White Paper

Reimagining Voice in the Driving Seat

Recognition Technologies® & ARM® for the Next Generation Automobile

Contents

1 Introduction

The modern automotive cockpit is sophisticated and loaded with smart technologies; however, it has little value if the driver finds it difficult to use. Existing smart in-vehicle infotainment (IVI) interfaces have greatly improved the cockpit experience, but voice to date has not delivered on its potential to bridge the gap between hands-free, in-vehicle control and safer driving [1]. Motivation from users for in-vehicle voice interfaces has never been higher. According to the 2016 KPCB Internet Trends Report, over 60% of respondents use voice when their hands are occupied [2]. In the U.S., 36% of respondents reported that the car was their primary setting for voice usage [2]. Yet existing systems are problematic as they are known to be unreliable, which has resulted in user abandonment and a broad based level of dissatisfaction. In 2015, the J.D. Power Initial Quality Study found that the rate of complaints for in-car voice recognition systems is nearly four times the rate of reported problems with transmissions [3]. Drivers as well as passengers seek on-demand, real-time voice controls that match high accuracy with easy operability and reliability. Therefore, a new model of voice operation is required.

> *"Voice-activated command systems and their software often are badly outdated or unreliable, leading to a tide of customer complaints and research questioning how safe they really are."*
> ***The Wall Street Journal*** [4]

Large vocabulary speech recognition engines offer a much improved solution over existing limited vocabulary engines. Utilizing advanced techniques such as deep neural networks, large vocabulary speech recognition engines provide superior learning and recognition capabilities and allow for the complex patterns of human speech to be accurately captured and recognized. Porting such an engine to mobile processors allows automakers to integrate a large vocabulary engine that facilitates natural speech input and leverages on-board, off-line functionality for critical control services. This provides an attractive solution for automakers whose customers are seeking unconstrained natural speech recognition with improved accuracy and application scope.

Comparing continuous speech recognition with existing in-vehicle voice systems is like speaking with a native speaker versus a non-native speaker. In order to have the highest probability of success when speaking with a non-native speaker, it is often necessary to speak more slowly, use common words or phrases and repeat often. This process is similar to existing in-vehicle voice systems. Continuous speech recognition, on the other hand, is similar to speaking with a native speaker in that it allows for a much broader and more natural interaction between both the user and the system. By expanding the vocabulary, the user in this system is free to converse as if having a conversation with a native speaker.

It is clear that an in-vehicle voice advancement is needed that resolves existing reliability issues and is able to fully operate in off-line mode. This white paper discusses a complete voice solution that best fits the needs of the next generation automotive cockpit.

2 Why a Larger Vocabulary is Better for In-Vehicle Voice Control

In a world where more than 6,000 languages are spoken, speech is the basis of everyday life [5]. Speech is a fundamental form of human connection. It allows humans to communicate, articulate, vocalize, recognize, understand, and interpret. Within speech, each person has a unique vocabulary. This is a set of words within a language that is understood and familiar to that person. Data collected by researchers from an independent American-Brazilian research project [6] found that native English-speaking adults understood an average of 22,000 to 32,000 vocabulary words and learned about one word a day. Non-native English-speaking adults knew an average range of 11,000 to 22,000 English words and learned about 2.5 words a day. Most current embedded in-vehicle recognition systems use a vocabulary size of less than 10,000 words and usually use at any point a much narrower context-based vocabulary, which is considered to be a limited vocabulary [7]. However, the words that comprise a native English-speaking

adult's vocabulary can vary greatly from that of a non-native English-speaking adult, partly due to accents and dialects. Dialects are distinct from languages as a dialect is a manifestation of a language that differs from the original in terms of pronunciation, vocabulary, and grammar. For example, in the U.S. alone, dialects range from the basic three – New England, Southern, and Western/General America – to more than 24 others [8]. Therefore, accents and dialects increase the vocabulary size needed for a recognition system to be able to correctly capture and process a wide range of speakers within a single language.

To begin to understand the complex task that a speech recognition system is faced with it is worth assessing the size of the targeted language. In 2010, researchers from Harvard University and Google [9] estimated that a total of 1,022,000 words comprise the English language, with that number increasing by approximately 8,500 new words per year. However, in many cases these figures include different iterations of the same word, many of which are archaic in modern English. It is generally believed that there are more than 150,000 words in current use [10]. When comparing this number to the average vocabulary size of a native English-speaking adult, it is clear that a recognition system with a 10,000 word or less vocabulary is not large enough. This type of limited vocabulary system may result in the user's vocabulary being unrecognized when outside the frame of reference.

When combining the average users' vocabulary with dialect and accent variations, the result is a large vocabulary foundation of more than 150,000 words. Large vocabulary, continuous speech recognition systems promotes the freedom of speaking to an unconstrained system and offers a significant performance increase and a much improved solution over existing constrained systems that use restrictive grammar, limited vocabulary, keyword and phrase spotting. With the continually improving computing power and compact size of mobile processors, large vocabulary engines that promote the use of natural speech are now available for the automotive market. The footprint for such an engine has been optimized and downsized to offer an attractive option for automakers whose customers are seeking unconstrained, natural language interfaces with improved accuracy and embedded on a mobile processor.

3 ARM® in Automotive Infotainment Systems

With more than 85% of infotainment systems and many other applications, such as dashboard and body built with ARM®-based chips, today's automotive experience is founded on ARM technology. A common architecture across all electronics and support from a leading tools ecosystem enables car makers and their suppliers to rapidly innovate in both hardware and software. These innovations are transforming safety, reducing energy consumption, and improving driver experiences.

IVI is the front line in offering a better user experience (UX) to drivers and passengers, for a more comfortable and efficient journey. The automotive industry is keen to provide a seamless UX to a customer base who already enjoy a polished mobile UX every day. More computing performance will be required in IVI to support not only better hardware, such as multiple larger displays with higher resolution, Heads Up Displays (HUD) and better surround speaker systems, but also software to control these systems more easily. Thanks to new regulation to restrict the use of mobile phones while driving, gesture and voice commands will increasingly become the interface of choice in IVI. Speech recognition, in particular, is already entering our daily life in the mobile and consumer space (e.g. Amazon Alexa®, Google Home®). Natural language processing is becoming an important UX requirement. ARM will support next generation IVI with higher performance and power-efficiency processors and power-optimization technology to provide a significant boost to these rich, context-aware experiences.

ARM Cortex®-A applications processors are already widely used in today’s infotainment systems. The majority of current infotainment systems run on Cortex-A7 and Cortex-A15 central processing unit (CPU) cores. Next generation systems will use Cortex-A53, Cortex-A57, and Cortex-A72. Cortex-A72 is the highest performance CPU from ARM

(3.5x performance of Cortex-A15), and is applicable for enterprise infrastructure, mobile, consumer, and automotive. It brings very compelling single-threaded performance, which in automotive is of particular interest for real-time voice recognition. Furthermore, the requisite performance is less than 1/3rd the power consumption of competitor solutions. Cortex-A53 is a very power-efficient processor with a 40% performance increase over its predecessor, the Cortex-A7. Architecturally, the Cortex-A53 is fully compatible with the Cortex-A57 and Cortex-A72 and supports the ARM big.LITTLE™ configurations for these high performance cores.

ARM big.LITTLE processing is a power-optimization technology where high-performance ARM CPU cores are combined with the most efficient cores in a single chip to deliver peak-performance capacity, higher sustained performance, and increased parallel processing performance, at significantly lower average power. The latest big.LITTLE software and platforms can save 75% of the CPU energy consumption in low to moderate performance scenarios and can increase the performance by 40% in highly threaded workloads. Automotive is known as a heat and power constrained environment where this kind of energy management is very applicable. Infotainment systems encompass audio, radio, video processing, navigation, and many interrupts from the vehicle's controller area network (CAN). LITTLE processors can filter out and handle the low-intensity tasks, leaving the big processors to utilize all available resources for tasks that require high performance. ARM is engaging leading operating system vendors such as QNX®, Automotive Grade Linux® (AGL), GENIVI® and Android® on optimizing software to best take advantage of the big.LITTLE processor configuration.

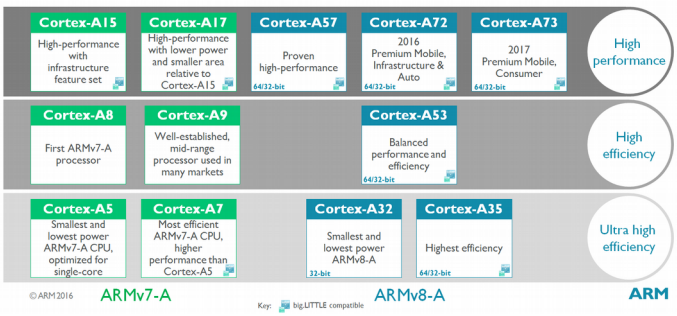

Figure 1. ARM® Cortex-A Portfolio

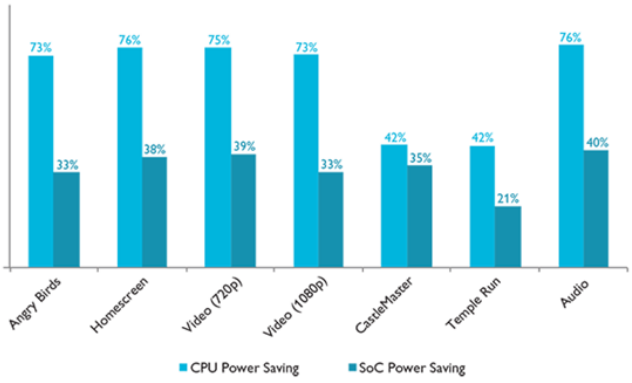

Figure 2. Measured power savings on a Cortex-A15 MP4-Cortex-A7 MP4 big.LITTLE MP SoC relative to a Cortex-A15 MP4 SoC

4 Combining Speaker & Speech Recognition: A Complete Automotive Voice Solution

In a speech recognition application, it is not the voice of the individual that is being recognized but the contents of his/her speech. However, in order to achieve highly accurate, in-vehicle voice control it is highly preferable to identify who the speaker is and segment the turns of each speaker throughout the interaction.

Once the audio from the speaker(s) is captured by the in-car microphones and converted into a mathematical representation of sound, commonly known as a waveform, the combination of text-independent speaker recognition and large vocabulary speech recognition allows for in-vehicle voice control versatility and the achievement of acceptable accuracies.

Effective speaker recognition calls for the segmentation of the audio stream, detection and/or tracking of speakers, and identification of those speakers. The recognition engine provides fusion functionality, combining the scores of both the speaker and speech engines and leading to a fused result that is used to make decisions more readily. The engine also provides the results of the individual engines in conjunction with the fused results so that any generic In-Vehicle Infotainment client application

may utilize the information in a way that would suite the practical application, such as real-time user control directives.

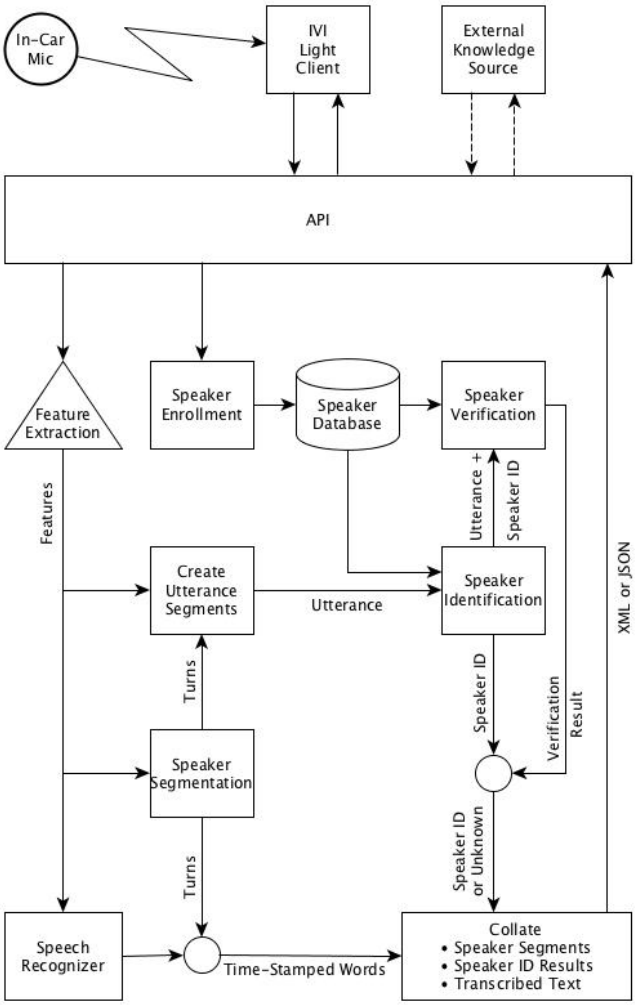

Figure 3. Diagram of a full IVI speaker diarization system including speech transcription and connectivity to cloud-based services (See Figure. 4)

In combining both speaker and speech recognition for the cockpit environment, the RecoMadeEasy® engine returns a collation of the following:

1. Speaker segmentation of incoming audio stream
2. Identification of in-vehicle speaker(s)
3. Verification of in-vehicle speaker(s)
4. Transcription of the in-vehicle audio stream
5. Posting of the above collated results in either Extensible Markup Language (XML) or JavaScript Object Notation (JSON) via an Application Programming Interface (API) to Human Machine Interface (HMI)

RecoMadeEasy® is comprised of multiple speaker recognition branches, both simple and compound. Simple branches that are self contained consist of speaker verification, speaker identification, and speaker and event classification. Compound branches that utilize one or more of the simple manifestations with added techniques consist of speaker segmentation and time-stamping, speaker detection and speaker tracking. Most speaker recognition systems are limited to text dependencies such as fixed text expectations, pre-defined text prompts or and/or language restrictions. RecoMadeEasy® is a purely text-independent and language-independent system that only relies on the vocal tract characteristics of the speaker and makes no assumption about the context of the speech. This indicates that an individual may enroll her/his voice into the system in one language and be identified or verified in a completely different language. This is useful for being able to handle verification and identification processes across any number or languages. In fact, the speaker recognition engine has been used for more than 100 languages.

In a typical in-vehicle usage case, there are often multiple speakers including the driver and passengers who want to interact with the voice system. In order to fully and accurately transcribe the interaction, it is necessary to know the speaker of each statement. When there is an enrolled model for each speaker, and prior to identifying the active speaker, the audio of that speaker is segmented and separated from adjoining speakers. In the absence of a prior model, the system labels a new speaker as an “Unknown” speaker.

As the audio is being segmented and separated, open-set identification is conducted. The audio stream segment is compared against all in-vehicle enrolled speaker models and the speaker ID of the model with the closest match is returned. In addition to matching the audio segment against known models, it is also compared to rejection models which are used to return an Unknown label for a speaker who has not been enrolled in the system before – this constitutes the “open-set” aspect of the identification.

In the case of verification, the engine is used to verify the identity of users based on the vocal characteristics by receiving a random user prompt, to be evaluated by the speech recognition engine. The provided ID of the user is then used to retrieve the model for that person from a database and the speech signal of the speaker is compared against the existing speaker model to verify.

In parallel to the execution of the numerous speaker recognition functions, the audio stream is passed to the speech recognition engine with the loaded vocabulary to produce the recognized text. The engine handles different user dialects and accents and generates a full speech to text transcription. The speech recognition engine uses a streaming interface, where the recognizer in the form of listeners and the client both run on the embedded device. Any light generic client capable of using a websocket interface may stream audio/video in any codec that is supported by GStreamer-1.0, including MP3, Ogg Vorbis, Free Lossless Audio Codec (FLAC), MP4, Pulse Code Modulation (PCM) or other codecs such as those supported by a standard Waveform Audio File Format (WAVE), to one such listener and get back real-time results of the transcript with optional alternative results, including likelihood scores.

Based on experience, close to 30% of the audio frames in a normal audio segment are silence frames [11]. In speech recognition, the extraneous silence segments will produce spurious nonsense words by taking leaps through different arcs of Hidden Markov Models. RecoMadeEasy® includes silence segmentation as a standard component via the use of energy thresholding, advanced features and model based techniques. In this way, the algorithm is used to estimate the threshold along the time line. Different thresholds may be established automatically to identify silence or non-speech signals adaptively. By eliminating silence/non-speech, the accuracy of the transcription is increased and reduces un-necessary processing energy and, in some cases, bandwidth utilization.

For the engine to function at its full potential and to allow in-vehicle users to speak naturally and be understood – even in a far-field, noisy environment, pre-processing techniques are integrated to help improve the quality of the audio input to the recognition system. The microphone peripheral converts automobile noise sources both inside and outside the vehicle to signal. These include tire and wind noise while the vehicle is in motion, engine noise, and noise produced by the car radio, fan, windscreen wipers, horn and turn signals. The engine leverages audio input pre-processing, which runs on the ARM® NEON DSP and radically improves the accuracy by bettering the quality of voice signals passed to the speech recognition engine for processing.

RecoMadeEasy® is a combination of both engines into a single interface engine, capable of performing many tasks on audio streams. For example, a full diarization of an audio file may be done using the engine by segmenting the audio file into different portions where uniform events happened such as where an individual speaks or non-speech events happen. In addition, the speaker identification is performed on each segment to provide the identity of the speakers in the stream. Non-speech events in the stream are also tagged using the classification capability of the speaker recognition portion of the engine. In parallel, the speech recognition engine provides a full real-time transcript of the audio in the stream. The engine is capable of handling many different dialects and accents in a single large-vocabulary transcription engine, while simultaneously processing audio segmentation and speaker identification. The result is an embedded engine that can process a real-time audio stream that includes multiple speakers. The embedded engine also tags each segment with the identity of the speaker and transcribes the conversation while providing timestamps for every turn in the transcript.

Furthermore, due to the large overlap of underlying modeling techniques between speaker and speech recognition systems, being developed by Recognition Technologies, both engines share all necessary C++ libraries, reducing duplication of loaded share libraries. This is in contrast to less desirable alternate scenarios where independent speaker recognition and speech recognition engines may be used. This optimal usage of resources such as memory and processing are essential in an embedded environment.

Features of the RecoMadeEasy® Engine

- Speaker Recognition: Fast Match for Large Populations (Speaker Identification, Verification, Classification, Detection, Tracking and Segmentation)
- Speech Recognition: 150,000-200,000 Word Vocabularies
- Custom Vocabularies can be easily created and used in combination with the main vocabulary
- Full Feature Control though Configuration Files
- Multiple Interfaces – C++ SDK, Graph-Based Interface, Web Services
- Handles Client Streaming – Integrating through WebSocket and native interprocess communication and supporting GStreamer 1.0 supported media
- Native Multi-Threading Support
- Default Configuration Evaluated and Adjusted Frequently through Benchmarking
- Inclusion and Support of Benchmarking Scripts and Environments
- Built-in Multiuser Support with Full Configuration Independence Across Users
- Linked with ffmpeg and GStreamer libraries for extensive codec support and streaming

5 Leaning into the Future: Embedded Voice and the Connected Car

Automakers are developing next generation cars that integrate a variety of new technologies that will make cars more digitally connected. Advanced cloud-based technologies that rely on distributed computing, such as natural language processing, artificial intelligence, machine learning in combination with speech recognition are commonplace. Connected services such as intelligent personal assistants (e.g. Amazon Alexa®, Google Home®) that provide natural language interfaces for conversational voice interaction are now widely used. These services are a combination of cloud-based speech recognition and natural language processing, where audio voice data is captured and transmitted over a wireless network. However, in real-time automotive cockpit environments, the use of these cloud-based technologies presents significant system and user concerns. According to VDC Research, major threats against connected cars fall into two categories: safety and data privacy[12]. All connected devices are at risk of some form of attack. Transmitting user voice data over a wireless network presents serious user safety and privacy concerns. Attackers, able to penetrate either the car or cloud provider ends, would be able to access users' personal biometric information. Once biometric information is compromised, it cannot be changed like a password. RecoMadeEasy® only stores statistical models about the biometric and not the biometric itself, therefore, the user's biometric cannot be compromised in this way. From a network standpoint, the transmitting of audio voice data per second to the cloud is over 300 times larger than text data per second, and the cumulative volume of voice data transmitted to the cloud presents serious ongoing bandwidth and latency concerns when scaled.

Cloud-based speech recognition systems are a great solution when full, uninterrupted and secure access to a cloud server is guaranteed. However, in real-time automotive cockpit environments, this is not the case and cars have to be able to reliably perform at task even in the presence of service interruption. In order to limit the number of potential points of failure, localized processing is highly preferred over centralized servers wherein each vehicle is capable of collecting and processing its own voice data. This is achieved by combining fully optimized and downsized embedded speech recognition with the continuously improving computing power and compact size of ARM® mobile processors. However, rather than simply writing off cloud-based technologies for automotive use and reverting back to embedded technologies, a hybrid model is desirable that exemplifies the positive benefits of both models. It is achievable for both models to work in conjunction with one another to extend system functionality, increase system efficiency and provide user security, privacy and protection.

Cloud-based speech recognition systems have two streaming limitations to address when transmitting voice audio data over a wireless network. First, the data rate or the size of the audio file per second of data is transmitted on average at 128 kbps, meaning that each one-second chunk of user voice audio

comprises about 128 kilobits of data. This would be achieved using Ogg Vorbis, MP3, or other lossy compression techniques. Second, the bandwidth or the connection speed to the cloud controls the automotive cockpit application's ability to transmit audio to the cloud. Cloud providers have to pay for their bandwidth usage and have to balance high quality performance with network bandwidth costs. Since voice audio data is the most dominating part of the data, the data rate and the bandwidth per line would be essentially the same. Therefore, in order to calculate the approximate bandwidth requirements, the number of concurrent lines that need to be supported are multiplied by the data rate.

In contrast, transmitting textual data on the slowest voice modem is minimal. For example, if we take the fastest speaker in the world who speaks at a rate of 586 words per minute, it would come to less than 10 words per second. An average word has 4.5 letters, therefore, this would make it 45 bytes per second or about 0.36 kilobits per second. It should be noted that the average person does not speak that fast. Despite taking into consideration the required JSON or VoiceXML tags, the bandwidth needed to transmit remains realistic.

When comparing the transmitting of audio voice data versus text data to the cloud, it becomes quite clear that performing speech recognition processing locally is an obvious solution for optimizing bandwidth consumption and providing significant bandwidth savings. The textual and audio voice data bandwidth levels differ by two orders of magnitude. In this hybrid model, mission critical concerns such as user security, privacy and protection are addressed by processing user speech natively on the device as well as ensuring continuous availability. Non-mission critical dimensions, such as natural language processing, can be processed in the cloud and uses only low bandwidth, textual data as the mode of bilateral transmission. Therefore, by prioritizing and modularizing both mission and non-mission critical processes by running RecoMadeEasy® speaker and speech recognition as an embedded process, we are able to extend system functionality, increase system efficiency such as bandwidth optimization and provide critical user security, privacy and protection.

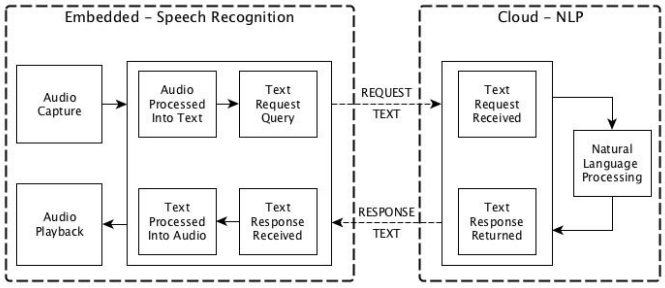

Figure 4. A hybrid model for embedded speech recognition and cloud-based natural language processing

6 Off The Line Performance: Real-time Speaker Segmentation, Enrollment, Identification, Verification & Speech Transcription

Speaker Enrollment, Verification and Identification

In order to fully and accurately diarize any interaction, it is necessary to know the speaker of each statement. An enrollment model is generated for each speaker based ideally on 60 seconds of speaker audio data, however the shortest enrollment time-frame may be as little as eight seconds. For open-set speaker identification, audio stream segments are compared against all enrolled speaker models and the claimed speaker ID of the model with the closest match is returned. In addition to matching the audio segment against known models, it is compared to rejection models which are used to return an Unknown label for a speaker who has not been enrolled in the system before. However, it should be noted that the RecoMadeEasy® Speaker and Speech Recognition engines do not require a claimed speaker ID to be passed to the system in order for the speaker segmentation, identification, and speech transcription to be performed.

In order for speaker verification to be evaluated by the engine, the engine requires two pieces of information; the first is a claimed ID and the second is an audio segment that is used to verify the identity

of the speaker based on the vocal characteristics. The provided ID of the user is then used to retrieve the model for that person from a database and the speech signal of the speaker is compared against the initial enrollment speaker model to verify.

The performance results, as detailed in all of the below figures, were generated using a medley of TED talks. The medley is comprised of eight TED speakers, including male and female and native and non-native English speakers. The total audio stream is one minute and twenty-five seconds with each speaker turn lasting approximately ten seconds. Code optimization of the engine has been performed that aims at meeting prime requirements for embedded systems such as timing accuracy, code size efficiency, low memory usage, time pressure, reliability, and recognition performance.

The stacked area chart in Figure 5 shows the memory usage for (i) speaker enrollment, (ii) simple speaker identification, and (iii) simple speaker verification, which both use a limited rejection mechanism (e.g. less than 10 speaker models), (iv) complex speaker identification, and (v) complex speaker verification, which are both based on a much larger range (e.g. more than 1000 rejection models).

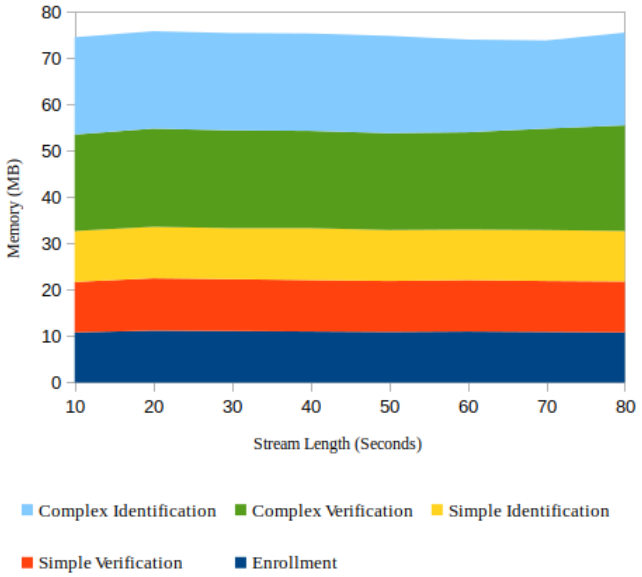

Figure 5. Speaker Recognition Memory Usage: the following numbers were computed on an ARM® Cortex-A53, Quad Core 1.5GHz processor

Speaker Segmentation

Automatic segmentation is elementary to the practical realization of speaker and speech recognition systems. A typical in-vehicle voice interaction contains speech and non-speech signals from a variety of sources, including clean speech, speech over music, speech over ambient noise, speech over speech, etc. The segmentation challenge is to be able to separate the speech produced by different speakers and other non-speech segments. Most speech recognizers will break down if they are presented with music instead of speech. Therefore, it is essential to the accuracy of the recognition system to separate the non-speech signals from recognizable speech.

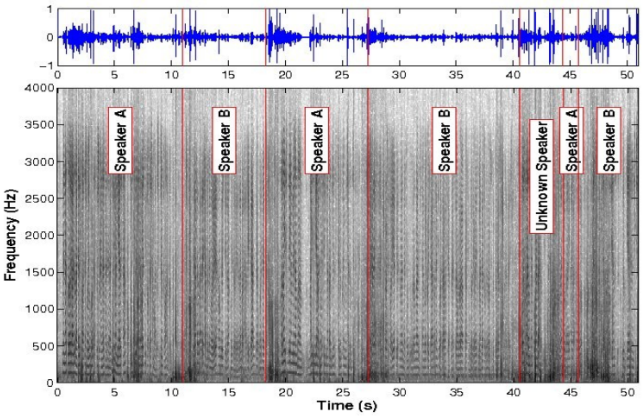

Figure 6. Speaker Open-Set Segmentation Results

Speaker Diarization, Identification, Tracking, and Speech Transcription

Prior to identifying the active speaker, the audio of that speaker is segmented and separated from adjoining speakers. In the absence of a prior model, the system labels a new speaker as an “Unknown” speaker. In parallel to the execution of the numerous speaker recognition functions, the audio stream is passed to the speech recognition engine with a vocabulary size of 150,000-200,000 words to produce the recognized text. The engine handles different user dialects and accents and generates a full speech to text transcription.

The line chart in Figure 7 shows the resident memory usage for the speaker diarization process in combination with concurrent processes. This includes

(i) speaker enrollment and (ii) segmentation, segment identification and speech transcription.

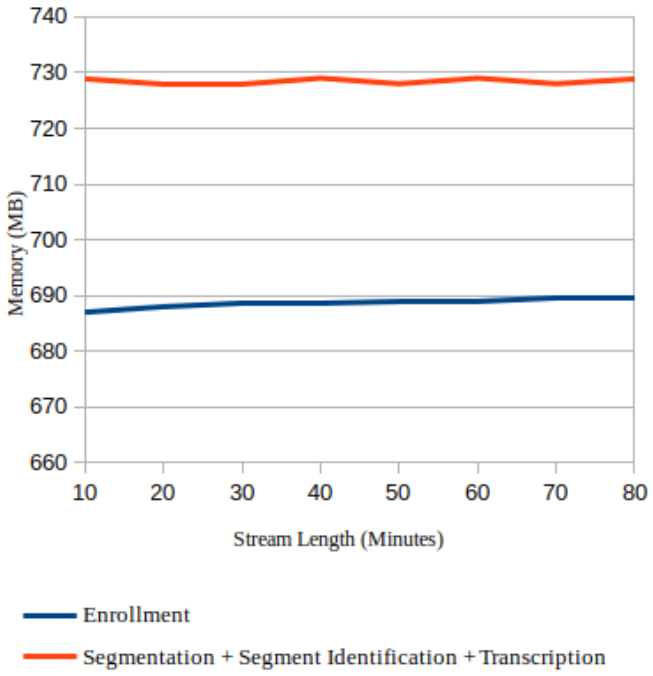

Figure 7. Speaker Diarization Performance: the following numbers were computed on an ARM® Cortex-A53, Quad Core 1.5GHz processor

The scatter chart in Figure 8 shows the error rates for (i) the speaker recognition process; this includes segmentation and segment identification, and (ii) the speech recognition process; this includes complete speech transcription using a 150,000-200,000 word vocabulary.

The speech recognition word error rate – or how frequently the engine transcribes a word incorrectly is calculated as the number of errors (replacement + deletions + insertions) divided by the total number of words during the speaker turn or segment. A human transcribed ground truth was established pre-test that allows for an accurate post-test comparison to establish a precise word error rate for the whole transcription process. As stated previously, the medley is comprised of eight male and female speakers. There are three native speakers whose nationalities include American and British, and five non-native speakers whose nationalities include French, Italian, Ghanaian, Saudi Arabian and Chinese. Each speaker turn lasts approximately ten seconds and is free form speech, which includes applause and non-sense words. Throughout the medley, there are both silence and non-speech noise frames that the engine eliminates in order to increase the accuracy of the transcription.

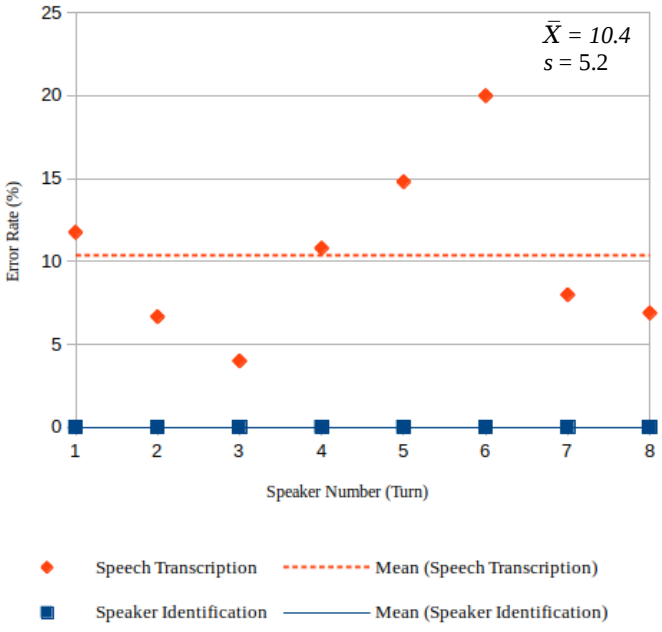

Figure 8. Speaker and Speech Recognition Error Rate: the following numbers were computed on an ARM®Cortex-A53, Quad Core 1.5GHz processor

7 Recognition Technologies® SDK

The RecoMadeEasy® C++ SDK is a series of dynamic libraries that contain all necessary components for In-Vehicle Infotainment voice integration. The standard package contains tailored and tested installation packages for currently available ARM® Cortex-A 64 bit based processors. The package is tailored to automotive-focused Linux operating systems and serves to enable natural language voice interfaces quickly and easily. An API is also included with the SDK that enables engineers to get started easily and without any work or costs for development.

RecoMadeEasy® C++ SDK and API

- Unified API and SDK for both engines
  - Simple and Short API Header
  - Fully Configurable Through Simple Configuration Files
- Small Application and Libraries
  - C++ code for Client Application
  - C++ code for Standalone Application
  - Speaker Recognition, 16 Libraries
  - Speech Recognition, 10 Libraries
- Complete Error and Warning Handling and Propagation to the Main Application
- Very Small Memory Footprint
  - Each Speaker Model is only 204kB

The RecoMadeEasy® SDK can be obtained by contacting Recognition Technologies: automotive@recognitiontechnologies.com

8 Summary

Voice recognition engines, like automotive engines, are not all built equal. Combining speaker and large vocabulary speech recognition systems offers a significant increase in accuracy, performance, and application scope over existing systems that have limited vocabulary, keyword, and phrase spotting. Combining this type of engine with the continuously improving computing power and compact size of ARM® processors makes continuous natural speech interfaces accessible for in-vehicle purposes. The RecoMadeEasy® SDK serves as a complete automotive cockpit voice solution that runs natively on automotive-focused Linux operating systems. Tier One companies can integrate the series of dynamic libraries and offer car makers a fully embedded and customizable voice interaction experience for their IVI products.

9 References

[1] "2016 U.S. Initial Quality Study." J.D. Power, 22 June 2016.
[2] Meeker, Mary. "2016 KPCB Internet Trends Report." Kleiner Perkins Caufield & Byers, 1 June 2016.
[3] Murtha, Philly. "2015 Initial Quality Study: Top 10 Reported Problems." J.D. Power, 1 July 2015.
[4] White, Joseph B. "The Hassle of ‘Hands Free’ Car Tech." Wall Street Journal, 23 Nov. 2014.
[5] "Ethnologue: Languages of the World."Ethnologue Web. Ed. Raymond G. Gordon. SIL International, 2005.
[6] Little, Regina. "Vocabulary of Native and Non-native Speakers: The Lifelong Pursuit of Language Learning." Cyracom, 3 Nov. 2013.
[7] Ivanecký, J., & Mehlhase, S. "Today’s Challenges for Embedded ASR."Mathematical and Engineering Methods in Computer Science. Vol. 8934. Springer, 2014. 16-29.
[8] Wolfram, W., & Schilling-Estes, N. “American English: Dialects and Variation.” Oxford: Basil Blackwell, 1998.
[9] Michel, Jean-Baptiste et al. “Quantitative Analysis of Culture Using Millions of Digitized Books.” Science, 2011. 176–182.
[10] "How Many Words Are There in the English Language?" Oxford English Dictionary, May 2016.
[11] Beigi, Homayoon. “Fundamentals of Speaker Recognition.” Springer, 2011
[12] https://www.vdcresearch.com

10 Authors

Mark Sykes - Recognition Technologies, Inc.
Homayoon Beigi - Recognition Technologies, Inc.
Soshun Arai - ARM Limited.