Vol.2011-SLP-88 No.7
2011/10/28

Speech Recognition in the Car: Challenges and Success factors – The Ford SYNC Case

Bart Baeyens[†] Hisayuki Murakami[‡]

Deploying a successful speech application in a car presents many challenges like noise environment, limited embedded computing resources, man-machine interaction, etc. Several recent systems such as Ford SYNC have been very well received in the market in the last few years. In this presentation, we review the main issues faced while implementing such a system as well as the success factors behind today's solutions.

1. Ford SYNC

Nuance and Ford have formed a strategic partnership to bring to market an innovative voice user interface optimized for vehicle in-cabin integration. This voice recognition solution, a key feature in Ford's renowned SYNC system[1], is changing the way drivers interact with in-car navigation and entertainment units and use digital media, portable music players and mobile phone devices in their vehicles.

Both Nuance and Ford recognized the importance in designing a user interface that is adaptable and more intuitive than previous generation in-car systems. Now, driver interaction with the entertainment unit, such as music search, and driver communications, such as placing/answering phone calls, is simpler as the SYNC system maximizes ease-of-use and minimizes distractions posed in traditional systems' manual input and visual confirmation requirements for use.

A market leader in innovation of a better in-car driver experience, Ford was the first auto manufacturer to integrate in-car voice command and control functionality across its fleet – from the luxury vehicle segments, through to affordable mass market vehicles further driving broad consumer adoption of speech in the car. Today, more than 4.5 million Ford vehicles have been equipped with Nuance technology spanning multiple technologies, applications, platforms and languages. The original SYNC system in 2007 was equipped with speech technology for music search, voice dialing and message reading. The next generation SYNC system that hit the market in 2010 added voice destination entry, satellite radio search and climate system command & control to the system. All these voice applications are built on top of the core technologies speech recognition, text-to-speech and acoustic echo cancellation.

2. Deploying Speech in a car

The car is a challenging environment to deploy speech recognition. In order to cope with the noise coming from the car, the road and the entertainment system, a typical speech system has the following characteristics:

- Directive microphone pointed at the driver position. This makes sure that the signal to noise ratio of the incoming speech signal is as high as possible.
- Push to talk button. The speech interaction is initiated by the driver by means of pushing a button. This starts the dialog which is guided by prompts. Each prompt ends with a beep after which the driver can speak the next command.

[†] Nuance Communications International BVBA.
[‡] Nuance Communications Japan K.K.

- Muting the entertainment system. The speech features in Ford SYNC are an integral part of the entertainment system. This allows muting the music output whenever voice commands are expected.

The car remains however a noisy and challenging environment, even with applying these basic techniques. Different driving speeds, varying road conditions, wipers, air-conditioning are examples of noise sources. For coping with these classes of noise, the VoCon engine[2] applies a set of techniques that make it more robust in noise:

- Channel normalization for handling different microphone characteristics and cabin acoustics.
- Built-in noise cancellation algorithm specifically designed for speech recognition.
- Adaptation to the levels and characteristic of the noise.
- Noise robust voice activity detection
- Explicit modeling of non speech sounds to handle non-stationary noises like wipers, etc.
- Acoustic model training with automotive recorded speech

In order to improve the speech recognition performance, the system also adapts the acoustic model towards the driver as he speaks. This results in a better match after a couple of seconds of speech and is particularly helpful for non-native people or strong regional accents. The second generation of SYNC deploys a fully unsupervised speaker adaptation system that automatically adapts to the user. It will detect speaker changes when they occur and reset the system to its initial state before starting the adaptation process again. This technique is completely transparent for the user and is compatible with the automotive use-case of multiple drivers for one car.
Next to the challenging environment, the automotive systems do have limited computing power and memory to run the system. The available resources need to be shared with other applications running on the same platform. The first generation of the Ford's SYNC computer was designed in cooperation with the in-car unit supplier and is built around a 400 MHz Freescale i.MX31L processor with an ARM 11 CPU core. It runs the Microsoft Auto operating system. The new generation has updated the processor to a 600 MHz Freescale i.MX51 processor with an ARM Cortex A8 core. The VoCon engine is designed for this class of processors and provides a set of tuning parameters to find the optimal trade-off between accuracy and speed for the given deployment. The tuning of the system has been critical to its success in the market. With databases that got recorded in the car specifically for the Ford application, the parameters have been tuned to their optimal value.

3. What can I say?

One of the biggest challenges in deploying speech recognition solutions is modeling correctly what the user is going to say. Users can have wrong expectations about what the system can understand and do and therefore speak phrases the system cannot understand. A lot of users do not return to the speech function after first failure. It is therefore absolutely necessary to design the device to be robust against variations in user language. The caricature of the system replying “I did not understand what you said, please repeat” has to be avoided. Ford SYNC deploys a couple of technologies to achieve this.

The grammar design is such that it contains a wide variation for specific commands. An example is the specification of the phone on which to call a contact. The home, office and mobile fields can be spoken in a variety different ways (like at home, at the office, at work, in office, etc.). By adding this variability, the chances that the grammar actually models what the user is going to say increases substantially. This variation is complemented with allowing more commands at the main menu. The second generation of Ford SYNC increased the number of commands from 100 to 10,000 at the start of the system. Users can now directly say “call John Smith” instead of first having to navigate to the phone menu by saying “Phone”.

Specific processing is performed on user data like address books and music titles. For every entry multiple phonetic transcriptions are generated that model spoken variants of the name. Next to these phonetic alternatives, also orthographic alternatives are generated. For the phone system the contact names are split in first name and last name and the system models 3 variants: full name, first name only or last name only. The processing of music titles is more complex and generate partial orthographies for titles where symbols like (), [], etc. are encountered. Also abbreviations like ft. or vol. are handled in a domain specific manner.

A lot of titles in people’s music collections are in a language different from the native language of the user. Multi-lingual solutions are important especially for music selection by voice. The first version of SYNC already deployed music selection in the native language and English for Canadian French and Mexican Spanish versions. The new generation has improved this technology and can roll it out in more geographical locations and with a wider language support. The technology behind this is the multi-lingual phonetization system of VoCon. The component has language identification built-in and applies the phonetic rules of the identified language.

Vol.2011-SLP-88 No.7
2011/10/28

A natural speech application in the automotive world is address entry by voice. Because of its complexity and its huge number of possibilities, it is one of the hardest as well. The Ford SYNC system contains so-called one-shot destination entry in which the user can say a building number, street name and city in one single command. This increases the naturalness of the interaction compared to waiting for the system prompts to enter the city, street and house number. The recognition task however quickly runs into the millions of individual streets. VoCon has search technology based on Finite State Transducers that is able to decode an address utterance with high accuracy within seconds.

One-shot entry provides a natural way of entering an address, but like with music titles and address books, data preparation is performed on the geographic information in order to better model what users typically say. Users tend to shorten street names like North Rodeo Drive by removing all or part of the prefixes and suffixes. During the development of Ford SYNC, tools are used to preprocess the geographical databases and make the orientation prefixes and suffixes and the street suffixes optional parts of the sentence. Different processing is needed for different languages and geographies.

Another technique to improve the user experience and currently deployed in OnStar systems[3] is called Natural Language Understanding. If the system fails to find a good match in the regular – grammar based – speech recognition applications, it falls back to statistical based recognizer followed by a semantic classification engine. This makes sure that sentences like “<cough> I would like to listen to <euh> Michael Jackson” still get routed correctly but also that sentences like “It’s hot today, isn’t it?” get the response “I think you want to do something with the climate control. Possible commands are….”

4. Conclusion

Deploying successful speech recognition in the car involves many challenges. It is important to understand the automotive environment and the expectations of the user in order to model the system in an optimal way. Ford SYNC deploys many of the state-of-the art techniques in speech recognition and speech interface design to help overcome these challenges.

References

1) Ford SYNC systems: http://www.ford.com/technology/sync/
2) Nuance VoCon 3200 Speech Recognition Engine:
http://www.nuance.com/for-business/by-product/automotive-products-services/vocon3200/index.htm
3) OnStar systems: http://www.onstar.com/web/portal/home