A soft whisper heard clearly: Getting automatic speech recognition right

Steve Taranovich - February 13, 2017

I have been seeing some innovative development in microphone and speech recognition. Consumers want to speak commands to their automobiles, mobile devices, and wearables, but ambient noise can get those messages wrong. Automatic speech recognition (ASR) and natural language processing (NLP) for systems like Siri, Google Now, Alexa, Cortana, etc., work pretty well in a quiet home, but our real-world environment surrounds us with a great deal of noise.

Most system designs developed to mitigate ambient noise will analyze the speech and noise and try to enhance the speech and suppress the noise. Seems reasonable. But this technique distorts the voice signal; this is a "physics" approach which suppresses noise signals and boosts voice signals – but this inherently introduces distortions that speech engines cannot process. Let's take a look at where most people use their phones and wearables (**Figure 1**).

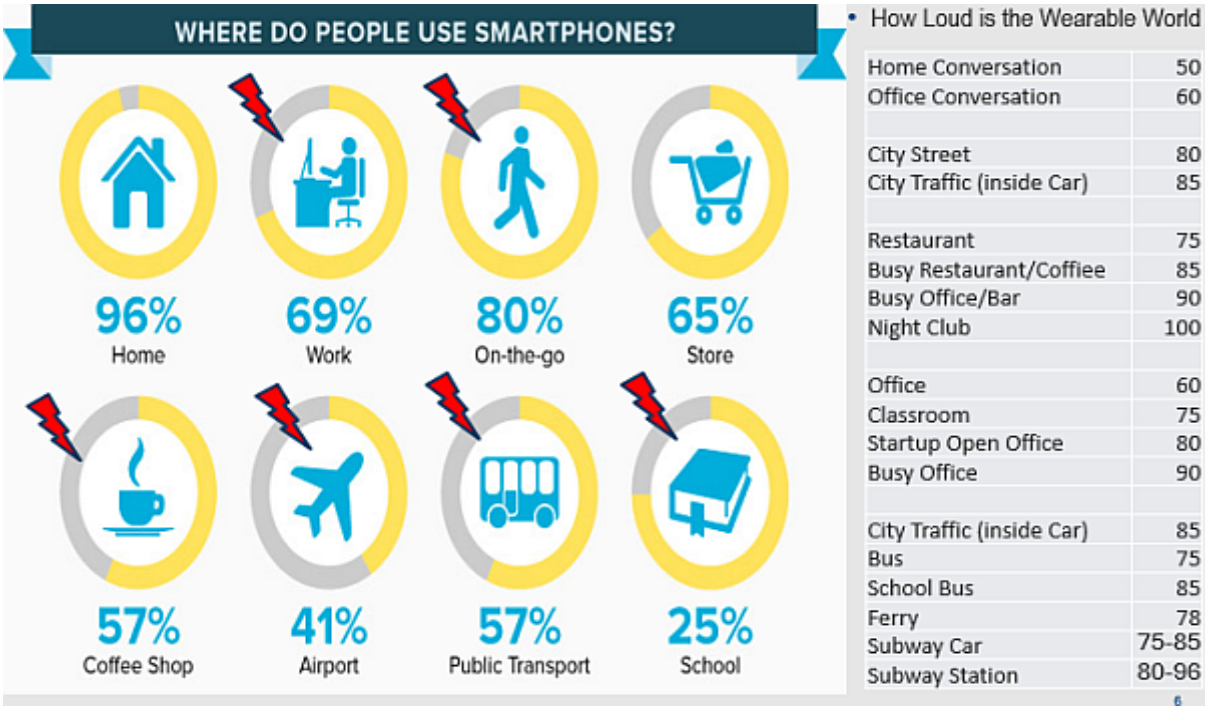

Figure 1 Where people use phones and wearables and noise levels in the real world (Image courtesy of Kopin)

Why do we want to use voice commands? @KCPB, a Venture capital firm tells us (**Figure 2**).

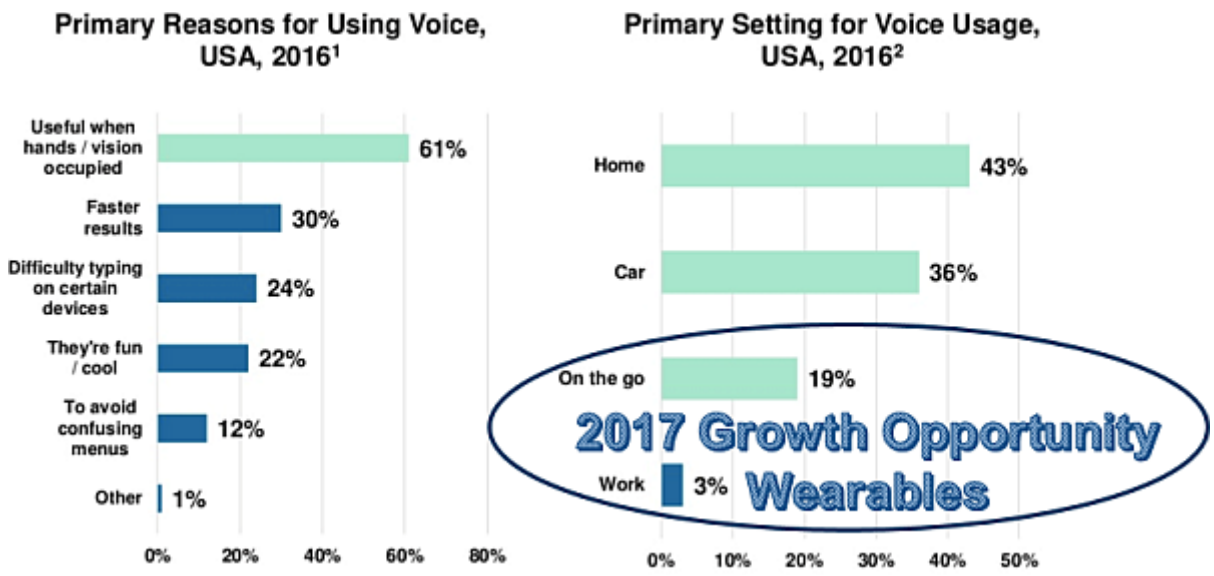

Figure 2 Hands- and vision-free interaction is the main reason consumers want to use voice in the home, car, or just on-the-go. (Image courtesy of @KCPB)

Speech recognition is presently at about 95% accuracy, but experts, like Andrew Ng, Chief Scientist at Baidu, say that going to 99% would be a game-changer. Accuracy is a primary goal and the secondary goal is latency (Who wants to wait 10 seconds to get a response from your system?) for explosive use of voice recognition by consumers. In 1970, Machine speech recognition was only 10s of words. Fast-forward to 2016 and 7-8 million words were recognizable with 90% accuracy in a low noise environment, according to Google.

Kopin's Whisper Voice Interface IC takes a different approach to noise with artificial intelligence (AI). They sample the acoustic environment 16,000 times per second, perform a dynamic analysis of noise and voice activity, and use their Voice Extraction Filter to "extract" voice without distortion once the parameters are tuned for the device and the application.

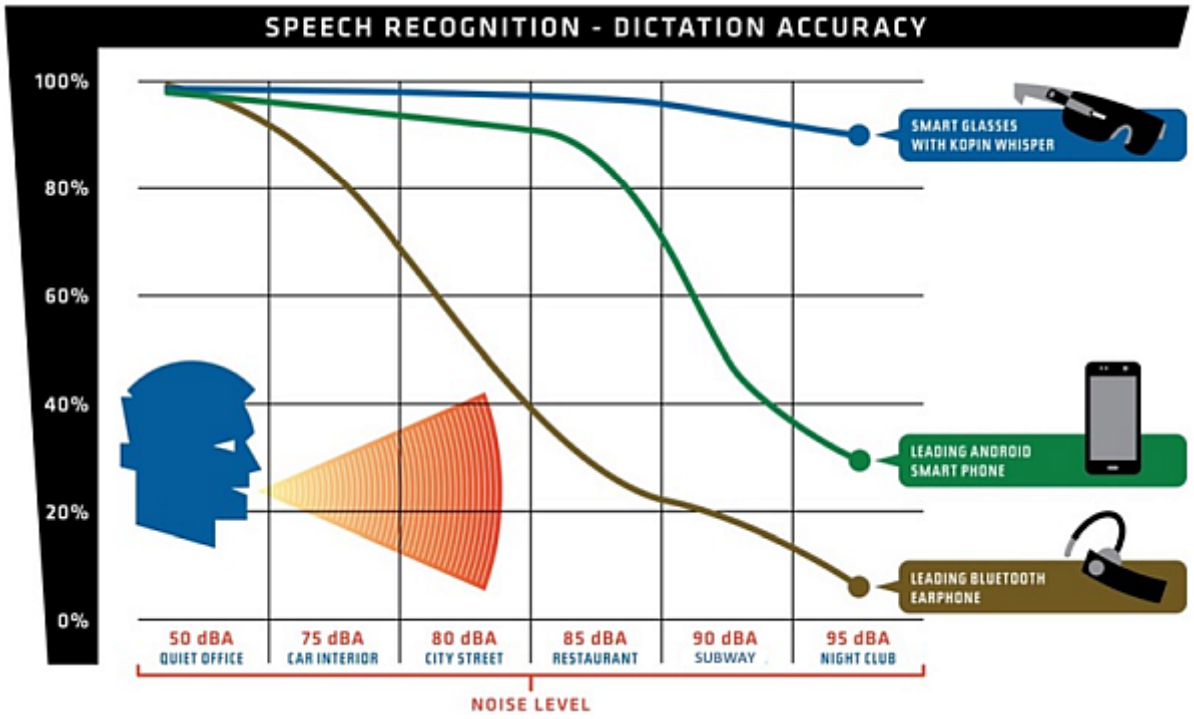

Figure 3 Voice recognition accuracy in a real world noise environment is demonstrated here. The chart compares the performance of smart glasses with the Whisper chip against the ASR and noise cancellation technologies found in two popular devices: a leading Bluetooth earphone and a leading smartphone. (Image courtesy of Kopin)

As can be seen in **Figure 3**, the Whisper chip's performance remains consistent as noise levels increase, the earphone's performance begins degrading at 75 decibels (the amount of noise associated with a car interior or dishwasher) while the smartphone's ASR performance starts to drop at approximately 85 decibels (or the amount of noise associated with restaurant).

The voice chip

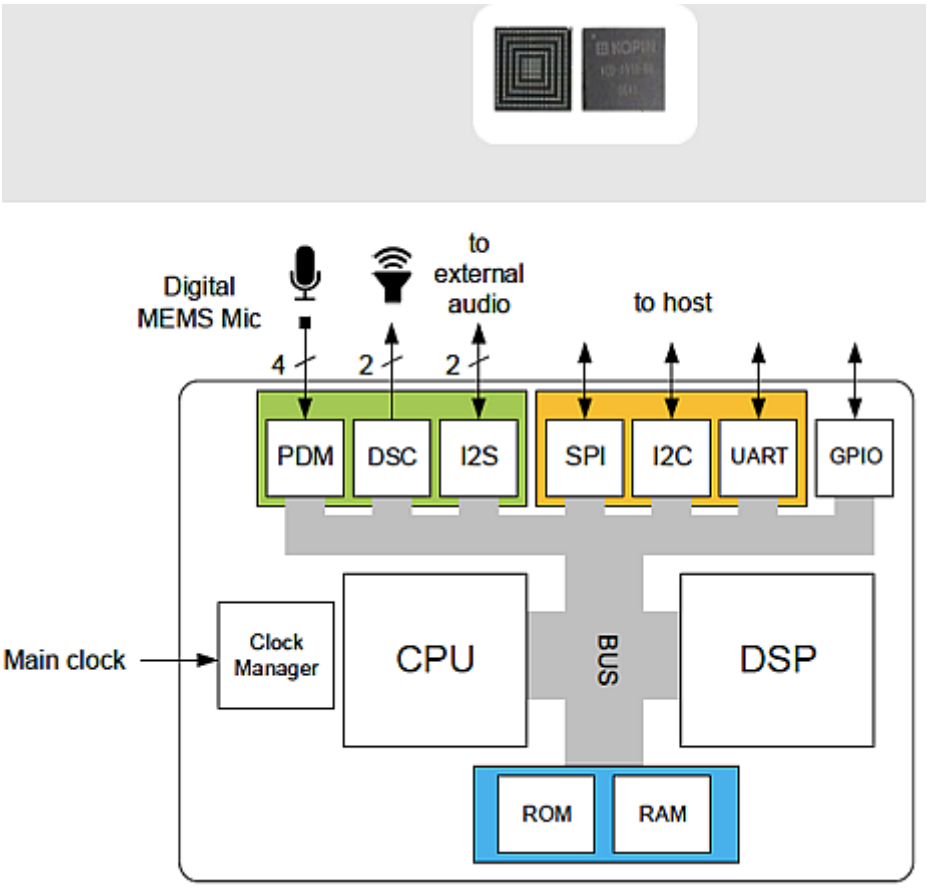

Figure 4 Logistically the Whisper Chip is simple to implement, and is situated between the microphones and speech engine. It also works with the leading operating systems, processors, and speech recognition engines.

I really like this Whisper Voice IC which is one of the best ways I have seen for voice recognition applications so far in the industry. The Voice Extraction technique lends to just about zero distortion to the voice signal. The adaptive voice detection architecture allows for “listening” which adjusts to the environmental noise level. The Whisper Voice Chip can be tuned for mid- and far-field applications (up to 5 meters distance from microphone).

Some of the audio processing capabilities of this solution are:

- Microphone Balancing
- Beam Forming
- Noise Cancellation
- Voice Activity Detection
- 16 kHz Sampling Rate
- 48 kHz PWM

The system separates speech from the noise without creating non-linear distortion in the voice signal. The most common cause of poor voice quality from a device and of ASR failures is non-linear distortion, whether using speech recognition software on the host device or cloud processing. The IC needs less than 10 mW in operation which is important to the life of battery-operated portables and

wearables.

Since this architecture is an all-digital solution, it can replace the codec–there is no ADC or DAC needed. Another nice advantage is that the chip provides efficient front-end audio processing, enabling less processing and power demand on the device's host processor.

Digital microphone inputs can handle up to four microphones and there are two digital speaker outputs. The chip's compact size of only 4×4 mm keeps the board footprint low especially in portables and wearables.

Since voice seems to be the most natural way for humans to communicate with the 'machine,' I give my endorsement for this solution which helps make voice recognition a reality and a practical solution (because of excellent performance, low power, and low cost which will bring consumers to use more portables and wearables coupled with a good microphone design). My vote in the microphone arena is Vesper.

Also see:

- Wake up and listen: Vesper quiescent-sensing MEMS device innovation
- SmartEverything and the rise of the microphone array
- Microphones: A sound technology choice for communication and control
- Basic principles of MEMS microphones
- Audio experts on microphone levels and pressure zone mics