Vietnamese Speech-based Question Answering over Car Manuals

Tin Duy Vo
VinAI Research
Hanoi, Vietnam
v.tinvd12@vinai.io

Manh Tien Luong
VinAI Research
Hanoi, Vietnam
v.manhlt3@vinai.io

Duong Minh Le
VinAI Research
Hanoi, Vietnam
v.duonglm1@vinai.io

Hieu Minh Tran
VinAI Research
Hanoi, Vietnam
v.hieutm4@vinai.io

Nhan Tri Do
VinAI Research
Hanoi, Vietnam
v.nhandt21@vinai.io

Tuan-Duy H. Nguyen
VinAI Research
Hanoi, Vietnam
v.duynht1@vinai.io

Thien Hai Nguyen
VinAI Research
Hanoi, Vietnam
v.thiennh7@vinai.io

Hung Hai Bui
VinAI Research
Hanoi, Vietnam
v.hungbh1@vinai.io

Dat Quoc Nguyen
VinAI Research
Hanoi, Vietnam
v.datnq9@vinai.io

Dinh Quoc Phung
VinAI Research
Hanoi, Vietnam
v.dinhpq2@vinai.io

ABSTRACT

This paper presents a novel Vietnamese speech-based question answering system QA-CarManual that enables users to ask car-manual-related questions (e.g. how to properly operate devices and/or utilities within a car). Given a car manual written in Vietnamese as the main knowledge base, we develop QA-CarManual as a lightweight, real-time and interactive system that integrates state-of-the-art technologies in language and speech processing to (i) understand and interact with users via speech commands and (ii) automatically query the knowledge base and return answers in both forms of text and speech as well as visualization. To our best knowledge, QA-CarManual is the first Vietnamese question answering system that interacts with users via speech inputs and outputs. We perform a human evaluation to assess the quality of our QA-CarManual system and obtain promising results.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**; • **Information systems → Retrieval tasks and goals**.

ACM Reference Format:
Tin Duy Vo, Manh Tien Luong, Duong Minh Le, Hieu Minh Tran, Nhan Tri Do, Tuan-Duy H. Nguyen, Thien Hai Nguyen, Hung Hai Bui, Dat Quoc Nguyen, and Dinh Quoc Phung. 2022. Vietnamese Speech-based Question Answering over Car Manuals. In *27th International Conference on Intelligent User Interfaces (IUI '22 Companion), March 22–25, 2022, Helsinki, Finland.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3490100.3516525

1 INTRODUCTION

As of 2015, two million passenger cars were registered in Vietnam and car sales have been growing at double-digit rates each year.[1] For safety reasons, car owners have to fully understand all features within a car and be competent to control the car before driving on a road. Appropriately manipulating all devices and/or utilities of a car, however, requires an expensive amount of time to carefully read hundreds of pages of the car's manual. This leads to the fact that drivers often ignore reading the manual. They alternatively either learn by trial and error or refer to other resources such as the Internet or colleagues, which might cause seriously negative consequences when relying on inappropriate resources. As a result, there is a need to develop an application to quickly return correct answers specifically mentioned in the car manual given the drivers' query utterances.

Recently, the performance of models in Automatic Speech Recognition (ASR), Natural Language Understanding (NLU) and Text-To-Speech (TTS) has been pushed to a cutting-edge boundary, which is relatively competitive with the human level. Our goal is to employ modern ASR, NLU and TTS approaches to build an application that communicates with users by taking speech commands as inputs, and acoustically and visually returning answers in a dialog fashion.

In this paper, we present QA-CarManual—a Vietnamese speech-based question answering system to answer car-manual-related questions. Our QA-CarManual is a user-friendly and interactive interface system that integrates recent state-of-the-art approaches in ASR, NLU and TTS. Given an input speech-based question w.r.t.

IUI '22 Companion, March 22–25, 2022, Helsinki, Finland

ACM ISBN 978-1-4503-9145-0/22/03.
https://doi.org/10.1145/3490100.3516525

[1] https://en.wikipedia.org/wiki/Transport_in_Vietnam

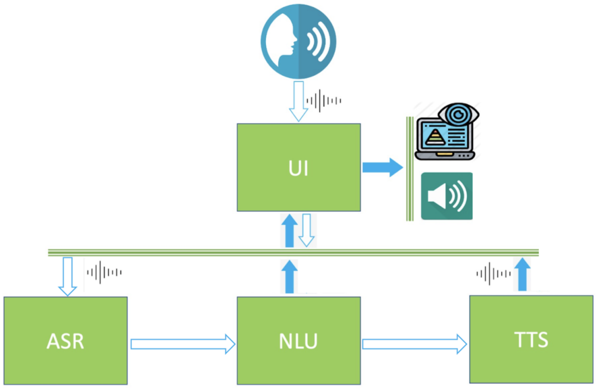
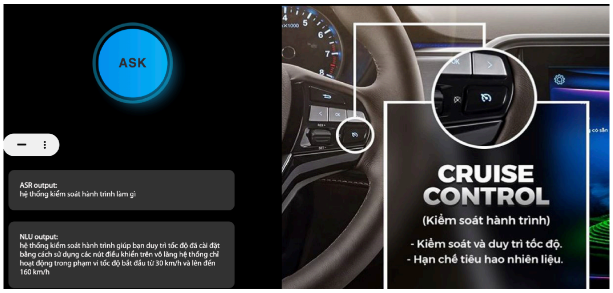


Figure 1: Our system's overall architecture.


ASK
CRUISE CONTROL
(Kiểm soát hành trình)
- Kiểm soát và duy trì tốc độ.
- Hạn chế tiêu hao nhiên liệu.

ASR output:
hệ thống kiểm soát hành trình làm gì

NLU output:
hệ thống kiểm soát hành trình giúp bạn duy trì tốc độ đã cài đặt bằng cách sử dụng các nút điều khiển trên vô lăng hệ thống chỉ hoạt động trong phạm vi tốc độ bắt đầu từ 30 km/h và lên đến 160 km/h

Figure 2: Our system's UI on web interface.

a car manual knowledge base, our system returns both text- and speech-based answers coupled with extra visualization information, e.g. either image or video. To the best of our knowledge, QA-CarManual is the first question answering system for Vietnamese, that takes speech query inputs and generates speech answer outputs. Experiments show that QA-CarManual obtains promising human evaluation results.

2 OUR QA-CARMANUAL SYSTEM

Our system consists of four components, including User Interface (UI) and three backend components of ASR, NLU and TTS, as shown in Figure 1. The UI component is an intermediate component that carries interactions between the user and the backend components. In particular, UI takes a query utterance from the user as input and passes it to the ASR component. Then ASR converts the input utterance from speech to query text. The NLU component takes the query text output from ASR as input to search in the knowledge base (KB) both text and visualization forms of a corresponding answer. The text form of the answer is then fed as input into the TTS component that generates a synthesized speech output. Finally, UI responds to the user by showing both the text and visualization forms of the answer outputted from NLU as well as the synthesized speech outputted from TTS.

2.1 The UI component

Figure 2 shows our system's UI. In order to interact with the system, the users need to press and hold the "Ask" button when inputting their command utterances. After releasing the "Ask" button, the backend is activated to process the utterances. UI then displays the backend's output answers in both forms of text and speech as well as visualizations to the users. For example, a driver could ask a Vietnamese question "chức năng của hệ thống kiểm soát hành trình là gì" (what is cruise control?), and our system produces an answer "hệ thống kiểm soát hành trình giúp bạn duy trì tốc độ đã cài đặt bằng cách sử dụng các nút điều khiển trên vô lăng, hệ thống chỉ hoạt động trong phạm vi tốc độ bắt đầu từ 30km/h và lên đến 160km/h" (Cruise control is a system that allows drivers to maintain a set speed without using the accelerator, by using the control elements on the steering wheel; the system is functional at speed range starting at 30 km/h and up to 160 km/h) in both forms of text and speech. In addition, our system also displays an image that visually locates the cruise control button in the car and summaries the function of the cruise control.

2.2 The ASR component

We employ Conformer-CTC [3] that is a variant of the Conformer model [2] with the Connectionist Temporal Classification loss [1] to develop our ASR component. We train Conformer-CTC using our in-house 4000-hour dataset augmented by noise injection and intensity adjustment approaches, and obtain the word error rate at about 8% on our internal test set. For inference, we further incorporate a 6-gram Byte-Pair-Encoding-based (BPE-based) language model (LM) [4] into the decoder to enhance the ASR performance. The 6-gram BPE-based LM is a statistical model which describes a probability distribution over sequences of six contiguous subwords. Technically, a BPE-based LM can be used with beam search decoders on top of Conformer-CTC to output more accurate candidates as: $\textsc{score}_{final} = \textsc{score}_{ASR} + \alpha * \textsc{score}_{LM} + \beta * \textsc{seq}_{length}$, where $\textsc{score}_{ASR}$ and $\textsc{score}_{LM}$ are the scores predicted by Conformer-CTC and the 6-gram BPE-based LM, while α and β are mixture weights representing the importance of the LM and the penalty term w.r.t. the sequence length ($\textsc{seq}_{length}$), respectively.

2.3 The NLU component

The NLU component requires a car manual KB. Given a Vietnamese car manual of a popular car brand, we create the KB of all possible triplets (textual question, textual answer, visual answer). Firstly, we automatically scan through the car manual to hierarchically extract section headings, content blocks and images as possible triplets of (textual question, textual answer, visual answer). Secondly, we employ 3 annotators to manually inspect each triplet and correct misaligned triplets, resulting in about 2000 triplets. For those extracted triplets where the visual answer is missing, we use an image as the default visual answer. Finally, we expand our triplet corpus up to 6000 triplets by making use of a synonym lexicon. The 6000-triplet corpus is referred to as our KB.

Our NLU consists of 3 modules Preprocessing, Answer Retrieval and Answer Selection. The Preprocessing module normalizes the input query text, performs lexicon-based query type classification, removes part-of-speech tag-based non-informative words in the

Table 1: Quality Assessment. TTS: 1–Poor; 2–Fair; 3–Good; 4–Very good; 5–Excellent.

	Pipeline		ASR		NLU		TTS
Correct	448	Insertion	11	Correct	448	Score 4-5	400
Relevant	21	Deletion	0	Relevant	21	Score 3	137
Incorrect	91	Substitution	30	Incorrect	91	Score 1-2	23
Accuracy	80.0%	Error Rate	7.3%	Accuracy	80.0%	Avg. score	4.04

input query and outputs a cleaned query text. Here, VnCoreNLP [8] is used to segment the input text into words and assign a part-of-speech tag to each word. The Answer Retrieval employs the BM25 algorithm [7] to compute ranking scores between the input cleaned query and the question-answer text pairs in all triplets from the KB and returns a list of top 5 ranked triplets. Here, a ranking score between a query and each triplet in our KB is computed as: $\textsc{score} = \text{BM25}_{qq}+\text{BM25}_{qa}$, where BM25_{qq} and BM25_{qa} represent the BM25 scores of the input cleaned query against the triplet's question and answer texts, respectively. From the list of 5 candidate triplets returned by the Answer Retrieval module, the Answer Extraction module filters out triplets whose ranking score is smaller than a pre-defined threshold or question has a different type against the input query's type. From a filtered list of remaining triplets, Answer Extraction outputs a pair of textual and visual answers from the triplet having the highest score. If the filtered list is empty, a default pair of "no answer" and "no image" is returned.

2.4 The TTS component

The TTS component first translates the answer text into phonemes based on their pronunciation and text normalization rules. Our TTS employs Glow-TTS [5] to predict mel-spectrogram from input phonemes. Here, we modify the input of the Glow-TTS model to be well-fitted with Vietnamese by using the Vietnamese phoneme dictionary. Our TTS then uses HiFi-GAN [6] to generate efficient and high-fidelity speech synthesis from the predicted mel-spectrogram.

3 EVALUATION

To evaluate the quality of our QA-CarManual system, we conduct a human-based manual evaluation, involving 7 examiners. Each examiner self-proposes 80 car-manual-related queries, speaks each query to our system and provides feedback w.r.t the performance of each of our system's components ASR, NLU and TTS as well as the whole system (i.e. Pipeline). We collect feedback from a total of 560 queries and show quality assessment results in Table 1. We find that the query-level error rate for ASR is at 7.3%, equivalent to 41 queries, including 11 queries accounted for the error type of Insertion and 30 queries accounted for the error type of Substitution. We obtain an accuracy of 80% for NLU, i.e. 448 input queries have correct textual answers (here, we also find that 21 output answers partially contain relevant information to 21 input queries). We achieve an average rating score of examiners for TTS at 4.04/5.0, which means a "very good" quality. The system as a whole obtains a promising accuracy at 80%, similar performance as of NLU, which satisfies users' expectations to a substantial extent.

4 CONCLUSION

In this paper, we have introduced QA-CarManual— the first Vietnamese question answering system that interacts with users via speech inputs and outputs. Our system contains a UI and three backend components ASR, NLU and TTS, incorporating modern approaches in language and speech processing to produce satisfying answers to the users' input car-manual-related queries. The human evaluation shows that our system obtains a promising performance.

REFERENCES

[1] Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, Vol. 148. 369–376.

[2] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Interspeech*. 5036–5040.

[3] Pengcheng Guo, Florian Boyer, Xuankai Chang, Tomoki Hayashi, Yosuke Higuchi, Hirofumi Inaguma, Naoyuki Kamo, Chenda Li, Daniel Garcia-Romero, Jiatong Shi, Jing Shi, Shinji Watanabe, Kun Wei, Wangyou Zhang, and Yuekai Zhang. 2021. Recent Developments on Espnet Toolkit Boosted By Conformer. In *ICASSP*. 5874–5878.

[4] W. Hu, Y. Luo, J. Meng, Z. Qian, and Q. Huo. 2020. A Study of BPE-based Language Modeling for Open Vocabulary Latin Language OCR. In *ICFHR*. 133–138.

[5] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 2020. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. In *NeurIPS*, Vol. 33. 8067–8077.

[6] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In *NeurIPS*.

[7] Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (2009), 333–389.

[8] Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. VnCoreNLP: A Vietnamese Natural Language Processing Toolkit. In *NAACL Demonstrations*. 56–60.