November 2019

Automatic Identification Of Spoken Language In Audio Clips

Omar Abdelaziz

Alex Greaves

Automatic Identification Of Spoken Language In Audio Clips

ABSTRACT

Audio transcription systems typically require that the language being spoken be specified explicitly such that a language-specific transcription technique can be employed. In cases where the spoken language is not explicitly indicated, a possible approach is to process the audio via all available transcription techniques and choose the transcription associated with the highest confidence. Such an approach does not scale to support a large number of natural languages and is computationally expensive. This disclosure describes automatic identification of the natural language being spoken in audio input. The audio is processed using a trained machine learning model to output a language code corresponding to the language being spoken. Such a two-step approach, with a language identification step preceding the transcription step, enables supporting a large number of natural languages without incurring computational costs, latencies, and inaccuracies of employing multiple transcription techniques in parallel.

KEYWORDS

- Language detection
- Language identification
- Speech input
- Virtual assistant
- Voice assistant
- Audio transcription
- Speech-to-text
- Smart speaker

BACKGROUND

Systems that generate transcripts of spoken text from audio clips use speech processing techniques to transcribe the spoken words. Typically, a separate technique is provided for each natural language. Therefore, audio transcription systems typically require that the language being spoken be specified explicitly such that a language-specific transcription technique can be employed. In cases where the spoken language is not explicitly indicated, a possible approach is to process the audio via all available transcription techniques and choose the transcription associated with the highest confidence. Alternatively, or in addition, the correct spoken language can be identified from examining how a virtual assistant would respond to each of the generated transcription. The language of the transcription that results in a valid response is chosen as the likely language being spoken in the audio.

While such approaches may be acceptable when the set of possible natural languages in speech input is limited to just a few languages, it does not scale for supporting the processing of audio clips that might contain any one of a large number of natural languages. Moreover, when dealing with a large set of natural languages, the approach is computationally expensive and not highly accurate for correctly identifying the natural language being spoken.

DESCRIPTION

This disclosure describes automatic identification of the natural language being spoken in audio input. The audio is processed using a trained machine learning model to output a language code corresponding to the language being spoken.

The detected language is compared with the default language set for the system, e.g., a virtual assistant that responds to spoken queries or other systems that take speech input. If the languages match, the transcription technique for the system default language is to interpret the

speech query. If the languages do not match, transcription is performed by choosing the transcription technique corresponding to the language detected by the language identification model. Such a two-step approach, with a language identification step preceding the transcription step, enables supporting a large number of natural languages without incurring computational costs, latencies, and inaccuracies of employing multiple transcription techniques in parallel.

With appropriate user permissions, audio from a large corpus of online videos can serve as labeled training data to train a machine learning model to perform language identification. For example, such videos are already marked with the language being spoken in the clips which serve as labels. Alternatively, or in addition, labeled training data for the language identification model can include audio input used with permission from active users of voice assistants, e.g., provided via smartphones, smart speakers, or other devices. As such systems receive multiple queries from active users, the natural language spoken is typically known with high confidence and can be used as a label in the training process.

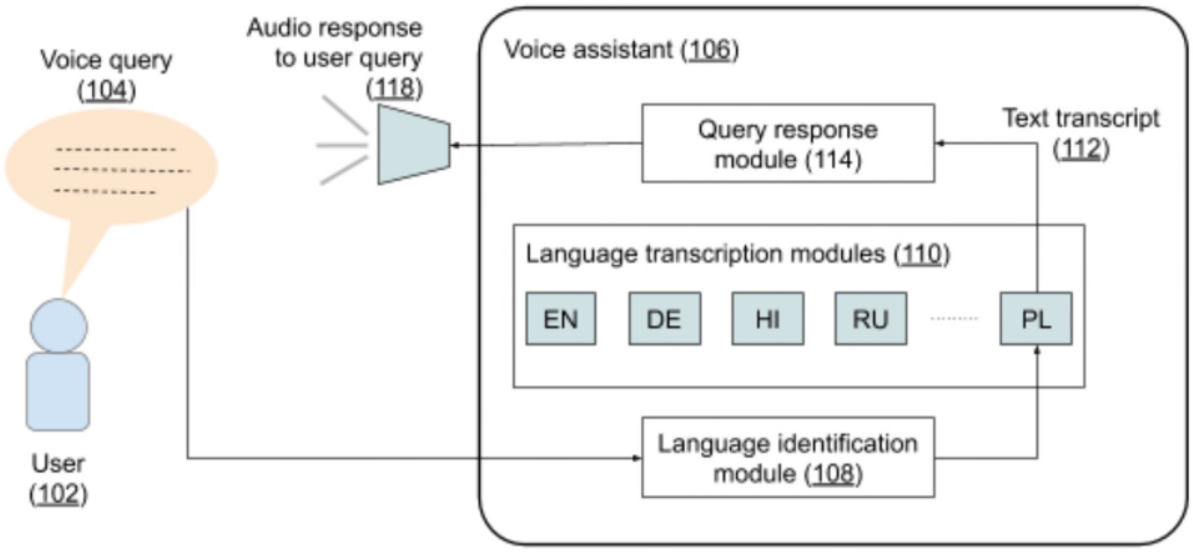

Fig. 1: Virtual assistant with automatic identification of spoken language

Fig. 1 illustrates a virtual assistant application that performs automatic identification of spoken language in a voice query, per the techniques described herein. As illustrated in Fig. 1, a user (102) issues a voice query (104) to a voice assistant device (106). The query is processed using a language identification module (108) that includes a trained machine learning model that acts as a language classifier. The model detects that the voice query is in Polish and accordingly, provides the language code PL which is used to select the corresponding language transcription module (110) from the set of available transcription modules, e.g., modules for English (EN), German (DE), Hindi (HI), Russian (RU), etc.

The Polish language transcription module analyzes the voice query and generates a corresponding text transcript (112). The transcript is provided to a query response module (114) which generates and provides an audio response (118) to the user query. The voice assistant can deliver the audio response in Polish, the language of the query.

With user permission, the techniques described above can be utilized in any systems that involve audio input, such as voice assistants, translation services, content distribution platforms, etc. The dynamic handling of language selection in audio input and output can enhance user experience compared to the current approach, e.g., in which the audio input is assumed to be one of the system default languages. If the user permits, the described techniques can be extended to help automatically select the default language for devices or apps, e.g., based on the language spoken at a time of device or app setup.

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 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 automatic identification of the natural language being spoken in audio input. The audio is processed using a trained machine learning model to output a language code corresponding to the language being spoken. Such a two-step approach, with a language identification step preceding the transcription step, enables supporting a large number of natural languages without incurring computational costs, latencies, and inaccuracies of employing multiple transcription techniques in parallel. An audio response can be provided to the user in the same language as that of the user query. The dynamic handling of language selection in audio input and output can significantly enhance user experience.