INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Sequence-to-Sequence Models Can Directly Translate Foreign Speech

Ron J. Weiss[1], *Jan Chorowski*[1], *Navdeep Jaitly*[2*], *Yonghui Wu*[1], *Zhifeng Chen*[1]

[1]Google Brain
[2]Nvidia

{ronw,chorowski}@google.com, njaitly@nvidia.com, {yonghui,zhifengc}@google.com

Abstract

We present a recurrent encoder-decoder deep neural network architecture that directly translates speech in one language into text in another. The model does not explicitly transcribe the speech into text in the source language, nor does it require supervision from the ground truth source language transcription during training. We apply a slightly modified sequence-to-sequence with attention architecture that has previously been used for speech recognition and show that it can be repurposed for this more complex task, illustrating the power of attention-based models.

A single model trained end-to-end obtains state-of-the-art performance on the Fisher Callhome Spanish-English speech translation task, outperforming a cascade of independently trained sequence-to-sequence speech recognition and machine translation models by 1.8 BLEU points on the Fisher test set. In addition, we find that making use of the training data in both languages by multi-task training sequence-to-sequence speech translation and recognition models with a shared encoder network can improve performance by a further 1.4 BLEU points.
Index Terms: speech translation, sequence-to-sequence model

1. Introduction

Sequence-to-sequence models were recently introduced as a powerful new method for translation [1, 2]. Subsequently, the model has been adapted and applied to various tasks such as image captioning [3, 4], pose prediction [5], and syntactic parsing [6]. It has also led to a new state of the art in Neural Machine Translation (NMT) [7]. The model has also recently achieved promising results on automatic speech recognition (ASR) even without the use of language models [8, 9, 10, 11]. These successes have only been possible because the sequence-to-sequence (seq2seq) models can accurately model very complicated probability distributions. This makes it possible to apply this model even in situations where a precise analytical model is difficult to intuit.

In this paper we show that a single sequence-to-sequence model is powerful enough to translate audio in one language directly into text in another language. Using a model similar to Listen Attend and Spell (LAS) [9, 10] we process log mel filterbank input features using a recurrent encoder. The encoder features are then used, along with an attention model, to build a conditional next-step prediction model for text in the target domain. Unlike LAS, however, we use the text in the translated domain as the target – the source language text is not used.

Conventionally, this task is performed by pipelining results from an ASR system trained on the source language with a machine translation (MT) system trained to translate text from the source language to text in the target language. However, we motivate an end-to-end approach from several different angles.

First, by virtue of being an *end-to-end* model, all the parameters are jointly adjusted to optimize the results on the final goal. Training separate speech recognition and translation models may lead to a situation where models perform well individually, but do not work well together because their error surfaces do not compose well. For example typical errors in the ASR system may be such that they exacerbate the errors of the translation model which has not been trained to see such errors in the *input* during training. Another advantage of an end-to-end model is that, during inference, a single model can have lower latency compared to a cascade of two independent models. Additionally, end-to-end models have advantages in low resource settings since they can directly make use of corpora where the audio is in one language while the transcript is in another. This can arise, for example, in videos that have been captioned in other languages. It can also reduce labeling budgets since speech would only need to be transcribed in one language. In extreme cases where the source language does not have a writing system, applying a separate ASR system would require first standardizing the writing system – a very significant undertaking [12, 13].

We make several interesting observations from experiments on conversational Spanish to English speech translation. As with LAS models, we find that the model performs surprisingly well without using independent language models in either the source or target language. While the model performs well without seeing source language transcripts during training, we find that we can leverage them in a multi-task setting to improve performance. Finally we show that the end-to-end model outperforms a cascade of independent seq2seq ASR and NMT models.

2. Related work

Early work on *speech translation* (ST) [14] – translating audio in one language into text in another – used lattices from an ASR system as inputs to translation models [15, 16], giving the translation model access to the speech recognition uncertainty. Alternative approaches explicitly integrated acoustic and translation models using a stochastic finite-state transducer which can decode the translated text directly using Viterbi search [17, 18].

In this paper we compare our integrated model to results obtained from cascaded models on a Spanish to English speech translation task [19, 20, 21]. These approaches also use ASR lattices as MT inputs. Post et al. [19] used a GMM-HMM ASR system. Kumar et al. [20] later showed that using a better ASR model improved overall ST results. Subsequently [21] showed that modeling features at the boundary of the ASR and the MT system can further improve performance. We carry this notion much further by defining an end-to-end model for the entire task.

Other recent work on speech translation does not use ASR. Instead [22] used an unsupervised model to cluster repeated audio patterns which are used to train a bag of words translation model. In [23] seq2seq models were used to align speech with translated text, but not to directly predict the translations. Our work is most similar to [24] which uses a LAS-like model for ST

[*] Work done at Google.

http://dx.doi.org/10.21437/Interspeech.2017-503

on data synthesized using a text-to-speech system. In contrast, we train on a much larger corpus composed of real speech.

3. Sequence-to-sequence model

We utilize a sequence-to-sequence with attention architecture similar to that described in [1]. The model is composed of three jointly trained neural networks: a recurrent *encoder* which transforms a sequence of input feature frames $x_{1..T}$ into a sequence of hidden activations, $h_{1..L}$, optionally at a slower time scale:

$$h_l = \text{enc}(x_{1..T}) \tag{1}$$

The full encoded input sequence $h_{1..L}$ is consumed by a *decoder* network which emits a sequence of output tokens, $y_{1..K}$, via next step prediction: emitting one output token (e.g. word or character) per step, conditioned on the token emitted at the previous time step as well as the entire encoded input sequence:

$$y_k = \text{dec}(y_{k-1}, h_{1..L}) \tag{2}$$

The dec function is implemented as a stacked recurrent neural network with D layers, which can be expanded as follows:

$$o_k^1,\ s_k^1 = d^1(y_{k-1}, s_{k-1}^1, c_{k-1}) \tag{3}$$

$$o_k^n,\ s_k^n = d^n(o_k^{n-1}, s_{k-1}^n, c_k) \tag{4}$$

where d^n is a long short-term memory (LSTM) cell [25], which emits an output vector o^n into the following layer, and updates its internal state s^n at each time step.

The decoder's dependence on the input is mediated through an *attention* network which summarizes the entire input sequence as a fixed dimensional context vector c_k which is passed to all subsequent layers using skip connections. c_k is computed from the first decoder layer output at each output step k:

$$c_k = \textstyle\sum_l \alpha_{kl} h_l \tag{5}$$

$$\alpha_{kl} = \text{softmax}(a_e(h_l)^T a_d(o_k^1)) \tag{6}$$

where a_e and a_d are small fully connected networks. The α_{kl} probabilities compute a soft alignment between the input and output sequences. An example is shown in Figure 1.

Finally, an output symbol is sampled from a multinomial distribution computed from the final decoder layer output:

$$y_k \sim \text{softmax}(W_y[o_k^D, c_k] + b_y) \tag{7}$$

3.1. Speech model

We train seq2seq models for both end-to-end speech translation, and a baseline model for speech recognition. We found that the same architecture, a variation of that from [10], works well for both tasks. We use 80 channel log mel filterbank features extracted from 25ms windows with a hop size of 10ms, stacked with delta and delta-delta features. The output softmax of all models predicts one of 90 symbols, described in detail in Section 4, that includes English and Spanish lowercase letters.

The encoder is composed of a total of 8 layers. The input features are organized as a $T \times 80 \times 3$ tensor, i.e. raw features, deltas, and delta-deltas are concatenated along the 'depth' dimension. This is passed into a stack of two convolutional layers with ReLU activations, each consisting of 32 kernels with shape $3 \times 3 \times$ depth in time $\times$ frequency. These are both strided by 2×2, downsampling the sequence in time by a total factor of 4, decreasing the computation performed in the following layers. Batch normalization [26] is applied after each layer.

This downsampled feature sequence is then passed into a single bidirectional convolutional LSTM [27, 28, 10] layer using a 1×3 filter (i.e. convolving only across the frequency dimension within each time step). Finally, this is passed into a stack of three bidirectional LSTM layers of size 256 in each direction, interleaved with a 512-dimensional linear projection, followed by batch normalization and a ReLU activation, to compute the final 512-dimensional encoder representation, h_l.

The decoder input is created by concatenating a 64-dimensional embedding for y_{k-1}, the symbol emitted at the previous time step, and the 512-dimensional attention context vector c_k. The networks a_e and a_d used to compute c_k (see equation 6) each contain a single hidden layer with 128 units. This is passed into a stack of four unidirectional LSTM layers with 256 units. Finally the concatenation of the attention context and LSTM output is passed into a softmax layer which predicts the probability of emitting each symbol in the output vocabulary.

The network contains 9.8m parameters. We implement it with TensorFlow [29] and train using teacher forcing on minibatches of 64 utterances. We use asynchronous stochastic gradient descent across 10 replicas using the Adam optimizer [30] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-6}$. The initial learning rate is set to 10^{-3} and decayed by a factor of 10 after 1m steps. L2 weight decay is used with a weight of 10^{-6}, and beginning from step 20k, Gaussian weight noise with std of 0.125 is added to all LSTM weights and decoder embeddings. We tuned all hyperparameters to maximize performance on the Fisher/dev set.

We decode using beam search with rank pruning at 8 hypotheses and a beam width of 3, using the scoring function proposed in [7]. We do not utilize any language models. For the baseline ASR model we found that neither length normalization nor the coverage penalty from [7] were needed, however it was helpful to permit emitting the end-of-sequence token only when its log-probability was 3.0 greater than the next most probable token. For speech translation we found that using length normalization of 0.6 improved performance by 0.6 BLEU points.

3.2. Neural machine translation model

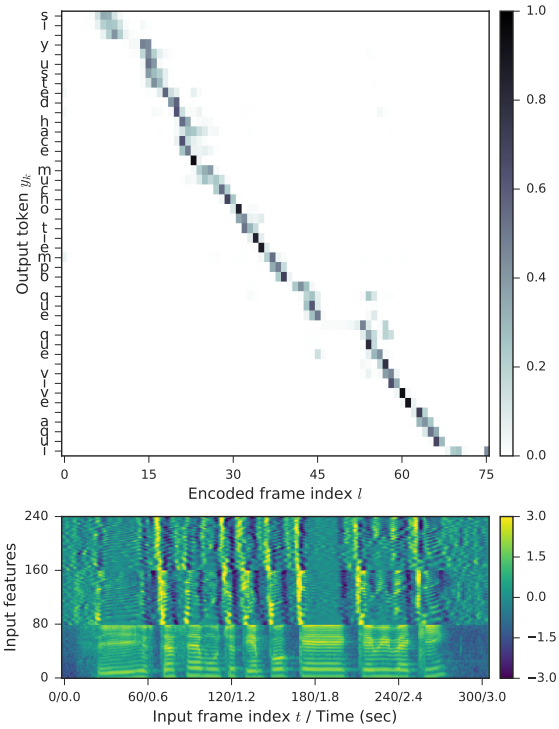
We also train a baseline seq2seq text machine translation model following [7]. To reduce overfitting on the small training corpus we significantly reduce the model size compared to those in [7].

The encoder network consists of four encoder layers (5 LSTM layers in total). As in the base architecture, the bottom layer is a bidirectional LSTM and the remaining layers are all unidirectional. The decoder network consists of 4 stacked LSTM layers. All encoder and decoder LSTM layers contain 512 units. The attention network uses a single hidden layer with 512 units. We use the same character-level vocabulary for input and output as the speech model described above emits.

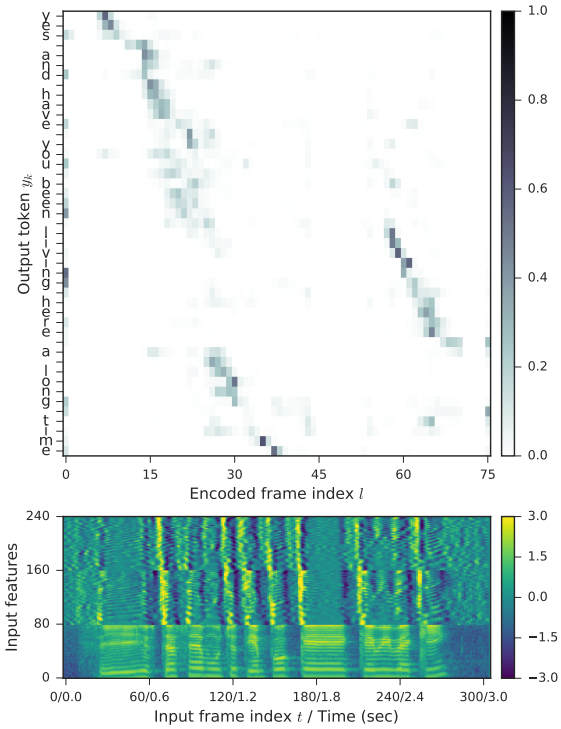
As in [7] we apply dropout [31] with probability 0.2 during training to reduce overfitting. We train using SGD with a single replica. Training converges after about 100k steps using minibatches of 128 sentence pairs.

3.3. Multi-task training

Supervision from source language transcripts can be incorporated into the speech translation model by co-training an auxiliary model with shared parameters, e.g. an ASR model using a common encoder. This is equivalent to a multi-task configuration [32]. We use the models and training protocols described above with these modifications: we use 16 workers that randomly select a model to optimize at each step, we introduce weight noise after 30k steps, and decay the learning rate after 1.5m overall steps.



(a) *Spanish speech recognition decoder attention.*



(b) *Spanish-to-English speech translation decoder attention.*

Figure 1: *Example attention probabilities α_{kl} from a multi-task model with two decoders. The ASR attention is roughly monotonic, whereas the translation attention contains an example of word reordering typical of seq2seq MT models, attending primarily to frames $l = 58 - 70$ while emitting "living here". The recognition decoder attends to these frames while emitting the corresponding Spanish phrase "vive aqui". The ASR decoder is also more confident than the translation attention, which tends to be smoothed out across many input frames for each output token. This is a consequence of the ambiguous mapping between Spanish sounds and English translation.*

4. Experiments

We conduct experiments on the Spanish Fisher and Callhome corpora of telephone conversations augmented with English translations from [19]. We split training utterances according to the provided segment annotations and preprocess the Spanish transcriptions and English translations by lowercasing and removing punctuation. All models use a common set of 90 tokens to represent Spanish and English characters, containing lowercase letters from both alphabets, digits 0-9, space, punctuation marks,[1] as well as special start-of-, end-of-sequence, and unknown tokens.

Following [19, 20, 21], we train all models on the 163 hour Fisher train set, and tune hyperparameters on Fisher/dev. We report speech recognition results as word error rates (WER) and translation results using BLEU [33] scores computed with the Moses toolkit[2] `multi-bleu.pl` script, both on lowercase reference text after removing punctuation. We use the 4 provided Fisher reference translations and a single reference for Callhome.

4.1. Tuning decoder depth

It is common for ASR seq2seq models to use a shallow decoder, generally comprised of one recurrent layer [8, 9, 10]. In contrast, seq2seq NMT models often use much deeper decoders, e.g. 8 layers in [7]. In analogy to a traditional ASR system, one may think of the seq2seq encoder behaving as the acoustic model while the decoder acts as the language model. The additional complexity of the translation task when compared to monolingual language modeling motivates the use of a higher capacity decoder network. We therefore experiment with varying the depth of the stack of LSTM layers used in the decoder for speech translation and find that performance improves as the decoder depth increases up to four layers, see Table 1.

Table 1: *Varying number of decoder layers in the speech translation model. BLEU score on the Fisher/dev set.*

Num decoder layers D				
1	2	3	4	5
43.8	45.1	45.2	45.5	45.3

Despite this intuition, we obtained similar improvements in performance on the ASR task when increasing decoder depth, suggesting that tuning the decoder architecture is worth further investigation in other speech settings.

4.2. Tuning the multi-task model

We compare two multi-task training strategies: one-to-many in which an *encoder* is shared between speech translation and recognition tasks, and many-to-one in which a *decoder* is shared between speech and text translation tasks. We found the first strategy to perform better. We also found that performing updates

[1] Most are not used in practice since we remove punctuation aside from apostrophes from the training targets.

[2] `http://www.statmt.org/moses/`

Table 2: *Varying the number of shared encoder LSTM layers in the multi-task setting. BLEU score on the Fisher/dev set.*

Num shared encoder LSTM layers			
3 (all)	2	1	0
46.2	45.1	45.3	44.2

Table 3: *Speech recognition model performance in WER.*

	Fisher			Callhome	
	dev	dev2	test	devtest	evltest
Ours[3]	25.7	25.1	23.2	44.5	45.3
Post et al. [19]	41.3	40.0	36.5	64.7	65.3
Kumar et al. [21]	29.8	29.8	25.3	–	–

more often on the speech translation task yields the best results. Specifically, we perform 75% of training steps on the core speech translation task, and the remainder on the auxiliary ASR task.

Finally, we vary how much of the encoder network parameters are shared across tasks. Intuitively we expect that layers near the input will be less sensitive to the final classification task, so we always share all encoder layers through the conv LSTM but vary the amount of sharing in the final stack of LSTM layers. As shown in Table 2 we found that sharing all layers of the encoder yields the best performance. This suggests that the encoder learns to transform speech into a consistent interlingual subword unit representation, which the respective decoders are able to assemble into phrases in either language.

4.3. Baseline models

We construct a baseline cascade of a Spanish ASR seq2seq model whose output is passed into a Spanish to English NMT model. Our seq2seq ASR model attains state-of-the-art performance on the Fisher and Callhome datasets compared to previously reported results with HMM-GMM [19] and DNN-HMM [21] systems, as shown in Table 3. Performance on the Fisher task is significantly better than on Callhome since it contains more formal speech, consisting of conversations between strangers while Callhome conversations were often between family members.

In contrast, our MT model slightly underperforms compared to previously reported results using phrase-based translation systems [19, 20, 21] as shown in Table 4. This may be because the amount of training data in the Fisher corpus is much smaller than is typically used for training NMT systems. Additionally, our models used characters as training targets instead of word- and phrase-level tokens often used in machine translation systems, making them more vulnerable to e.g. spelling errors.

4.4. Speech translation

Table 5 compares performance of different systems on the full speech translation task. Despite not having access to source language transcripts at any stage of the training, the end-to-end model outperforms the baseline cascade, which passes the 1-best Spanish ASR output into the NMT model, by about 1.8 BLEU points on the Fisher/test set. We obtain an additional improvement of 1.4 BLEU points or more on all Fisher datasets in the multi-task configuration, in which the Spanish transcripts are used for additional supervision by sharing a single encoder

[3] Averaged over three runs.

Table 4: *Translation BLEU score on ground truth transcripts.*

	Fisher			Callhome	
	dev	dev2	test	devtest	evltest
Ours	58.7	59.9	57.9	28.2	27.9
Post et al. [19]	–	–	58.7	–	27.8
Kumar et al. [21]	–	65.4	62.9	–	–

Table 5: *Speech translation model performance in BLEU score.*

	Fisher			Callhome	
Model	dev	dev2	test	devtest	evltest
End-to-end ST[3]	46.5	47.3	47.3	16.4	16.6
Multi-task ST / ASR[3]	48.3	49.1	48.7	16.8	17.4
ASR→NMT cascade[3]	45.1	46.1	45.5	16.2	16.6
Post et al. [19]	–	35.4	–	–	11.7
Kumar et al. [21]	–	40.1	40.4	–	–

sub-network across independent ASR and ST decoders. The ASR model converged after four days of training (1.5m steps), while the ST and multitask models continued to improve, with the final 1.2 BLEU point improvement taking two more weeks.

Informal inspection of cascade system outputs yields many examples of compounding errors, where the ASR model makes an insertion or deletion that significantly alters the meaning of the sentence and the NMT model has no way to recover. This illustrates a key advantage of the end-to-end approach where the translation decoder can access the full latent representation of the speech without first collapsing to an n-best list of hypotheses.

A large performance gap of about 10 BLEU points remains between these results and those from Table 4 which assume perfect ASR, indicating significant room for improvement in the acoustic modeling component of the speech translation task.

5. Conclusion

We present a model that directly translates speech into text in a different language. One of its striking characteristics is that its architecture is essentially the same as that of an attention-based ASR neural system. Direct speech-to-text translation happens in the same computational footprint as speech recognition – the ASR and end-to-end ST models have the same number of parameters, and utilize the same decoding algorithm – narrow beam search. The end-to-end trained model outperforms an ASR-MT cascade even though it never explicitly searches over transcriptions in the source language during decoding.

While we can interpret the proposed model's encoder and decoder networks, respectively, as acoustic and translation models, it does not have an explicit concept of source transcription. The two sub-networks exchange information as abstract high-dimensional real valued vectors rather than discrete transcription lattices as in traditional systems. In fact, reading out transcriptions in the source language from this abstract representation requires a separate decoder network. We find that jointly training decoder networks for multiple languages regularizes the encoder and improves overall speech translation performance. An interesting extension would be to construct a multilingual speech translation system following [34] in which a single decoder is shared across multiple languages, passing a discrete input token into the network to select the desired output language.

6. References

[1] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of ICLR*, 2015.

[2] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, 2014, pp. 3104–3112.

[3] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.

[4] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention." in *Proceedings of ICML*, vol. 14, 2015, pp. 77–81.

[5] G. Gkioxari, A. Toshev, and N. Jaitly, "Chained predictions using convolutional neural networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 728–743.

[6] O. Vinyals, Ł. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, "Grammar as a foreign language," in *Advances in Neural Information Processing Systems*, 2015, pp. 2773–2781.

[7] 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," *arXiv preprint arXiv:1609.08144*, 2016.

[8] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems*, 2015, pp. 577–585.

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

[10] Y. Zhang, W. Chan, and N. Jaitly, "Very deep convolutional networks for end-to-end speech recognition," in *Proceedings of ICASSP*, 2017.

[11] J. Chorowski and N. Jaitly, "Towards better decoding and language model integration in sequence to sequence models," *arXiv preprint arXiv:1612.02695*, 2016.

[12] S. Bird, L. Gawne, K. Gelbart, and I. McAlister, "Collecting bilingual audio in remote indigenous communities," in *Proceedings of COLING*, 2014.

[13] A. Anastasopoulos, D. Chiang, and L. Duong, "An unsupervised probability model for speech-to-translation alignment of low-resource languages," *arXiv preprint arXiv:1609.08139*, 2016.

[14] F. Casacuberta, M. Federico, H. Ney, and E. Vidal, "Recent efforts in spoken language translation," *IEEE Signal Processing Magazine*, vol. 25, no. 3, 2008.

[15] H. Ney, "Speech translation: Coupling of recognition and translation," in *Proceedings of ICASSP*, vol. 1. IEEE, 1999, pp. 517–520.

[16] E. Matusov, S. Kanthak, and H. Ney, "On the integration of speech recognition and statistical machine translation." in *Proceedings of Interspeech*, 2005, pp. 3177–3180.

[17] E. Vidal, "Finite-state speech-to-speech translation," in *Proceedings of ICASSP*, vol. 1. IEEE, 1997, pp. 111–114.

[18] F. Casacuberta, H. Ney, F. J. Och, E. Vidal, J. M. Vilar, S. Barrachina, I. Garcıa-Varea, D. Llorens, C. Martınez, S. Molau *et al.*, "Some approaches to statistical and finite-state speech-to-speech translation," *Computer Speech & Language*, vol. 18, no. 1, pp. 25–47, 2004.

[19] 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 *Proceedings of IWSLT*, 2013.

[20] G. Kumar, M. Post, D. Povey, and S. Khudanpur, "Some insights from translating conversational telephone speech," in *Proceedings of ICASSP*. IEEE, 2014, pp. 3231–3235.

[21] G. Kumar, G. W. Blackwood, J. Trmal, D. Povey, and S. Khudanpur, "A coarse-grained model for optimal coupling of ASR and SMT systems for speech translation." in *Proceedings of EMNLP*, 2015, pp. 1902–1907.

[22] S. Bansal, H. Kamper, A. Lopez, and S. Goldwater, "Towards speech-to-text translation without speech recognition," *arXiv preprint arXiv:1702.03856*, Feb. 2017.

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

[24] 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 *NIPS Workshop on End-to-end Learning for Speech and Audio Processing*, 2016.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of ICML*, 2015, pp. 448–456.

[27] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Advances in Neural Information Processing Systems*, 2015, pp. 802–810.

[28] I. Bogun, A. Angelova, and N. Jaitly, "Object recognition from short videos for robotic perception," *CoRR*, vol. abs/1509.01602, 2015. [Online]. Available: http://arxiv.org/abs/1509.01602

[29] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2016.

[30] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of ICLR*, 2015.

[31] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[32] M.-T. Luong, Q. V. Le, I. Sutskever, O. Vinyals, and L. Kaiser, "Multi-task sequence to sequence learning," in *Proceedings of ICLR*, 2016.

[33] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[34] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado *et al.*, "Google's multilingual neural machine translation system: Enabling zero-shot translation," *arXiv preprint arXiv:1611.04558*, 2016.