Low-Resource Speech-to-Text Translation

Sameer Bansal¹, Herman Kamper², Karen Livescu³, Adam Lopez¹, Sharon Goldwater¹

¹School of Informatics, University of Edinburgh, UK
²Stellenbosch University, South Africa
³Toyota Technological Institute at Chicago, USA

{sameer.bansal, sgwater, alopez}@inf.ed.ac.uk, kamperh@sun.ac.za, klivescu@ttic.edu

Abstract

Speech-to-text translation has many potential applications for low-resource languages, but the typical approach of cascading speech recognition with machine translation is often impossible, since the transcripts needed to train a speech recognizer are usually not available for low-resource languages. Recent work has found that neural encoder-decoder models can learn to directly translate foreign speech in high-resource scenarios, without the need for intermediate transcription. We investigate whether this approach also works in settings where both data and computation are limited. To make the approach efficient, we make several architectural changes, including a change from character-level to word-level decoding. We find that this choice yields crucial speed improvements that allow us to train with fewer computational resources, yet still performs well on frequent words. We explore models trained on between 20 and 160 hours of data, and find that although models trained on less data have considerably lower BLEU scores, they can still predict words with relatively high precision and recall—around 50% for a model trained on 50 hours of data, versus around 60% for the full 160 hour model. Thus, they may still be useful for some low-resource scenarios.

Index Terms: speech translation, low-resource speech processing, encoder-decoder models

1. Introduction

Conventional systems for speech-to-text translation [1] typically pipeline automatic speech recognition and machine translation, and since both of these applications require large training sets, these systems are available for only a tiny fraction of the world’s highest-resource languages. But speech-to-text translation could be especially valuable in low-resource scenarios, for example in documentation of unwritten or endangered languages [2–6]; or in crisis relief, where emergency workers might need to respond to calls or requests in a foreign language [7]. These applications have motivated recent research on low-resource speech translation trained on a (potentially) cheaper resource: speech paired with its translation, with no intermediate transcriptions.¹

¹There is also recent work [8–10] using multitask learning to learn both translation and transcription models, showing improvements on the individual tasks. We focus here on the scenario without transcriptions.

Initial work studied speech-to-text alignment without translation [11, 12], or focused on translating a few keywords using heuristic methods with just a few hours of training data [13, 14].

Recently, recurrent encoder-decoder models have been used to develop end-to-end speech-to-text translation models, which have been tested in high-resource settings on synthesized speech [15], audiobooks [10, 16], and a large dataset of conversational telephone speech [9]. So far, these neural models have been shown to produce high-quality translations with substantial resources—typically more than 100 hours of translated audio, from which models are trained for many days on multiple GPUs. But in our scenarios of interest, we expect to have less data, less time, and fewer computational resources. How will neural models perform in such low-resource settings?

In this paper we perform an extensive investigation of the effects of training end-to-end speech-to-text translation models with limited resources. We implement a model inspired by the state-of-the-art architecture of Weiss et al. [9], but modify it to permit training in reasonable time on a single GPU (§2). The biggest change, compared to [9] (and also [8, 10]), is to use word-level decoding instead of character-level. We show that word-level models can be trained much faster than character-level models and still obtain comparable precision and recall; the tradeoff for this speed improvement is that they struggle to translate infrequent word types, leading to a drop in overall accuracy as measured by BLEU and METEOR.

We investigate the model’s performance as we gradually reduce training data from the full 160 hours to as little as 20 hours. With only 50 hours of training data, our model still produces accurate translations for short utterances, with precision and recall around 50%. Although translation quality is much worse with 20 hours of data, precision and recall are still around 30%, which may be useful in low-resource applications. The 50-hour model trains in less than three days on a single GPU.

2. Speech-to-Text Model

Following Weiss et al. [9], we combine convolutional neural network (CNN) and recurrent neural network (RNN) components to build an encoder-decoder model with attention, but we modify the system (Table 1) so that we can train even our larger models in 3-5 days on a single GPU.

2.1. Speech encoder

Raw speech input is converted to Mel filterbank features computed every 10ms. Weiss et al.’s [9] model uses 240-dimensional input speech features, consisting of 80 filters stacked with
We use the Fisher Spanish speech dataset [25]: a multispeaker dataset with about 4.5 hours of speech split into 4K utterances each. We train using a single model copy. Although Weiss et al.'s model [9] is trained using asynchronous stochastic gradient descent (ASGD) across 10 replicas; we train using teacher-forcing [23] ratio of 0.8. Our code is implemented in Chainer [24].

2.3. Optimizer
We use cross-entropy as the loss function. We regularize with dropout [21], with a ratio of 0.5 over the embedding and the LSTM layers [22], and an L2 penalty of 0.0001. We use a teacher-forcing [23] ratio of 0.8. Our code is implemented in Chainer [24]. Weiss et al.'s [9] model is trained using asynchronous stochastic gradient descent (ASGD) across 10 replicas; we train using a single model copy. Although Weiss et al.'s model benefited greatly from adding Gaussian noise to the weights during training (personal communication), we were unable to replicate this benefit and did not use noise injection.

3. Experiments
3.1. Experimental setup
We use the Fisher Spanish speech dataset [25]: a multispeaker corpus of telephone calls in a variety of Spanish dialects recorded in realistic noise conditions. The English translations were collected through crowdsourcing, as described in [26], and are used to train all models. There are four English references per utterance for the development and test sets, and one per utterance for the training set. We only use one of the two development sets (dev, not dev2). The training set comprises 160 hours of speech, split into 140K utterances; the development and test sets have about 4.5 hours of speech split into 4K utterances each.

3.2. Evaluation
In order to understand different aspects of model behavior, we evaluate with several metrics: BLEU [28], METEOR [29], and unigram precision/recall on the Fisher dev set, using all 4 human references. BLEU measures how well a predicted translation matches a set of references based on a modified n-gram precision; it does not compute recall, and instead uses a brevity penalty to account for mismatch in reference and predicted lengths. METEOR computes both precision and recall and combines them via a harmonic mean, with greater weight given to recall. The final score is computed using a set of parameters, tuned for individual languages to correlate well with human judgments.

We lower-case and remove punctuation from the English translations and tokenize the text using NLTK [27]. This gives about 17K training word types and 1.5M tokens. There are about 300 out-of-vocabulary (OOV) word types (400 tokens) in the dev set, out of 40K tokens.

We first train a model using the entire 160 hours of labeled training data. To understand the impact of training data size on translation quality, we further train models using smaller subsets: 80, 50, 25, 20 hours of data, selected at random, from the entire training data. We use the same set of hyper-parameters—tuned for 160hrs—for all models. With these model parameters and training setup, we are able to train an epoch—a complete pass through the entire 160 hours of training data—in about 2 hours on a single Titan X (or equivalent) GPU.

3.3. Results and discussion
Figures 1 and 2 show the BLEU, METEOR, and precision/recall scores on the dev set for each model as we change the amount of training data. Table 2 shows the BLEU scores on the test set.

---

2 Code available at: https://github.com/0xSameer/speech2text/tree/seq2seq

3 https://www.nltk.org/api/nltk.tokenize.html#nltk.tokenize.word_tokenize

4 BLEU and precision are computed using multi-bleu.pl from the MOSES toolkit [30], which computes 4-gram BLEU.

---
Word vs. character models. Weiss et al.’s [9] character-level model achieves close to human performance. Our word-level model, trained on 160 hours of data, converges to a BLEU score of 29.5 in about five days, a much lower score than Weiss et al.’s. One reason for this discrepancy is our different architecture and training setup (§2 and Table 1), which allowed us to train our models on our available computing resources.

Training the character-level model takes nearly twice as long per epoch (4 hours for 160hrs of data) as the word-level model (2 hours). Our character-level model also has much smaller performance gains per epoch. To speed up character-level model training, we truncate utterances longer than 15 seconds (20 seconds for word models), reducing training time to 3 hours per epoch. Figure 3 compares character-level models trained in this way to word-level models for two training set sizes.

One of the main benefits of character models is their ability to gracefully handle OOV or infrequent words. On the dev set, the Weiss et al. model predicts about 130 word types that have 25–100 tokens; our models predict about 150 word types for the 160hrs model and Weiss et al.’s [9] model, for words of different frequencies. We only consider content word types—words that are more than five characters long and are not in the NLTK stopword lists. The word-level model recall drops rapidly for medium frequency words, and for rare word types it has almost 0% recall. From this, we see that the main benefit of the character-level model is in handling of rare words, rather than previously unseen words.

Training considerations. Regularization parameters are critical to model performance. Figure 4 shows that increasing the L2 weight decay to a rate of $10^{-4}$ from $10^{-6}$ improved BLEU by about 2 points. Even though we use a high L2 penalty and dropout ratio, the models can overfit; the training loss continues to decrease, and we use early stopping based on dev set BLEU.

We also tried using batch normalization [31] at each CNN layer, and layer normalization [32] at each LSTM layer; but these did not have any noticeable impact on training performance.

Other design considerations. Using a bi-directional LSTM encoder does not seem to have an effect for the largest training sets, but improves BLEU by about one point in the ≤50 hour cases.
predicted words do not carry much of the meaning.

**Neural models in the extreme low-data setting.** One may wonder whether, for very low-data settings, neural models still outperform older non-neural models at all. While we did not directly compare to a non-neural model, one indication is that, in the lowest-data settings, BLEU and METEOR are much worse but unigram precision and recall are still in the 25-35% range. These results compare very favorably to the 2-10% precision/recall reported by [14], who used a heuristic speech-to-text translation system trained on the CALLHOME corpus (also about 20 hours of Spanish conversational telephone speech with English text translations). So, it appears that even in a very low-resource setting with a model that is not state-of-the-art, the neural approach significantly outperforms previous non-neural models.

## 4. Conclusion

We performed a thorough analysis of a neural end-to-end speech-to-text translation model, with a specific focus on how performance is affected when using limited computational resources and limited amounts of data. We also showed the effects of a number of architectural design decisions using several performance metrics. While word-level models fall behind previously proposed character-level models when trained on around 160 hours of translated speech, our word-level models can be trained much faster and give reasonable performance on smaller training sets. Although translation quality drops, models trained on only 20 hours of translated speech achieve precision and recall of around 30% for content words. This could still be useful in search applications in severely low-resource settings. We believe that our extensive analyses in this work will contribute to better decision-making for architectural choices in computation- and data-limited settings.

In future work we aim to consider sub-word modelling [33], which could balance the trade-off between training costs and translation performance. In addition, we plan to try speech features that are targeted to low-resource multi-speaker settings [34, 35] and speaker normalization [36].

## 5. Acknowledgements

We thank Ron Weiss and Jan Chorowski for sharing their translation output, and Kenneth Heafield for giving access to GPUs. This work was supported in part by a James S. McDonnell Foundation Scholar Award and a Google faculty research award.
6. References