Multimodal Approaches for Natural Language Processing in Medical Data

Oleh Basystiuk[a], Nataliia Melnykova[a]

[a] *Lviv Polytechnic National University, Lviv, 79000, Ukraine*

Abstract

Nowadays, Artificial Intelligence became widespread and deeply integrated into our life routines. One of the most interesting and fast-growing technology in the Artificial Intelligence is speech recognition, and it's a part of the multimodal data concept, which includes voice, audio, and text data. The paper overview the possibilities of multimodal approaches for natural language processing problem based on audio data in the field of medical data. Generally, there are there main concepts: Sequence-to-Sequence, Deep Neural Networks based on Hidden Markov Model, and Connectionist Temporal Classification based on the End-to-End model. In this research we review the possibility to utilize natural language processing methods in the medical sphere, to increase the overall time efficiency of medical workers, and optimize mechanical work related to fulfilling information or transforming it from audio to text data. Furthermore, it was realized a comparative analysis of the existing approaches, to select the most advanced and reliable for building robust multimodal audio-to-text systems and conducting future research. This research in the future could be utilized to create wide range Speech-to-Text models for specific medical fields which will improve speech translation tasks, reduce workload and improvise the time efficiency of medical workers.

Keywords 1

Speech-to-Text, speech recognition, deep neural network, sequence-to-sequence, connectionist temporal classification.

1. Introduction

Speech-to-Text is a very popular technology that is widely adopted and used in a nowadays environments and business applications. AI-based programs increase the effectiveness of many fields, which are related to audio and text domains, such as journalistic, jurisprudence, support, entertainment, etc. It can also influence and stimulate the medical field, especially patient care, including supplementary information and detection, audio interpretation, computer integration and text classification. Healthcare-related software developers work with healthcare services to develop new ML-based programs to deliver integrated solutions that help lead the healthcare industry in the next few years [1-3].

Today, speech-to-text systems model the work of an interpreter. Their effectiveness depends on the language's ability to understand the grammar, recognize patterns, and are often closely related to the domain in which it is intended to be used, since each domain has its own terminology and limitations. In audio-to-text transformation, the main aspect units are not a single word, but the whole sentences or phraseological units, to explaining idea of it in general. Only by using them we could handle complex ideas be expressed in existing data chunk [3]. As a result of this approach, we faced multiple specific interfaces that were built differently and they are doing good only in the limited area of tasks, so the main problem of the next iteration of audio-to-text product development is as least to create a unique expert level of

IDDM'2022: 5th International Conference on Informatics & Data-Driven Medicine, November 18–20, 2022, Lyon, France
EMAIL: oleh.a.basystiuk@lpnu.ua (A. 1); nataliia.i.melnykova@lpnu.ua (A. 2);
ORCID: 0000-0003-0064-6584 (A. 1); 0000-0002-2114-3436 (A. 2);

systems for different fields, and in this research, we will review the medical sphere, select areas where multimodal data are used and analyze ways how we could increase its handling productivity, moreover receive extra benefits from data summary and analyzation after all.

In a previous conducted study [6-8], wide range of possible use cases for speech recognition systems was overviewed. Moreover, it was showcased the open APIs interface benefits, and ability of building up a unified interface which are based on recurrent neural networks.

This paper aims to review the medical sphere, select areas where multimodal data are used and analyze ways how we could increase its handling productivity, compare a wider range of existing solutions, including linear, nonlinear, and multimodal data, and select the most effective way to implement speech recognition systems to utilize them in future as a platform for medical speech recognition system proposition. Moreover, evaluate the most time-effective and accurate approach, that could be utilize in a wide range of future research, related to audio-to-text and speech-to-text domain.

The main contributions of this paper:

1. Recognize possible use cases of audio-to-text utilization in medical sphere, prepared data models for future data collection and based on them, create method to cover various data sources.
2. Analyze existing architecture to recognize their time complexity and productivity in wide range of proposed cases.
3. We explored different methods to solve speech recognition tasks; conducted evaluation of the performance on potentially different size of vocabulary, namely, 2k, 5k, 10k.
4. The evaluation and analysis of speech recognition performance is carried out using different language modeling units and approaches, such as deep neural networks (DNNs), connectionist temporal classification (CTC) and sequence-to-sequence (Seq2seq). These language models help explore the impact of context-independent and contextual recognition models in the medical field.
5. Speech recognition performance has been compared and better results have been proposed for future implementation as a platform basement for future solutions in the medical sphere predefined in the initial sections of the article.

The paper is organized as follows: different types of well-known speech-to-text methods are discussed in Section 2. Description of how these technologies can be used in the medical field and overview of the value chain of proposed speech recognition methods are presented in Section 3. The experimental setup and results presented in Section 4, Result overview and discussion provided in Section 5. Conclusions and future work ideas presented in Section 6.

2. Medical Speech Recognition Value Chain

We first performed a systematic analysis of the literature, and then we complemented and validated our findings with expert interviews in order to provide a thorough overview of the pertinent obstacles and potential of speech-to-text recognition in the medical field. Scientific research uses systematic literature reviews as a key method for effectively combining available data and addressing a particular research question. It is unclear how extent machine learning will actually influence the medical industry, as some scholars have suggested application cases at the prototype level or a theoretical construct [4-5]. For this reason, we confirm and supplement the opportunities and challenges identified in the literature review from a practical perspective. Expert interviewing is considered a qualitative-empirical research method to enhance understanding, generate new collect concrete data, and based on them insights [9-10]. Moreover, include expert interviews and literature review, to identify benefits as increased potency and effectiveness and present them in Fig. 1.

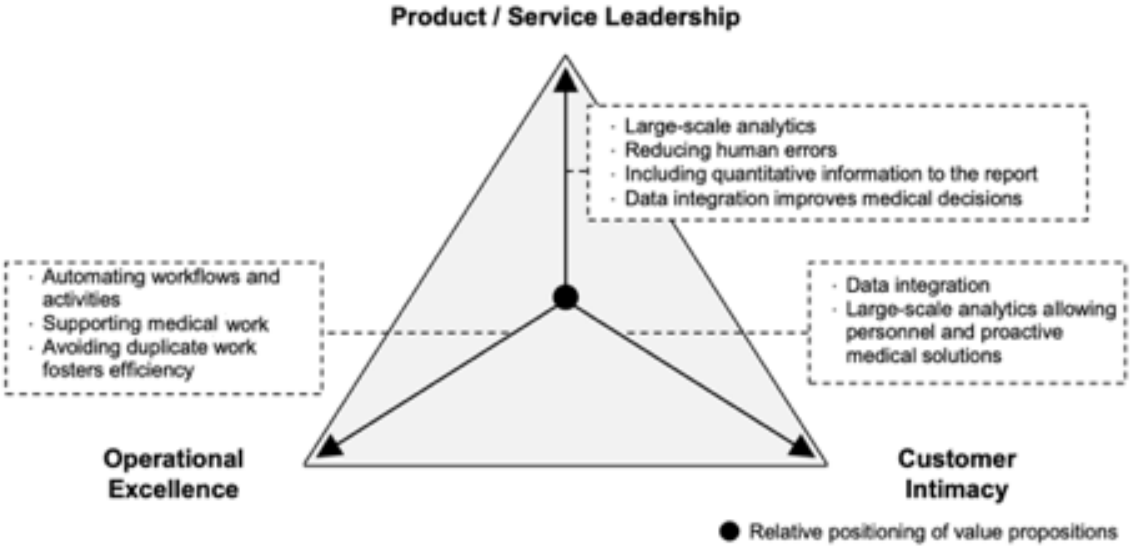

Figure 1: Value chain of Medical Speech Recognition

3. Concepts overview

After detailed elaboration of the list of possible areas of application of audio-to-text transformations in medicine and the formation of a clear list of requirements and the future structure of datasets with which it is necessary to work, it is worth moving on to the issue of choosing one of the approaches for recognition in order to build an integrated solution for a wide range of applications based on it in the future medical tasks that were listed in the previous section. The most popular approaches to converting audio information into text include:

1. Speech recognition systems use deep neural networks (DNNs), namely the hybrid DNN based on Hidden Markov Model system. The hybrid system beats traditional Gaussian model systems dramatically on a variety of big vocabulary continuous voice recognition tasks by combining the strengths of deep neural networks learning with sequential modeling capabilities, based on Markov model approach. By contrasting a variety of system configurations, we represent general example of training systems and highlight the essential elements of the systems at the Figure 1.

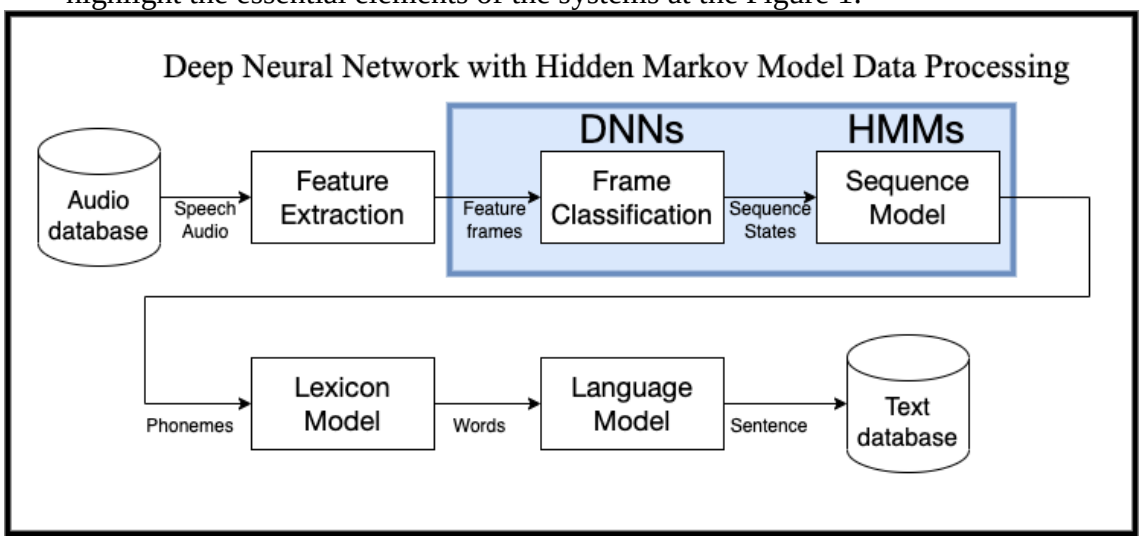

Figure 2: Deep Neural Network with HMM (Hidden Markov Model) flow

2. Recurrent neural networks' output layer for connectionist temporal classification (CTC). CTC was created primarily for temporal classification jobs, or sequence labeling issues where the alignment between the inputs and the target labels is unclear, as its name suggests. Contrary to the hybrid approach covered in the previous chapter, CTC uses a single neural network to represent every feature of the sequence and does not call for the network to be merged with a hidden Markov model [11]. The label sequence can be extracted from the network outputs without the need for training data.

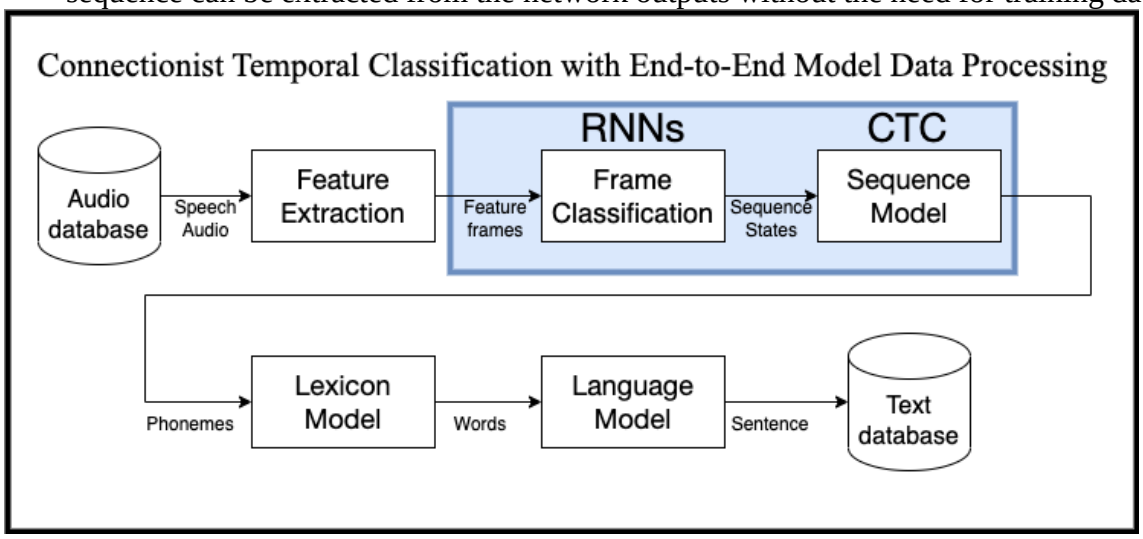

Figure 3: CTC (Connectionist Temporal Classification) with end-to-end approach flow

3. One of the deep learning models, called Seq2Seq (sequence-to-sequence) have excelled in tasks like machine translation and text summarization. According to Seq2Seq models, decoder's attention layers can only access the words that come before a certain word in the input, whereas the encoder's attention layers can access every word in the original phrase. The presence of connections allows RNN to memorize and reproduce the entire sequence of reactions to one stimulus. Initially, the sequence is sent into the encoder, which is made up of RNNs, who then creates a final embedding at the end of the sequence. The Decoder receives this and utilizes it to forecast a sequence [12]. After each prediction, it uses the prior hidden state to predict the following instance of the sequence.

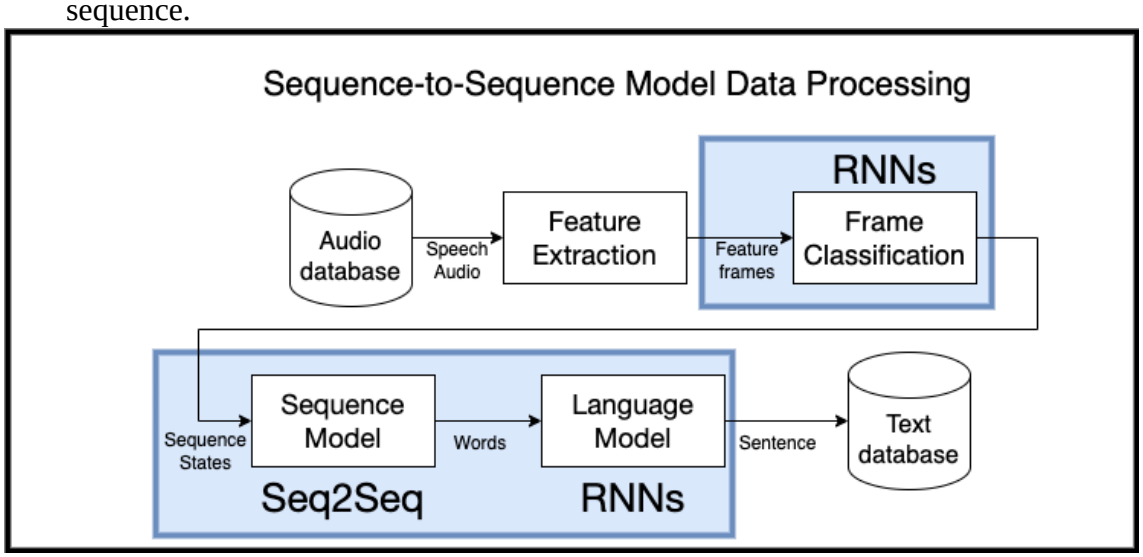

Figure 4: Sequence-to-Sequence model flow

We need to undertake study in order to acquire the maximum accuracy and time efficiency score for our particular situation based on the ways described above. The most effective method was our Sequence-to-Sequence based on the Recurrent Neural Network approach when examining the libs used for building machine learning approaches, in the prior research. The most widely used machine learning (ML) libraries for an RNN-based method to language translation are Tensorflow, Keras, and PyTorch.

4. Results

After learning and decoding, results of the experiments were divided into three main categories and described: Acoustic modeling using symbols based on verbal comparison of results with training samples, checking for correctness of grammatical structures, and expert discussion by experts. The hybrid RNN with sequence-to-sequence model uses a combination of two models. RNN models combined with position-based attention decoders also help networks accelerate learning. Each training description and the results of their decoding were presented as a model on phoneme, symbolic, grammatical analysis. We discuss the results of each inter-sequence model separately, and take the character-based results as a basis for future research.

In character-based tests, we used word-based RNNs to test their recognition performance across different test vocabulary sizes. The training is done with different vocabulary sizes. 2.5k, 5k, 10k, decoded using RNN word level. These vocabulary dimensions are part of the training dataset used in the thematical medical articles and dictionaries. Word sequences are generated easily in the word-based models via selecting the following corresponding vertices. All predictions generated with the corresponding vocabulary size are scored within the 5k score test range. The results were evaluated using the sequence RNN method for character error rate (CER) and word error rate (WER). The minimum results entered with a datasets size of 10K and the summary for different datasets are shown in Table 1.

Table 1

Sequence-to-sequence model results

Dataset Size	CER	WER
2500	32.04%	41.06%
5000	28.19%	39.30%
1000	25.42%	36.60%

Multimodal data augmentation techniques can help make low-resource languages competitive in sequence-to-sequence methods by scaling the training datasets. A character is a linguistic unit used during character-based sequence-to-sequence character-based language modeling. The minimum CER and WER received during experiments were 24.32% and 41.06%. WER increased when compared to the word results above. Unlike WER, CER is reduced by context independence. The results support a continuation of experiments with other datasets generated by segmentation algorithms that took into account context, ecpesially relevant to the medical field.

We compared letter-based WER with phoneme-based WER using different vocabularies based on the most common words. The performance of RNN attention methods was slightly better than the DNN based on Hidden Markov Model or as Connectionist Temporal Classification based on end-to-end labels. These test medical datasets are found most effective using the Seq2Seq transformation algorithm, and transferring all the data as one paragraph or sentence, to take into account general meaning of it, and do not concentrate on world based translation.

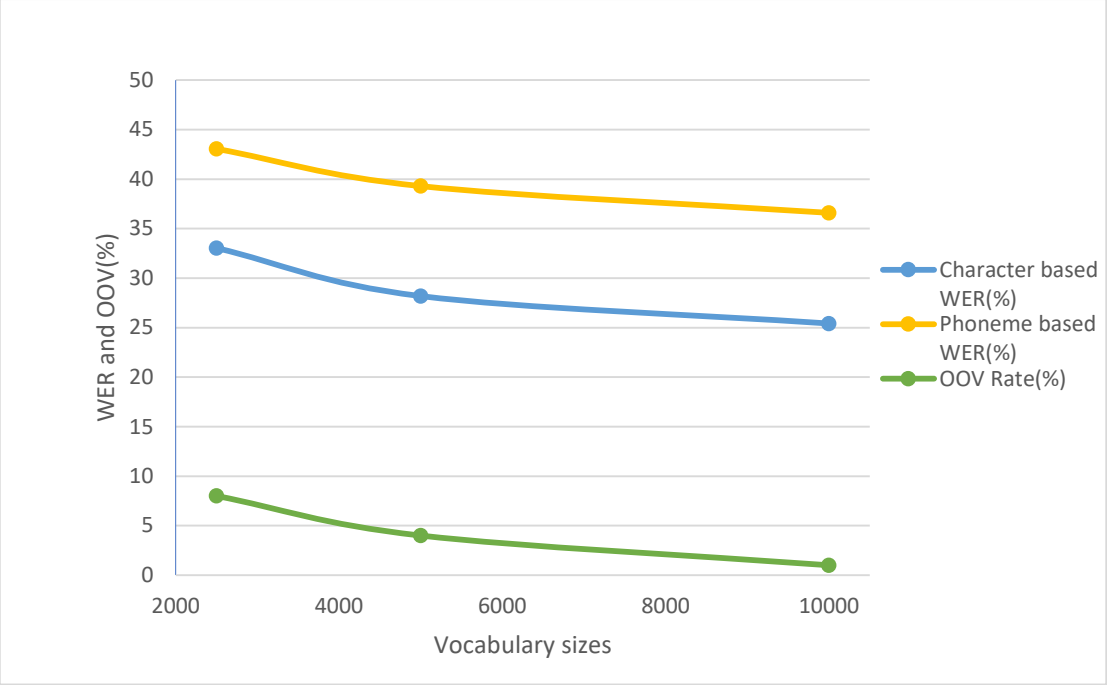

Figure 5: Visualization of table 1

5. Discussion

Speech-to-text approaches and developing and scaling fast nowadays, so we need to scale to other fields and make them more reliable and accurate in a wide range of fields, moreover think about the future integrity of conducted research. This research is a basement for future series of research, so here we conducted the initial step of predefining fields and tasks, which future platforms will need to solve. We also, analyze the techniques for training neural networks that are not just applicable to machine translation. Based on section 3 we realized that utilization of the hybrid approach based on recurrent networks with a sequence-to-sequence model, in our opinion, will provide optimal results and will provide high time efficiency with a good accuracy rate. RNNs are now one of the most popular technologies utilized in audio-to-text translation, and we anticipate upgrades in this area in the near future.

6. Conclusion

In this research, we present the results of our investigation into Multimodal Approaches for Medical Speech Recognition. Primarily, we begin developing data models for predetermined data, the medical industry, and use cases that will be applied throughout the value chain. We choose a method for audio-to-text transformation using recurrent neural networks based on seq2seq model algorithms as a result. With the help of libraries, we can make the most of this audio data by extracting features from these multimodal data using techniques like speech recognition. These data can be used for a variety of tasks after being converted to text utilizing the Natural Language Processing method. This study will be the basis for future research series, so here we have taken the first step to predefine the areas and tasks that future platforms will need to solve. We also analyze techniques for training neural networks as well as machine translation. Moreover, to decrease the error rate, reduce latency with no harm for performance is an important topic for the future model. We also plan to research and propose an expert-level system for hybrid language translation in the medical field.

7. References

[1] Dong Yu, Li Deng. "Automatic Speech Recognition: A Deep Learning Approach" Springer Longon, 2015. DOI: 10.1007/978-1-4471-5779-3.

[2] Ivan Izonin, et. al., "The Combined Use of the Wiener Polynomial and SVM for Material Classification Task in Medical Implants Production", International Journal of Intelligent Systems and Applications (IJISA), Vol.10, No.9, pp.40-47, 2018. DOI: 10.5815/ijisa.2018.09.05

[3] Vitaly Yakovyna, Natalya Shakhovska, "Software failure time series prediction with RBF, GRNN, and LSTM neural networks", Procedia Computer Science 207(4):837-847, DOI:10.1016/j.procs.2022.09.139.

[4] Ivan Izonin, et. al., "A GRNN-based Approach towards Prediction from Small Datasets in Medical Application", Procedia Computer Science. 184, 2021, pp. 242–249.

[5] Nataliya Shakhovska, et. al.: "The Developing of the System for Autimatic Audio to Text Conversion", IT&AS'2021: Symposium on Information Technologies and Applied Sciences, March 5–6, 2021, Bratislava, Slovak Republic.

[6] Nataliya Shakhovska, et. al. "Big Data analysis in development of personalized medical system", The 10th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN), 160, 229-234. (2019)

[7] Nataliya Melnykova et. al. (2019). "Using big data for formalization the patient's personalized data". Paper presented at the Procedia Computer Science, 155 624-629.

[8] Eshete Derb Emiru et. al. (2021) "Improving Amharic Speech Recognition System Using Connectionist Temporal Classification with Attention Modeland Phoneme-Based Byte-Pair-Encodings", Information, 12: 1-22.

[9] Nataliya Boyko, et. al.: "Usage of Machine-based Translation Methods for Analyzing Open Data in Legal Cases". In: Proc. of the CybHyg-2019, Kyiv, Ukraine, November 30, 2019, pp. 328–338. CEUR-WS.org.

[10] Berezsky O., Dubchak L., Batryn N., Datsko T., Berezska K., Pitsun O., Batko Y. Fuzzy System For Breast Disease Diagnosing Based On Image Analysis. Proceedings of the II International Workshop Informatics & Data- Driven Medicine (IDDM 2019). Lviv, Ukraine. 11-13 November, 2019.

[11] Berezsky O., Verbovyy S., Pitsun O. Hybrid Intelligent information techology for biomedical image processing. Proceedings of the IEEE International Conference «Computer Science and Information Technologies» CSIT'2018, Lviv. Ukraine, 11-14 September, 2018. P. 420-423.

[12] Zoryana Rybchak, et. al. "Analysis of methods and means of text mining". ECONTECHMOD, 6(2), 2017, pp. 73-78.

[13] GitHub Repository "Speech recognition algorithms performance evaluation". https://github.com/obasys/speech-recognition-algorithms-performance-evaluation. (accessed Jun. 7, 2022)