*INTERSPEECH 2021: Show & Tell Contribution*
30 August – 3 September, 2021, Brno, Czechia

# Automatic Radiology Report Editing through Voice

*Manh Hung Nguyen*[1], *Vu Hoang*[1], *Tu Anh Nguyen*[1], *Trung H. Bui*[2]

[1]VinBrain LLC, Vietnam
[2]Independent Researcher, USA
{v.hungngm, v.vuhoang, v.tuna17}@vinbrain.net, bhtrung@gmail.com

## Abstract

We present a system that allows radiologists to edit the radiology report through their voices. This is a function in our bigger system at VinBrain LLC that uses AI algorithms to assist radiologists with chest x-ray diagnosis, the system can suggest the abnormalities, then bases on the radiologist's confirmations or conclusions to automatically generate the report using predefined templates. We then allow the radiologist to freely edit the report using voice. The system combines two components, the first is the Speech Recognition System (SRS), and the second is the Natural Language Understanding System (NLUS) that executes the user's command. The user can delete, modify or add an arbitrary whole sentence. In addition, we successfully developed an SRS for such a non-mainstream language as Vietnamese and adapted it for the radiology domain.

**Index Terms**: radiology reporting, speech recognition, voice recognition, natural language understanding

## 1. Introduction

Speech recognition technology has been used in radiology reporting for a long time [1][2]. It has proved positive effects on report producing productivity [2][3][4]. Normally, when radiologists using this technology, they dictate the whole report, edit it if necessary and accept it. Some common report sentences may be dictated case by case. In our system, after the radiologist confirms and concludes the abnormalities in the image, the report will be generated automatically based on predefined templates. Just in case the radiologist wants to change some sentences, we designed a special module so they can edit it freely by voice. Our module consists of two submodules as shown in Fig. 1. The first one is the Speech Recognition System (SRS) which is adapted to the radiology domain. We successfully developed an SRS for such a non-mainstream language as Vietnamese. The second one is the Natural Language Understanding System (NLUS), which receives the transcribed text and parses it to execute the user's request. We discuss the details of these submodules in the following sections.

## 2. The report editing module

This module helps the radiologist to change, add or delete a whole sentence using voice. The user's voice is recorded and transcribed into text. The text will be analyzed to understand the user command and extract the relevant content. We do not strict the user to a limited set of command templates, so the user can naturally express the commands.

### 2.1. The speech recognition system

We adapt the architecture of the QuartzNet [5] for Vietnamese. We first preprocess the audio signal and input it into the QuartzNet model. The preprocessing transforms the time series audio signal, which is a sequence of sound pressure over time, to mel spectrogram, which shows the evolution of the mel frequency spectrum in time. The QuartzNet we used consists of a 1D conv layer, followed by 5 groups of blocks and 3 additional 1D conv layers. Each block has 5 identical modules, each module consists of a depthwise conv layer, a pointwise conv layer, a normalization layer and a ReLU layer. There is a residual connection between blocks. Each block is repeated 3 times. The model is trained with the Connectionist Temporal Classification (CTC) loss [6]. We set the number of output channels of the last layer corresponding to the number of characters in our character set for Vietnamese, which is 145.

For the dataset, we used an internal dataset for Vietnamese consists of about 11,000 hours of speech in general language and additional 95 hours of speech for radiology. The first dataset is recorded by most of the employees in our corporation, VinGroup, the largest conglomerate in Vietnam. There is a total of about 44,000 voices, the representative for all regional areas, gender, and a large range of age. The second dataset is recorded by hiring tens of students.

Finally, we achieved 3.69% Word Error Rate (WER) for the Vietnamese radiology language in a held-out test set. There are 37 voices in this dataset separated from the voices used in training with diverse ages, genders, and regions. The total recorded files are 1,413.

### 2.2. The natural language understanding system

The NLUS does two tasks. The first is an intent classification which classifies the command into addition, deletion, modification, or unknown. The second task is sequence tagging which tags each token in the command with whether it expresses the intent, position, or content.

#### *2.2.1. Data generation*

For training both tasks, we borrowed the multi-task learning with shared weight. We generate commands using templates for each type of command with a dataset of about 340,000 sentences extracted from our internal dataset of practical radiology reports. For each action, we generate all the possible ways in Vietnamese to express it. Each expression manner is a template, we then fill in the content into the template to create a full command. The content to be filled is the actual content that the user wants to input, and the position of the sentence in the report that the user wants to edit. The examples of equivalent Table 1. We listed out 5,760 templates for addition, 5,880 templates for modification, and 152 templates for deletion. The original sentences are added for the unknown class.

#### *2.2.2. Intent classification*

There are 3 types of commands provided as mentioned above. For addition, deletion and modification, the system will make

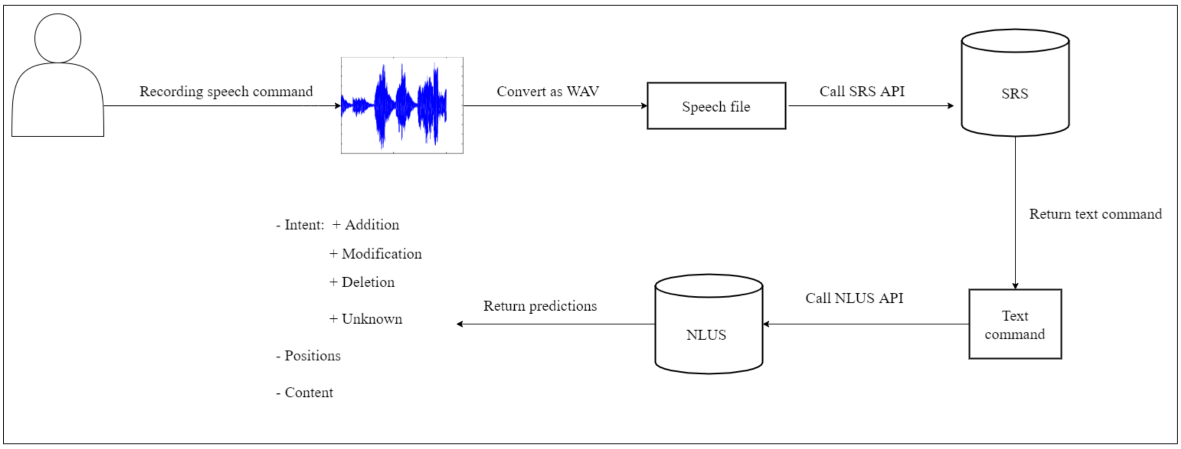

Figure 1: *Overview of the report editing module.*

| Command | Content | Intent | Position |
|---|---|---|---|
| Add normal mediastinum after the first sentence | Normal mediastinum | Addition | After the first sentence |
| Change the third sentence to enlarged cardiomegaly | Enlarged cardiomegaly | Modification | The third sentence |
| Delete the fourth sentence | | Deletion | The fourth sentence |

Table 1: *Examples of commands*

the corresponding action. If the intent is unknown, there is no action executed. Our trained model will classify each input command into one of these classes.

### 2.2.3. *Sequence tagging*

This task tags each word in a command whether it's used to express the intention, the content, or the position. The actual content is then extracted, the actual position in the report is gotten by rule from the position words. Based on the extracted content, position, and the classification of intent, the action will be taken.

### 2.2.4. *Model architecture*

We use a pre-trained BERT-based model for Vietnamese, PhoBert [7], to jointly learn two tasks. The [CLS] token representation is fed into a softmax layer for the classification task. Each other token representation is also classified into tags, then go through a CRF layer to get the final tagging sequence. The loss of both tasks is combined to train the whole network. We achieved the perfect performance on both tasks in the generated dataset.

## 3. Conclusion and future work

In this paper, we described a system for assisting radiologists on report editing using speech recognition and natural language understanding technologies for Vietnamese. We hope to roll out this function to help boost the productivity of the radiologists in Vietnam.

## 4. References

[1] L. H. Schwartz, P. Kijewski, H. Hertogen, P. S. Roossin, and R. A. Castellino, "Voice recognition in radiology reporting." *American Journal of Roentgenology*, vol. 169, no. 1, pp. 27–29, Jul. 1997, publisher: American Roentgen Ray Society. [Online]. Available: https://doi.org/10.2214/ajr.169.1.9207496

[2] A. Al-Aiad, A. K. Momani, Y. Alnsour, and M. Alsharo, "The Impact of Speech Recognition Systems on The Productivity and The Workflow in Radiology Departments: A Systematic Review," *AMCIS 2020 TREOs*, vol. 62.

[3] D. R. Williams, S. K. Kori, B. Williams, S. J. Sackrison, H. M. Kowalski, M. G. McLaughlin, and B. S. Kuszyk, "JOURNAL CLUB: Voice Recognition Dictation: Analysis of Report Volume and Use of the Send-to-Editor Function," *American Journal of Roentgenology*, vol. 201, no. 5, pp. 1069–1074, Oct. 2013, publisher: American Roentgen Ray Society. [Online]. Available: https://doi.org/10.2214/AJR.10.6335

[4] I. Hammana, L. Lepanto, T. Poder, C. Bellemare, and M.-S. Ly, "Speech Recognition in the Radiology Department: A Systematic Review," *Health Information Management Journal*, vol. 44, no. 2, pp. 4–10, 2015, _eprint: https://doi.org/10.1177/183335831504400201. [Online]. Available: https://doi.org/10.1177/183335831504400201

[5] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, "Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions," 2019.

[6] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural 'networks," vol. 2006, 01 2006, pp. 369–376.

[7] D. Q. Nguyen and A. T. Nguyen, "Phobert: Pre-trained language models for vietnamese," *CoRR*, vol. abs/2003.00744, 2020. [Online]. Available: https://arxiv.org/abs/2003.00744