Using Voicebox-based Synthetic Speech for ASR Adaptation

Hira Dhamyal[1,2], Leda Sari[2], Vimal Manohar[2], Nayan Singhal[2], Chunyang Wu[2], Jay Mahadeokar[2], Matt Le[2], Apoorv Vyas[2], Bowen Shi[2], Wei-Ning Hsu[2], Suyoun Kim[2], Ozlem Kalinli[2]

[1]Carnegie Mellon University
[2]Meta

hyd@andrew.cmu.edu, ledasari@meta.com

Abstract

Automatic Speech Recognition (ASR) model training requires large amounts of paired data, i.e. audio/text pairs. However, such paired data is expensive to collect and even harder to annotate as opposed to using unpaired text data. With increasingly better speech synthesis models, we can now generate natural-sounding speech and utilize large amounts of unpaired text. In this paper, we use the Voicebox model for speech synthesis. Firstly, we assess synthetic speech quality by comparing the amount of synthetic speech required to obtain the same ASR performance as real speech. We find that in noisy settings 10 times more synthetic data than real data is required to achieve equal performance whereas in clean settings, only 7 times more is needed. Secondly, we explore the improvements in the ASR performance brought by the acoustic variability and lexical variability from the unpaired text and synthesized speech. We find that having both acoustic and lexical variability is better than either one individually. Having lexical variability is better on average than acoustic variability when there are smaller amounts of unpaired text, however, acoustic variability becomes more important as the amount of unpaired text increases.

Index Terms: synthetic data, ASR, automatic speech recognition utomatic speech recognition, ASR, unpaired data, text injection, synthetic data, out-of-domain, unpaired text data

1. Introduction

Automatic Speech Recognition (ASR) models require large amounts of paired training data (speech/text pairs) especially when being used in production systems. Such large amounts are expensive and time-consuming to obtain. However unpaired text data is commonly available; i.e. text without corresponding speech. Therefore, it is highly useful to be able to utilize this unpaired text to train ASR models.

Many previous studies have explored training ASR models using only unpaired text data. For example, [1] uses synthesized speech for the unpaired text using Text-To-Speech (TTS) models, [2] trains ASR models on speaker personalized TTS models, [3] proposes text-injection methodologies, i.e., using unpaired text to update parts of the model, [4–6] propose using unpaired text data to train a separate language model which is then utilized in ASR decoding and scoring modules.

In this paper, we adopt a speech synthesis-based approach. Instead of using conventional TTS systems, we make use of a generative speech model, that can perform multiple speech-related tasks, i.e., the Voicebox model [7] to synthesize speech from the unpaired text. Voicebox is a state-of-the-art model for speech synthesis, speech editing, denoising, and other speech tasks. It is trained using masked input speech where the goal is to generate the masked portions of the speech with the guidance of the surrounding audio and the text transcript.

When using synthetic data, the natural question that arises is: *'how good is the speech synthesis for Automatic Speech Recognition?'*. We address this question by training multiple ASR models with increasing hours of synthetic data and measuring how many hours of synthetic data are needed to match the performance of real audio/text pairs.

The next important question that we address in this paper is regarding the *source of the unpaired text data that is synthesized*. We analyze different sources that the textual data can originate from. The text data can either come from the same corpus multiple times or it can come from a different one. First, if the data comes from the same corpus, the audio/text pairs used downstream for ASR training would only differ in the audio part, and hence the data only introduces acoustic variability. Second, if the data comes from a different corpus, the audio/text pairs used for ASR training would differ both in their lexical and acoustic properties. The third setting is where the unpaired text is not synthesized at all, rather the speech representations for unpaired text are generated by averaging over existing paired data as opposed to synthesis, following the strategy described in [3]. This text data comes from a different corpus. In this case, we only change the text part and hence introduce only lexical variability in the data.

In this paper, we investigate the following: (1) how much synthetic data is needed to match the performance of real data on ASR, and (2) the impact of lexical and acoustic variability in unpaired text/synthetic audio pairs on the ASR Model performance.

We find that for the noisy audio test set, 10 times more synthetic data is required to match the performance of real data; and for the clean audio data test set 7 times more synthetic data is required. Secondly, having more acoustic and lexical variability improves ASR performance than just having acoustic variability by a relative 11.2% in the clean setting and by 10.3% in the noisy setting. Acoustic and lexical variability improves ASR performance more than just having lexical variability by a relative 7.9% in the clean setting and by 9.6% in the noisy setting. Having lexical variability is 3.1% relatively better than having acoustic variability in the clean setting and 0.6% relatively better in the noisy setting, however in higher training data setting, acoustic variability is more beneficial than having lexical variability. As expected, the largest gain is achieved when the unpaired text data is from a new source while adding both lexical and acoustic variability to the existing paired data.

2. Related Work

Many studies have tackled the problem of using unpaired text data in training ASR models. For example [1, 8] uses

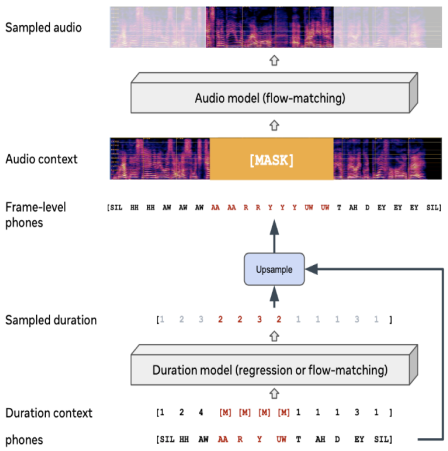

Figure 1: *Overview of the VoiceBox Model [7]*

Tacotron2 [9] TTS model for synthesis and augments with available paired data. Out of these, [1] explores prosodic variation in the synthesis speech by using Global Style Token (GST) to expand prosodic variation, and [8] explores acoustic variation introduced by the different speakers using the synthetic speech generation conditioned speaker representations. In our work, the acoustic diversity is not only conditioned on the speaker but also on the prosody, speech rate, and other acoustic variables. Secondly, in [8] there is no idea of how much augmenting acoustically diverse data relates to the training data scaling, while we are able to relate the two. In [8], lexical variation is investigated by adding new utterances sampled from a language model, which are then synthesized which not only introduces lexical but also acoustic variability, whereas we carefully only investigate lexical variability by using the text-only injection method described in [3]. In addition, compared to the studies, we use one of the latest speech generative models (Voicebox) to synthesize speech and compare its performance to real speech.

Other studies use unpaired text to update the model parameters instead of synthesizing speech. For example, text-injection [3, 10] use averaged audio representations, from the available audio samples, paired with the unpaired text to train the model. In [11–13], the authors use the unpaired text data to pre-train the decoder part of a transformer-based encoder-decoder model, treating it as a language model. Other works like [4, 5, 14] train a separate language model with the unpaired text data, which is then used for scoring the outputs of the decoder of the ASR model. MAESTRO [15] uses paired data to learn the embedding space for the audio and text modalities, followed by a shared embedding space. The unpaired modalities, audio and text separately are then used to update their respective modalities. These studies have shown that making use of unpaired text data helps reduce the WER of an ASR model.

3. Models

3.1. VoiceBox

In this work, for the unpaired text we generate the speech using the VoiceBox model [7]. This model consists of an audio model and a duration model, both of these models are trained using Conditional Flow Matching [16]. It is a non-auto regressive generative model, capable of performing speech editing, denoising, style transfer, and diverse sampling. We denote this model as $\mathcal{V}$.

Name	Paired Data (hr)	Unpaired Text (hr)	Voicebox	Source of Unpaired Data
Baseline	10-100	0	-	-
S	0	100-1000	Yes	Librispeech
A	100	960-2860	Yes	Librispeech
L	100	960-2860	No	Libri-Text
L + A	100	960-2860	Yes	Libri-Text

Table 1: *Different experimental settings used in our work.*

3.2. RNN-T ASR Model

For the ASR models, we use the RNN-T model architecture. We denote this model as $\mathcal{M}$. All implementations used an in-house extension of the PyTorch-based [17] *fairseq* toolkit. We used 80-dimensional log Mel filterbank features that are first projected to 128, then spliced and stacked to 512 dimensions, reducing the sequence length by 4x. The encoder consists of Emformer blocks with 4 attention heads and a 1024-dimensional feed-forward layer. The decoder network contains one 256-dimensional LSTM layers with layernorm and dropout. Both the encoder and decoder outputs are projected to 768 dimensions before passing to an additive joiner, which contains a linear layer with 4097 output BPE units. We use the Adam optimizer, and tri-stage learning rate scheduler, with a peak learning rate of 5×10^{-5}. The model is fine-tuned for 20 epochs and the final model is used for evaluation.

4. Experimental Setup

4.1. Dataset

We use Librispeech [18] which contains both speech and its transcripts. The training subset contains 960 hours of speech, from multiple different speakers, a collection of books read by non-professional speakers. The training, dev, and test subsets in the corpus are divided into portions, namely 'other' and 'clean' based on whether the audios are noisy or clean. The test subset consists of a total of 10.5 hours of data, similarly divided into 'other' and 'clean' subsets.

We utilize the LM corpus provided in LibriSpeech [19] [Libri-Text] which contains 800M tokens and has a vocabulary size of 200k from 14.5k public books from Project Gutenberg. This corpus only contains text, with no associated speech.

For the baseline ASR $\mathcal{M}$ and Voicebox $\mathcal{V}$ model training, we use the in-house video ASR data which consists of 14K-hour manually transcribed social media videos. This is a collection of public and de-identified English videos and contains a diverse range of speakers, accents, topics, and acoustic conditions.

4.2. Experimental Settings

The Voicebox model $\mathcal{V}$ is pre-trained on the in-house video ASR dataset which consists of public social media audio/text pairs. Similarly, $\mathcal{M}$ is also pre-trained on the video ASR dataset and later finetuned on Librispeech or Libri-Text based on the experiment. Since the data for pre-training and testing, video data vs. Librispeech or LibriText, are very different, this is considered an out-of-domain setting. Table 1 summarizes the different experimental settings.

Real Data Baseline: We finetune the ASR model $\mathcal{M}$ on 10, 50, and 100 hours of randomly selected real data pairs from Librispeech train corpus. Each additional data selected contains the previous subset as part of it.

Synthetic Data: We finetune $\mathcal{M}$ on K hours of synthetic data, where K ranges from 100 to 1000 hours. In each experiment,

the model is started from the seed model and fine-tuned on the data. This experiment is labeled as S.

Acoustic Variability: To incorporate only acoustic variability into the ASR Model, the additional unpaired text data, used for speech synthesis, is selected from the same data that the model $\mathcal{M}$ has already seen; i.e. Librispeech train. Since no new lexical variability is being included, the model only sees acoustic variability in the training data. This experiment is labeled as A.

Lexical Variability: We use the J-AT model [3] for this experiment. The model has the same architecture as $\mathcal{M}$. The model estimates average audio embeddings from the paired audio/text in the training data. It pairs these averaged audio embeddings with the unpaired text and trains the ASR model. The unpaired text comes from a new corpus, i.e. Libri-Text. We use this strategy to introduce only lexical variability into the data, since no audio data is being used, there is no acoustic variability. This experiment is labeled as L.

Lexical and Acoustic Variability: To incorporate both acoustic and lexical variability in the ASR model, the additional unpaired text data, used for speech synthesis, is selected from a new corpus; such that the model $\mathcal{M}$ has not seen that data before. Since this text data includes previously unseen words, lexical variability is increased, and in turn acoustic variability increases. This experiment is labeled as L + A. Note that $\mathcal{V}$ can generate varying speech for the same text, where the prosody, speaker's gender, speaking style, or speaking rate may vary. For our purposes, we change the random seed of the model $\mathcal{V}$ to introduce variability in speech given the same text, since we do not have explicit control over these variables.

5. Results and Observations

5.1. Synthetic Speech vs Real Speech

Figure 2 shows the WER decrease when more data is subsequently added to the ASR training. For the test-other subset in Librispeech, it can be observed that with 100 hours of synthetic data, we can beat the performance of 10 hours of real data. To reach the performance of 50 hours of real data, we need about 500 hours of synthetic data and for 100 hours of real data, 1K hours of synthetic data is needed. Therefore about 10 times more synthetic data than real data is needed to achieve similar ASR performance. Similarly, for the cleaner audios in test-clean, about 7 times more synthetic data than real data is required to achieve similar ASR performance.

If we assume the presence of little paired data, i.e. 100 hours, then by using an additional 960 hours of synthetic data, we can beat the performance of 100 hours of real data on test-other (this can be seen in the next section). This is encouraging to see that with little paired data, we can improve the performance by using additional synthetic data. This reduces the real paired data required and therefore makes the ASR fine-tuning for a new domain a relatively cheaper process. For the topline experiment, using all 960 hours of Librispeech-train, we can achieve WER of 11.03 and 4.78 on test-other and test-clean respectively.

Please note that the Librispeech test data is out of domain for the initial ASR model, and this model is fine-tuned only for a small number of iterations. Hence, the WER reported here is much higher than the SoTA numbers which are obtained from a more common setting which involves training only on Librispeech [20, 21]. We try the above experiment multiple times with randomly sampling the hours of data under each experiment and our observations remain consistent.

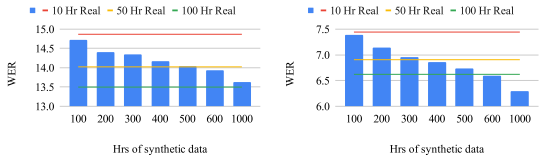

Figure 2: *WER comparison when different hours of synthetic data is included into training. The horizontal lines indicate the WER when only 10, 50 and 100 hours of real data is used in training. Left chart and right chart show the results on test-other and test-clean respectively.*

5.2. Lexical vs Acoustic variability in Synthetic Speech

In the following experiments, we use 100 hours of paired data from Librispeech train. In the first experiment, the remaining 860 hours of Librispeech train are synthesized and added into training. In successive experiments, additional unpaired text is selected and used for speech synthesis (100 hours of audio). This is then included in the training data. We control the lexical and acoustic variability incorporated in the training data by controlling for the source of the additional text/synthesized speech pairs.

To add only acoustic variability (experiment denoted as 'A'), we select an additional 100 hours of data from Librispeech-train. This selected data is already present in the training data in its synthesized form; another version of this synthesis is added to the training data. The Voicebox model introduces variability in the speaker's gender, speaking style, speech rate, and therefore having two synthetic versions of the text, we introduce only acoustic variability into the data.

To add only lexical variability and no acoustic (experiment denoted as 'L'), the additional 100 hours of data is selected from Libri-Text. We use the J-AT [3] strategy as explained before where the unpaired text is paired with averaged audio embeddings and used for model training. Since only new text is included and no audio is used, only lexical variability is incorporated into the model.

To add both acoustic and lexical variability (experiment denoted as 'L + A'), similarly as above, additional 100 hours of data is selected from Libri-Text. This selected data is synthesized and included in the training. Since the text and synthesized speech are new each time, both acoustic and lexical variability are introduced in the data. Figure 3 compares the three different settings.

Firstly, it can be observed that with both lexical and acoustic variability L + A, the performance is better in both test-other and test-clean, as compared to when only acoustic, A, or lexical variability, L, is present individually. Furthermore, when more synthetic data is included in training, the performance consistently improves for the L + A setting, slightly in the case of only A and worsens in the case of lexical variability L. And although, having lexical variability is better in both test-other and test-clean, with the increment in training data in the low-resource data settings, having more acoustic variability turns out to be better than only lexical variability (that is owing to the worsening performance of only lexical variability experiment).

Finally, it can be observed that it is better to use the Voicebox model to synthesize speech (L) than to use the unpaired text as-is (exp A). On average, using Voicebox leads to a 9.6% relative WER improvement in test-other and a 8.0% relative improvement in test-clean. At best, in the high-resource data experiments, the synthetic data setting leads to more improvement, namely 19.4% in test-other and 19.8% in test-clean.

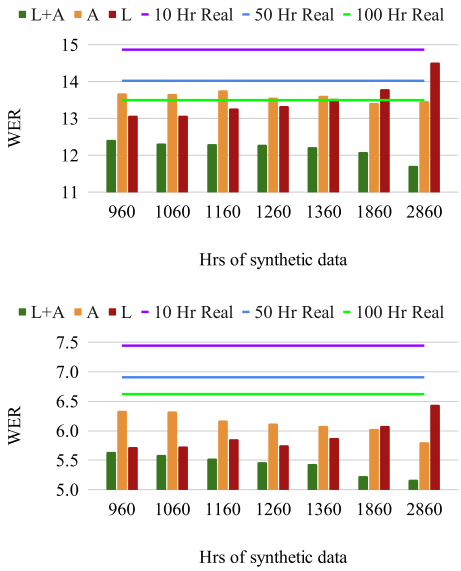

Figure 3: *WER comparison when synthetic data is added into training. The* L + A *shows when both linguistic and acoustic variability is added. The* A *and* L *show when only acoustic and linguistic variability is present respectively. The top and bottom charts show the results of test-other and test-clean respectively. The three horizontal lines show the performance of the model trained on 10, 50, and 100 hours of real data respectively.*

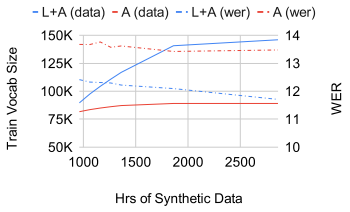

Figure 4: *Training vocab size and the WER on test other in two experiments, i.e.* L + A *and* A*, as a function of training hours of synthetic data used in ASR.*

5.3. Analysis

To analyze the impact of the lexical variability of synthesized speech on WER, we compare how WER decreases with the increase in vocabulary size of the training data. Figure 4 shows the training vocabulary size and the WER on test-other subset in two experiments, i.e. L + A and A. We can observe that there is a negative correlation between the training vocab size and the WER; i.e. as the training vocab size increases the WER decreases. For test-other and test-clean, the Pearson correlation between WER and training vocab size is -0.8, and -0.9 respectively. This observation is similar to that reported in [22], where new topic words were introduced in the training vocabulary and an improvement in WER was observed.

To evaluate the acoustic variability, one of the metrics that can be analyzed is the silence, graphemes, and word durations of the training data. Figure 5 shows the box plot of silence durations in the synthesized speech used in experiments L + A, L, and the real speech (these silences do not include the beginning and end silences). We observe that the silence duration in the L + A experiment is higher on average than the ones

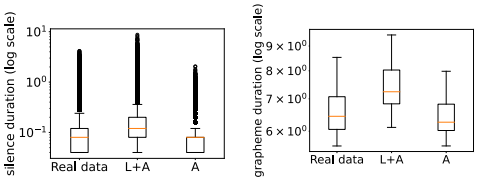

Figure 5: *Box plot comparing distribution of silences and graphemes in the two synthesized datasets and the real data.*

in A experiment and the real data. This is also observed for the grapheme durations. It has been reported previously that slower speech rates actually lead to better ASR performance [23], which is what we observe in these experiments as well.

Similarly, we analyze the word durations for the experiments. Figure 6 shows the qq-plot for the average word duration for data used in the experiment L + A compared with A, data used in A compared with the real data, and the data used in experiment L + A compared with real data. It can be observed that the data used in experiment L + A, has durations most similar to the ones of the real data (lies closer to the 45-degree line). This also explains why experiment L + A has better performance than experiment A, and this could be explained by the increasing lexical and acoustic variability from the text coming from a new corpus.

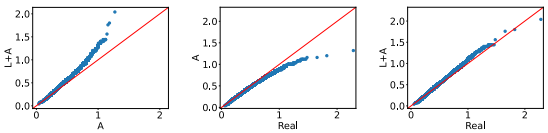

Figure 6: *qq-plot for the average word duration for two sets of synthesized data with compared with each other and the real data. Left-to-right:* L + A *vs.* A*,* L + A *vs. real,* A *vs. real.*

6. Conclusion

In this paper, we use Voicebox to synthesize speech from unpaired text and use the resulting synthetic speech to train ASR models. We establish benchmarks for the Voicebox-based synthesized speech. We find that 10 times and 7 times more synthetic speech than real speech/text paired data is required in noisy and clean settings, respectively, to get matching performance. Secondly, we explore the benefits from using lexically and acoustically diverse synthetic speech as augmentation in training data. We find that having both lexical and acoustic variability is better than just acoustic variability and lexical individually. Furthermore, we find that in the case when lower unpaired text data is available, having more lexical variability is better than only acoustic variability, however as more unpaired text becomes available (training data size increases), having more acoustic variability is better. Overall, generated speech from Voicebox models, using various seeds, leads to diverse speech samples with acoustic and prosodic variability in the speech. Therefore we believe that our conclusions based on the Voicebox synthesized speech are generalizable to other methods like speech perturbation. For future work, we aim to extend the comparison to other TTS and speech generative models, like AudioBox [24].

7. References

[1] B. G. J. Li, R. Gadde and V. Lavrukhin, "Training neural speech recognition systems with synthetic speech augmentation," in *arXiv preprint arXiv:1811.00707*, 2018.

[2] K. Yang, T.-Y. Hu, J.-H. R. Chang, H. S. Koppula, and O. Tuzel, "Text is all you need: Personalizing asr models using controllable speech synthesis," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[3] S. Kim, K. Li, L. Kabela, R. Huang, J. Zhu, O. Kalinli, and D. Le, "Joint audio/text training for transformer rescorer of streaming speech recognition," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022, pp. 5717–5722.

[4] S. Kim, Y. Shangguan, J. Mahadeokar, A. Bruguier, C. Fuegen, M. L. Seltzer, and D. Le, "Improved neural language model fusion for streaming recurrent neural network transducer," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7333–7337.

[5] Z. Meng, W. Wang, R. Prabhavalkar, T. N. Sainath, T. Chen, E. Variani, Y. Zhang, B. Li, A. Rosenberg, and B. Ramabhadran, "Jeit: Joint end-to-end model and internal language model training for speech recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[6] M. K. Baskar, L. Burget, S. Watanabe, R. F. Astudillo *et al.*, "Eat: Enhanced asr-tts for self-supervised speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6753–6757.

[7] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar *et al.*, "Voicebox: Text-guided multilingual universal speech generation at scale," *arXiv preprint arXiv:2306.15687*, 2023.

[8] A. Rosenberg, Y. Zhang, B. Ramabhadran, Y. Jia, P. Moreno, Y. Wu, and Z. Wu, "Speech recognition with augmented synthesized speech," in *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*. IEEE, 2019, pp. 996–1002.

[9] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[10] T. N. Sainath, R. Prabhavalkar, A. Bapna, Y. Zhang, Z. Huo, Z. Chen, B. Li, W. Wang, and T. Strohman, "Joist: A joint speech and text streaming model for asr," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 52–59.

[11] C. Gao, G. Cheng, R. Yang, H. Zhu, P. Zhang, and Y. Yan, "Pre-training transformer decoder for end-to-end asr model with unpaired text data," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6543–6547.

[12] Z. Meng, Y. Gaur, N. Kanda, J. Li, X. Chen, Y. Wu, and Y. Gong, "Internal language model adaptation with text-only data for end-to-end speech recognition," *arXiv preprint arXiv:2110.05354*, 2021.

[13] Z. Meng, T. Chen, R. Prabhavalkar, Y. Zhang, G. Wang, K. Audhkhasi, J. Emond, T. Strohman, B. Ramabhadran, W. R. Huang *et al.*, "Modular hybrid autoregressive transducer," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 197–204.

[14] M. Zeineldeen, A. Glushko, W. Michel, A. Zeyer, R. Schlüter, and H. Ney, "Investigating methods to improve language model integration for attention-based encoder-decoder asr models," *arXiv preprint arXiv:2104.05544*, 2021.

[15] Z. Chen, Y. Zhang, A. Rosenberg, B. Ramabhadran, P. Moreno, A. Bapna, and H. Zen, "Maestro: Matched speech text representations through modality matching," 2022.

[16] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *The Eleventh International Conference on Learning Representations*, 2022.

[17] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[19] "Librispeech language models, vocabulary and g2p models," https://openslr.org/11/, accessed: 2023-07-31.

[20] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 244–250.

[21] K. Kim, F. Wu, Y. Peng, J. Pan, P. Sridhar, K. J. Han, and S. Watanabe, "E-branchformer: Branchformer with enhanced merging for speech recognition," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 84–91.

[22] R. Gody and D. Harwath, "Unsupervised fine-tuning data selection for asr using self-supervised speech models," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[23] Y. Meng, Y.-H. Chou, A. T. Liu, and H.-y. Lee, "Don't speak too fast: The impact of data bias on self-supervised speech models," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3258–3262.

[24] A. Vyas, B. Shi, M. Le, A. Tjandra, Y.-C. Wu, B. Guo, J. Zhang, X. Zhang, R. Adkins, W. Ngan *et al.*, "Audiobox: Unified audio generation with natural language prompts," *arXiv preprint arXiv:2312.15821*, 2023.