>
<td>ST + ASR-PT</td>
<td>20.67</td>
<td>25.96</td>
<td>32.24</td>
<td>20.84</td>
<td>23.27</td>
<td>24.83</td>
<td>19.94</td>
<td>13.96</td>
<td></td>
</tr>
<tr>
<td>Cascade</td>
<td>22.52</td>
<td>27.92</td>
<td>34.53</td>
<td>24.02</td>
<td>26.74</td>
<td>27.57</td>
<td>22.61</td>
<td>16.13</td>
<td></td>
</tr>
<tr>
<td>ST + AFS</td>
<td>21.57</td>
<td>26.78</td>
<td>33.34</td>
<td>22.60</td>
<td>24.68</td>
<td>25.90</td>
<td>20.80</td>
<td>14.90</td>
<td></td>
</tr>
<tr>
<td>ST + AFS</td>
<td>22.38</td>
<td>27.04</td>
<td>33.43</td>
<td>23.35</td>
<td>25.05</td>
<td>26.34</td>
<td>21.08</td>
<td>14.70</td>
<td></td>
</tr>
<tr>
<td>TemporalST + AFS</td>
<td>84.4%</td>
<td>84.5%</td>
<td>83.2%</td>
<td>84.9%</td>
<td>84.4%</td>
<td>84.7%</td>
<td>84.8%</td>
<td>84.7%</td>
<td></td>
</tr>
<tr>
<td>Sparsity RateST + AFS</td>
<td>85.1%</td>
<td>84.5%</td>
<td>84.7%</td>
<td>83.9%</td>
<td>83.5%</td>
<td>85.1%</td>
<td>84.8%</td>
<td>84.7%</td>
<td></td>
</tr>
<tr>
<td>Speedup↑</td>
<td>ST + AFS</td>
<td>1.38×</td>
<td>1.35×</td>
<td>1.50×</td>
<td>1.34×</td>
<td>1.54×</td>
<td>1.43×</td>
<td>1.59×</td>
<td>1.31×</td>
</tr>
<tr>
<td></td>
<td>ST + AFS</td>
<td>1.37×</td>
<td>1.34×</td>
<td>1.50×</td>
<td>1.39×</td>
<td>1.42×</td>
<td>1.26×</td>
<td>1.46×</td>
<td>1.37×</td>
</tr>
</tbody>
</table>

Table 2: Performance over 8 languages on MuST-C dataset. *: results reported by the ESPNet toolkit (Watanabe et al., 2018), where the hyperparameters of beam search are tuned for each dataset.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Model</th>
<th>En-Fr</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU↑</td>
<td>Bérard et al. (2018)</td>
<td>13.40</td>
</tr>
<tr>
<td></td>
<td>Watanabe et al. (2018)</td>
<td>16.68</td>
</tr>
<tr>
<td></td>
<td>Liu et al. (2019a)</td>
<td>17.02</td>
</tr>
<tr>
<td></td>
<td>Wang et al. (2019)</td>
<td>17.05</td>
</tr>
<tr>
<td></td>
<td>Wang et al. (2020)</td>
<td>17.66</td>
</tr>
<tr>
<td>ST</td>
<td>14.32</td>
<td></td>
</tr>
<tr>
<td>ST + ASR-PT</td>
<td>17.05</td>
<td></td>
</tr>
<tr>
<td>Cascade</td>
<td>18.27</td>
<td></td>
</tr>
<tr>
<td>ST + AFS</td>
<td>18.33</td>
<td></td>
</tr>
<tr>
<td>ST + AFS</td>
<td>18.56</td>
<td></td>
</tr>
<tr>
<td>TemporalST + AFS</td>
<td>84.7%</td>
<td></td>
</tr>
<tr>
<td>Sparsity RateST + AFS</td>
<td>83.5%</td>
<td></td>
</tr>
<tr>
<td>Speedup↑</td>
<td>ST + AFS</td>
<td>1.84×</td>
</tr>
<tr>
<td></td>
<td>ST + AFS</td>
<td>1.78×</td>
</tr>
</tbody>
</table>

Table 3: Performance on LibriSpeech En-Fr.

Speech Translation Pioneering studies on ST used a cascade of separately trained ASR and MT systems (Ney, 1999). Despite its simplicity, this approach inevitably suffers from mistakes made by ASR models, and is error prone. Research in this direction often focuses on strategies capable of mitigating the mismatch between ASR output and MT input, such as representing ASR outputs with lattices (Saleem et al., 2004; Mathias and Byrne, 2006; Zhang et al., 2019a; Beck et al., 2019), injecting synthetic ASR errors for robust MT (Tsvelkov et al., 2014; Cheng et al., 2018) and differentiable cascade modeling (Kano et al., 2017; Anastasopoulos and Chiang, 2018; Sperber et al., 2019).

In contrast to cascading, another option is to perform direct speech-to-text translation. Duong et al. (2016) and Bérard et al. (2016) employ the attentional encoder-decoder model (Bahdanau et al., 2015) for E2E ST without accessing any intermediate transcriptions. E2E ST opens the way to bridging the modality gap directly, but it is data-hungry, sample-inefficient and often underperforms cascade models especially in low-resource settings (Bansal et al., 2018). This led researchers to explore solutions ranging from efficient neural architecture design (Karita et al., 2019; Di Gangi et al., 2019; Sung et al., 2019) to extra training signal incorporation, including multi-task learning (Weiss et al., 2017; Liu et al., 2019b), submodule pretraining (Bansal et al., 2019; Stoian et al., 2020; Wang et al., 2020), knowledge distillation (Liu et al., 2019a), meta-learning (Indurthi et al., 2019) and data augmentation (Kocabiyikoglu et al., 2018; Jia et al., 2019; Pino et al., 2019). Our work focuses on E2E ST, but we investigate feature selection which has rarely been studied before.

Speech Feature Selection Encoding speech signals is challenging as acoustic input is lengthy, noisy and redundant. To ease model learning, previous work often selected features via downsampling techniques, such as convolutional modeling (Di
Gangi et al., 2019) and fixed-rate subsampling (Lu et al., 2015). Recently, Zhang et al. (2019b) and Na et al. (2019) proposed dynamic subsampling for ASR which learns to skip uninformative features during recurrent encoding. Unfortunately, their methods are deeply embedded into recurrent networks, hard to adapt to other architectures like Transformer (Vaswani et al., 2017). Recently, Salesky et al. (2020) have explored phoneme-level representations for E2E ST, but this requires non-trivial phoneme recognition and alignment.

Instead, we resort to sparsification techniques which have achieved great success in NLP tasks recently (Correia et al., 2019; Child et al., 2019; Zhang et al., 2020). In particular, we employ $L_0$Drop (Zhang et al., 2020) for AFS to dynamically retain informative speech features, which is fully differentiable and independent of concrete encoder/decoder architectures. We extend $L_0$Drop by handling both temporal and feature dimensions with different gating networks, and apply it to E2E ST.

6 Conclusion and Future Work

In this paper, we propose adaptive feature selection for E2E ST to handle redundant and noisy speech signals. We insert AFS in-between the ST encoder and a pretrained, frozen ASR encoder to filter out uninformative features contributing little to ASR. We base AFS on $L_0$Drop (Zhang et al., 2020), and extend it to modeling both temporal and feature dimensions. Results show that AFS improves translation quality and accelerates decoding by $\sim1.4\times$ with an average temporal sparsity rate of $\sim84\%$. AFS successfully narrows or even closes the performance gap compared to cascading models.

While most previous work on sparsity in NLP demonstrates its benefits from efficiency and/or interpretability perspectives (Zhang et al., 2020), we show that sparsification in our scenario – E2E ST – leads to substantial performance gains.

In the future, we will work on adapting AFS to simultaneous speech translation.

Acknowledgments

We would like to thank Shucong Zhang for his great support on building our ASR baselines. IT acknowledges support of the European Research Council (ERC Starting grant 678254) and the Dutch National Science Foundation (NWO VIDI 639.022.518). This work has received funding from the European Union’s Horizon 2020 Research and Innovation Programme under Grant Agreement No 825460 (ELITR). Rico Sennrich acknowledges support of the Swiss National Science Foundation (MUTAMUR; no. 176727).

References


