June 08, 2018

Matching language and accent in virtual assistant responses

Jatin Matani

Philippe Gervais

Marcos Calvo

Sandro Feuz

Thomas Deselaers

Matching language and accent in virtual assistant responses

ABSTRACT

Traditionally a voice-controlled virtual assistant interacts with a user in a neutral, accent-free way. A virtual assistant that matches the language and accent of the user can be more appealing. This disclosure uses machine-learning models to match the language and accent of a virtual assistant to that of the user who commands it.

KEYWORDS

- digital assistant
- virtual assistant
- natural language processing
- speech accent
- language dialect
- speaker identification
- text-to-speech (TTS)

BACKGROUND

A voice-controlled virtual assistant, e.g., implemented in consumer devices such as smartphones, wearables, home speakers, appliances, etc. typically responds in neutral accents. This makes the responses sound somewhat sterile, and leads to human-assistant interaction that does not feel natural. However, a virtual assistant that matches the language and accent of the user can be more appealing. This is particularly true in households where multiple languages and/or regional accents are spoken, e.g., British vs. American accented English, Northern vs. Southern vs. Swiss German, etc.

DESCRIPTION

This disclosure uses machine-learning models to match the language and accent of a virtual assistant to that of the user who commands it.

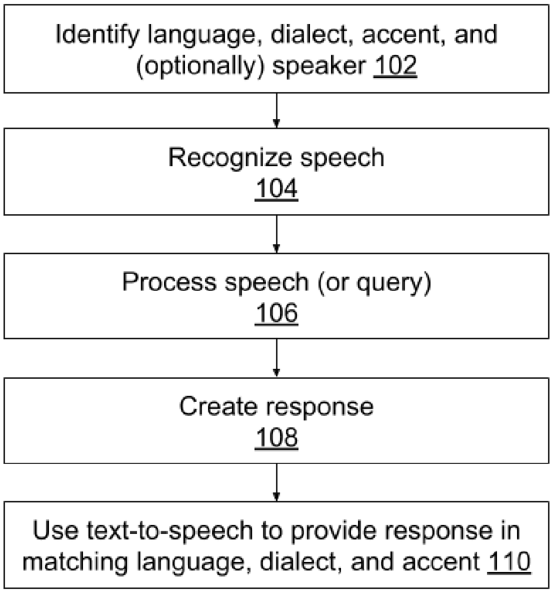
Identify language, dialect, accent, and (optionally) speaker 102
Recognize speech 104
Process speech (or query) 106
Create response 108
Use text-to-speech to provide response in matching language, dialect, and accent 110

Fig. 1: Matching language, dialect, and accent of the virtual assistant to that of the user

Fig. 1 illustrates matching language, dialect and accent of the virtual assistant to the user who commands it. The technique to match the virtual assistant speech to that of the user is implemented upon user permission. Permission is obtained for analysis of user speech data and users are provided with options to modify the permission at any time.

Upon detecting a user-issued command, the virtual assistant identifies, using a machine-learning model that takes as input the user’s speech, the speaker (optional), and the language, dialect and accent (102). The problems of speaker, language, dialect, and accent identification can be treated independently, e.g., using separate classifiers for each. Alternately, the problem can be treated as a joint prediction problem such that speaker identification assists language, dialect, or accent detection, and vice-versa. The machine-learning model can be implemented as a multi-layer neural network, e.g., a long short-term memory (LSTM) neural network. Other types of recurrent neural networks, and/or convolutional neural networks, and techniques such as support vector machines, random forests, boosted decision trees, etc., can also be used to implement recognition of language, dialect, or accent.

The virtual assistant recognizes the speech or query (104) and processes it (106) in the language, dialect, and accent of the speaker. The virtual assistant generates a response (108) in response to the detected query. The accent, dialect, and language detected may inform the response of the virtual assistant. The processes that lead to creation of the response in the appropriate language can be directly integrated into the output processor of the virtual assistant. Alternately, such processes can be implemented in a preprocessor that translates standard assistant responses into accented responses with words selected or modified as appropriate to reflect the detected dialect.

The virtual assistant uses a text-to-speech (TTS) synthesizer to provide a spoken response (110) in a language, dialect, and accent that matches that of the user. Each language-accent pair is generally served by a dedicated TTS synthesizer. However, for some accents, it may be feasible to create an accented output from an unaccented output using signal processing techniques.

In this manner, the techniques of this disclosure provide more natural responses to users of voice-controlled assistants. In turn, this can lead to a better user experience and retention. Furthermore, the system is designed in a way to customize existing virtual assistant implementations in a manner that obviates rebuilding the entire assistant for each language/accent pair. Rather, specific components of the virtual assistant that pertain to customized language/accent responses are built. Alternately, the entire virtual assistant, including machine-learning models, can be modified for each language-accent pair.

CONCLUSION

Traditionally a voice-controlled virtual assistant interacts with a user in a neutral, accent-free way. A virtual assistant that matches the language and accent of the user can be more appealing. This disclosure uses machine-learning models to match the language and accent of a virtual assistant to that of the user who commands it.