Renesas / Cyberon Speech Recognition

Endpoint Embedded Voice User Interface Solution

Georgios Flamis, Senior Staff System Design Engineer, Renesas Electronics Corporation

Vincent d'Agostino, Founder D'Agostino Industries Group

Introduction

Traditional voice or speech recognition technology is based on a trained model with specific words or phrases. Natural language processing is word-order independent. This requires large computing power to run the real time data through a neural network. The Cyberon approach is different, allowing the algorithms to run on a small, general purpose MCU.

Background

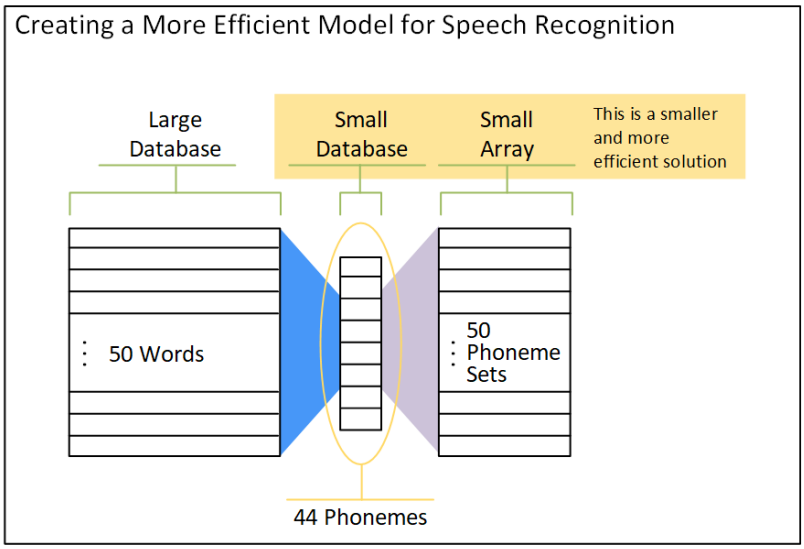
The formulation of language, or the sounds associated with language, differs among languages. The smallest unit of linguistic sound is called a phoneme, and different languages have different numbers of them. English, for example, contains 44 phonemes to form al the words in the language. Italian has 32, and the traditional Hawaiian language uses only 14.

Let's assume a language (or command set) has 200 words. Those words could be broken down into the 44 phonemes (assuming English) and memorized as sequences of phonemes that form the words instead.

Each phoneme could be “tokenized” (given a numeric code) and the code combinations can be used to form the language. Since sound takes up a large amount of storage, but tokens do not, this could be a very efficient way to store and process commands from a language.

Processing Tokens

Tokens can be very easy to process, and thus can have simpler algorithms that run efficiently on smaller devices like Microprocessors (MCUs). This is the same philosophy as tokenizing an interpreted language (for example, Basic).



Traditional voice recognition using learned speech also produces tokens. Processing these tokenized words, and the extension of that to natural language processing (which, in its simplest form, deals with reordering, missing, or added words in sentences) still requires, at minimum, an A-Class core on a device. Often, it’s pushed up to the cloud for processing.

Processing the tokens for the phonemes that define the sounds of a language, (and thus the words) in a given (and expected) order simplifies the process and can be accomplished in a reasonably small MCU. It is this approach that Cyberon has adopted that creates an algorithm so efficient it can run even on an RA2 series (M0/M23 class) device.

Regional differences can cause slightly different pronunciations of phonemes, which we refer to as dialects. Any processing algorithm must account for multiple ways a phoneme can be pronounced.

By restricting the syntax of the command set (in other words, requiring command words to be vocalized in a certain order – equivalent to traversing a tree, which is a standard practice in designing non VoiceUI interfaces) the solution becomes even simpler.

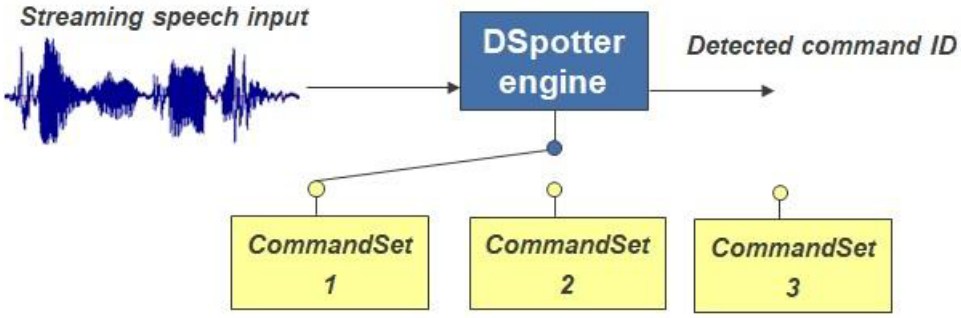
Tools

Instead of training an algorithm on the specific words required for functionality, Cyberon has built a library of phonemes and the combinations of them that create words/phrases. The database of the words contains alternate pronunciations of common phrases that are pronounced multiple ways. The tool that uses this

database to break down words into phonemes and then tokenize them for use in your program is called the *DSpotter Modeling Tool*.

DSpotter asks you to type in your command word or phrase in one of 44 supported languages (you can only use one language at a time in designing the VoiceUI, but multiple command sets can be swapped out at runtime to change languages). It then breaks the word down to phonemes. If there are multiple pronunciations, the tool creates multiple phoneme combinations for a given word or phrase.

These command sets and supporting data for recognition are built into a binary file that the user includes in the project along with the Cyberon library packaged for runtime. The libraries and the binary are used together to support the recognition of the designed speech commands.



Using the Library

The DSpotter tool allows the creation of command groups which can be logically connected by the user program to create a UI with different levels, so a multiple-level command can be issued (for example: “Set” followed by “High” or Low”). Each command in a group has its own index, as does each command within a level.

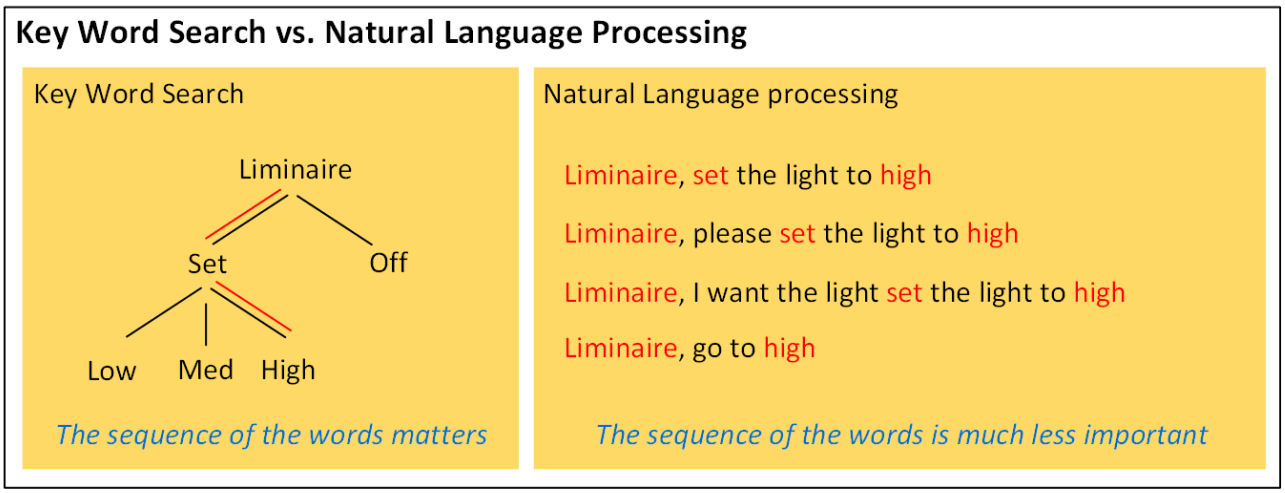
The library processes incoming sound from the microphones and is in search of matching phonemes. When it finds a match, it returns with the index number and the group number. The main application code must just create a hierarchical switch statement to process the command words/phrases as they come.

The library is small enough that it can fit on an RA2 series device. The command set and number of supported languages can grow as the memory size is increased as you migrate the design from RA2 to the RA6 Voice families. (RA6: 2 MB Flash / 640 kB SRAM; RA4: 1 MB Flash / 128 kB SRAM; RA2: 256 kB Flash / 32 kB SRAM)

Limitations

The Renesas / Cyberon solution can be categorized as speech recognition, not voice recognition. If you need speaker dependent commands, then this is not a right fit. For most product designs, however, speaker-independence is actually preferred.

This solution is not natural language processing. If the “tree” gets lost, it will reset back to the top level keywork. A good suggestion is to put some kind of indication on the UI (an indication on a screen, a separate LED, et cetera) to indicate when the processor “thinks” it’s at the top level of the command set.



Turnkey RA Voice User Reference Kit

The easy-to-use hardware platform for Voice User Interface solutions is based on the Renesas RA family of 32-bit Arm® Cortex®-M MCUs providing scalable options for a variety of VUI applications. It demonstrates how easy it is to control your system with a simple voice command interface without extensive coding experience or in-house expertise.

Watch the Renesas Voice UI Hands-On Lab for an overview of the hardware platform and sample projects. The lab covers building an enhanced working solution that recognizes your own voice commands to trigger the corresponding operation using the DSpotter Modeling Tool. Visit the VUI solution page to get started.

Conclusion

In conclusion, the Renesas / Cyberon solution is a powerful yet simple way to add voice commands to an existing or new product in a very short time. The tools are well supported through Renesas, and the partnership ensures designers do not have to pay up front for a tools license because Renesas has built it into the chip price for a “pay as you go” model.

Reference

- Cyberon DSpotter application page
- Voice & Face Recognition solution page
- Voice UI Hands-On Lab

Corporate Headquarters

TOYOSU FORESIA, 3-2-24 Toyosu, Koto-ku, Tokyo 135-0061, Japan
https://www.renesas.com

Contact Information

For further information on a product, technology, the most up-to-date version of a document, or your nearest sales office, please visit:
https://www.renesas.com/contact-us

Trademarks

Doc Number: R01WP0015EU0100