INTEGRATION OF A KALDI SPEECH RECOGNIZER INTO A SPEECH DIALOG SYSTEM FOR AUTOMOTIVE INFOTAINMENT APPLICATIONS

The Kaldi toolkit enables the research and development of speech recognition technologies and applications. It supports classical GMM-HMM approaches and hybrid approaches like TDNN-HMM. The toolkit provides recipes for different corpora. It is possible to train a speech recognition model for the English librispeech corpus out of the box. Additionally, the toolkit offers built-in implementations to decode Kaldi models.

The speech dialog system for our research is part of the human-machine-interface (HMI) modeling tool EB GUIDE. It is used in the automotive industry, in particular to create infotainment systems. The dialog flow is modeled using a state machine. Each state represents a *talk* to specify a dialog step. The tool provides different command and prompt *spidgets* (speech gadgets) to handle speech recognition and speech output within a talk. The specified dialog flow and user interface can be simulated for different target platforms.

In this paper we investigate if Kaldi can be integrated into the speech dialog system and if it is suitable to be used in an automotive context. We implemented a Kaldi recognition module that is able to run on 64-bit Windows and Linux platforms. It provides a socket interface to enable control of the recognizer by an external application like our speech dialog system. A new plugin for the speech dialog system was developed to integrate the recognition module into the dialog flow. It provides also a new command spidget to specify an intent and to configure Kaldi parameters and the used Kaldi model. The ability to load, unload and preload Kaldi models helps to handle limited resources of the target system.

Since Kaldi is only suitable to perform speech to text decoding, an additional semantic analysis is needed. Therefore the spidget provides the possibility to execute a dynamic time warping method to match the recognized text with one of the spidget's example sentences. It is also possible to use alternative methods to look for the exact match of a single phrase or the occurrence of a set of single words in the recognition result. The combination of the methods with multiple active spidgets using the same Kaldi model allows the specification and execution of simple command and control systems. A demonstrator based on an infotainment scenario is developed to show the implemented solutions.

Finally an evaluation of Kaldi together with the dynamic time warping method is performed in an automotive context. The aim of the method is to reduce the sentence error rate (SER) in a specific talk state. The evaluation is executed using our HMI modeling tool. The librispeech TDNN-HMM model is used by Kaldi to perform the speech to text recognition. A corpus containing phrases from typical infotainment scenarios is applied for the evaluation of the SER. The applied method was able to reduce the SER from 34% to 3%.

The Kaldi speech recognizer is suitable for automotive command and control scenarios and recognizes intents robustly. For use cases based on dynamic vocabulary, a named entity extraction needs to be added in future.

About the Authors

T. Ranzenberger; C. Hacker
Elektrobit Automotive GmbH
Erlangen Germany
thomas.ranzenberger@elektrobit.com
christian.hacker@elektrobit.com

F. Gallwitz
Technische Hochschule Nürnberg
Nuremberg Germany
florian.gallwitz@th-nuernberg.de