Comparing Human-Labeled and AI-Labeled Speech Datasets for TTS

Johannes Wirth and René Peinl
Institute of Information Systems at Hof University, Germany

Johannes.wirth.3@iisys.de
Rene.peinl@iisys.de

Abstract: As the output quality of neural networks in the fields of automatic speech recognition (ASR) and text-to-speech (TTS) continues to improve, new opportunities are becoming available to train models in a weakly supervised fashion, thus minimizing the manual effort required to annotate new audio data for supervised training. While weak supervision has recently shown very promising results in the domain of ASR, speech synthesis has not yet been thoroughly investigated regarding this technique despite requiring the equivalent training dataset structure of aligned audio-transcript pairs. In this work, we compare the performance of TTS models trained using a well-curated and manually labeled training dataset to others trained on the same audio data with text labels generated using both grapheme- and phoneme-based ASR models. Phoneme-based approaches seem especially promising, since even for wrongly predicted phonemes, the resulting word is more likely to sound similar to the originally spoken word than for grapheme-based predictions. For evaluation and ranking, we generate synthesized audio outputs from all previously trained models using input texts sourced from a selection of speech recognition datasets covering a wide range of application domains. These synthesized outputs are subsequently fed into multiple state-of-the-art ASR models with their output text predictions being compared to the initial TTS model input texts. This comparison enables an objective assessment of the intelligibility of the audio outputs from all TTS models, by utilizing metrics like word error rate and character error rate. Our results not only show that models trained on data generated with weak supervision achieve comparable quality to models trained on manually labeled datasets, but can outperform the latter, even for small, well-curated speech datasets. These findings suggest that the future creation of labeled datasets for supervised training of TTS models may not require any manual annotation but can be fully automated.

Keywords: Text-to-Speech, Dataset generation, Pseudo labeling

1. Introduction

In the last few years, text-to-speech (TTS) technology has made remarkable progress, reaching previously unseen levels of naturalness and intelligibility. State-of-the-art (SOTA) systems consisting of neural networks have shown that they are capable of producing human-like speech that is almost indistinguishable from recordings of real human voices. Despite these remarkable results, the process of further developing and improving such systems becomes more demanding, especially when it comes to generation of new data.

The output quality of TTS systems strongly depends on large, high-quality datasets that are thoroughly labeled and well curated. However, the generation of such data represents a significant bottleneck, as it mostly involves manual labeling and quality assurance. For this reason, high-quality labeled speech data is scarce. This is especially true for non-English languages. While similar domains like automatic speech recognition (ASR) already benefit from techniques like pseudo labeling with ASR, TTS has not yet been equally considered regarding this method, despite speech data for TTS generally being of higher quality and thus also being easier to correctly be transcribed by ASR systems.

This work investigates the quality of pseudo-labeling TTS datasets using several speech recognition systems that involve both grapheme-based and phone-based approaches. The method described hereafter involves training SOTA TTS models on a dataset labeled by both humans and the previously mentioned ASR systems. We then conduct objective evaluations of intelligibility on model outputs using ASR. The results of these evaluations suggest that AI-labeled speech datasets are not only equally capable of labeling speech data with an accuracy comparable to human annotators, but in several cases can surpass the quality of human labeling. Our findings support the potential of pseudo-labeling as a powerful tool to improve the efficiency and effectiveness of TTS dataset generation.

2. Related Work

TTS systems have shown impressive progress since deep learning techniques have been applied to this challenge from 2016 on (Oord et al., 2016). The generation pipeline originally consisted of a Mel generator like the autoregressive Tacotron 2 (Shen et al., 2018) and TransformerTTS (Li et al., 2019), or the non-autoregressive like Glow-TTS (Kim et al., 2020) and Flow-TTS (Miao et al., 2020) and a vocoder to generate waveforms like Hifi-GAN (Kong et al., 2020) and WaveGrad (Chen et al., 2020). Later on, so called end-to-end TTS systems came into the focus that directly generate audio waveforms based on phoneme input like VITS (Kim et al., 2021) and JETS (Lim et al., 2022).

Most of these models were trained on the famous LjSpeech dataset (Ito, 2017), a single-speaker dataset with 24h of speech from the LibriVox speaker Linda Johnson. Single speaker data provided the best speech quality with a limited amount of training data. Multi-speaker datasets like VCTK (Yamagishi et al., 2019) which was recorded in studio-quality and TTS datasets curated from audiobooks such as M-AILABS (Solak, 2019), LibriTTS (Zen et al., 2019) and HiFi-TTS (Bakhturina et al., 2021) did not yield the same quality as single-speaker models although they contain far more data (586h in case of LibriTTS).

Only recently, large-scale TTS systems were developed, that are trained on thousands of hours of speech, that show clear advantages over single speaker models, especially in terms of prosody (Jiang et al., 2024). The required datasets are not human-labeled anymore but pseudo-labeled by ASR systems, e.g. LibriLight (Kahn et al., 2020), with its' nearly 60k hours of speech, which is used e.g. by VALL-E (Wang et al., 2023), MEGA-TTS 2 and NaturalSpeech 3 (Ju et al., 2024). The quality of the pseudo-labels is unknown, but due to low word error rates (WERs) of SOTA ASR models, especially for read speech with low noise, it seems good enough to produce high quality synthetic speech. However, this cannot be taken for granted, as the multi-lingual LibriSpeech dataset (MLS) (Pratap et al., 2020) for example has been shown to contain systematic errors in German language, e.g. for numbers and some text in the transcript, that was actually not spoken (Wirth and Peinl, 2022). In general, the quality of commercial and open source ASR systems have significantly advanced, besides English also in German (Wirth and Peinl, 2022). Although WERs can grow as high as 30% for challenging datasets like the multilingual TEDx Corpus (Salesky et al., 2021) (German subset), for read speech, good recording quality and low noise settings, you can expect WERs below 5% (e.g. M-AILBS: 4.28%, Voxforge: 3.36%, HUI: 1.89% with Conformer-Transducer large). For ASR systems, using pseudo labels has already been shown to provide benefits. (Hwang et al., 2022) demonstrated that a 600M parameter Conformer Transducer model delivers as good performance on ASR for voice search and conversations when being trained on over a million hours of pseudo labels in English language. (Zhang et al., 2024) use manually labeled data first (57k hours) and show that further training on up to 520k hours of pseudo-labeled speech reduces WER from 8.7% to 7.6%.

Despite the success and proven benefits of pseudo-labeling for the training of ASR models, this approach has not yet been adopted and evaluated in the field of TTS, which places even higher demands on audio quality and audio transcript alignment than ASR as well as phonemization of transcripts. Although transcript phonemization is not a necessity to train TTS models, phonemized texts enable more fine-grained control over the generation of synthesized speech outputs, as characters/graphemes can be pronounced differently depending on the context, while phonemes can be pronounced unambiguously.

In this work, we propose a phone-based ASR system to do pseudo-labeling and show, that it not only outperforms grapheme-based SOTA ASR systems (with algorithmic phonemization) for that task, but TTS models trained on the pseudo-labeled phones also produce better results than the one trained on the human-labeled reference data on multiple test sets.

3. Approach

To evaluate the quality of AI-based labeling for TTS datasets, a well-annotated and manually quality-assured speech dataset is selected to serve as a gold standard for subsequent comparisons. The audio recordings from this set are fed into a SOTA ASR model to create a variant of the original dataset with labels generated by the ASR model. Both the human-labeled and AI-labeled dataset variants are then phonemized and used to train two separate models with identical TTS architectures and hyperparameters, ensuring comparability.

Next, a test dataset consisting of text inputs is used to generate speech outputs from both TTS models. These speech outputs are then processed by another speech recognition system to produce transcript predictions. The predicted transcripts are compared to the original text inputs, and the WER is calculated for each model. By comparing the WERs, it is possible to determine how the performance of a TTS model trained on AI-labeled data compares to those trained on human-labeled data. While this type of evaluation does not substitute typical metrics for speech synthesis quality, such as mean opinion score (MOS), it effectively measures the intelligibility of synthesized speech output without the need for survey setups and participant recruitment (Peinl and Wirth, 2022). Although WER cannot be utilized to evaluate subjective speech factors such as the naturalness of the synthesized voice, the pace of speech or the pleasantness of the voice, as would be possible by means of a time-consuming evaluation of these factors with listener surveys and MOS-based ranking, it serves as a reasonable basis for determining the quality of synthesized speech in the scope of this work as general intelligibility of speech outputs is the decisive factor for the practical usability of a TTS system.

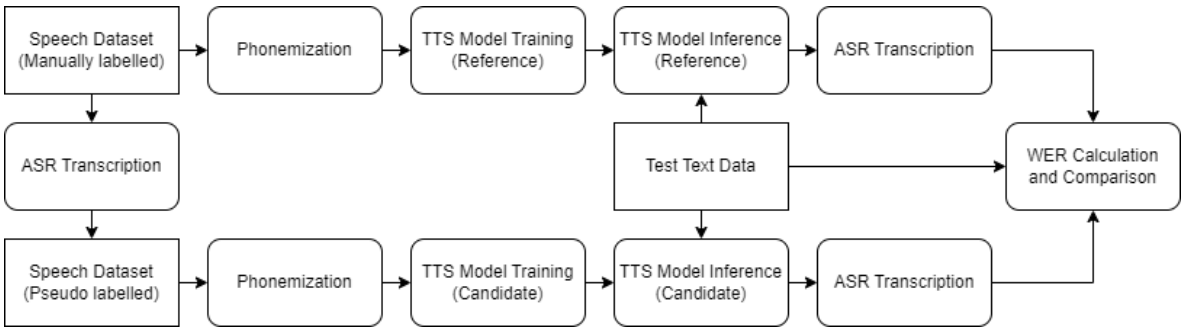

Figure 1: Flow Diagram of the approach used for comparison of human- and AI-labeled speech datasets for TTS systems. Entities are depicted in sharp corners, processes in round corners

As can be seen in Figure 1, the described approach is only illustrated for the comparison of a single pseudo-labeled dataset with an original dataset. The following sections present an implementation of this method with multiple pseudo-labeled datasets created using different ASR models.

4. Experimental Setup

The following subsections explain the specific measures established and models used to implement the previously described approach in detail.

4.1 Model Training

As reference dataset, the German single speaker dataset “Thorsten Voice” (Müller and Kreutz, 2022) was chosen, since it is of manageable size (~11hrs of audio-transcript pairs), has high audio quality and was labeled as well as quality-assured completely manually. In addition, multiple freely available TTS models have already been trained on this dataset, proving its high quality (Müller, 2024). Two different ASR models were used to pseudo-label the audio recordings of this dataset: The freely available Whisper v3 Large (Radford et al., 2022) model without utilizing the language detection mechanism (German was statically set as the transcription language) and an in-house model based on a SOTA Fast-Conformer Transducer Large architecture (Rekesh et al., 2023), which in contrast to Whisper uses a phone-based tokenizer and was trained on phone sequences from multiple languages (publication in preparation). This model does not include an explicit language detection mechanism but was trained with the aim of verbatim conversion of speech input into phoneme output, regardless of the input language, also enabling the model to create robust predictions with mixed language speech input. The phone representations are created using the international phonetic alphabet (IPA), whereas the grapheme-based Whisper model generates transcript predictions in plain readable text. The labels of the original dataset and the dataset pseudo-labeled using the Whisper model were further phonemized to IPA using a custom-built pipeline which, in addition to number and symbol normalization, also includes named entity recognition and language detection in order to ensure correct phonemization of loanwords from other languages. This is especially important for the German language, as it borrows many words and phrases from English and French. Punctuation was also removed from all transcripts to exclude it as a factor in the output quality of the models. The step of phonemization was skipped for the phone-based ASR model as it already creates outputs in the desired IPA format.

In the following step, each previously described dataset was used to train a VITS (Kim et al., 2021) model using the Espnet toolkit (Hayashi et al. 2020). Each TTS model was trained for 400 epochs, using an Nvidia RTX 6000 Ada graphics card.

4.2 Evaluation

Speech outputs were generated for each model to assess the quality of the TTS models pre-trained with the various dataset versions. These were then used as inputs for SOTA ASR models, and generated transcript predictions were compared with the original texts using WER to obtain an objective metric for the intelligibility of the TTS models.

Input texts were taken from the labels of the test splits of various speech datasets from diverse domains, including BAS HEMPEL (“BAS HEMPEL,” 2017), ASR Bundestag (Wirth and Peinl, 2023), Mozilla Commonvoice (Ardila et al., 2020), the Multilingual TEDx Corpus, M-AILABS, MLS, the Spoken Wikipedia Corpus (Baumann et al., 2019), the test split of the Thorsten Voice dataset used for training, Tuda-De (Radeck-Arneth et al., 2015), Voxforge (Voxforge.org, n.d.) and Voxpopuli (Wang et al., 2021). Since M-AILABS contains heterogeneous

speech data, on the one hand political speeches by former German chancellor Angela Merkel and on the other hand books read by single speakers, this dataset was further divided into the partial sets “M-AILABS” (excluding Merkel) and ”ASR Merkel“. For multilingual datasets only the German subsets were used.

Both Whisper V3 large, which was previously utilized to create a dataset variation with pseudo labels as well as a Conformer Transducer model pre-trained on German speech data provided by Nvidia (“STT De Conformer-Transducer Large,” 2022) were used to transcribe the synthesized speech outputs.

5. Results

The evaluation of TTS model outputs was carried out for every 100 training epoch checkpoints for all models; only the evaluation results for the epoch 400 checkpoints are listed below, as all models performed best at this stage, in comparison to scores of previous checkpoints.

Table 1: ASR evaluation results based on word error rate for the TTS models trained on the original dataset as well as pseudo labeled with Whisper v3 large and a phone-based ASR model (Whisper v3 large evaluation)

Dataset	Original	Phone Labelling	Whisper Labelling
HEMPEL	**11.82%**	11.95%	12.48%
ASR Bundestag	11.57%	**11.32%**	12.43%
ASR Merkel	6.76%	**6.37%**	7.51%
Commonvoice	18.63%	**17.40%**	20.82%
German TED	**12.54%**	14.54%	15.30%
M-AILABS	19.25%	**14.40%**	20.27%
MLS	14.26%	**13.46%**	15.90%
SWC	22.52%	**21.57%**	24.94%
Thorsten	**23.62%**	23.89%	25.78%
Tuda	11.63%	**11.15%**	14.29%
Voxforge	15.48%	**13.22%**	15.30%
Voxpopuli	**7.64%**	7.69%	9.45%
Mean	14.64%	**13.91%**	16.21%

Table 2: Evaluation results based on word error rate for the TTS models trained on the original dataset as well as pseudo labeled with Whisper v3 large and a phone-based ASR model (Conformer Transducer DE evaluation)

Dataset	Original	Phone Labelling	Whisper Labelling
HEMPEL	12.14%	**12.04%**	12.12%
ASR Bundestag	10.56%	**9.41%**	12.05%
ASR Merkel	5.76%	**5.74%**	6.36%
Commonvoice	15.30%	**14.29%**	16.79%
German TED	**12.57%**	14.32%	15.40%
M-AILABS	**11.47%**	11.56%	16.07%
MLS	9.92%	**9.72%**	11.93%
SWC	18.34%	**18.22%**	22.43%
Thorsten	**18.23%**	21.81%	24.42%
Tuda	**8.43%**	10.18%	12.58%
Voxforge	11.47%	**11.30%**	13.28%
Voxpopuli	**5.23%**	6.29%	7.48%
Mean	**11.62%**	12.07%	14.24%

As can be seen in Table 1 as well as Table 2, the TTS model trained on the pseudo-labeled dataset using a phone-based model exceeds the model trained on the dataset labeled by humans in most cases. Depending on the ASR model used for evaluation, these differences vary with the Whisper v3 results being slightly in favor of phone-based pseudo-labeling and the Conformer Transducer evaluation in favor of human labeling, which is

mainly attributed to the different datasets used to train these models. It can also be seen that all TTS models show comparatively small differences in terms of intelligibility in direct comparison, with the exception of few outliers.

While the model trained on original labels performed better on the test sets for Tedx, Thorsten Voice and Voxpopuli by small margins, it was outperformed in at least one model evaluation on all other test datasets. The model trained on pseudo-labeled data generated with Whisper v3 large did not achieve the best results on any test dataset, but its performance is not significantly worse across the board.

6. Interpretation

A direct comparison between TTS models trained on human- and AI-labeled data shows that all presented dataset labeling methods lead to very similar model performance levels. Notably, the model utilizing labels generated by the phone-based ASR system surpasses the one with original labels as well as the other variation trained on labels created with Whisper on most test datasets. This observation confirms the hypothesis that high-quality labeling of audio data with AI is possible and can even increase the quality of speech datasets based on the ASR model used to generate pseudo labels. We assume that the better model quality is not due to incorrect human labeling of the original dataset used, but due to the ability of the phone-based speech recognition model to represent verbatim spoken sounds better than is possible with grapheme-to-phoneme conversion. While the phonetic representation of most words is generated correctly by phonemization frameworks such as Phonemizer (Bernard and Titeux, 2021), it does not always correspond to what is actually spoken. Often, spoken words are shortened, concatenated or stressed differently depending on the context, which can lead to a greater discrepancy between actual spoken words in audio recordings and the phonetic transcript if a phonemization framework is used. This effect would be further reinforced by other factors such as spoken dialects. However, the potential cause of incorrect labeling described here can be avoided by using a phone- instead of a grapheme-based ASR model for pseudo-labeling and thus skipping the phonemization step in data preprocessing.

The high quality of the labels generated by the phone-based ASR model offers the additional benefit of its implicit multilingual pre-training. This eliminates the need to detect the language of the audio input or the necessity of maintaining multiple high-quality speech recognition models in different languages.

The difference in output quality between the two models trained on pseudo labels can be attributed to the rather logical than verbatim transcript outputs of the Whisper model in combination with the previously described induction of errors through phonemization. Since dataset labels used for speech synthesis should contain exactly what is actually said within a recording and not meaningful renditions, the phone-based model trained on verbatim output performs better on almost every test dataset in both evaluations. Conversely, this core finding may imply that phone-based speech recognition performs worse than grapheme-based speech recognition in the area of practically applied ASR, since in this area contextually correct transcription is in most cases more relevant than the exact transcription of what was said (including interfering factors such as repeated words due to stuttering and filler words), however, this was not investigated in the context of this work.

7. Conclusion and Future Work

In this work, it was shown that AI-based labeling of TTS datasets is not only a solid alternative to human annotation but is clearly the better option in the scope of the described approach and experiments. This underlines the effectiveness of modern ASR systems in generating high-quality labels that meet the requirements of TTS data preparation.

Based on these findings, the creation of new datasets could be fully automated, which can drastically reduce the development cycles of new models and increase the quality of speech synthesis systems in shorter iterations, enabling researchers to focus on model optimization instead of data quality assurance. In addition, since the best ASR model for pseudo labeling in the experiments conducted is phone-based and language-agnostic, the suggested and implemented approach is not only suitable for languages that have strongly pretrained ASR models but could also be used for data preparation in the scope of low-resource languages.

For future work, it is planned to further train the phone-based ASR model, which has achieved the best results for dataset labeling in the conducted experiments. By additionally improving the labeling quality, it will be possible to leverage large quantities of freely available, high-quality audio recordings. These recordings will be pseudo-labeled using the enhanced, phone-based ASR model and can subsequently be utilized to train a multi-

speaker state-of-the-art TTS model, specifically focusing on advancing the landscape of TTS for the German language and benefiting both researchers and end-users alike.

The phone-based ASR model including the custom built phonemization pipeline will be made freely available as soon as the accompanying research paper has been published in order make all presented results replicable.

References

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M., Weber, G., 2020. Common Voice: A Massively-Multilingual Speech Corpus, in: Proceedings of the 12th Language Resources and Evaluation Conference (Pp. 4218-4222).

Bakhturina, E., Lavrukhin, V., Ginsburg, B., Zhang, Y., 2021. Hi-Fi Multi-Speaker English TTS Dataset. arXiv preprint arXiv:2104.01497.

BAS HEMPEL, 2017.

Baumann, T., Köhn, A., Hennig, F., 2019. The Spoken Wikipedia Corpus collection: Harvesting, alignment and an application to hyperlistening. Language Resources and Evaluation 53, 303–329.

Bernard, M., Titeux, H., 2021. Phonemizer: Text to Phones Transcription for Multiple Languages in Python. Journal of Open Source Software 6, 3958. https://doi.org/10.21105/joss.03958

Chen, N., Zhang, Y., Zen, H., Weiss, R.J., Norouzi, M., Chan, W., 2020. WaveGrad: Estimating gradients for waveform generation. arXiv preprint arXiv:2009.00713.

Hwang, D., Sim, K.C., Huo, Z., Strohman, T., 2022. Pseudo Label Is Better Than Human Label.

Ito, K., 2017. The LJ Speech Dataset [WWW Document]. URL https://keithito.com/LJ-Speech-Dataset (accessed 5.3.21).

Jiang, Z., Liu, J., Ren, Y., He, J., Ye, Z., Ji, S., Yang, Q., Zhang, C., Wei, P., Wang, C., Yin, X., Ma, Z., Zhao, Z., 2024. Mega-TTS 2: Boosting Prompting Mechanisms for Zero-Shot Speech Synthesis. https://doi.org/10.48550/arXiv.2307.07218

Ju, Z., Wang, Y., Shen, K., Tan, X., Xin, D., Yang, D., Liu, Y., Leng, Y., Song, K., Tang, S., Wu, Z., Qin, T., Li, X.-Y., Ye, W., Zhang, S., Bian, J., He, L., Li, J., Zhao, S., 2024. NaturalSpeech 3: Zero-Shot Speech Synthesis with Factorized Codec and Diffusion Models, in: ICML.

Kahn, J., Rivière, M., Zheng, W., Kharitonov, E., Xu, Q., Mazaré, P.-E., Karadayi, J., Liptchinsky, V., Collobert, R., Fuegen, C., Likhomanenko, T., Synnaeve, G., Joulin, A., Mohamed, A., Dupoux, E., 2020. Libri-Light: A Benchmark for ASR with Limited or No Supervision. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 7669–7673. https://doi.org/10.1109/ICASSP40776.2020.9052942

Kim, J., Kim, S., Kong, J., Yoon, S., 2020. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. arXiv preprint arXiv:2005.11129.

Kim, J., Kong, J., Son, J., 2021. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech.

Kong, J., Kim, J., Bae, J., 2020. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. Advances in Neural Information Processing Systems 33.

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

Lim, D., Jung, S., Kim, E., 2022. JETS: Jointly Training FastSpeech2 and HiFi-GAN for End to End Text to Speech. https://doi.org/10.48550/arXiv.2203.16852

Miao, C., Liang, S., Chen, M., Ma, J., Wang, S., Xiao, J., 2020. Flow-TTS: A Non-Autoregressive Network for Text to Speech Based on Flow, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 7209–7213.

Müller, T., 2024. Deutsche AI TTS-Stimme Kostenlos in Hoher Qualität ohne Cloud [WWW Document]. Thorsten Voice. URL https://www.thorsten-voice.de/kostenloses-deutsches-text-to-speech-tts/ (accessed 8.27.24).

Müller, T., Kreutz, D., 2022. ThorstenVoice Dataset 2022.10. https://doi.org/10.5281/zenodo.7265581

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

Peinl, R., Wirth, J., 2022. Quality Assurance for Speech Synthesis with ASR, in: Intelligent Systems Conference (IntelliSys 2022). Amsterdam, Netherlands.

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., Collobert, R., 2020. Mls: A large-scale multilingual dataset for speech research. arXiv preprint arXiv:2012.03411.

Radeck-Arneth, S., Milde, B., Lange, A., Gouvêa, E., Radomski, S., Mühlhäuser, M., Biemann, C., 2015. Open Source German Distant Speech Recognition: Corpus and Acoustic Model, in: Král, P., Matoušek, V. (Eds.), Text, Speech, and Dialogue, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 480–488. https://doi.org/10.1007/978-3-319-24033-6_54

Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., 2022. Robust Speech Recognition via Large-Scale Weak Supervision. https://doi.org/10.48550/ARXIV.2212.04356

Rekesh, D., Koluguri, N.R., Kriman, S., Majumdar, S., Noroozi, V., Huang, H., Hrinchuk, O., Puvvada, K., Kumar, A., Balam, J., Ginsburg, B., 2023. Fast Conformer with Linearly Scalable Attention for Efficient Speech Recognition. https://doi.org/10.48550/arXiv.2305.05084

Salesky, E., Wiesner, M., Bremerman, J., Cattoni, R., Negri, M., Turchi, M., Oard, D.W., Post, M., 2021. The multilingual tedx corpus for speech recognition and translation. arXiv preprint arXiv:2102.01757.

Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., 2018. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions, in: 2018 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4779–4783.

Solak, I., 2019. The M-AILABS speech dataset [WWW Document]. URL https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/

STT De Conformer-Transducer Large, 2022.

Voxforge.org, n.d. Free Speech... Recognition (Linux, Windows and Mac) - voxforge.org.

Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S., Wei, F., 2023. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers.

Wang, C., Rivière, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., Dupoux, E., 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. arXiv preprint arXiv:2101.00390.

Wirth, J., Peinl, R., 2023. ASR Bundestag: A Large-Scale political debate dataset in German. https://doi.org/10.48550/arXiv.2302.06008

Wirth, J., Peinl, R., 2022. ASR in German: A Detailed Error Analysis. arXiv preprint arXiv:2204.05617.

Yamagishi, J., Veaux, C., MacDonald, K., 2019. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92). https://doi.org/10.7488/ds/2645

Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R.J., Jia, Y., Chen, Z., Wu, Y., 2019. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. arXiv:1904.02882 [cs, eess].

Zhang, K., Chkhetiani, L., Ramirez, F.M., Khare, Y., Vanzo, A., Liang, M., Martin, S.R., Oexle, G., Bousbib, R., Peyash, T., Nguyen, M., Pulliam, D., Donato, D., 2024. Conformer-1: Robust ASR via Large-Scale Semisupervised Bootstrapping.