Can Synthetic Speech Improve End-to-End Conversational Speech Translation?

Bismarck Bamfo Odoom[1] bodoom1@jhu.edu
Nathaniel R. Robinson[1] nrobin38@jhu.edu
Elijah Rippeth[2] erip@cs.umd.edu
Luis Tavarez-Arce[3] ltavare1@jhu.edu
Kenton Murray[1] kenton@jhu.edu
Matthew Wiesner[1] wiesner@jhu.edu
Paul McNamee[1] mcnamee@jhu.edu
Philipp Koehn[1] phi@jhu.edu
Kevin Duh[1] kevinduh@cs.jhu.edu
[1] Johns Hopkins University, Baltimore, MD, USA
[2] University of Maryland, College Park, MD, USA
[3] SCALE 2023 Workshop Participant

Abstract

Conversational speech translation is an important technology that fosters communication among people of different language backgrounds. Three-way parallel data in the form of source speech, source transcript, and target translation is usually required to train end-to-end systems. However, such datasets are not readily available and are expensive to create as this involves multiple annotation stages. In this paper, we investigate the use of synthetic data from generative models, namely machine translation and text-to-speech synthesis, for training conversational speech translation systems. We show that adding synthetic data to the training recipe increasingly improves end-to-end training performance, especially when limited real data is available. However, when no real data is available, no amount of synthetic data helps.

1 Introduction

The growing globalization of our society requires effective technologies that foster communication among individuals of varying language backgrounds. Speech translation is an important technology that fosters everyday communication among individuals from different language backgrounds, bridging cultural and linguistic barriers. The technology has improved dramatically in recent years thanks to deep learning, but most gains have been demonstrated on formal settings such as parliamentary speeches, prepared monologues, and university lectures. Informal conversations pose significant challenges due to the lack of training data. Conversations deviate from formal written language and include informal expressions, slang, overlapping speech, incomplete sentences, varying intonation, pace, and emotion, which are typically not present in standard speech translation datasets. To capture these nuances, conversational datasets are essential for training models to understand and translate real-life spoken language accurately. However, creating this type of data usually involves individuals talking on the telephone for hours about various topics, followed by multiple annotation stages involving segmenting the long-form speech into chunks, transcribing the various chunks, and then translating them into the target language. Executing these tasks is tedious, time-consuming, and expensive.

This motivates a new approach of utilizing synthetic data from generative models. Generative

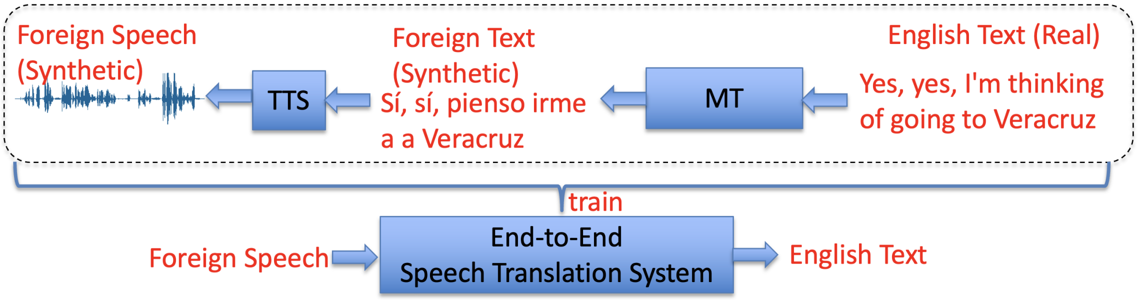

Figure 1: Overview of approach in generating three-way parallel data (foreign speech, foreign text, english text) for training end-to-end conversational speech translation systems.

models present an attractive solution by being able to produce large amounts of synthetic conversational speech quickly in a flexible and cost-effective manner. When used as a data augmentation technique, this synthetic data can potentially improve the performance of speech translation systems in conversational speech domains.

Over the years, machine translation has seen significant advances. Machine translation (MT) models utilizing the transformer architecture when trained on large data sets across multiple languages generalize better and can generate adequate and fluent translations in near real-time. In addition, text-to-speech synthesis (TTS) has attained significant strides resulting in high-quality synthesized voices that closely mimic human speech. In this study, we leverage the advances from MT and TTS to show that a synthetic conversational speech dataset that is easier and cheaper to create can be used for the task of conversational end-to-end speech translation. To do this, we create a dataset of synthetic speech by back-translating monolingual text from the target language to the source language, and then generating the speech in the source language speech from the back-translated text using a TTS system (Figure 1).

We seek to answer the following questions:

1. Does incorporating synthetic data into the training recipe help end-to-end training for conversational speech translation?
2. How do we use synthetic data effectively for end-to-end conversational speech translation?
3. Can synthetic data be used in place of real data for conversational end-to-end speech translation?

2 Background and Related Work

Synthetic Data The use of synthetic data has been extensively studied for text-based machine translation. Sennrich et al. (2016a) shows that generating synthetic source sentences from target monolingual data through back-translation helps boost neural machine translation performance. Amin et al. (2021) investigate the use of synthetic data for training RNN-T ASR models via a multi-stage training pipeline with continual learning. Rossenbach et al. (2020) show that training attention-based ASR systems on synthetic data leads to huge improvements in word-error-rate (WER). Rossenbach et al. (2021) compare the benefits of training with synthetic data for four ASR architectures, namely - attention encoder-decoder (AED), hybrid ASR, CTC, and monotic RNN-T. Fang and Feng (2023) train a target-to-unit model to map the target text to source speech units (Lee et al., 2022). They then utilize a unit vocoder to map the source units into a waveform. Robinson et al. (2022); Karakasidis et al. (2023) expand these augmentation methods to low-resource and accented ASR, respectively.

Text-to-Speech Synthesis The task of text-to-speech synthesis (TTS) is to generate an output speech corresponding to an input transcript. Early techniques such as formant synthesis used the source-filter model for intelligibility but lacked naturalness. Modern neural synthesis methods, such as Tacotron2 (Shen et al., 2018), TransformerTTS (Li et al., 2019), FastSpeech (Ren et al., 2019) and VITS (Kim et al., 2021) simplify the pipeline and deliver high-quality voice output by leveraging deep learning. Tacotron2 employs an autoregressive decoder with attention mechanisms, while TransformerTTS replaces RNNs with Transformers for faster training. FastSpeech optimizes the process by using non-autoregressive methods, addressing the speed limitations of previous models. VITS employs a conditional variational autoencoder augmented with normalizing flows and an adversarial training process which enhances the quality of synthesized speech.

Speech Translation Speech translation research has seen a revival in recent years. For example, the IWSLT 2023 campaign showcased a variety of tasks, including multilingual speech translation, speech-to-speech translation, low-resource speech translation, automatic dubbing or subtitling, and simultaneous speech translation (Agarwal et al., 2023). Both cascaded systems consisting of speech recognition and machine translation components as well as end-to-end direct speech translation systems have been explored. End-to-end systems can be trained with a combination (Babu et al., 2022) of 2-fold parallel data or via multi-task learning (Radford et al., 2023). In the majority of cases, the training data for these systems come from TED talks, university lectures, conference presentations, European parliamentary speeches. These are prepared, public talks which exhibit different characteristics from the informal multiparty conversations of interest here.

3 Data Creation Methods

We use the term three-way parallel data to refer to the aligned source speech, source transcript, and target translation that is necessary for end-to-end speech translation model training. We first describe the manually-created 3-way parallel data (referred to as **Real** in subsequent sections) used for baseline models. We then explain the synthetically generated 3-way parallel data (which we refer to as **Synth** in subsequent sections) used for data augmentation.

3.1 Real 3-way Parallel Data

We use the Fisher-Callhome Spanish-English dataset (Post et al., 2013), a three-way conversational telephone speech dataset consisting of Spanish speech, Spanish transcript, and English text translation for our experiments. While we argue that there is a lack of manually-created conversational data, this dataset is a rare exception: it is an immensely large dataset by academic research standards, created by crowdsourcing translations of an existing transcribed speech recognition dataset. The reason we chose this dataset is that is enables us to perform data ablation experiments to understand how much real data is needed in a data augmentation setup.

For preprocessing, we resample the audio to 16kHz and apply speed perturbation (0.9, 1.0, 1.1). The audio is transformed into a 80 dimensional log-filterbank and we apply specaugment (Park et al., 2019) with bi-cubic time-warping. We use byte-pair-encoding (BPE) tokenization (Sennrich et al., 2016b) with a vocabulary of size 4000.

3.2 Synthetic 3-way Parallel Data

To create the synthetic speech-text pairs, we use over 500,000 lines of conversational-style text in English. This text was collected from the English translations of various conversational speech datasets (Ansari et al., 2020; Song et al., 2014). We back-translate (Sennrich et al., 2016a) this text using the `nllb-200-1.3B`[1] multilingual machine translation model (Team et al., 2022) into Spanish. Spanish speech is synthesized by feeding the back-translated text into the VITS [2] text-to-speech system. Specifically, we use the VITS model trained on CSS10 Spanish (Park and Mulc, 2019) then apply voice conversion using freevc24[3]. The target speakers used for voice conversion are the speakers for the original files; future work is to explore more diversity in speakers by sampling in speaker embedding space (Jia et al., 2019).

To illustrate the whole pipeline with a concrete example, we begin with a Callhome Chinese file spoken by speaker A: First, we translate the English text portion to Spanish text. Second, we synthesize a generic Spanish voice using VITS. Finally, we apply voice conversion with speaker A as the target speaker, generating a Spanish voice that sounds like the original Chinese speaker A. This

[1] https://huggingface.co/facebook/nllb-200-1.3B

[2] https://github.com/coqui-ai/TTS

[3] https://github.com/OlaWod/FreeVC

Corpus	Lang	#Hours		
		Train	Dev	Test
Real: Spanish Fisher/Callhome	Sp-En	170	9.3	4.5/1.8
Synth: backtranslation + TTS	Sp-En	520	-	-

Table 1: Dataset statistics showing the number of hours of both real and synthetic speech. We use both the Fisher and the Callhome test sets.

procedure is repeated independently for each file that we wish to add to the augmentation dataset. This process yielded about 520 hours of synthetic speech in Spanish. The resulting dataset consists of synthetic speech in Spanish, back-translated text in Spanish (transcript), and the English text (translation). We apply the same preprocessing techniques in Section 3.1 and refer to this dataset as **Synth** in subsequent sections.

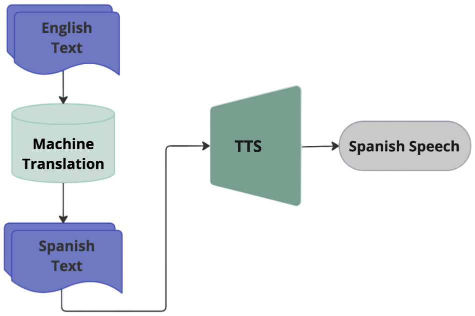

Figure 2: The synthetic data generation pipeline. We collect over 500,000 lines of English conversational style text and translate them into Spanish. We pass the back-translated text into a text-to-speech synthesis (TTS) system to synthesize Spanish speech from the transcript.

3.2.1 Quality of Synthetic Speech

We use the NISQA model (Mittag et al., 2021) to analyze the speech quality. The model consists of a convolutional neural network with a self-attention and attention-pooling block. The model predicts the mean opinion score which is a common metric used to measure the quality of TTS generated speech. The model predicts a mean opinion score of 4.29 out of 5 signifying that the synthesized speech is of high quality. Text references are not available to measure the translation quality (e.g. BLEU) of the synthetic text used to generate the synthetic speech. However, we know that NLLB is generally a strong model for this language pair; while domain differences may degrade text translation, a manual check of a small subset of translations reveals that they mostly preserve the semantics.

4 Speech Translation Model

4.1 Model Architecture

The speech encoder is based on the conformer architecture (Gulati et al., 2020), which combines the strengths of convolutional neural networks (CNNs) and Transformer models to handle the speech input efficiently. We use 8 conformer blocks with 16 attention heads within its multi-head self-attention modules, enabling the model to focus on different segments of the input sequence concurrently. Each Conformer block contains feed-forward networks with 2048 linear units. We use relative positional encodings and relative self-attention mechanisms, the swish activation function is used, and dropout of 0.1.

The text decoder is a Transformer model (Vaswani et al., 2017) featuring 8 blocks with 2048 linear units each. The ReLU activation function is used and a dropout rate of 0.1. The total number of trainable parameters is 38.7M. We initialize all models from scratch and train on 2 NVIDIA V100 32GB GPUs. All models are trained for 50 epochs with batch size of 64.

4.2 Data Augmentation Scheme

We perform a simple data augmentation scheme: concatenating the **Real** data in Section 3.1 and the **Synth** data as one training set. The training objective treats samples from both datasets in the same way, with no specific up-sampling or down-sampling. More advanced methods are conceivable, such as pre-training on **Synth** and fine-tuning **Real** and modifying the training objective to treat real and synthetic data differently. In this work, we focused on the simple data concatenation, with experiments focusing on different data proportions, to more easily study the impact of synthetic data.

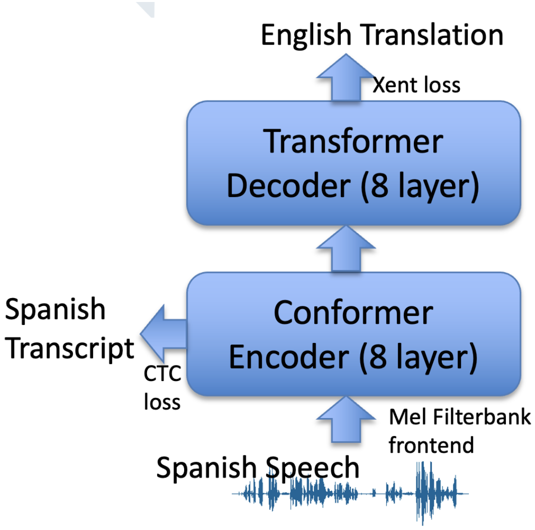

Figure 3: Model architecture. The frontend extracts log-mel filterbank features from the speech. The encoder consists of 8 conformer blocks and the decoder is a transformer decoder featuring 8 transformer blocks.

5 Experiments and Results

This section outlines the various experiments and results of this study. We use the ESPnet (Watanabe et al., 2018) toolkit for all experiments. We evaluate all systems with BLEU (Papineni et al., 2002), as implemented by SacreBLEU[4] (Post, 2018).

5.1 Training on only real data

Firstly, we train multiple systems on only real speech. These systems are treated as baselines. We train with 5, 10, 20, 50, 100, and 170 hours of **Real** speech. As shown in Table 2, the performance continues to improve as we scale the size of the dataset.

5.2 Training with both real and synthetic data

Here, we investigate whether incorporating **Synth** data into our training recipe helps end-to-end training. We do this by training the same system end-to-end on a dataset comprising both real and synthetic speech.

5.2.1 Fixing amount of real speech and varying amount of synthetic speech

In many circumstances, due to the expensive nature of collecting conversational speech data, practitioners have a fixed budget of data to train models on. We mimic this situation by fixing the amount of real speech in the training recipe and then progressively increasing the amount of synthetic data in the training recipe. We simulate the low data resource case where we have less than 10 hours of real speech, the mid-data resource case, where we have about 50 hours of real speech, and the relatively high resource data case where we have over a hundred hours of real speech. As shown in Table 2, there is a consistent improvement in model performance with increased synthetic speech when the amount of real speech remains fixed. This suggests that given a fixed amount of real training speech, including synthetic speech improves performance. Robinson et al. (2022) corroborated this trend for ASR.

5.2.2 How much improvement do we get?

We observe that when there is a small amount of real speech, including a large amount of synthetic speech can lead to substantial improvements in model performance as compared to when there is a higher amount of real speech. This is particularly useful for low-data resource scenarios.

From Table 2, we observe the cases where there are 5 and 170 hours of **Real** and **Synth** data, respectively. Adding these 170 hours of **Synth** results in +9.1 and +9.0 BLEU over the baseline trained on only real speech for Fisher and CallHome test sets, respectively. Increasing the amount of synthetic speech in the training recipe to 340 hours results in +10.7 and +9.7 BLEU compared to the baseline. Further increasing the amount of synthetic speech to 520 hours results in +14.7 and +11.8 BLEU. For the high-resource case with 170 hours of real speech, adding 170 hours of synthetic speech results in +1.4 and +2.3 BLEU on Fisher and CallHome test sets, respectively. Doubling the amount of synthetic speech results in +1.1 and +2.0 BLEU, and increasing to 520 hours gives +1.5 and +2.0 BLEU. This suggests that when there is already a large amount of real speech in training, including more synthetic speech does not provide significant additional benefits. The real data likely captures most relevant vari-

[4] Signature:BLEU+case:mixed+nrefs:1+tok:13a+smooth:exp+version:2.3.1

Training data (hours)		BLEU	
Real	**Synth**	**Fisher ↑**	**Callhome ↑**
0	0	0	0
	170	0	0
	340	0	0
	520	0	0
5	0	0.6	0.7
	170	9.7	9.7
	340	11.3	10.4
	520	**15.3**	**12.5**
10	0	0.7	1.1
	170	12.2	12.5
	340	13.9	13.4
	520	**17.0**	**14.9**
20	0	5.8	5.6
	170	16.7	15.3
	340	17.6	16.0
	520	**19.3**	**16.4**
50	0	14.3	12.3
	170	21.9	17.6
	340	22.1	18.2
	520	**22.9**	**19.0**
100	0	21.7	17.0
	170	24.5	20.3
	340	25.1	20.0
	520	**24.9**	**20.2**
170	0	25.3	20.1
	170	26.7	**22.4**
	340	26.4	22.1
	520	**26.8**	22.1

Table 2: BLEU scores of systems trained on varying amounts of the Real and Synth. When Synth is 0 the system was trained on only Real. When real is 0, the system was trained on only Synth.

ations, and the synthetic data may not add much new information.

5.3 Training on only synthetic data

To explore an extreme scenario, we conducted experiments where no real speech was included in the training setup. Instead, models were trained only on **Synth**. This approach of relying solely on synthetic data for training poses domain adaptation challenges. The models must generalize from the synthetic training environment, which may not fully capture the nuances and variations present in real speech. Consequently, we observed performance discrepancies when these models, trained only on synthetic data, were applied to real speech.

5.3.1 Does training on only synthetic data work?

When evaluated on the **Real** test set, the model obtains a BLEU score of 0, signifying a complete lack of generalizability due to the absence of real-world data during training. However, we obtain up to 30 BLEU on our best system when we evaluate on **Synth-Fisher** and **Synth-CallHome**, which are versions of the real test set where the input speech is synthesized using the same TTS system. (See Table 6.) This shows that the systems trained on only synthetic data do not generalize outside the synthetic data domain (though it suggests speech translation models trained only on synthetic data could theoretically be paired with voice conversion to accomplish speech translation, by converting real voices to synthetic voices before inference). See §5.3.3 for more on this analysis on this trend.

5.3.2 Bridging the generalization gap

Our experiments show that this generalization gap is mitigated by incorporating a small amount of real data into the training recipe. This helps the model generalize beyond the synthetic domain. When trained on only the synthetic data, the system cannot model the noisy channel effects introduced, as

Spanish Trancript	**English Translation**	**Synth-170 Translation**	**Synth-170-Real-5 Translation**
Sí, eso es para eso, de seguro. No importa.	Yes, that's what's for, sure. It doesn't matter.	Uh uh	Yes, it's for Suhur, it's not matter
Y qué estudia, mama, qué están estudiando.	And what's she studying, mom, what career.	Uh uh	And that's all, mom, who's studying
mmm sí eso pasa aquí en Estados Unidos acá pa- casi demandan a la empresa	hmmm, if that happens here in the United States, they, they would sue the company	Uh uh	And if that happens here is a company in Canada

Table 3: Example translations to show how adding a small amount of real data to a synthetic training recipe helps the model generalize beyond the synthetic domain. Synth-170 is the system trained on 170 hours of synthetic data. Synth-170-Real-5 is the system trained on 170 hours of **Synth** data and 5 hours of **Real**

Spanish Trancript	**English Translation**	**Synth-340 Translation**	**Synth-340-Real-5 Translation**
Sí, eso es para eso, de seguro. No importa.	Yes, that's what's for, sure. It doesn't matter.	ah ah	yes, for that, of course, it doesn't matter
Y qué estudia, mama, qué están estudiando.	And what's she studying, mom, what career.	no no	and what was my mom? What are you studying?
mmm sí eso pasa aquí en Estados Unidos acá pa- casi demandan a la empresa	hmmm, if that happens here in the United States, they, they would sue the company	ah no	and that happens here in the United States, send me a company

Table 4: Example translations to show how adding a small amount of real data to a synthetic training recipe helps the model generalize beyond the synthetic domain. Synth-340 is the system trained on 340 hours of synthetic data. Synth-340-Real-5 is the system trained on 340 hours of **Synth** data and 5 hours of **Real**

Spanish Trancript	**English Translation**	**Synth-520 Translation**	**Synth-520-Real-5 Translation**
Sí, eso es para eso, de seguro. No importa.	Yes, that's what's for, sure. It doesn't matter.	And, and	yes, that's for sure, they don't matter
Y qué estudia, mama, qué están estudiando.	And what's she studying, mom, what career.	in, in, in, in	And that's it, mom, what's she studying?
mmm sí eso pasa aquí en Estados Unidos acá pa- casi demandan a la empresa	hmmm, if that happens here in the United States, they, they would sue the company	Right, right, In	hmmm, that happens here in the United States, they would sue the company

Table 5: Example translations to show how adding a small amount of real data to a synthetic training recipe helps the model generalize beyond the synthetic domain. Synth-520 is the system trained on 520 hours of synthetic data. Synth-530-Real-5 is the system trained on 520 hours of **Synth** data and 5 hours of **Real**

Dataset	Hours	Synth-Fisher ↑	Synth-CallHome ↑
Synth	5	0.0	0.0
	10	0.1	0.4
	20	0.2	0.3
	50	4.5	5.5
	100	2.3	1.6
	170	5.8	6.9
	340	11.4	8.5
	520	**30.7**	**26.0**

Table 6: BLEU scores for training on varying hours of only synthetic speech and testing on synthetic speech testsets

the data was collected from telephone conversations. Introducing a small amount of real data likely helps the system model the acoustic mismatch. The performance increases further if the amount of synthetic data is increased. For example, in Table 2, given **Synth** amounts 170, 340, and 520 hours, when we add 5 hours of **Real** data to each, the 520-hour recipe does best. We display example outputs for these systems compared with those trained on no **Real** data in Tables 3, 4, and 5.

5.3.3 Inference on Synthetic Data

As mentioned in §5.3.1, models trained on only synthetic speech do not generalize to real speech. We look deeper by evaluating the performance of the models on **Synth-CallHome** and **Synth-Fisher**, the synthetic versions of the real test sets. We use this as a proxy to examine the claim that training on only synthetic data may not generalize to real test sets. As shown in Table 6, models trained on only synthetic data perform well on synthetic test sets (inputs that match the acoustic conditions of their training data), though they cannot perform at all for real test sets.

6 Takeaways

We summarize our findings here: (1) When there is a small amount of real speech available, including a large amount of synthetic speech leads to higher performance gains in end-to-end training; (2) When there is a large amount of real data available, including synthetic data leads to minimal performance gains; (3) Training only on synthetic speech data does not generalize outside of the synthetic domain; (4) To generalize outside the synthetic data domain, some amount of real speech has to be present in the training recipe (5) When there is no real data present, no amount of synthetic data helps.

7 Conclusion

We investigated whether using synthetic data generated from backtranslation and text-to-speech synthesis for end-to-end conversational speech translation improves performance. Incorporating synthetic data into a conversation speech translation training recipe helps improve the overall system's performance especially when there is limited real speech available. When models are trained on only synthetic data, we find that models do not generalize beyond their training domain. This mismatch between the synthetic training data and real-world data leads to suboptimal performance when models trained on only synthetic data are applied to real speech. This highlights the importance of incorporating at least some real speech data during training to bridge the domain gap effectively. In the case where no real speech is available, no amount of synthetic data helps.

There are several open questions worth examining as future work:

- What happens if a self-supervised pre-trained speech encoder like wav2vec (Baevski et al., 2020) is incorporated into the model? Would it be more or less robust to synthetic data?

- What happens if TTS quality is much lower, which is likely in lower-resource languages? (For example, it would be instructive to repeat the experiments with other languages.)

- Would the conclusions change if we examine more advanced augmentation besides simple concatenation of **Real** and **Synth**?

- Could a model trained only on synthetic speech be paired with voice conversion to accomplish speech translation?

References

Agarwal, M., Agrawal, S., Anastasopoulos, A., Bentivogli, L., Bojar, O., Borg, C., Carpuat, M., Cattoni, R., Cettolo, M., Chen, M., Chen, W., Choukri, K., Chronopoulou, A., Currey, A., Declerck, T., Dong, Q., Duh, K., Estève, Y., Federico, M., Gahbiche, S., Haddow, B., Hsu, B., Mon Htut, P., Inaguma, H., Javorský, D., Judge, J., Kano, Y., Ko, T., Kumar, R., Li, P., Ma, X., Mathur, P., Matusov, E., McNamee, P., P. McCrae, J., Murray, K., Nadejde, M., Nakamura, S., Negri, M., Nguyen, H., Niehues, J., Niu, X., Kr. Ojha, A., E. Ortega, J., Pal, P., Pino, J., van der Plas, L., Polák, P., Rippeth, E., Salesky, E., Shi, J., Sperber, M., Stüker, S., Sudoh, K., Tang, Y., Thompson, B., Tran, K., Turchi, M., Waibel, A., Wang, M., Watanabe, S., and Zevallos, R. (2023). FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In Salesky, E., Federico, M., and Carpuat, M., editors, *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Amin, F., Yang, W., Liu, Y., Barra-Chicote, R., Meng, Y., Maas, R., and Droppo, J. (2021). Synthasr: Unlocking synthetic data for speech recognition. *arXiv preprint arXiv:2106.07803*.

Ansari, E., Axelrod, A., Bach, N., Bojar, O., Cattoni, R., Dalvi, F., Durrani, N., Federico, M., Federmann, C., Gu, J., Huang, F., Knight, K., Ma, X., Nagesh, A., Negri, M., Niehues, J., Pino, J., Salesky, E., Shi, X., Stüker, S., Turchi, M., Waibel, A., and Wang, C. (2020). FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34.

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., and Auli, M. (2022). XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech 2022*, pages 2278–2282.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations.

Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Fang, Q. and Feng, Y. (2023). Back translation for speech-to-text translation without transcripts. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4567–4587, Toronto, Canada. Association for Computational Linguistics.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Hu, T.-Y., Armandpour, M., Shrivastava, A., Chang, J.-H. R., Koppula, H., and Tuzel, O. (2021). Synt++: Utilizing imperfect synthetic data to improve speech recognition. *arXiv preprint arXiv:2110.11479*.

Jia, Y., Johnson, M., Macherey, W., Weiss, R. J., Cao, Y., Chiu, C.-C., Ari, N., Laurenzo, S., and Wu, Y. (2019). Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7180–7184.

Karakasidis, G., Robinson, N., Getman, Y., Ogayo, A., Al-Ghezi, R., Ayasi, A., Watanabe, S., Mortensen, D. R., and Kurimo, M. (2023). Multilingual tts accent impressions for accented asr. In *International Conference on Text, Speech, and Dialogue*, pages 317–327. Springer.

Kim, J., Kong, J., and Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *arXiv preprint arXiv 2106.06103*.

Kong, J., Kim, J., and Bae, J. (2020). Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*, pages 17022–17033.

Lee, A., Chen, P.-J., Wang, C., Gu, J., Popuri, S., Ma, X., Polyak, A., Adi, Y., He, Q., Tang, Y., Pino, J., and Hsu, W.-N. (2022). Direct speech-to-speech translation with discrete units. In Muresan, S., Nakov, P., and

Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3327–3339, Dublin, Ireland. Association for Computational Linguistics.

Li, N., Liu, S., Liu, Y., Zhao, S., and Liu, M. (2019). Neural speech synthesis with transformer network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:6706–6713.

Li, Y. A., Han, C., Raghavan, V. S., Mischler, G., and Mesgarani, N. (2023). Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models.

Mittag, G., Naderi, B., Chehadi, A., and Möller, S. (2021). Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Interspeech 2021*. ISCA.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*. ISCA.

Park, K. and Mulc, T. (2019). Css10: A collection of single speaker speech datasets for 10 languages. *Interspeech*.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Post, M., Kumar, G., Lopez, A., Karakos, D., Callison-Burch, C., and Khudanpur, S. (2013). Improved speech-to-text translation with the fisher and callhome Spanish-English speech translation corpus. In *Proceedings of the 10th International Workshop on Spoken Language Translation: Papers*.

Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. (2019). Fastspeech: Fast, robust and controllable text to speech. In *Advances in Neural Information Processing Systems*.

Robinson, N., Ogayo, P., Gangu, S., Mortensen, D. R., and Watanabe, S. (2022). When is tts augmentation through a pivot language useful? In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2022, pages 3538–3542.

Rossenbach, N., Zeineldeen, M., Hilmes, B., Schlüter, R., and Ney, H. (2021). Comparing the benefit of synthetic training data for various automatic speech recognition architectures. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 788–795.

Rossenbach, N., Zeyer, A., Schlüter, R., and Ney, H. (2020). Generating synthetic audio data for attention-based speech recognition systems. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7069–7073.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Improving neural machine translation models with monolingual data. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural machine translation of rare words with subword units. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y., and Wu, Y. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783.

Song, Z., Strassel, S., Lee, H., Walker, K., Wright, J., Garland, J., Fore, D., Gainor, B., Cabe, P., Thomas, T., Callahan, B., and Sawyer, A. (2014). Collecting natural SMS and chat conversations in multiple languages: The BOLT phase 2 corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1699–1704, Reykjavik, Iceland. European Language Resources Association (ELRA).

Team, N., Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., and Wang, J. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Advances in neural information processing systems. volume 30.

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Enrique Yalta Soplin, N., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A., and Ochiai, T. (2018). ESPnet: End-to-end speech processing toolkit. In *Proceedings of Interspeech*, pages 2207–2211.