Interspeech 2018
2-6 September 2018, Hyderabad

Leveraging translations for speech transcription in low-resource settings

Antonios Anastasopoulos, David Chiang

Department of Computer Science and Engineering
University of Notre Dame, IN, USA
aanastas@nd.edu, dchiang@nd.edu

Abstract

Recently proposed data collection frameworks for endangered language documentation aim not only to collect speech in the language of interest, but also to collect translations into a high-resource language that will render the collected resource interpretable. We focus on this scenario and explore whether we can improve transcription quality under these extremely low-resource settings with the assistance of text translations. We present a neural multi-source model and evaluate several variations of it on three low-resource datasets. We find that our multi-source model with shared attention outperforms the baselines, reducing transcription character error rate by up to 12.3%.
Index Terms: neural multi-source models, speech transcription, endangered languages

1. Introduction

For many low-resource and endangered languages, speech data is easier to obtain than textual data. Oral tradition has historically been the main medium for passing cultural knowledge from one generation to the next, and an estimated 43% of the world's languages is still unwritten [1]. Traditionally, documentary records of endangered languages are created by highly trained linguists in the field. However, modern technology has the potential to enable creation of much larger-scale resources. Recently proposed frameworks [2] propose collection of bilingual audio, rendering the resource interpretable through translations. New technologies have been developed to facilitate collection of spoken translations [3] along with speech in an endangered language, and there already exist recent examples of parallel speech collection efforts focused on endangered languages [4, 5]. The translation is usually in a high-resource language that functions as a *lingua franca* of the area, as it is common for members of an endangered-language community to be bilingual.

Since speech transcription is a costly and slow process, automatically producing transcriptions has the potential to significantly speed up documentation. We focus on this language documentation scenario and explore methods that learn from a small number of transcribed speech utterances along with their translations. We use the neural attentional model [6] and experiment with extensions that take both speech utterances and their translations as input sources. We assume that the translations are in a high-resource language that can be automatically transcribed; therefore, in our experiments, the translation input is text instead of speech. We also explore different parameter-sharing methods across the attention mechanisms.

We experiment on three diverse low-resource language pairs. One is Ainu, a severely endangered language, with translations in English. We also experiment on a recently collected speech corpus of Mboshi [7], with translations in French. Lastly, we evaluate our models on Spanish-English, using the CALLHOME dataset.

Our proposed multi-source model that employs a shared attention mechanism outperforms the baselines in almost all cases. In Mboshi, we find that our model reduces character error rates (CER) by 1.2 points. In Spanish, we observe a reduction of 4.6 points in CER over the strongest baseline, and more than 14.4 points over a speech-only baseline. In Ainu, although our multi-source model doesn't reduce the overall CER, we show that it actually is beneficial in the cases where the single-source speech transcription model has greatest difficulty.

This work was generously supported by NSF Award 1464553.

2. Model

Our models are based on a sequence-to-sequence model with attention [6]. In general, this type of model is composed of three parts: a recurrent encoder, the attention, and a recurrent decoder (see Figure 1a).[1]

Let $\mathbf{X^1} = \mathbf{x}^1_1 \dots \mathbf{x}^1_N$ be a sequence of speech frames, $\mathbf{X}^2 = \mathbf{x}^2_1 \dots \mathbf{x}^2_M$ a sequence of translation characters, and $\mathbf{Y} = \mathbf{y}_1 \dots \mathbf{y}_K$ be a sequence of the target characters of the transcription. A *single-source* speech recognition model attempts to model $P(\mathbf{Y} \mid \mathbf{X}^1)$, while a *single-source* translation model would model $P(\mathbf{Y} \mid \mathbf{X}^2)$.

A *multi-source* model can jointly model $P(\mathbf{Y} \mid \mathbf{X}^1, \mathbf{X}^2)$, and thus we need two encoders (see Figure 1b). One encoder transforms the input sequence of speech frames $\mathbf{x}^1_1 \dots \mathbf{x}^1_N$ into a sequence of input states $\mathbf{h}^1_1 \dots \mathbf{h}^1_N$:

$$\mathbf{h}^1_n = \mathrm{enc}^1(\mathbf{h}^1_{n-1}, \mathbf{x}^1_n) \tag{1}$$

and the second encoder transforms the translation character sequence $\mathbf{x}^2_1 \dots \mathbf{x}^2_M$ into another sequence of input states $\mathbf{h}^2_1 \dots \mathbf{h}^2_M$:

$$\mathbf{h}^2_m = \mathrm{enc}^2(\mathbf{h}^2_{m-1}, \mathbf{x}^2_m). \tag{2}$$

An attention mechanism transforms the two sequences of input states into a sequence of *concatenated context vectors* via two matrices of *attention weights*:

$$\mathbf{c}_k = \left[\textstyle\sum_n \alpha^1_{kn}\mathbf{h}^1_n \quad \sum_m \alpha^2_{km}\mathbf{h}^2_m\right]. \tag{3}$$

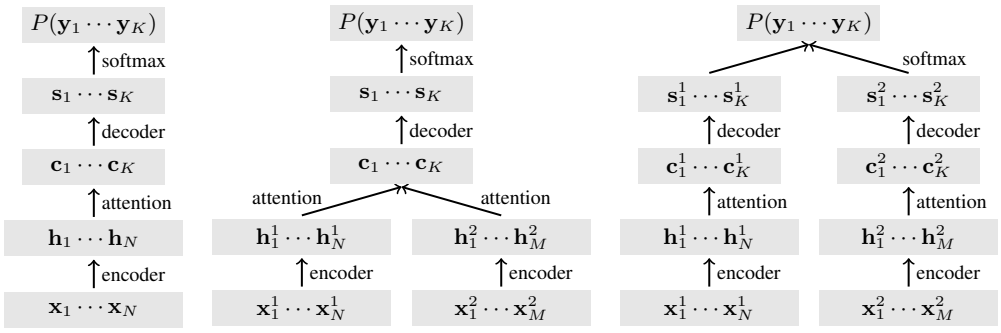
Finally, the decoder computes a sequence of *output states* from which a probability distribution over output characters can be computed:

$$\mathbf{s}_k = \mathrm{dec}(\mathbf{s}_{k-1}, \mathbf{c}_k, \mathbf{y}_{k-1}) \tag{4}$$

$$P(\mathbf{y}_k) = \mathrm{softmax}(\mathbf{s}_k). \tag{5}$$

[1] For simplicity, we have assumed only a single layer for both the encoder and decoder. It is possible to use multiple stacked RNNs; typically, the output of the encoder(s) and decoder(s) ($\mathbf{c}_n$ and $P(\mathbf{y}_k)$, respectively) would be computed from the top layer only.

10.21437/Interspeech.2018-2162



(a) single-source (b) multi-source (c) coupled ensemble

Figure 1: *Source-side variations on the standard attentional model. In the standard* single-source *model, the decoder attends to a single encoder's states. In our proposed* multisource *setup, we have two input sequences encoded by two different encoders, and attention mechanisms provide two context to the decoder. Note that for clarity's sake there are dependencies not shown.*

The attention mechanisms produce the attention weights with the following computations, as in [8], with $\mathbf{v}^1$, $\mathbf{v}^2$, $\mathbf{W}^s_{\alpha^1}$, $\mathbf{W}^s_{\alpha^2}$, $\mathbf{W}^h_{\alpha^1}$, and $\mathbf{W}^h_{\alpha^2}$ being parameters to be learnt:

$$\alpha^1_{kn} = \text{softmax}(\mathbf{v}^1 \tanh(\left[\mathbf{W}^s_{\alpha^1}\mathbf{s}_{k-1}; \mathbf{W}^h_{\alpha^1}\mathbf{h}^1_n\right])) \quad (6)$$

$$\alpha^2_{km} = \text{softmax}(\mathbf{v}^2 \tanh(\left[\mathbf{W}^s_{\alpha^2}\mathbf{s}_{k-1}; \mathbf{W}^h_{\alpha^2}\mathbf{h}^2_m\right])). \quad (7)$$

Since both attention mechanisms provide context to the same decoder, we can tie the computation of the weights so that the two mechanisms share the $\mathbf{v}$ and $\mathbf{W}^s_\alpha$ parameters . We refer to this version as *tied* attention mechanism:

$$\alpha^1_{kn} = \text{softmax}(\mathbf{v} \tanh(\left[\mathbf{W}^s_{\alpha}\mathbf{s}_{k-1}; \mathbf{W}^h_{\alpha^1}\mathbf{h}^1_n\right])) \quad (8)$$

$$\alpha^2_{km} = \text{softmax}(\mathbf{v} \tanh(\left[\mathbf{W}^s_{\alpha}\mathbf{s}_{k-1}; \mathbf{W}^h_{\alpha^2}\mathbf{h}^2_m\right])). \quad (9)$$

If the two encoders share the same output size for their $\mathbf{h}^1$ and $\mathbf{h}^2$ vectors, then the two attentions could further share the $\mathbf{W}^h_\alpha$ parameters. This effectively merges them into one, *shared* attention mechanism, so that:

$$\alpha^1_{kn} = \text{softmax}(\mathbf{v} \tanh(\left[\mathbf{W}^s_{\alpha}\mathbf{s}_{k-1}; \mathbf{W}^h_{\alpha}\mathbf{h}^1_n\right])) \quad (10)$$

$$\alpha^2_{km} = \text{softmax}(\mathbf{v} \tanh(\left[\mathbf{W}^s_{\alpha}\mathbf{s}_{k-1}; \mathbf{W}^h_{\alpha}\mathbf{h}^2_m\right])). \quad (11)$$

Furthermore, another line of work that is pertinent to our work is based in *ensembling*. Traditionally, ensembles refer to models that have been trained on similar data for the similar task but their predictions are combined at inference time, usually achieving better performance than a single model.

In our case, we explore ensembles of a transcription and a translation model. In traditional ensembling, the models are trained separately. Recently, however, *coupled ensembles* were shown to outperform simple ensembles [9]. In the *coupled ensemble* setting (see Figure 1c), the two models are trained jointly, albeit they don't share any parameters.

The two decoder outputs are averaged right before the softmax layer, in order to produce a single output probability distribution. It was shown [9] that this approach works better than combining the two predictions after the softmax layer:

$$\mathbf{s}^1_k = \text{dec}^1(\mathbf{s}^1_{k-1}, \mathbf{c}^1_k, \mathbf{y}_{k-1}) \quad (12)$$

$$\mathbf{s}^2_k = \text{dec}^2(\mathbf{s}^2_{k-1}, \mathbf{c}^2_k, \mathbf{y}_{k-1}) \quad (13)$$

$$P(\mathbf{y}_k) = \text{softmax}(\frac{\mathbf{s}^1_k + \mathbf{s}^2_k}{2}). \quad (14)$$

3. Related Work

The *speech translation* problem has been traditionally approached by feeding the output of a speech recognition system into a Machine Translation (MT) system. Speech recognition uncertainty was integrated with MT by using speech output lattices as input to translation models [10, 11]. A sequence-to-sequence model for speech translation without transcriptions has been introduced [12], but was only evaluated on alignment. Synthesized speech data were translated in [13] using a model similar to the Listen Attend and Spell model [14], while a larger-scale study [15] used an end-to-end system for translating audio books between French and English. Sequence-to-sequence models to both transcribe Spanish speech and translate it in English have also been proposed [16], by jointly training the two tasks in a multitask scenario with two decoders sharing the speech encoder. This model was extended by us [17], with the translation decoder receiving information both from the speech encoder and the transcription decoder.

Multi-source models have also been studied, for statistical MT [18] and neural MT [19]. The two inputs are the same sentence in two languages, and the model is trained on trilingual text. Multi-source ensembles for MT have also been explored [20], where the two ensemble components were trained separately, using text input in different languages.

The only previous work that operates under similar assumptions to ours, that is, having access only to translations of the train and test utterances and no other parallel data, is `LatticeTM` [21], a model that composes word lattices (the result of ASR) with the weighted finite-state transducer of a translation model that is jointly learned. They showed consistent reductions in word error rate (WER) on two datasets, including CALLHOME. Note that since they use word lattices to represent the speech recognition output and do not attempt acoustic modeling, their setting is easier than ours. Our approach, instead, operates directly on the speech signal.

Table 1: *Multi-Source models achieve lower Character Error Rates (CER) on all three target languages, even in extremely low resource settings (Ainu, Mboshi). In Spanish, we observe an 8.4% reduction in CER.*

Source	Transcription CER		
	Ainu	Mboshi	Spanish
speech (audio)	40.7	29.8	52.0
translation (text)	74.9	68.2	44.6
coupled ensemble	**40.6**	36.8	42.2
multi-source	46.0	37.5	41.6
+tied	41.4	32.6	**37.6**
+shared	**40.6**	**28.6**	38.7

Table 2: *Evaluating the output with word-level BLEU, multi-source models significantly improve upon the baselines on higher-resource settings (Spanish). On Ainu, the best model performs on par with the very competitive baseline.*

Source	Transcription BLEU	
	Ainu	Spanish
speech (audio)	**28.92**	9.41
translation (text)	5.89	14.73
coupled ensemble	26.99	16.94
multi-source	24.03	17.59
+tied	26.95	**20.82**
+shared	**28.57**	19.47

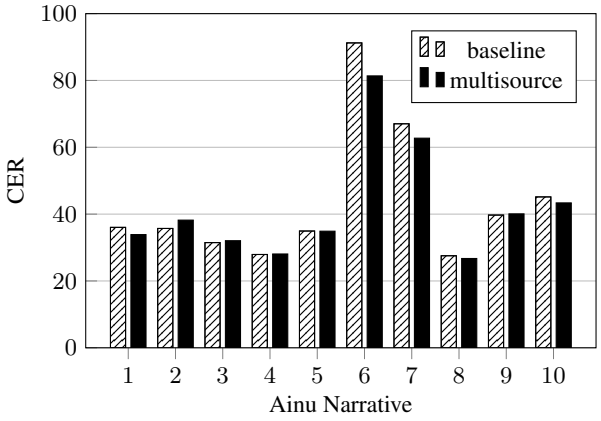


Figure 2: *Character Error Rates of the best baseline system and our best multisource system for each Ainu narrative. The gains of using the translations are apparent in the cases that are harder for a speech-only system: narratives 6 and 7 are more sung than narrated, rendering them harder to transcribe.*

4. Experiments

4.1. Implementation

Our models are implemented in DyNet [22].[2] We use a dropout of 0.2, and train using Adam with initial learning rate of 0.0002 for up to 300 epochs. The hidden layer and the attention size are 512 units. The acoustic encoder employs a 3-layer speech encoding scheme [12]. The first bidirectional layer receives the audio sequence in the form of 39-dimensional Perceptual Linear Predictive (PLP) features [23] computed over overlapping 25ms-wide windows every 10ms. The second and third layers consist of LSTMs with hidden state sizes of 128 and 512 respectively. Each layer encodes every second output of the previous layer. The translation encoder uses a bi-directional LSTM layer. The input and output character embedding size is set to 32.

For testing, we select the model with the best performance on dev. At inference time, we use a beam size of 4 and the beam score includes length normalization [24] with a weight of 0.8, which has been found to work well for low-resource neural Machine Translation [25].

4.2. Data

We evaluate all our models on three diverse datasets.

Mboshi-French: Mboshi (Bantu C25 in the Guthrie classification) is a language spoken in Congo-Brazzaville, without standard orthography. We use a corpus of 5517 parallel utterances (about 4.4 hours of audio) collected from three native speakers using the LIG-Aikuma app for the BULB project [5, 7]. The corpus provides non-standard grapheme transcriptions (produced by linguists to be close to the language phonology) as well as French translations. We sampled 100 segments from the training set to be our dev set, and used the original dev set (514 utterances) as our test set.

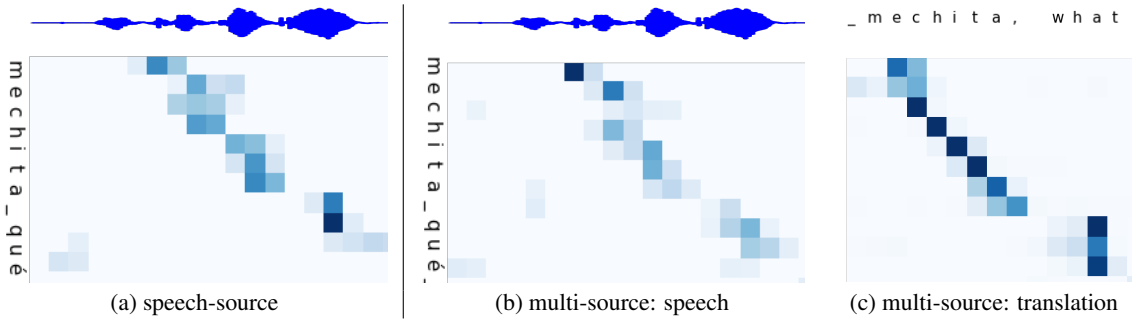
Ainu-English: Hokkaido Ainu is the sole surviving member of the Ainu language family and is generally considered a language isolate. As of 2007, only ten native speakers were alive. The Glossed Audio Corpus of Ainu Folklore provides 10 narratives (about 2.5 hours of audio), transcribed at the utterance level, glossed, and translated in Japanese and English.[3] All narratives were collected from the same speaker; the audio quality, though, is quite low, as the recordings were performed in an often noisy environment. Furthermore, the narratives have a distinct prosodic quality to them: at least two of them are more sung than narrated, rendering the dataset even harder to work with. This is further addressed in Section 4.3.

Since there does not exist a standard train-dev-test split, we employ a cross validation scheme for evaluation purposes. In each fold, one of the 10 narratives becomes the test set, with the previous one (mod 10) becoming the dev set, and the remaining 8 narratives becoming the training set. The models for each of the 10 folds are trained and tested separately. On average, for each fold, we train on about 2000 utterances; the dev and test sets consist of about 270 utterances. In Section 4.3 we report results on the concatenation of all folds, but also provide a breakdown of the performance in each fold.

Spanish-English: Spanish is obviously neither an endangered nor a low-resource language, but we pretend that it is one, by not making use of any Spanish resources like language models or pronunciation lexicons. We use the Spanish CALLHOME corpus (LDC96S35), which consists of telephone conversations between Spanish native speakers based in the US and their relatives abroad (about 20 total hours of audio) with more than 240 speakers. We use translations created by [26] and keep the original train-dev-test split, with the training set comprised of 80 conversations and dev and test of 20 conversations each. Unlike the other two datasets, there is no speaker overlap between the train and the test set.

[2] Our code is available online at `https://bitbucket.org/antonis/dynet-multisource-models`.

[3] `http://ainucorpus.ninjal.ac.jp/corpus/en/`



(a) speech-source (b) multi-source: speech (c) multi-source: translation

Figure 3: *Attentions on a speech sample from the dev set, that includes a proper name ("Mechita") unseen during training. The multi-source model (using a shared attention mechanism) receives informative context from the translation so as to produce the output.*

4.3. Results

The character error rates of the baseline and our multi-source models are presented in Table 1. In the two extremely low-resource datasets, Ainu and Mboshi, we find that the speech-only baselines are quite competitive. The very small number of speakers (one and three, respectively) makes the speech transcription task easier. In contrast, the translation task is much harder with so little data, as is also confirmed by the poor performance of the translation-only models in Ainu and Mboshi.

In the relatively higher-resource CALLHOME dataset, the translation-only model outperforms the speech-only one, with an improvement of 7.4 points in CER, a 12.3% reduction. This is most likely due to the lack of speaker overlap across the training and the test set. In this setting, the acoustic modelling part is harder than encoding the translation sentence.

For completeness, in Table 2 we also report word-level BLEU scores [27], the most common evaluation metric for Machine Translation. A higher BLEU almost always translates to lower WER. We only report results on Ainu and Spanish as Mboshi does not have standardized word segmentation rules. The BLEU scores reinforce our previous analysis. In the Ainu dataset, where a translation-only model is impossible to train (as outlined by the BLEU score of about 5.9), the multi-source model with *shared* attention mechanism performs on par with the speech-only model, as does the coupled ensemble model. In the CALLHOME dataset, our multi-source model with the *tied* attention mechanism achieves significantly higher BLEU scores than the translation-only or the coupled ensemble models. In addition, our best model's WER on the CALLHOME test set is 53.0, outperforming `LatticeTM` that achieved a WER of 56.2, despite the fact that our model operates in a harder setting, trained and tested directly on speech.

The performance of each fold of cross-validation for Ainu is shown in Figure 2. For each narrative, it compares the speech-only baseline system with our best multi-source system. The overall performance of the speech-only single-source model and our best model is similar with a CER of 40.7 and 40.6 respectively. A possible reason is that all the Ainu stories are narrated by the same speaker, making it a generally easier task for a speech recognition system. But we also see that in the cases where speech transcription is harder, translation information does help. Namely, narratives 6 and 7 are *sung*, making them harder to transcribe with a speech-only system trained on spoken data, as indicated by the higher error rates: 91.2 and 67.0, respectively. The multi-source models achieve noticeable improvements of 9.9 and 4.3 points on these narratives.

We further quantify the effect of the different sharing mechanisms for the attentions. Using word-level forced alignments on the CALLHOME dataset [12] we can evaluate the accuracy of the attention. Treating the forced alignments as reference, we compute the percentage of the weights of the attention over the speech source that fall within the boundaries of the forced alignment spans. Note that the forced alignments naturally include noise, so they should be treated as a "silver standard." However, they can still provide indications that could reveal the effect of parameter sharing.

We computed the average sum of this *attention accuracy* by forced decoding on the CALLHOME development set. We find that the average sum for the speech single-source model is almost 71%, a value similar to the average sums of the attention accuracy of the coupled ensemble and the multi-source model that employs no sharing mechanism. Instead, the attention accuracy of the model with the shared mechanism is almost 75%. The model with tied attentions, which achieves the best results on CALLHOME, has an attention accuracy of 76%.

Figure 3 presents the attention weights over a sample taken from the development set, produced by forced decoding. The segment includes an out-of-vocabulary word, the name *Mechita*, never seen during training. The attention weights over the speech source with the single-source model (3a) are not too different from the weights of the multi-source model with tied attentions (3b). However, the multi-source model in this case takes advantage of the translation and receives most of its context from the text source (3c), as the attention weights over the characters of the name are quite high (albeit, off-by-one, as often is the case in neural attention-based translation).

5. Conclusion

We presented multi-source neural architectures that receive an audio segment and its translation and produce a character-level transcription in low-resource settings. We showed that providing the translation as an additional input signal is beneficial to the transcription task, as our models outperform the single-source baselines. Furthermore, we find that sharing the decoder and the attention parameters leads to lower character error rate over either a coupled ensemble architecture or simple attention mechanisms without parameter sharing.

These results will hopefully lead to new tools for endangered language documentation. Projects like the BULB project that aim to collect about 100 hours of audio with translations, stand to benefit from our approach, since it would be impractical to manually transcribe this much audio for many languages. We hope that this work will provide a concrete basis for leveraging translations in a language documentation pipeline.

6. References

[1] M. P. Lewis, G. F. Simons, C. D. Fennig *et al.*, *Ethnologue: Languages of the world*. Dallas, TX: SIL International, 2009, vol. 16.

[2] S. Bird, L. Gawne, K. Gelbart, and I. McAlister, "Collecting bilingual audio in remote indigenous communities," in *Proc. COLING*, 2014. [Online]. Available: http://www.aclweb.org/anthology/C14-1096

[3] S. Bird, F. R. Hanke, O. Adams, and H. Lee, "Aikuma: A mobile app for collaborative language documentation," in *Proc. of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 2014. [Online]. Available: http://www.aclweb.org/anthology/W14-2201

[4] D. Blachon, E. Gauthier, L. Besacier, G.-N. Kouarata, M. Adda-Decker, and A. Rialland, "Parallel speech collection for under-resourced language studies using the LIG-Aikuma mobile device app," in *Proc. SLTU (Spoken Language Technologies for Under-Resourced Languages)*, vol. 81, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050916300448

[5] G. Adda, S. Stüker, M. Adda-Decker, O. Ambouroue, L. Besacier, D. Blachon, H. Bonneau-Maynard, P. Godard, F. Hamlaoui, D. Idiatov *et al.*, "Breaking the unwritten language barrier: The BULB project," *Procedia Computer Science*, vol. 81, pp. 8–14, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050916300370

[6] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. ICLR*, 2015. [Online]. Available: http://arxiv.org/abs/1409.0473

[7] P. Godard, G. Adda, M. Adda-Decker, J. Benjumea, L. Besacier, J. Cooper-Leavitt, G.-N. Kouarata, L. Lamel, H. Maynard, M. Mueller *et al.*, "A very low resource language speech corpus for computational language documentation experiments," 2017, arXiv:1710.03501. [Online]. Available: http://arxiv.org/abs/1710.03501

[8] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1412–1421.

[9] A. Dutt, D. Pellerin, and G. Quénot, "Coupled ensembles of neural networks," *arXiv preprint arXiv:1709.06053*, 2017.

[10] H. Ney, "Speech translation: Coupling of recognition and translation," in *Proc. ICASSP*, vol. 1, 1999.

[11] E. Matusov, S. Kanthak, and H. Ney, "On the integration of speech recognition and statistical machine translation," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[12] L. Duong, A. Anastasopoulos, D. Chiang, S. Bird, and T. Cohn, "An attentional model for speech translation without transcription," in *Proc. NAACL HLT*, 2016, pp. 949–959.

[13] A. Bérard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," in *Proc. NIPS Workshop on End-to-end Learning for Speech and Audio Processing*, 2016. [Online]. Available: https://arxiv.org/abs/1612.01744

[14] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. ICASSP*. IEEE, 2016, pp. 4960–4964.

[15] A. Bérard, L. Besacier, A. C. Kocabiyikoglu, and O. Pietquin, "End-to-end automatic speech translation of audiobooks," *arXiv preprint arXiv:1802.04200*, 2018.

[16] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly transcribe foreign speech," in *Proc. INTERSPEECH*, 2017. [Online]. Available: https://arxiv.org/abs/1703.08581

[17] A. Anastasopoulos and D. Chiang, "Tied multitask learning for neural speech translation," in *Proc. NAACL HLT*, 2018, to appear.

[18] F. J. Och and H. Ney, "Statistical multi-source translation," in *Proceedings of MT Summit*, vol. 8, 2001, pp. 253–258.

[19] B. Zoph and K. Knight, "Multi-source neural translation," in *Proc. NAACL-HLT*. Association for Computational Linguistics, 2016, pp. 30–34. [Online]. Available: http://www.aclweb.org/anthology/N16-1004

[20] E. Garmash and C. Monz, "Ensemble learning for multi-source neural machine translation," in *Proc. COLING*, 2016, pp. 1409–1418.

[21] O. Adams, G. Neubig, T. Cohn, and S. Bird, "Learning a translation model from word lattices." in *Proc. INTERSPEECH*, 2016, pp. 2518–2522.

[22] G. Neubig, C. Dyer, Y. Goldberg, A. Matthews, W. Ammar, A. Anastasopoulos, M. Ballesteros, D. Chiang, D. Clothiaux, T. Cohn *et al.*, "DyNet: The dynamic neural network toolkit," 2017, arXiv:1701.03980. [Online]. Available: http://arxiv.org/abs/1701.03980

[23] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[24] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, arXiv:1609.08144. [Online]. Available: https://arxiv.org/abs/1609.08144

[25] T. Q. Nguyen and D. Chiang, "Transfer learning across low-resource related languages for neural machine translation," in *Proc. IJCNLP*, 2017. [Online]. Available: https://arxiv.org/abs/1708.09803

[26] M. Post, G. Kumar, A. Lopez, D. Karakos, C. Callison-Burch, and S. Khudanpur, "Improved speech-to-text translation with the Fisher and Callhome Spanish-English speech translation corpus," in *Proc. IWSLT*, 2013. [Online]. Available: http://www.cs.jhu.edu/~post/papers/post2013improved.pdf

[27] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proc. ACL*. Association for Computational Linguistics, 2002, pp. 311–318.