Technical Disclosure Commons

Defensive Publications Series

August 2021

# Automated Audio Generation for Testing Voice Interface Devices

Amit Singhal

Yao-Jen Chang

Srikrish Subramanian

Xinchen Wang

# Automated Audio Generation for Testing Voice Interface Devices


## ABSTRACT

Testing of voice interface devices across multiple languages and locales is difficult due to factors such as the lack of availability of native speakers, inconsistency of human speech samples across languages, difficulty in scaling the number of human-provided query samples, etc. This disclosure describes the use of automated translation and text-to-speech generation technologies to obtain machine-generated audio in various languages. A set of query strings is translated into multiple languages and a text-to-speech synthesizer generates a consistent set of audio samples. The generated audio samples can be used to test voice interface devices.



## KEYWORDS

- Voice interface
- Spoken command
- Internationalization (i18n)
- Localization
- Text-to-speech (TTS)
- Machine translation
- Virtual assistant
- Smart speaker
- Smart appliance
- Smart home
- Car infotainment
- Automated test framework


## BACKGROUND

Voice interface devices such as smart speakers, smartphones, car dashboard systems, smart home devices, smart appliances etc. are tested for multiple queries in multiple languages. Due to the difficulty of finding native speakers in certain languages, it can be challenging to test the internationalization (i18n) of voice interface devices and to verify their performance across multiple languages and locales.

Even when native human speakers of such languages are found, e.g., in the test and development community, it is unclear if such speakers speak widespread dialects/accents or lesser-known regional variants. The few native human speakers are inadequate to meet the scale of testing, which can include thousands of queries over tens of languages. It is also difficult to maintain consistency of audio quality across all i18n locales with different people recording for each locale. Additionally, even when human-recorded audio is available, privacy requirements make it difficult to use them for automated testing of voice interface devices. For example, to protect the privacy of human speakers that record their speech for testing purposes, a mechanism is provided for them to delete such recordings as needed, which adds to the complexity of test and development.

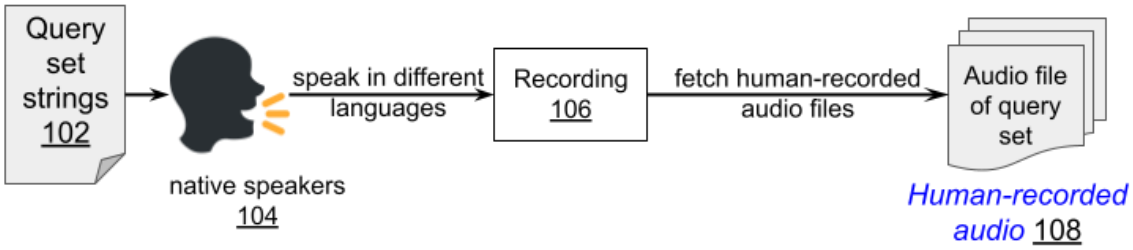


**Fig. 1: Traditional procedure for generating i18n test files for voice interface devices**

Fig. 1 illustrates a traditional procedure for generating i18n test files for voice interface devices. A set of query strings (e.g., ‘turn the AC on,’ ‘take me to the nearest coffee shop,’ etc.; 102) is spoken by native speakers (104) of various languages. Their utterances are recorded (106), e.g., converted to audio files. During automated testing of voice interface devices, such human-recorded audio (108) is fetched and serves as test input to the voice interface devices.

DESCRIPTION

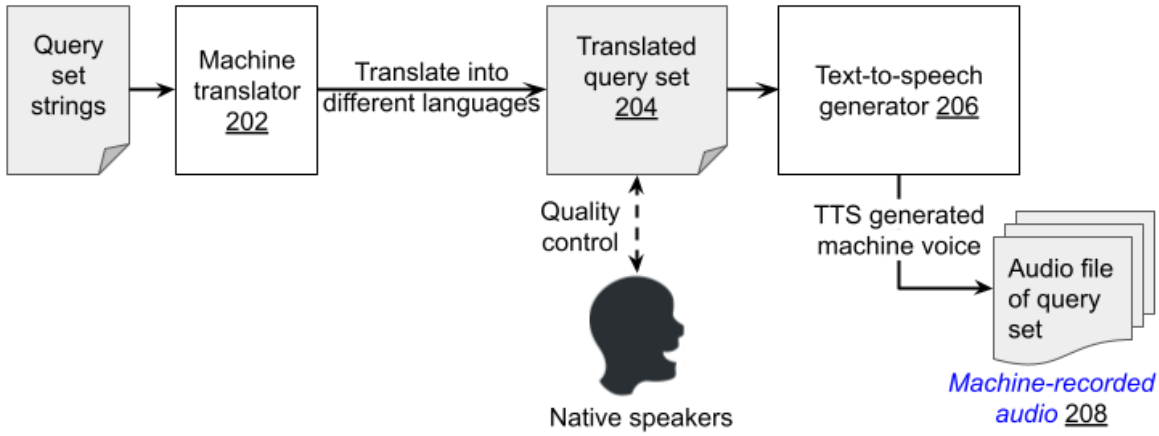


**Fig. 2: Automated audio generation for testing voice interface devices in multiple languages**

Fig. 2 illustrates automated audio generation for testing voice interface devices for multiple languages, per the techniques of this disclosure. A set of query strings is translated into multiple languages using a machine translator (202) to generate a translated query set (204). Samples from the translated query set are sent to native speakers to test for accuracy and for corrections as necessary. A text-to-speech generator (206) generates machine-recorded audio (208), which can be fetched to serve as test input to voice interface devices.

The techniques generate synthesized audio files of quality consistent across languages and locales. Inconsistencies typically found in human-recorded audio files, e.g., audio volume, pronunciation (accent) across different speakers, speed of audio, etc. are absent or controllable. Automated generation scales well to the requirements of generating thousands of test queries across a large number of query domains and languages. For example, voice queries from domains such as local commands (e.g., “turn up the volume,” “set temperature to 72 degrees,” “set timer for 25 minutes,” etc.), media-related commands (e.g., “play Frank Sinatra songs,” “resume podcast A,” etc.), information-seeking (e.g., “tomorrow’s weather,” “stock A price,”

"my next appointment,") etc. need to be adequately represented in the dataset of test queries, and in all the supported languages, such that the dataset is sufficiently representative of practical use of the voice interface.

Since the test queries are machine-synthesized audio, the techniques reduce costs that are otherwise incurred when deploying native human speakers. Privacy concerns associated with the use of recordings generated by human speakers are circumvented. Machine-synthesized audio files can be retained for testing as long as testing procedures warrant. The techniques apply generally to the optimization and training of automatic speech recognizers for multiple languages under real-world conditions. For example, background noise of differing types can be inserted into the machine-synthesized audio to simulate such conditions.

Further to the descriptions above, a user may be 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 social network, social actions or activities, profession, 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 may be 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 may have 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 the use of automated translation and text-to-speech generation technologies to obtain machine-generated audio in various languages. A set of query strings is translated into multiple languages and a text-to-speech synthesizer generates a consistent set of audio samples. The generated audio samples can be used to test voice interface devices.