Postprint of: Czyżewski A., Optimizing Medical Personnel Speech Recognition Models Using Speech Synthesis and Reinforcement Learning, Journal of the Acoustical Society of America, Vol. 154, iss. 4 (2023), pp. A202–A203, DOI: 10.1121/10.0023271

Optimizing Medical Personnel Speech Recognition Models Using Speech Synthesis and Reinforcement Learning

Andrzej Czyżewski

Gdańsk University of Technology ETI Faculty, Multimedia Systems Department

e-mail: ac@pg.edu.pl

Abstract: Text-to-Speech synthesis (TTS) can be used to generate training data for building Automatic Speech Recognition models (ASR). Access to medical speech data is because it is sensitive data that is difficult to obtain for privacy reasons; TTS can help expand the data set. Speech can be synthesized by mimicking different accents, dialects, and speaking styles that may occur in a medical language. Reinforcement Learning (RL), in the context of ASR, can be used to optimize a model based on specific goals. A model can be trained to minimize errors in speech-to-text transcription, especially for technical medical terminology. In this case, the "reward" to the RL model can be negatively proportional to the number of transcription errors.

The paper presents a method and experimental study from which it is concluded that the combination of TTS and RL can enable the creation of a speech recognition model that is better suited to the specific needs of medical personnel, helping to expand the training data and optimize the model to minimize transcription errors. The learning process used reward functions based on Mean Opinion Score (MOS), a subjective metric for assessing speech quality, and Word Error Rate (WER), which evaluates the quality of speech-to-text transcription.

1. Objectives and Goals

As a result of the project, a solution will be developed and implemented with the use of which physicians will recall available diagnostic test results and clinical parameters of patients by voice,
fill in disease charts in the course of medical history in interactive mode, create descriptions and prescribe treatment as required. The system automatically generates templates for filling them
out, allowing data to be entered directly into widespread health care information systems, including data from the medical history and automatically structured descriptions of diagnostic results
that will be voice-editable and will enable dictation of test referrals, prescriptions and sick leave.
The cloud-based speech recognition system will be built on the basis of a corpus of Polish speech augmented with a dictionary of medical terms and commercial drug names for use by doctors
of various specialties, including radiologists, surgeons, doctors working in hospital emergency departments, and specialists providing medical advice. A variant of the solution enhanced with two-way communication, based on both speech recognition and speech synthesis, will be able to be used in situations where a doctor cannot operate a text editor manually. It is envisaged to build a
comprehensive speech corpus based on recordings made in typical acoustic conditions of doctors' offices and, in addition, in surgical masks, operating rooms, and in conditions that make
effective speech recognition difficult, i.e. in the presence of interference and simultaneous speech. For this purpose, an innovative acoustic probe will be developed to allow speaker separation.
Neural algorithms, trained using state-of-the-art methods for effective speech recognition of medical personnel in Polish, will be built and made available in cloud computing.

Text-to-Speech synthesis (TTS) can be used to generate training data for building Automatic Speech Recognition models (ASR). The model can be trained to minimize errors in speech-to-text transcription, especially for technical medical terminology. In this case, the "reward" to the RL model can be negatively proportional to the number of transcription errors. The paper presents a method and experimental study from which it is concluded that the combination of TTS and RL can enable the creation of a speech recognition model that is better suited to the specific needs of medical personnel, helping to expand the training data and optimize the model to minimize transcription errors. The learning process used reward functions based on Mean Opinion Score (MOS), a subjective metric for assessing speech quality, and Word Error Rate (WER), which evaluates the quality of speech-to-text transcription.

2. Data Foundations: Building the Pillars of Insight

Medical phrases from various fields such as oncology, pathomorphology, radiology, the emergency room, and surgical procedures, were recorded. In addition to medical phrases, our data set also includes information on drugs and their intake. An illustrative depiction of the acquisition process during surgery is provided in Figure 1, featuring the use of a sonic probe.

Fig 1. Placement of the probe in the procedure room of a clinical emergency department

Furthermore, experts across various medical disciplines actively contribute by adding recordings and their corresponding transcriptions to our server shown in Figure 2. Recognizing the inherent challenges of gathering data in dynamic environments like emergency rooms and during surgical procedures, we have innovated a state-of-the-art sonic probe. It was developed at the Multimedia Systems Department at the University of Technology, enables selective and directional gathering in noisy environment. The schematic diagram of our 6-channel acoustical probe is presented in Figure 3 below.

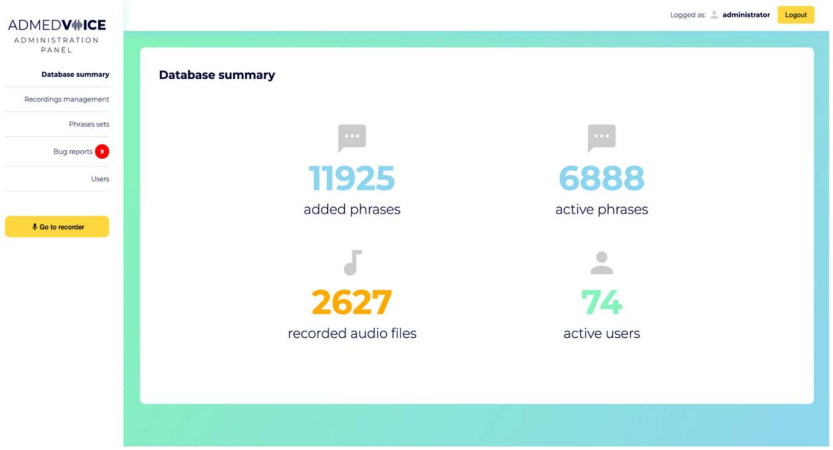

Fig 2. Server repository for data acquisition

In addition to the obtained recordings, our database includes synthesized speech created with the assistance of the SpeechGen system, a powerful tool known for generating natural and fluent speech. SpeechGen offers 19 different lectors in the Polish language with the flexibility to adjust tempo, pitch, and intonation.

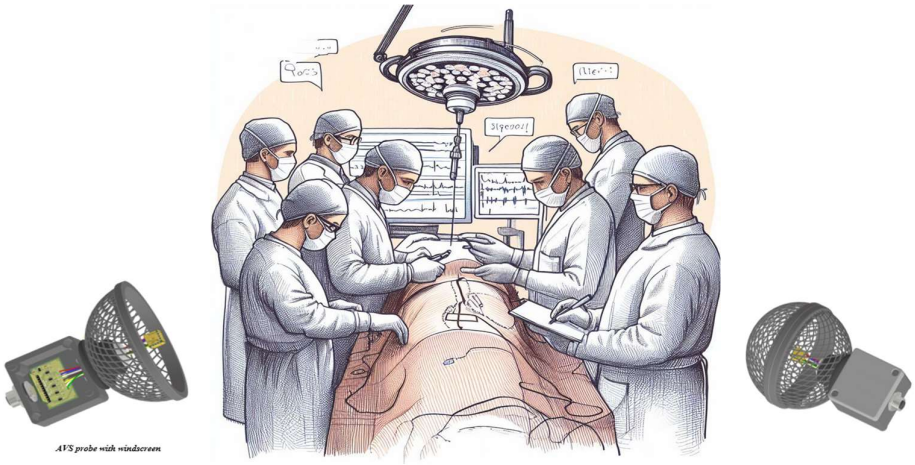

Fig 3. Diagram of the acoustic probe

3. Base model and training

The Whisper model by OpenAI, serves as the foundational tool in our research endeavor. Within the project scope, this model has been tailored to our specifications, leveraging the Transformer architecture in a full-scale sequence-to-sequence encoder-decoder setup. To meet the project's objectives, we opted for the utilization of a pre-trained small-scale model, encompassing 224 million parameters. This model underwent training for 680 thousand hours on labeled recordings. To apply the model in a medical setting, the model requires further training. The model is trained on pairs of audio recordings and corresponding text, where the length of the recording is standardized to 30 seconds. Shorter recordings are augmented with silence, while recordings longer than 30 seconds are truncated. The input audio is transformed into a mel-scale spectrogram during preprocessing.

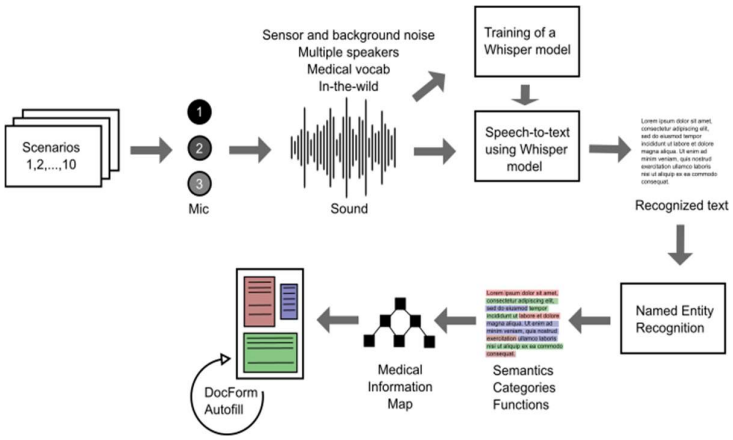

Fig 4. Block diagram of the presented system

The diagram illustrates a block schematic of a speech-to-text process with applications in the medical field, focusing on named entity recognition. Below is a description of the steps depicted in the schematic:

Scenarios: Initially, there are various scenarios labeled with numbers 1 to 10, representing different situations or data inputs to the system. These scenarios feed into a microphone.

Sound: The microphone collects audio containing background noise, multiple speakers, specialized medical vocabulary, and ambient sounds.

Whisper Model: The audio is then processed by the Whisper model, trained to convert speech to text under these conditions

Recognized Text: The text generated by the Whisper model appears as a block of text, likely a speech transcription.

Named Entity Recognition: The transcription is further processed by a named entity recognition module, which identifies and categorizes crucial information such as names and medical terms.

Medical Information Map: Information from the named entity recognition is used to create a medical information map, which may encompass semantics, categories, and functions related to medical data.

Document Form Autofill: At the end of the process, the recognized and organized information is used to automatically fill in medical forms.

This block schematic suggests that the entire process is an automated natural language processing (NLP) system aimed at facilitating medical documentation by automating the transcription and recognition of medical terminology.

4. Conclusions and Next Steps

Our research demonstrates the potential to expand training data and optimize the ASR model for medical professionals. We focused on acquiring high-quality data from diverse sources, including the use of innovative tools like the 6-channel sonic probe. In the future, we plan to further train the ASR model to better suit the needs of medical professionals and improve the accuracy of medical speech transcription. Our work aims to support healthcare professionals in Poland and contribute to enhancing patient care.

Acknowledgments

Polish National Center for Research and Development (NCBR) supported research in the project: "ADMEDVOICE- Adaptive intelligent speech processing system of medical personnel with the structuring of test results and support of therapeutic process", no. INFOSTRATEG4/0003/2022.

Literature

[1] Distance package documentation. https://pypi.org/project/Distance/, (access 07.05.2023).
[2] Jiwer package documentation. https://pypi.org/project/jiwer/, (access 07.05.2023).
[3] Speech-to-text: Automatic speech recognition - google cloud, (access 07.05.2023). https://cloud.google.com/speech-to-text.
[4] Speech to text − audio to text translation | microsoft azure.
https://azure.microsoft.com/en-us/products/cognitive-services/speechto- text/, (access 07.05.2023).
[5] Graves A. Sequence transduction with recurrent neural networks. https://arxiv.org/abs/1211.3711.
[6] Kevin Chu, Leslie Collins, and Boyla Mainsah, 'Using automatic speech recognition and speech synthesis to improve the intelligibility of cochlear implant users in reverberant listening environments', in

ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6929−6933, (2020).

[7] Li Fu, Xiaoxiao Li, Libo Zi, Zhengchen Zhang, Youzheng Wu, XiaodongHe, and Bowen Zhou, ' Incremental learning for end-to-end automatic speech recognition', in 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 320−327, (2021).

[8] Ayoub Ghriss, Bo Yang, Viktor Rozgic, Elizabeth Shriberg, and Chao Wang, 'Sentiment-aware automatic speech recognition pre-training for enhanced speech emotion recognition', in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7347−7351, (2022).

[9] Google. Google stt documentation. https://cloud.google.com/speechto- text/docs, (access 07.05.2023).

[10] Chiu Ch. Parmar N. Zhang Y. Yu J. Han W. Wang S. Zhang Z. Wu Y. Pang R. Gulati A., Qin J. Conformer: Convolution-augmented transformer for speech recognition. https://arxiv.org/abs/2005.08100.

[11] Casper J. Catanzaro B. Diamos G. Elsen E. Prenger R. Satheesh S. Sengupta S. Coates A. Ng A. Y. Hannun A., Case C. Deep speech: Scaling up end-to-end speech recognition. https://arxiv.org/abs/1412.5567.

[12] Shigeki Karita, Shinji Watanabe, Tomoharu Iwata, Marc Delcroix, Atsunori Ogawa, and Tomohiro Nakatani, 'Semi-supervised end-to-end speech recognition using text-to-speech and autoencoders', in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6166−6170, (2019).

[13] Bueno Ltd. Voice synthetiser speechgen.io. https://speechgen.io/pl/, (access 07.05.2023).

[14] Andrew Cameron Morris, Viktoria Maier, and Phil Green, 'From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition', in Proc. Interspeech 2004, pp. 2765−2768, (2004).

[15] OpenAI. Whisper, latest release. official github. https://github.com/openai/whisper/releases/tag/v20230314, (access 07.05.2023).

[16] Tao Xu Brockman G. McLeavey Ch. Sutskever I. Radford A., Kim W. J. Robust speech recognition via large-scale weak supervision. https://arxiv.org/abs/2212.04356.

[17] Audacity Team. Audacity documentation. https://www.audacityteam.org/, (access 07.05.2023).

[18] Punitha Vancha, Harshitha Nagarajan, Vishnu Sai Inakollu, Deepa Gupta, and Susmitha Vekkot, ' Word-level speech dataset creation for sourashtra and recognition system using kaldi', in 2022 IEEE 19th India Council International Conference (INDICON), pp. 1−6, (2022).

[19] Parmar N. Uszkoreit J. Jones L. Gomez A. N. Kaiser L. Polosukhin I. Vaswani A., Shazeer N. Attention is all you need. https://arxiv.org/abs/1706.03762.

[20] Wei Wang, Shuo Ren, Yao Qian, Shujie Liu, Yu Shi, Yanmin Qian, and Michael Zeng, 'Optimizing alignment of speech and language latent spaces for end-to-end speech recognition and understanding', in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7802−7806, (2022).