Politeness Detection in Speech for Human-Computer Interaction

Selma Yilmazyildiz Kayaarma[1], Sherik Lehal[1], and Hichem Sahli[1,2]

[1] Vrije Universiteit Brussel (VUB), Dept. of Electronics and Informatics (ETRO),
Pleinlaan 2, B-1050 Brussels, Belgium.
[2] Interuniveristy Microelectronics Center (IMEC), 3001 Leuven, Belgium.

Abstract. In this study, we demonstrate a deep learning based politeness detection system, using both speech features and transcriptions, which is implemented in a smart speaker interaction scenario.

1 Introduction

Smart speakers (like Amazon Echo, Google Home, etc.) are AI virtual assistants in the form of wireless speakers. Gartner reports that smart speakers are a first purchase for a connected home and predicts that 75 % of US houses will have smart speakers by 2020 [1].

In their current state, AI virtual assistants are configured to recognize imperative statements (e.g. 'Turn on the light!'). They always respond in the same way even if the users yell or scream at them. This type of interaction can set an inappropriate example to children and may lead them to imitate such language in normal conversations. This stands as a growing concern by many families. However, only limited research efforts ([2],[3]) have addressed this issue.

As a step to overcome this concern, the objective of this demonstration is incorporating polite manners in daily use of current voice assistants. The system first recognizes the politeness in user request through analyzing both the linguistic and acoustic parts of the request by bidirectional Long Short-Term Memory (BLSTM) [4] and Convolutional Recurrent Neural Network (CRNN) [5] based deep learning architecture and then responds accordingly.

2 System Architecture

Manners are encoded with two channels in speech: What is said (i.e. linguistic channel) and how it is said (i.e. acoustic channel). In detecting politeness as a manner, our system utilizes both of these channels. As can be seen from Fig. 1, our architecture considers speech transcriptions in the linguistic channel as well as corresponding acoustic speech features Mel Frequency Cepstral Coefficients (MFCC) - which together provide a deep neural network that incorporates both semantic relationships and the necessary low-level acoustic features required to distinguish among politeness/impoliteness accurately.

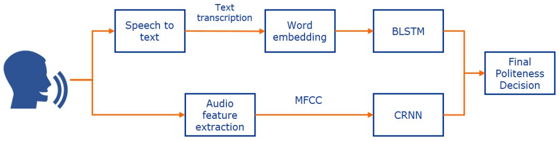

Fig. 1. General System Architecture

Fig. 2. Mycroft Mark 1

3 Demo Overview

We will demonstrate our system using Mycroft Mark 1 (see Fig. 2) which is an open-source smart speaker. The politeness detection is implemented as a custom skill on the device and will be activated by turning on the ”Polite Mode”.

The user will make various requests to Mycroft Mark 1 (e.g. asking time, weather, etc.) both politely and impolitely. The request is analyzed using the architecture described above and the response is customized based on the output of the politeness detection model. The system still fulfills the requests even if the user input is impolite. But it enhances its response with a comment, asking the user to request more politely next time or providing the feedback that the request wasn't made in an appropriate way. If the request is made in a polite way, its response includes also a rewarding comment to the user. (A demonstration video can be viewed at: https://youtu.be/uwINIBvTbqE).

4 Conclusion

In this study we have showcased a deep learning based politeness detection system, using both speech features and transcriptions. While this demo showcases the politeness detection in a smart speaker use case, especially considering the increasing popularity of AI enabled conversational interfaces, this capability can be applied in a broader range of human-computer interaction use cases.

References

1. Market Trends: Connected Home Adoption in the U.S. and the U.K., https://www.gartner.com/en/documents/3941863/market-trends-connected-home-adoption-in-the-u-s-and-the. Last accessed 09 Sep 2019
2. Bonfert, M., Spliethver, M., Arzaroli, R., Lange, M., Hanci, M., Porzel, R.: If you ask nicely: a digital assistant rebuking impolite voice commands. In: Proceedings of the International Conference on Multimodal Interaction, pp. 95–102. ACM (2018)
3. Biele, C., Jaskulska, A., Kopec, W., Kowalski, J., Skorupska, K., Zdrodowska, A.: How Might Voice Assistants Raise Our Children?. In: International Conference on Intelligent Human Systems Integration, pp. 162–167. Springer, Cham (2019)
4. Graves A, et al.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural networks, **18**(5-6), 602–610 (2005)
5. Sainath, Tara N., et al.: Convolutional, long short-term memory, fully connected deep neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4580–4584. IEEE (2015)