Technical Disclosure Commons

Defensive Publications Series

November 2021

# Improving Speech Recognition Quality Using Grammar Training Phrases

Amit Singhal

Yao Lin

Follow this and additional works at: https://www.tdcommons.org/dpubs_series

Recommended Citation
Singhal, Amit and Lin, Yao, "Improving Speech Recognition Quality Using Grammar Training Phrases", Technical Disclosure Commons, (November 28, 2021)
https://www.tdcommons.org/dpubs_series/4754

This work is licensed under a Creative Commons Attribution 4.0 License.
This Article is brought to you for free and open access by Technical Disclosure Commons. It has been accepted for inclusion in Defensive Publications Series by an authorized administrator of Technical Disclosure Commons.

# Improving Speech Recognition Quality Using Grammar Training Phrases

ABSTRACT

When a new voice feature is to be launched on a device with a voice interface, e.g., a digital assistant application, the natural language understanding (NLU) model is built using training data for the new feature. Speech biasing models are typically added to improve recognition accuracy for queries that are specific to the feature or contain non-common words. Such biasing models are often built using traffic logs, collected with user permission, after the initial release of the feature. However, this approach may not provide high speech recognition quality during product testing and initial launch. This disclosure describes techniques to improve the ASR quality of a new feature from the time of initial release and without relying on traffic logs. To that end, speech biasing models are built using grammar training phrases.

KEYWORDS

- Voice assistant
- Virtual assistant
- Voice interface
- Voice query
- Speech biasing
- Biasing model
- Automatic speech recognition (ASR)
- Natural Language Understanding
- Vehicle infotainment
- Smart speaker

## BACKGROUND

Automatic speech recognition (ASR) models are used to recognize voice commands or queries from users in hardware products such as smartphones, smart cars, smart speakers, as well as applications that enable speech interaction, e.g., digital assistant applications. When a new voice feature is to be launched on a device with a voice interface, e.g., via a digital assistant application, the natural language understanding (NLU) model is built with training data obtained for the new feature.

However, merely using the generic language model may result in the ASR quality not being good enough, especially for queries that contain non-common words that are specific to the feature. Speech biasing models are normally added to improve the accuracy of recognition in such cases. Typically, such biasing models are built based on user-permitted traffic logs, obtained after the initial release of the feature. However, such an approach can only be used after the new feature or device is launched and gets enough traffic/user feedback. Consequently, this approach cannot provide improvements in speech recognition quality during product testing and/or initial launch. This is a significant drawback for devices that have relatively long development, testing and roll out cycles such as cars with infotainment systems or other devices that are configured for voice input. This also poses difficulties in achieving consistent quality across different locales.

## DESCRIPTION

This disclosure describes techniques to improve the ASR quality of a voice interaction feature at the time of initial release, without relying on traffic logs. When new voice features are developed (e.g., for a digital voice assistant) and data is obtained to train the natural language understanding (NLU) grammar models, speech biasing models are also built using those training

phrases. These models are then used to enhance ASR quality. Biasing models are used to override the model weight of language words using various contextual signals, including previous words matched in a query. For example, using the phrase "seat heater" in a biasing model can override the language model weight of "heater" given that the previous word was "seat".

In addition, the capabilities and policy enforcement configured by the original equipment manufacturer (OEM) for a specific device can be used as contextual signals to customize and specifically include biasing models for features applicable for that device. Device capabilities can be sent directly to the speech recognition server. OEM configured policies can be saved in a database and provided by a server. These policies can be updated by the OEM at any time.

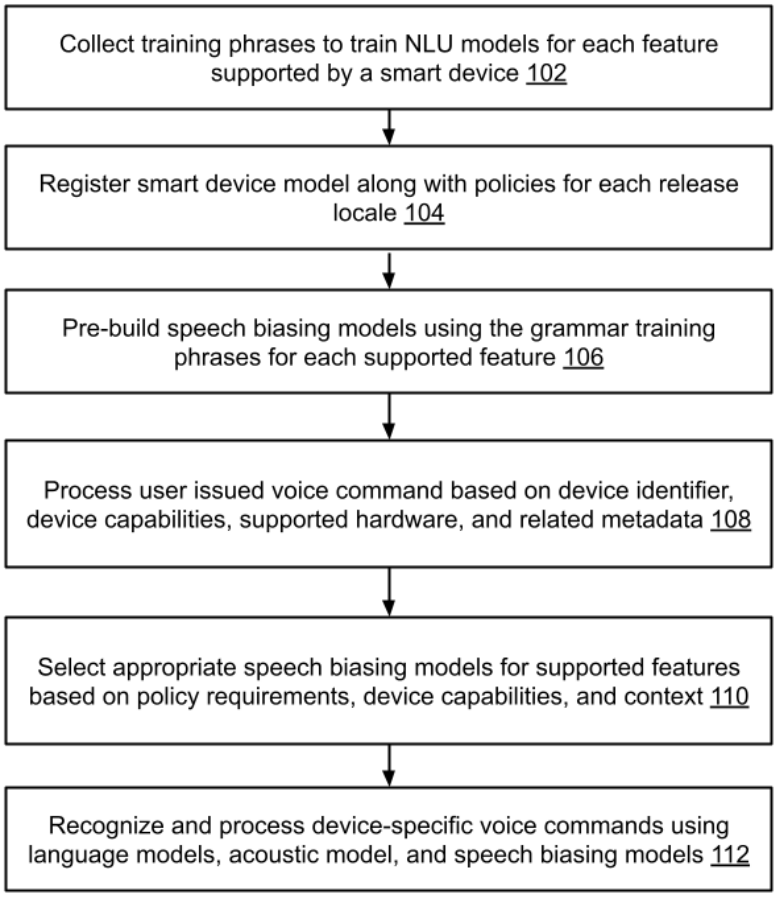


**Fig 1: Example process to improve ASR quality from initial release using biasing models**

Fig. 1 illustrates an example process for achieving improvement in automatic speech recognition, per techniques of this disclosure. To develop voice interaction features specific to a smart device that is capable of receiving voice input, training phrases are first collected to train NLU models for the different features (102). For example, the smart device may be an in-car infotainment system that receives and responds to spoken queries. The original equipment manufacturer (OEM) of such systems registers each of their device models with policies in each i18n locale that the device is targeted for release (104).

Speech biasing models are pre-built using the grammar training phrases for each supported feature (106). To process a user issued voice command, the device identifier (device type), device capabilities, supported hardware, and related metadata are obtained. These can be used to retrieve policy requirements, device capabilities and relevant context. Based on the retrieved information, speech biasing models for the supported features (110) can be selected. Device specific voice commands can then be recognized and processed using language models, acoustic models, and speech biasing models (112).

For example, to develop voice features for an in-car device, training phrases can be collected to train NLU models for various features. For example, the training phrases can be “fan to windshield and floor,” “Turn on the seat heater for the driver,” etc. The OEM can register each of their car models with policies in each applicable locale that the car model is targeted to release. This allows adjustment of different commands available in different locales where the car model is sold.

Without the use of speech biasing models, speech recognition can have a high error rate. For example, the spoken command “set the fan speed to high” may be recognized as “set the fan speed too high”; “seat heat” may be recognized as “seat cheat”; etc.

Through the feature development cycle, including testing, sufficient user traffic and feedback cannot be received and utilized to improve performance. The described techniques can be used to improve speech recognition quality for any feature provided via a voice interface by applying speech biasing in the ASR layer for that feature using appropriate training phrases. This can substantially improve the user experience at initial launch.

Further to the descriptions above, a user is provided with controls allowing the user to make an election as to both if and when systems, programs, or features described herein may enable the collection of user information (e.g., information about a user's voice queries, device type, locale/language, a user's preferences, or a user's current location), and if the user is sent content or communications from a server. In addition, certain data are treated in one or more ways before it is stored or used so that personally identifiable information is removed. For example, a user's identity may be treated so that no personally identifiable information can be determined for the user, or a user's geographic location may be generalized where location information is obtained (such as to a city, ZIP code, or state level) so that a particular location of a user cannot be determined. Thus, the user has control over what information is collected about the user, how that information is used, and what information is provided to the user.

## CONCLUSION

This disclosure describes techniques to improve the ASR quality of a new feature from the time of initial release and without relying on traffic logs. To that end, speech biasing models are built using grammar training phrases.