Synthetic AI Data Pipeline for Domain-Specific Speech-to-Text Solutions

Anderson Luiz Karl [1], Guilherme Sales Fernandes[1], Leonardo Augusto Pires [1], Yvens R. Serpa [1,3], Carlos Caminha [2]

[11]Audo Tecnologia e Saúde
[2]Universidade Federal do Ceará, UFC, Brasil
[3]Saxion University of Applied Sciences, Enschede, Netherlands

Abstract. *In this article, we propose a pipeline to fine-tune domain-specific Speech-to-Text (STT) models using synthetic data generated by Artificial Intelligence (AI). Our methodology eliminates the need for manually labelled audio data, which is expensive and difficult to obtain, by generating domain-specific data with a Large Language Model (LLM) combined with multiple Text-to-Speech (TTS) solutions. We applied our pipeline to the radiology domain and compared the results with different approaches based on the availability of domain-specific data, varying from the total absence of domain-specific data to the use of only domain-specific high-quality data (ground truth). Our performance improved the accuracy of the baseline by 40.19% and 10.63% for the WhisperX Tiny and Small models, respectively, which, although performed worse than the results from using the ground truth, shows that it is possible to achieve good results with minimal cost and effort. Finally, the result analysis shows a good insight into the amount of action necessary to achieve good results based on the availability of real data.*

1. Introduction

Automatic audio transcription, commonly referred to as Speech-to-Text (STT), has been a common practice for many work fields, such as health, justice, education, and business [Kumar 2024]. However, precision in recognizing and transcribing language is important to guarantee the correct and efficient use of the transcribed information. That is especially important in domain-specific applications, in which the use of technical terms and jargon increases the recognition and transcription challenge [Suh et al. 2024]. However, many of the typically available solutions for this problem are built on generic data. Due to that, their results are of lower quality when used in domain-specific scenarios [Chan et al. 2016].

A common approach to solving this issue is to build and refine solutions using domain-related contexts, vocabularies and other types of data [Huang et al. 2020]. Nowadays, it is standard to use generic AI models as the base for STT solutions and fine-tune these models with domain-specific data [Mak et al. 2024]. However, the fine-tuning process is expensive and requires a significant amount of data and effort [Hu et al. 2022]. For medical applications, for example, it is necessary to collect sensitive data, have health professionals check, correct and validate it, and guarantee its privacy and security in regard to the involved patients and personnel [Johnson et al. 2014].

Nevertheless, the need for high-quality STT solutions is evident in many work sectors. In Radiology, for example, it is a common practice to have physicians use STT

tools in their work practice to increase productivity over traditional transcription, the latter in which the professional records a report via voice to be later transcribed manually by another professional (usually without a medical background) [Hammana et al. 2015]. Any errors or delays in this process may result in possible harm and consequences to the patients and their treatments [Vorbeck et al. 2000]. Another common example is courts and judicial procedures, in which a large quantity of domain-specific texts is generated and often transcribed manually, resulting in expensive and inefficient processes [da Cruz et al. 2022].

In this context, this work proposes a low-cost pipeline for the training and fine-tuning of STT AI models when domain-specific data is required but not readily available. Our pipeline is based on the use of AI models to generate synthetic domain-specific data. For that, we have used a Large Language Model (LLM) to produce domain-specific content that simulates real use cases. Specifically for this work, we have explored the radiology domain, generating data for synthetic radiology reports using an LLM and a specific prompting approach. The synthetic data is then converted into audio files through Text-to-Speech (TTS) tools. Thus, the fine-tuning process is done entirely using synthetic data generated via AI. Additionally, due to the focus on being a low-cost solution, the results of this work were done by using inexpensive or freely available solutions. Simultaneously, this work also presents a comparison analysis of a range of possible final results depending on the availability of domain-specific data.

2. Related Work

Automatic audio transcription has been a fruitful research field in computer sciences over many years [Yu et al. 2010, Blackley et al. 2019]. Many of the traditional works in this field are focused on the inherited challenges of it, such as handling language subtleties, structure, and fluency [Gontier et al. 2021], and the limitations on the access of adequate datasets [Hu et al. 2022]. These challenges increase when dealing with domain-specific scenarios [Samarakoon et al. 2018].

In regards to datasets, the majority of works in the field use datasets in the English language [Casanova et al. 2022]. When working in scenarios with other languages, researchers must not only solve the recurrent STT challenges but also adapt their solutions, such as done by Gruzitis *et al.* [Gruzitis et al. 2022] which adapted their models to the Latvian language, and the work of Vivancos-Vincente *et al.* [Vivancos-Vicente et al. 2016] for Spanish and Portuguese. Alternatively, the work proposed by Casanova *et al.* [Casanova et al. 2022] shows an alternative to training models for different languages based on data augmentation from only one speaker for the targeted language, using cross-lingual voice conversion and multi-speaker TTS techniques.

Moreover, access to good domain-specific datasets is a challenge, and its production involves high costs with domain experts, data analysis, and validation. This problem is often faced with the use of synthetic data [Li et al. 2018, Rosenberg et al. 2019, Laptev et al. 2020, Huang et al. 2020, Yang et al. 2023]. However, synthetic data is frequently distant from real use cases due to the absence of mistakes and imperfections that are often common in human-made data, which makes it "too perfect" compared to real-world cases. This"perfection problem" is handled with the introduction of synthetic errors and imperfections, such as done by the Synt++ solution proposed by Hu *et al.*

[Hu et al. 2022], in which noise and random artefacts are introduced to the synthetic data generation so it more closely resembles real-life data.

Only recently the process of data synthesis using LLM have been explored, such as the work presented by Vásquez-Correa *et al.* [Vásquez-Correa et al. 2023], which generates domain-specific synthetic data through prompting to fine-tune an STT solution for the English, Spanish, and Basque languages. Silva *et al.* [Silva et al. 2024] also uses an LLM to generate synthetic data for a hardware failure prediction dataset. Their dataset was generated from problem categories and reports from major component manufacturers in the market.

Similarly, this work proposes a new approach to synthetic data based on prompting. The synthetic data is then converted into audio files through TTS algorithms and used to fine-tune a generic STT AI model. Our approach uses a simple and low-cost generic STT AI model as a means to prove its usefulness in scenarios with minimal resources. Moreover, this work presents a comparison analysis of results based on the availability of domain-specific data, varying from the total absence of domain-specific data (our solution) to the use of only domain-specific high-quality data (an ideal solution).

3. Methodology

3.1. Datasets

To validate the efficiency of our proposed pipeline, we used a dataset of manually labelled audio data from radiology professionals, which was divided into a set for training and another for testing. The training set included 98 audio files from two cisgender male radiologists with a total duration of 1 hour, 10 minutes and 8 seconds of audio. The testing dataset consisted of 82 audio files from the same two radiologists, with a total duration of 1 hour, 4 minutes and 21 seconds of audio. Both training and testing sets had an equal amount of audio files for the two radiologists, and all audio files were spoken in Portuguese. All audio files were recorded in real-world scenarios, including background noise from the respective workplaces, audio artefacts, and other common issues. This dataset constitutes our ground truth dataset, which was used to compare with the results from the other approaches explored.

3.2. Methods and Technologies

The transformers library by Hugging Faces [Vaswani 2017] was used to fine-tune the STT model, which was also configured for the Portuguese language. We opted for a traditional fine-tuning process using all of the available weights. For the inference, we have used the WhisperX model [Bain et al. 2023], which offers a quicker and more precise transcription, with the Ctranslate2 backend for better compatibility and reduced inference time. The main reason for using WhisperX was the presence of an internal Voice Activity Detection (VAD), which considerably reduces the hallucination tendencies and optimizes the use of VRAM [Koenecke et al. 2024].

We have used GPT-4o as the LLM to generate synthetic domain-specific radiology reports using a specific approach and prompts [Islam and Moushi 2024]. The synthetic reports were fed into TTS solutions to generate audio files for the fine-tuning process.

As TTS solutions, we have used the ElevenLabs solution[1], which is fairly low cost

[1]https://elevenlabs.io/

for its quality, and the Google Text-to-Speech[2]. Both tools allowed for a variety of intonations, speech styles, and variations, which helped to reduce the "perfection problem" often produced in synthetic data. Furthermore, the use of two TTS solutions improved the representation and diversity of speech patterns and accents.

3.3. Metrics

The Word Error Rate (WER) metric was used to assess the precision of the STT solutions [Ali and Renals 2018]. The WER metric is calculated by the ratio between the number of transcribed errors and the number of words originally spoken. These errors are classified as Substitutions (S), Insertions (I), and Deletions (D). The WER formula we used was: $\text{WER} = \frac{S+D+I}{N}$, where N is the number of words originally spoken.

4. Results

4.1. Proposed Pipeline

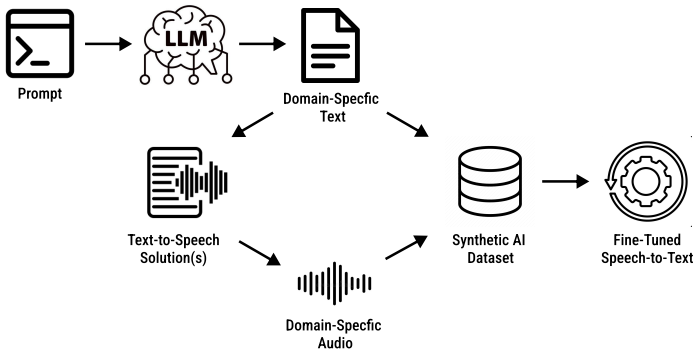

Figure 1. Proposed Pipeline.

As shown in Figure 1, the proposed pipeline aims to fine-tune an STT model using a set of synthetic domain-specific data. It starts with a specialist prompt for the LLM. This specialist prompt must consider specific terminology and domain-specific information to guarantee that the synthetic data closely resembles real-life data.

The LLM-generated synthetic data is fed into TTS solutions and converted into audio files. It is important to include variations in tone of voice and synthetic noise in this process to reduce the "perfection problem". Together, the LLM-generated synthetic data and its audio representation compose the AI-labelled dataset. This dataset is then used to fine-tune the STT model of choice.

4.2. AI-Labelled Dataset

GPT-4o was used as the LLM tool for the domain-specific synthetic data generation. For that, we first introduced the model to the radiology context and gave it a series of radiology specialities and exam types, such as *computer tomography* and *radiography*. Furthermore, to guarantee typical report-style phrasing, we instructed the LLM to create phrases and sentences in a progressive format, starting from normal descriptions, followed by potential

[2]https://cloud.google.com/text-to-speech

findings and specific diagnostics for those. Finally, the LLM was instructed not to include abbreviations and to provide the results in a JSON format without additional text. The prompt used can be seen in Figure 2.

You must generate {number_of_phrases} phrases in Portuguese that could be present in a {type_of_report} report made by a physician expert on a specific medical field you will be given as input. Generate the phrases and sentences following a logical chain of thought, starting from regular cases and progressing to possible findings and specific diagnostics related to the given context. Explore multiple phrase types, ranging from basic descriptions to detailed conclusions. Avoid using abbreviations, and every time you need to mention a specific term, use it in its most complete form (for example, use centimetres instead of cm and beats per minute instead of bpm).
Format the output: return a *JSON* object with the phrase list. Do not include any additional text before and after the *JSON*.
JSON output example:

```
{
  "phrases": [
    "O paciente apresenta ritmo cardíaco regular, com 72 batimentos por minuto.",
    "A imagem mostra um aumento moderado no tamanho do ventrículo esquerdo.",
    "Não há evidências de derrame pleural ou ascite."
  ]
}
```

Output only the *JSON* with the {number_of_phrases} phrases without additional texts.

Figure 2. Prompt used to generate domain-specific radiology texts. The example phrases and sentences are written in Portuguese to exemplify better the input we used.

As previously mentioned, we have used two TTS tools for the synthetic audio generation: *ElevenLabs* and *Google Text-to-Speech*. The use of both tools is meant to diversify the generated data with varying speaking patterns, rhythm, intonation and quality.

We generated 46 minutes and 43 seconds of audio using *ElevenLabs* in a total of 980 files. These files were equally split into five different male voices. As for the *Google Text-to-Speech*, we generated 58 minutes and 55 seconds of audio, again, in a total of 980 files, using only one male voice available. The dataset for the synthetically generated data is available in a GitHub repository[3].

Figure 3 (a) and (b) shows the audio length distribution for the synthetic dataset compared to the real, manually labelled data we had. As seen, the overall distribution is quite similar, while the synthetic data tends to be shorter, resulting in more files. The word cloud in Portuguese for both datasets can be seen in Figure 3 (c) and (d). Both datasets show domain-specific terms, with a greater presence of punctuation terms (commas, dots, etc) on the real dataset. Alternatively, the synthetic dataset has a higher presence of phrases such as “Não há” or “Há sinais” (meaning “There is no” and “There are signs of,” respectively in English), showing a tendency to repeat phrase structures with the same starting terms. The distribution of terms and times between TTS tools is fairly similar.

4.3. Analisys

Figure 4 shows the results for the WER metric for four different scenarios: a baseline (WhispherX without fine-tuning); WhispherX fine-tuned using the synthetic data; WhispherX using synthetic audio data generated from real radiology reports; WhispherX

[3]https://github.com/AtkLLM/AI-DrivenSpeechModel-Dataset

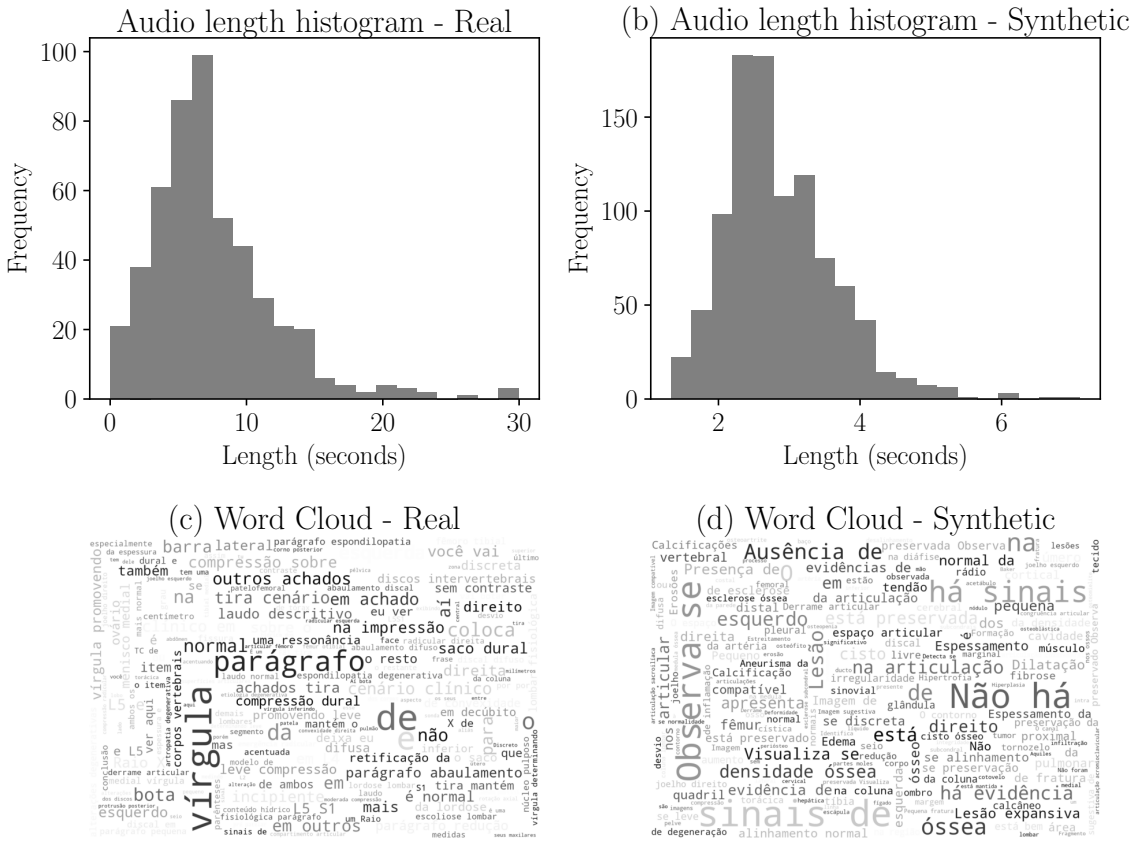

Figure 3. Histograms of audio length used: (a) histogram of the duration of manually labelled audio files used for training; (b) histogram of the duration of AI-generated audio files; (c) Word Cloud from Real Data; (d) Word Cloud from Synthetic Data;

fine-tuned using the ground truth (ideal scenario). The results are shown for both the WhisperX Small and Tiny versions, its smallest and more easily accessible forms.

In both cases, for the WhisperX Tiny and Small versions, the behaviour was fairly similar, with the Tiny version having considerably higher values for the WER compared to the Small version. The baseline, as expected, presented the highest WER value (104.13 and 37.35), while the ground truth version achieved the lowest one (40.11 and 23.11). It is worth mentioning that the ground truth Tiny version while performing significantly worse, achieved a similar WER when compared to the baseline Small version (40.11 compared to 37.35).

It is worth noting that the high WER value from the baseline WhisperX Tiny is the result of hallucination, which made the model include words that were not present in its results. That made the WER value rise above 100%.

Our proposed pipeline, which used only synthetic data for the fine-tuning process, achieved a WER of 62.28 and 33.38 for the Tiny and Small versions, respectively, which correspond to an improvement of 41.85% and 10.62%. These results were achieved using only the audio files generated by ElevenLabs, which performed better than the ones generated by the Google-TTS tool (WER equal to 70.96 and 33.94 for the Tiny and Small versions) and by a combination of both ElevenLabs and Google-TTS (WER equal to 66.75

and 33.94 for the Tiny and Small versions).

To exemplify a case in which there are some real-use data for the fine-tuning, we have tested using only real-case radiology reports (ignoring the LLM step) and producing the audio data from them using the same TTS tools mentioned previously. This new data was used to fine-tune both WhisperX Tiny and Small versions, achieving the WER of 56.24 and 30.76, respectively, which are 45.99% and 17.64% better than the baseline. For these results, we have only used audio data generated by ElevenLabs since it achieved better results in previous tests.

Our results show that it is possible to achieve better outcomes by using a completely synthetic approach. While it still performs worse compared to approaches with real-data approaches, it shows a promising approach that has plenty of room for experimentation and improvement and incurs a very low cost compared to generating a dataset with real data.

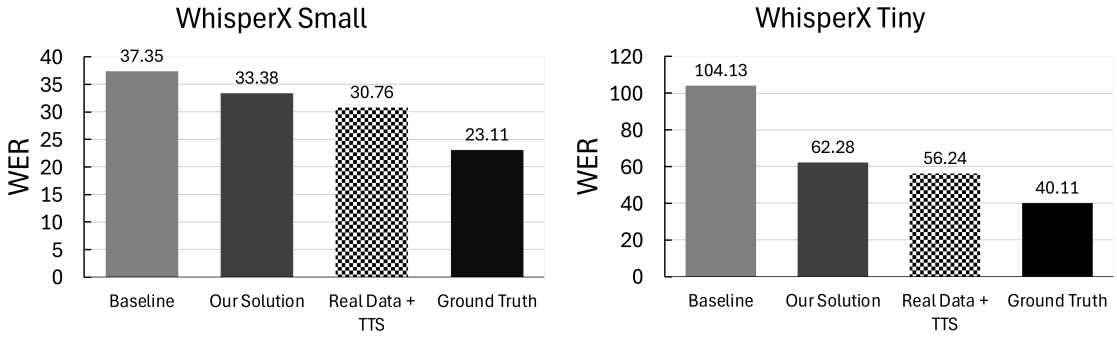

Figure 4. Results from the four approaches using both WhisperX Small and Tiny models. The WER metric is shown on the y-axis.

The ground truth results are, as expected, the best results with the lowest WER values for both models. However, it is also the most expensive approach with its caveats and challenges. Moreover, it is not unlikely that its best results are a consequence of some level of overfitting since the training and test data come from the same physicians using the same equipment in the same environments. On the other hand, the synthetic dataset was composed of a wider variety of voices and intonations that, while similar to the real ones in terms of context and intonation, are still fairly different. On that, the wider range of possible voices from the ElevenLabs tool might explain why it performed better than the Google-TTS tool. From our experiments, the Google-TTS tool tends to generate very clean and "perfect" robot-like audio files that are remote from real-use cases.

5. Conclusion

This work presented a pipeline for fine-tuning domain-specific STT solutions using synthetic data produced by a combination of LLM prompting and TTS tools. Our proposed pipeline produces good-quality synthetic data and overcomes the "perfect problem" by using TTS tools for a wider range of voices, intonation, and rhythm. Our findings show that our pipeline improves the results compared to a non-fine-tuned solution.

Given the results, we can also make assumptions based on the availability of real domain-specific data. As Figure 4 shows, and as expected, the more real data used, the

better the results. Yet, the difference between the use of some real data (using real-case reports data with TTS for audio generation) and 100% synthetic data is not significant (about 10% improvement for the Tiny model, and 8% improvement for the Small model, when comparing both approaches), indicating that in some cases, the synthetic-only approach might provide good enough results. Nevertheless, it is worth spending resources acquiring domain-specific knowledge and data, especially to produce a specialist LLM prompt required by our approach, but it will not necessarily reflect a significant improvement over the synthetic data.

Our choice of using WhisperX Tiny and Small models is focused on providing a low-cost solution for domain-specific scenarios. Higher WhisperX models are likely to provide better results, but they require expensive hardware and more resources for training. Besides that, higher models would require higher costs to host online for a production-ready solution. Considering our scenario, considerable investment would be required to host such a strategy for a single hospital with multiple simultaneous physicians working at the same time daily. Yet, our results indicate that, with the use of a ground truth dataset, it might be possible to improve a simpler model through fine-tuning to perform as well as a baseline better model, as we saw with the results from the ground truth fine-tuned Tiny model compared to the baseline Small model. In our preliminary tests, we found that the baseline WhisperX Medium model has a WER of 28.85, which is slightly higher than the ground truth fine-tuned WhisperX Small model we presented (23.11).

Besides operational costs, the complexity of the AI model used impacts its inference time (the time it takes to generate the output given the input). Simpler models, such as Tiny and Small, have a relative inference time significantly smaller than larger models [Bain et al. 2023]. For real-time settings, this is of major importance, such as the one explored in this study for radiology STT solutions.

As future work, our pipeline could be assessed for other domain-specific contexts, as well as more experimentation on the synthetic data variation that further approaches real-case scenarios, including the use of different accents, acoustic conditions, and background noise. The use of a more diverse ground truth set might also provide better insight into possible overfitting and more realistic results for fine-tuned models trained with it. It is not unlikely that a production-ready solution achieves a WER value closer to the results from our approaches than the current ground truth ones. Besides that, a longer audio ground truth dataset could surely provide better insights into our results since it was limited to a little over 1 hour long due to budget and time constraints.

Finally, a more fine-grained analysis of the balance between synthetic and real data could provide further insight into how much effort is needed to create hybrid approaches that more closely resemble real data, including the use of real audio instead of purely relying on TTS Tools. That might provide a great approach for fine-tuning STT models with a fraction of the usual associated costs when using high-quality ground truth sets.

6. Acknowledgements

The authors would like to thank the Brazilian Agency FUNCAP-CE for its financial support under the project NUP 31052.001303/2023-62. We would also like to thank Raiza Vaz for her help in building the ground truth database.

References

Ali, A. and Renals, S. (2018). Word error rate estimation for speech recognition: e-wer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 20–24.

Bain, M., Huh, J., Han, T., and Zisserman, A. (2023). Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Blackley, S. V., Huynh, J., Wang, L., Korach, Z., and Zhou, L. (2019). Speech recognition for clinical documentation from 1990 to 2018: a systematic review. *Journal of the American Medical Informatics Association*, 26(4):324–338.

Casanova, E., Shulby, C., Korolev, A., Junior, A. C., Soares, A. d. S., Aluísio, S., and Ponti, M. A. (2022). Asr data augmentation in low-resource settings using cross-lingual multi-speaker tts and cross-lingual voice conversion. *arXiv preprint arXiv:2204.00618*.

Chan, W., Jaitly, N., Le, Q., and Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 4960–4964. IEEE Press.

da Cruz, F. B., de Souza Britto, M. C., Moreira, G. M., and Junior, A. d. S. B. (2022). Robôs substituem juízes? o estado da arte da inteligência artificial no judiciário brasileiro. *Revista Antinomias*, 3(1):8–41.

Gontier, F., Serizel, R., and Cerisara, C. (2021). Automated audio captioning by fine-tuning bart with audioset tags. In *DCASE 2021-6th Workshop on Detection and Classification of Acoustic Scenes and Events*.

Gruzitis, N., Dargis, R., Lasmanis, V. J., Garkaje, G., and Gosko, D. (2022). Adapting automatic speech recognition to the radiology domain for a less-resourced language: the case of latvian. In *Intelligent Sustainable Systems: Selected Papers of WorldS4 2021, Volume 1*, pages 267–276. Springer.

Hammana, I., Lepanto, L., Poder, T., Bellemare, C., and Ly, M.-S. (2015). Speech recognition in the radiology department: a systematic review. *Health Information Management Journal*, 44(2):4–10.

Hu, T.-Y., Armandpour, M., Shrivastava, A., Chang, J.-H. R., Koppula, H., and Tuzel, O. (2022). Synt++: Utilizing imperfect synthetic data to improve speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7682–7686. IEEE.

Huang, Y., He, L., Wei, W., Gale, W., Li, J., and Gong, Y. (2020). Using personalized speech synthesis and neural language generator for rapid speaker adaptation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7399–7403. IEEE.

Islam, R. and Moushi, O. M. (2024). Gpt-4o: The cutting-edge advancement in multimodal llm. *Authorea Preprints*.

Johnson, M., Lapkin, S., Long, V., Sanchez, P., Suominen, H., Basilakis, J., and Dawson, L. (2014). A systematic review of speech recognition technology in health care. *BMC Med. Inform. Decis. Mak.*, 14(1):94.

Koenecke, A., Choi, A. S. G., Mei, K. X., Schellmann, H., and Sloane, M. (2024). Careless whisper: Speech-to-text hallucination harms. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1672–1681.

Kumar, Y. (2024). A comprehensive analysis of speech recognition systems in healthcare: Current research challenges and future prospects. *SN Computer Science*, 5.

Laptev, A., Korostik, R., Svischev, A., Andrusenko, A., Medennikov, I., and Rybin, S. (2020). You do not need more data: Improving end-to-end speech recognition by text-to-speech data augmentation. In *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 439–444. IEEE.

Li, J., Gadde, R., Ginsburg, B., and Lavrukhin, V. (2018). Training neural speech recognition systems with synthetic speech augmentation. *arXiv preprint arXiv:1811.00707*.

Mak, F., Govender, A., and Badenhorst, J. (2024). Exploring asr fine-tuning on limited domain-specific data for low-resource languages. *Journal of the Digital Humanities Association of Southern Africa (DHASA)*, 5.

Rosenberg, A., Zhang, Y., Ramabhadran, B., Jia, Y., Moreno, P., Wu, Y., and Wu, Z. (2019). Speech recognition with augmented synthesized speech. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 996–1002. IEEE.

Samarakoon, L., Mak, B., and Lam, A. Y. (2018). Domain adaptation of end-to-end speech recognition in low-resource settings. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 382–388. IEEE.

Silva, M. d. L. M., Mendonça, A. L. C., Neto, E. R. D., Chaves, I. C., Caminha, C., Brito, F. T., Farias, V. A. E., and Machado, J. C. (2024). Facto dataset: A dataset of user reports for faulty computer components. In *Anais do VI Dataset Showcase Workshop*, pages 1–12. SBC.

Suh, J., Na, I., and Jung, W. (2024). Improving domain-specific asr with llm-generated contextual descriptions.

Vásquez-Correa, J. C., Arzelus, H., Martin-Doñas, J. M., Arellano, J., Gonzalez-Docasal, A., and Álvarez, A. (2023). When whisper meets tts: Domain adaptation using only synthetic speech data. In *International Conference on Text, Speech, and Dialogue*, pages 226–238. Springer.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Vivancos-Vicente, P. J., Castejón-Garrido, J. S., Paredes-Valverde, M. A., Salas-Zárate, M. d. P., and Valencia-García, R. (2016). Ixhealth: A multilingual platform for advanced speech recognition in healthcare. In *Technologies and Innovation: Second International Conference, CITI 2016, Guayaquil, Ecuador, November 23-25, 2016, Proceedings 2*, pages 26–38. Springer.

Vorbeck, F., Ba-Ssalamah, A., Kettenbach, J., and Huebsch, P. (2000). Report generation using digital speech recognition in radiology. *European Radiology*, 10:1976–1982.

Yang, K., Hu, T.-Y., Chang, J.-H. R., Koppula, H. S., and Tuzel, O. (2023). 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)*, pages 1–5. IEEE.

Yu, D., Deng, L., and Dahl, G. (2010). Roles of pre-training and fine-tuning in context-dependent dbn-hmms for real-world speech recognition. In *Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning*. sn.