INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Direct speech-to-speech translation with a sequence-to-sequence model

Ye Jia[*], *Ron J. Weiss*[*], *Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, Yonghui Wu*

Google

{jiaye,ronw}@google.com

Abstract

We present an attention-based sequence-to-sequence neural network which can directly translate speech from one language into speech in another language, without relying on an intermediate text representation. The network is trained end-to-end, learning to map speech spectrograms into target spectrograms in another language, corresponding to the translated content (in a different canonical voice). We further demonstrate the ability to synthesize translated speech using the voice of the source speaker. We conduct experiments on two Spanish-to-English speech translation datasets, and find that the proposed model slightly underperforms a baseline cascade of a direct speech-to-text translation model and a text-to-speech synthesis model, demonstrating the feasibility of the approach on this very challenging task.

Index Terms: speech-to-speech translation, voice transfer, attention, sequence-to-sequence model, end-to-end model

1. Introduction

We address the task of speech-to-speech translation (S2ST): translating speech in one language into speech in another. This application is highly beneficial for breaking down communication barriers between people who do not share a common language. Specifically, we investigate whether it is possible to train model to accomplish this task directly, without relying on an intermediate text representation. This is in contrast to conventional S2ST systems which are often broken down into three components: automatic speech recognition (ASR), text-to-text machine translation (MT), and text-to-speech (TTS) synthesis [1–4].

Cascaded systems have the potential problem of errors compounding between components, e.g. recognition errors leading to larger translation errors. Direct S2ST models avoid this issue by training to solve the task end-to-end. They also have advantages over cascaded systems in terms of reduced computational requirements and lower inference latency since only one decoding step is necessary, instead of three. In addition, direct models are naturally capable of retaining paralinguistic and non-linguistic information during translation, e.g. maintaining the source speaker's voice, emotion, and prosody, in the synthesized translated speech. Finally, directly conditioning on the input speech makes it easy to learn to generate fluent pronunciations of words which do not need to be translated, such as names.

However, solving the direct S2ST task is especially challenging for several reasons. Fully-supervised end-to-end training requires collecting a large set of input/output speech pairs. Such data are more difficult to collect compared to parallel text pairs for MT, or speech-text pairs for ASR or TTS. Decomposing into smaller tasks can take advantage of the lower training data requirements compared to a monolithic speech-to-speech model, and can result in a more robust system for a given training budget. Uncertain alignment between two spectrograms whose underlying spoken content differs also poses a major training challenge.

In this paper we demonstrate *Translatotron*[1], a direct speech-to-speech translation model which is trained end-to-end. To facilitate training without predefined alignments, we leverage high level representations of the source or target content in the form of transcriptions, essentially multitask training with speech-to-text tasks. However no intermediate text representation is used during inference. The model does not perform as well as a baseline cascaded system. Nevertheless, it demonstrates a proof of concept and serves as a starting point for future research.

Extensive research has studied methods for combining different sub-systems within cascaded speech translation systems. [5, 6] gave MT access to the lattice of the ASR. [7, 8] integrated acoustic and translation models using a stochastic finite-state transducer which can decode the translated text directly using Viterbi search. For synthesis, [9] used unsupervised clustering to find F0-based prosody features and transfer intonation from source speech and target. [10] augmented MT to jointly predict translated words and emphasis, in order to improve expressiveness of the synthesized speech. [11] used a neural network to transfer duration and power from the source speech to the target. [12] transfered source speaker's voice to the synthesized translated speech by mapping hidden Markov model states from ASR to TTS. Similarly, recent work on neural TTS has focused on adapting to new voices with limited reference data [13–16].

Initial approaches to end-to-end speech-to-text translation (ST) [17, 18] performed worse than a cascade of an ASR model and an MT model. [19, 20] achieved better end-to-end performance by leveraging weakly supervised data with multitask learning. [21] further showed that use of synthetic training data can work better than multitask training. In this work we take advantage of both synthetic training targets and multitask training.

The proposed model resembles recent sequence-to-sequence models for voice conversion, the task of recreating an utterance in another person's voice [22–24]. For example, [23] proposes an attention-based model to generate spectrograms in the target voice based on input features (spectrogram concatenated with ASR bottleneck features) from the source voice. In contrast to S2ST, the input-output alignment for voice conversion is simpler and approximately monotonic. [23] also trains models that are specific to each input-output speaker pair (i.e. one-to-one conversion), whereas we explore many-to-one and many-to-many speaker configurations. Finally, [25] demonstrated an attention-based direct S2ST model on a toy dataset with a 100 word vocabulary. In this work we train on real speech, including spontaneous telephone conversations, at a much larger scale.

2. Speech-to-speech translation model

An overview of the proposed Translatotron model architecture is shown in Figure 1. Following [15, 26], it is composed of several separately trained components: 1) an *attention-based sequence-*

[*] Equal contribution.

[1] Audio samples are available at https://google-research.github.io/lingvo-lab/translatotron.

http://dx.doi.org/10.21437/Interspeech.2019-1951

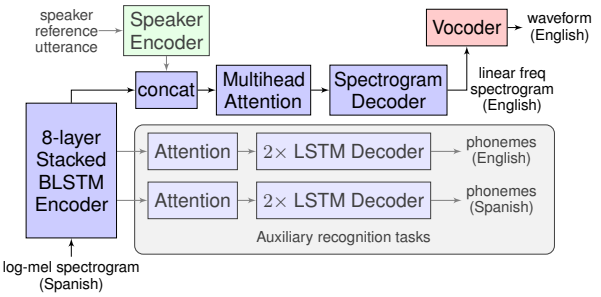

Figure 1: *Proposed model architecture, which generates English speech (top right) from Spanish speech (bottom left), and an optional speaker reference utterance (top left) which is only used for voice transfer experiments in Section 3.4. The model is multi-task trained to predict source and target phoneme transcripts as well, however these auxiliary tasks are not used during inference. Optional components are drawn in light colors.*

to-sequence network (blue) which generates target spectrograms, 2) a *vocoder* (red) which converts target spectrograms to time-domain waveforms, and, 3) optionally, a pretrained *speaker encoder* (green) which can be used to condition the decoder on the identity of the source speaker, enabling cross-language voice conversion [27] simultaneously with translation.

The sequence-to-sequence encoder stack maps 80-channel log-mel spectrogram input features into hidden states which are passed through an attention-based alignment mechanism to condition an autoregressive decoder, which predicts 1025-dim log spectrogram frames corresponding to the translated speech. Two optional auxiliary decoders, each with their own attention components, predict source and target phoneme sequences.

Following recent speech translation [21] and recognition [28] models, the encoder is composed of a stack of 8 bidirectional LSTM layers. As shown in Fig. 1, the final layer output is passed to the primary decoder, whereas intermediate activations are passed to auxiliary decoders predicting phoneme sequences.We hypothesize that early layers of the encoder are more likely to represent the source content well, while deeper layers might learn to encode more information about the target content.

The spectrogram decoder uses an architecture similar to Tacotron 2 TTS model [26], including pre-net, autoregressive LSTM stack, and post-net components. We make several changes to it in order to adapt to the more challenging S2ST task. We use multi-head additive attention [29] with 4 heads instead of location-sensitive attention, which shows better performance in our experiments. We also use a significantly narrower 32 dimensional pre-net bottleneck compared to 256-dim in [26], which we find to be critical in picking up attention during training. We also use reduction factor [30] of 2, i.e. predicting two spectrogram frames for each decoding step. Finally, consistent with results on translation tasks [19, 31], we find that using a deeper decoder containing 4 or 6 LSTM layers leads to good performance.

We find that multitask training is critical in solving the task, which we accomplish by integrating auxiliary decoder networks to predict phoneme sequences corresponding to the source and/or target speech. Losses computed using these auxiliary recognition networks are used during training, which help the primary spectrogram decoder to learn attention. They are not used during inference. In contrast to the primary decoder, the auxiliary decoders use 2-layer LSTMs with single-head additive attention [32]. All three decoders use attention dropout and LSTM zoneout regularization [33], all with probability 0.1. Training uses the Adafactor optimizer [34] with a batch size of 1024.

Table 1: *Dataset-specific model hyperparameters.*

	Conversational	Fisher
Num train examples	979k	120k
Input / output sample rate (Hz)	16k / 24k	8k / 24k
Learning rate	0.002	0.006
Encoder BLSTM	8×1024	8×256
Decoder LSTM	6×1024	4×1024
Auxiliary decoder LSTM	2×256	2×256
source / target input layer	8 / 8	4 / 6
dropout prob	0.2	0.3
loss decay	constant 1.0	$0.3 \rightarrow 0.001$ at 160k steps
Gaussian weight noise stddev	none	0.05

Since we are only demonstrating a proof of concept, we primarily rely on the low-complexity Griffin-Lim [35] vocoder in our experiments. However, we use a WaveRNN [36] neural vocoder when evaluating speech naturalness in listening tests.

Finally, in order to control the output speaker identity we incorporate an optional speaker encoder network as in [15]. This network is discriminatively pretrained on a speaker verification task and is not updated during the training of Translatotron. We use the *dvector V3* model from [37], trained on a larger set of 851K speakers across 8 languages including English and Spanish. The model computes a 256-dim speaker embedding from the speaker reference utterance, which is passed into a linear projection layer (trained with the sequence-to-sequence model) to reduce the dimensionality to 16. This is critical to generalizing to source language speakers which are unseen during training.

3. Experiments

We study two Spanish-to-English translation datasets: the large scale "conversational" corpus of parallel text and read speech pairs from [21], and the Spanish Fisher corpus of telephone conversations and corresponding English translations [38], which is smaller and more challenging due to the spontaneous and informal speaking style. In Sections 3.1 and 3.2, we *synthesize* target speech from the target transcript using a single (female) speaker English TTS system; In Section 3.4, we use real human target speech for voice transfer experiments on the conversational dataset. Models were implemented using the Lingvo framework [39]. See Table 1 for dataset-specific hyperparameters.

To evaluate speech-to-speech translation performance we compute BLEU scores [40] as an objective measure of speech intelligibility and translation quality, by using a pretrained ASR system to recognize the generated speech, and comparing the resulting transcripts to ground truth reference translations. Due to potential recognition errors (see Figure 2), this can be thought of as a lower bound on the underlying translation quality. We use the *16k Word-Piece* attention-based ASR model from [41] trained on the 960 hour LibriSpeech corpus [42], which obtained word error rates of 4.7% and 13.4% on the test-clean and test-other sets, respectively. In addition, we conduct listening tests to measure subjective speech naturalness mean opinion score (MOS), as well as speaker similarity MOS for voice transfer.

3.1. Conversational Spanish-to-English

This proprietary dataset described in [21] was obtained by crowdsourcing humans to read the both sides of a conversational Spanish-English MT dataset. In this section, instead of using the human target speech, we use a TTS model to synthesize target

Table 2: *Conversational test set performance. Single reference BLEU and Phoneme Error Rate (PER) of aux decoder outputs.*

Auxiliary loss	BLEU	Source PER	Target PER
None	0.4	-	-
Source	42.2	5.0	-
Target	42.6	-	20.9
Source + Target	42.7	5.1	20.8
ST [21] $\rightarrow$ TTS cascade	48.7	-	-
Ground truth	74.7	-	-

speech in a single female English speaker's voice in order to simplify the learning objective. We use an English Tacotron 2 TTS model [26] but use a Griffin-Lim vocoder for expediency. In addition, we augment the input source speech by adding background noise and reverberation in the same manner as [21].

The resulting dataset contains 979k parallel utterance pairs, containing 1.4k hours of source speech and 619 hours of synthesized target speech. The total target speech duration is much smaller because the TTS output is better endpointed, and contains fewer pauses. 9.6k pairs are held out for testing.

Input feature frames are created by stacking 3 adjacent frames of an 80-channel log-mel spectrogram as in [21]. The speaker encoder was not used in these experiments since the target speech always came from the same speaker.

Table 2 shows performance of the model trained using different combinations of auxiliary losses, compared to a baseline ST $\rightarrow$ TTS cascade model using a speech-to-text translation model [21] trained on the same data, and the same Tacotron 2 TTS model used to synthesize training targets. Note that the ground truth BLEU score is below 100 due to ASR errors during evaluation, or TTS failure when synthesizing the ground truth.

Training without auxiliary losses leads to extremely poor performance. The model correctly synthesizes common words and simple phrases, e.g. translating "hola" to "hello". However, it does not consistently translate full utterances. While it always generates plausible speech sounds in the target voice, the output can be independent of the input, composed of a string of nonsense syllables. This is consistent with failure to learn to attend to the input, and reflects the difficulty of the direct S2ST task.

Integrating auxiliary phoneme recognition tasks helped regularize the encoder and enabled the model to learn attention, dramatically improving performance. The target phoneme PER is much higher than on source phonemes, reflecting the difficulty of the corresponding translation task. Training using both auxiliary tasks achieved the best quality, but the performance difference between different combinations is small. Overall, there remains a gap of 6 BLEU points to the baseline, indicating room for improvement. Nevertheless, the relatively narrow gap demonstrates the potential of the end-to-end approach.

3.2. Fisher Spanish-to-English

This dataset contains about 120k parallel utterance pairs[2], spanning 127 hours of source speech. Target speech is synthesized using Parallel WaveNet [43] using the same voice as the previous section. The result contains 96 hours of synthetic target speech.

Following [19], input features were constructed by stacking 80-channel log-mel spectrograms, with deltas and accelerations. Given the small size of the dataset compared to that in Sec. 3.1, we found that obtaining good performance required significantly more careful regularization and tuning. As shown in Table 1, we used narrower encoder dimension of 256, a shallower 4-layer decoder, and added Gaussian weight noise to all LSTM weights as regularization, as in [19]. The model was especially sensitive to the auxiliary decoder hyperparameters, with the best performance coming when passing activations from intermediate layers of the encoder stack as inputs to the auxiliary decoders, using slightly more aggressive dropout of 0.3, and decaying the auxiliary loss weight over the course of training in order to encourage the model training to fit the primary S2ST task.

[2]This is a subset of the Fisher data due to TTS errors on target text.

Table 3: *Performance on the Fisher Spanish-English task in terms of 4-reference BLEU score on 3 eval sets.*

Auxiliary loss	dev1	dev2	test
None	0.4	0.6	0.6
Source	7.4	8.0	7.2
Target	20.2	21.4	20.8
Source + Target	24.8	26.5	25.6
Source + Target (1-head attention)	23.0	24.2	23.4
Source + Target (encoder pre-training)	30.1	31.5	31.1
ST [19] $\rightarrow$ TTS cascade	39.4	41.2	41.4
Ground truth	82.8	83.8	85.3

Table 4: *Naturalness MOS with 95% confidence intervals. The ground truth for both datasets are synthetic English speech.*

Model	Vocoder	Conversational	Fisher-test
Translatotron	WaveRNN	4.08 ± 0.06	3.69 ± 0.07
	Griffin-Lim	3.20 ± 0.06	3.05 ± 0.08
ST$\rightarrow$TTS	WaveRNN	4.32 ± 0.05	4.09 ± 0.06
	Griffin-Lim	3.46 ± 0.07	3.24 ± 0.07
Ground truth	Griffin-Lim	3.71 ± 0.06	-
	Parallel WaveNet	-	3.96 ± 0.06

Experiment results are shown in Table 3. Once again using two auxiliary losses works best, but in contrast to Section 3.1, there is a large performance boost relative to using either one alone. Performance using only the source recognition loss is very poor, indicating that learning alignment on this task is especially difficult without strong supervision on the translation task.

We found that 4-head attention works better than one head, unlike the conversational task, where both attention mechanisms had similar performance. Finally, as in [21], we find that pre-training the bottom 6 encoder layers on an ST task improves BLEU scores by over 5 points. This is the best performing direct S2ST model, obtaining 76% of the baseline performance.

3.3. Subjective evaluation of speech naturalness

To evaluate synthesis quality of the best performing models from Tables 2 and 3 we use the framework from [15] to crowdsource 5-point MOS evaluations based on subjective listening tests. 1k examples were rated for each dataset, each one by a single rater. Although this evaluation is expected to be independent of the correctness of the translation, translation errors can result in low scores for examples raters describe as "not understandable".

Results are shown in Table 4, comparing different vocoders where results with Griffin-Lim correspond to identical model configurations as Sections 3.1 and 3.2. As expected, using WaveRNN vocoders dramatically improves ratings over Griffin-Lim into the "Very Good" range (above 4.0). Note that it is most fair

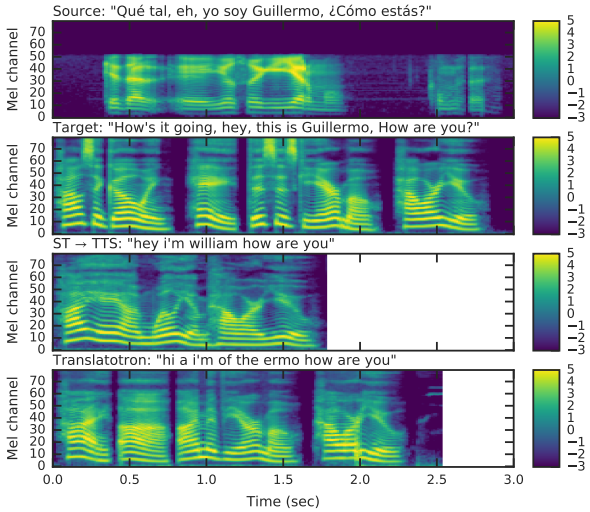

Figure 2: *Mel spectrograms of input (top, upsampled to 24 kHz) and WaveRNN vocoder output (bottom) waveforms from a Fisher corpus example, along with ASR transcripts. Note that the spectrogram scales are different to the model inputs and outputs. Corresponding audio is on the companion website.*

to compare the Griffin-Lim results to the ground truth training targets since they were generated using corresponding lower quality vocoders. In such a comparison it is clear that the S2ST models do not score as highly as the ground truth.

Finally, we note the similar performance gap between Translatotron and the baseline under this evaluation. In part, this is a consequence of the different types of errors made by the two models. For example, Translatotron sometimes mispronounces words, especially proper nouns, using pronunciations from the source language, e.g. mispronouncing the /ae/ vowel in "Dan" as /ah/, consistent with Spanish but sounding less natural to English listeners, whereas by construction, the baseline consistently projects results to English. Figure 2 demonstrates other differences in behavior, where Translatotron reproduces the input "eh" disfluency (transcribed as "a", between $0.4 - 0.8$ sec in the bottom row of the figure), but the cascade does not. It is also interesting to note that the cascade translates "Guillermo" to its English form "William", whereas Translatotron speaks the Spanish name (although the ASR model mistranscribes it as "of the ermo"), suggesting that the direct model might have a bias toward more directly reconstructing the input. Similarly, in example 7 on the companion page Translatotron reconstructs "pasejo" as "passages" instead of "tickets", potentially reflecting a bias for cognates. We leave detailed analysis to future work.

3.4. Cross language voice transfer

In our final experiment, we synthesize translated speech using the voice of the source speaker by training the full model depicted in Figure 1. The speaker encoder is conditioned on the ground truth target speaker during training. We use a subset of the data from Sec. 3.1 for which we have paired source and target recordings. Note that the source and target speakers for each pair are always different – the data was not collected from bilingual speakers. This dataset contains 606k utterance pairs, resampled to 16 kHz, with 863 and 493 hours of source and target speech, respectively; 6.3k pairs, a subset of that from Sec. 3.1, are held out for testing. Since target recordings contained noise, we apply the denoising and volume normalization from [15] to improve output quality.

Table 5 compares performance using different conditioning strategies. The top row transfers the source speaker's voice to the translated speech, while row two is a "cheating" configuration since the speaker embedding can potentially leak information about the target content to the decoder. To verify that this does not negatively impact performance we also condition on random target utterances in row three. In all cases performance is worse than models trained on synthetic targets in Tables 2 and 4. This is because the task of synthesizing arbitrary speakers is more difficult; the training targets are much noisier and training set is much smaller; and the ASR model used for evaluation makes more errors on the noisy, multispeaker targets. In terms of BLEU score, the difference between conditioning on ground truth and random targets is very small, verifying that content leak is not a concern (in part due to the low speaker embedding dimension). However conditioning on the source trails by 1.8 BLEU points, reflecting the mismatch in conditioning language between the training and inference configurations. Naturalness MOS scores are close in all cases. However, conditioning on the source speaker significantly reduces similarity MOS by 1.4 points. Again this suggests that using English speaker embeddings during training does not generalize well to Spanish speakers.

Table 5: *Voice transfer performance when conditioned on source, ground truth target, or a random utterance in the target language. References for MOS-similarity match the conditioning speaker.*

Speaker Emb	BLEU	MOS-naturalness	MOS-similarity
Source	33.6	3.07 ± 0.08	1.85 ± 0.06
Target	36.2	3.15 ± 0.08	3.30 ± 0.09
Random target	35.4	3.08 ± 0.08	3.24 ± 0.08
Ground truth	59.9	4.10 ± 0.06	-

4. Conclusions

We present a direct speech-to-speech translation model, trained end-to-end. We find that it is important to use speech transcripts during training, but no intermediate speech transcription is necessary for inference. Exploring alternate training strategies which alleviate this requirement is an interesting direction for future work. The model achieves high translation quality on two Spanish-to-English datasets, although performance is not as good as a baseline cascade of ST and TTS models.

In addition, we demonstrate a variant which simultaneously transfers the source speaker's voice to the translated speech. The voice transfer does not work as well as in a similar TTS context [15], reflecting the difficulty of the cross-language voice transfer task, as well as evaluation [44]. Potential strategies to improve voice transfer performance include improving the speaker encoder by adding a language adversarial loss, or by incorporating a cycle-consistency term [13] into the S2ST loss.

Other future work includes utilizing weakly supervision to scale up training with synthetic data [21] or multitask learning [19,20], and transferring prosody and other acoustic factors from the source speech to the translated speech following [45–47].

5. Acknowledgements

The authors thank Quan Wang, Jason Pelecanos and the Google Speech team for providing the multilingual speaker encoder, Tom Walters and the Deepmind team for help with WaveNet TTS, Quan Wang, Heiga Zen, Patrick Nguyen, Yu Zhang, Jonathan Shen, Orhan Firat, and the Google Brain team for helpful discussions, and Mengmeng Niu for data collection support.

6. References

[1] A. Lavie, A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavalda, T. Zeppenfeld, and P. Zhan, "JANUS-III: Speech-to-speech translation in multiple languages," in *Proc. ICASSP*, 1997.

[2] W. Wahlster, *Verbmobil: Foundations of speech-to-speech translation*. Springer, 2000.

[3] S. Nakamura, K. Markov, H. Nakaiwa, G.-i. Kikui, H. Kawai, T. Jitsuhiro, J.-S. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto, "The ATR multilingual speech-to-speech translation system," *IEEE Transactions on Audio, Speech, and Language Processing*, 2006.

[4] International Telecommunication Union, "ITU-T F.745: Functional requirements for network-based speech-to-speech translation services," 2016.

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

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

[7] E. Vidal, "Finite-state speech-to-speech translation," in *Proc. ICASSP*, 1997.

[8] F. Casacuberta, H. Ney, F. J. Och, E. Vidal, J. M. Vilar *et al.*, "Some approaches to statistical and finite-state speech-to-speech translation," *Computer Speech and Language*, vol. 18, no. 1, 2004.

[9] P. Aguero, J. Adell, and A. Bonafonte, "Prosody generation for speech-to-speech translation," in *Proc. ICASSP*, 2006.

[10] Q. T. Do, S. Sakti, and S. Nakamura, "Toward expressive speech translation: a unified sequence-to-sequence LSTMs approach for translating words and emphasis," in *Proc. Interspeech*, 2017.

[11] T. Kano, S. Takamichi, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "An end-to-end model for cross-lingual transformation of paralinguistic information," *Machine Translation*, pp. 1–16, 2018.

[12] M. Kurimo, W. Byrne, J. Dines, P. N. Garner, M. Gibson, Y. Guan, T. Hirsimäki, R. Karhila, S. King, H. Liang *et al.*, "Personalising speech-to-speech translation in the EMIME project," in *Proc. ACL 2010 System Demonstrations*, 2010.

[13] E. Nachmani, A. Polyak, Y. Taigman, and L. Wolf, "Fitting new speakers based on a short untranscribed sample," in *ICML*, 2018.

[14] S. O. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," in *Proc. NeurIPS*, 2018.

[15] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Proc. NeurIPS*, 2018.

[16] Y. Chen, Y. Assael, B. Shillingford, D. Budden, S. Reed, H. Zen, Q. Wang, L. C. Cobo, A. Trask, B. Laurie *et al.*, "Sample efficient adaptive text-to-speech," in *Proc. ICLR*, 2019.

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

[18] A. Bérard, L. Besacier, A. C. Kocabiyikoglu, and O. Pietquin, "End-to-end automatic speech translation of audiobooks," in *Proc. ICASSP*, 2018.

[19] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly translate foreign speech," in *Proc. Interspeech*, 2017.

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

[21] Y. Jia, M. Johnson, W. Macherey, R. J. Weiss, Y. Cao, C.-C. Chiu, N. Ari *et al.*, "Leveraging weakly supervised data to improve end-to-end speech-to-text translation," in *Proc. ICASSP*, 2019.

[22] A. Haque, M. Guo, and P. Verma, "Conditional end-to-end audio transforms," in *Proc. Interspeech*, 2018.

[23] J. Zhang, Z. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019.

[24] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanevsky, and Y. Jia, "Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation," in *Proc. Interspeech*, 2019.

[25] M. Guo, A. Haque, and P. Verma, "End-to-end spoken language translation," *arXiv preprint arXiv:1904.10760*, 2019.

[26] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang *et al.*, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, 2017.

[27] A. F. Machado and M. Queiroz, "Voice conversion: A critical survey," in *Proc. Sound and Music Computing*, 2010, pp. 1–8.

[28] C.-C. Chiu, T. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. Weiss, K. Rao *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *Proc. ICASSP*, 2018.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, 2017.

[30] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017.

[31] 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:1609.08144*, 2016.

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

[33] D. Krueger, T. Maharaj, J. Kramár, M. Pezeshki, N. Ballas, N. R. Ke, A. Goyal, Y. Bengio *et al.*, "Zoneout: Regularizing RNNs by randomly preserving hidden activations," in *Proc. ICLR*, 2017.

[34] N. Shazeer and M. Stern, "Adafactor: Adaptive learning rates with sublinear memory cost," in *Proc. ICML*, 2018, pp. 4603–4611.

[35] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[36] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. v. d. Oord, S. Dieleman *et al.*, "Efficient neural audio synthesis," in *Proc. ICML*, 2018.

[37] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, "Fully supervised speaker diarization," in *Proc. ICASSP*, 2019.

[38] M. Post, G. Kumar, A. Lopez, D. Karakos, C. Callison-Burch *et al.*, "Improved speech-to-text translation with the Fisher and Callhome Spanish–English speech translation corpus," in *Proc. IWSLT*, 2013.

[39] J. Shen, P. Nguyen, Y. Wu, Z. Chen *et al.*, "Lingvo: a modular and scalable framework for sequence-to-sequence modeling," 2019.

[40] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *ACL*, 2002.

[41] K. Irie, R. Prabhavalkar, A. Kannan, A. Bruguier, D. Rybach, and P. Nguyen, "Model unit exploration for sequence-to-sequence speech recognition," *arXiv:1902.01955*, 2019.

[42] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015.

[43] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche *et al.*, "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proc. ICML*, 2018.

[44] M. Wester, J. Dines, M. Gibson, H. Liang *et al.*, "Speaker adaptation and the evaluation of speaker similarity in the EMIME speech-to-speech translation project," in *ISCA Tutorial and Research Workshop on Speech Synthesis*, 2010.

[45] Y. Lee and T. Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," in *Proc. ICASSP*, 2019.

[46] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor *et al.*, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. ICML*, 2018.

[47] W.-N. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen *et al.*, "Hierarchical generative modeling for controllable speech synthesis," in *Proc. ICLR*, 2019.