

Validation of In-Vehicle Speech Recognition Using Synthetic Mixing

John Huber, Ranjani Rangarajan, An Ji, Francois Charette, Scott Amman, Joshua Wheeler, and
 Brigitte Richardson
 Ford Motor Company

ABSTRACT

This paper describes a method to validate in-vehicle speech recognition by combining synthetically mixed speech and noise samples with batch speech recognition. Vehicle cabin noises are prerecorded along with the impulse response from the driver's mouth location to the cabin microphone location. These signals are combined with a catalog of speech utterances to generate a noisy speech corpus. Several factors were examined to measure their relative importance on speech recognition robustness. These include road surface and vehicle speed, climate control blower noise, and driver's seat position. A summary of the main effects from these experiments are provided with the most significant factors coming from climate control noise. Additionally, a Signal to Noise Ratio (SNR) experiment was conducted highlighting the inverse relationship with speech recognition performance.

CITATION: Huber, J., Rangarajan, R., Ji, A., Charette, F. et al., "Validation of In-Vehicle Speech Recognition Using Synthetic Mixing," *SAE Int. J. Passeng. Cars – Electron. Electr. Syst.* 10(1):2017, doi:10.4271/2017-01-1693.

INTRODUCTION

In-Vehicle speech recognition robustness is a challenge to the automotive industry. Factors such as background noise level and microphone placement affect how well the system performs. Live hardware validation using on-road testing with subjects of various dialects and accents can be both costly and time consuming. In addition, it is arguably impractical to collect a statistically significant amount of on-road data to draw proper conclusions on the results. Running batch speech recognition can not only validate the recognition robustness of a particular vehicle model, but a greater understanding of the various factors that affect speech recognition can be achieved using this simulation.

COMPONENTS OF SYNTHETIC NOISY SPEECH

The creation of noisy speech utterances that properly represent the interior acoustics of an automobile requires some specific components. Along with pre-recorded speech, there are three other parts to the process. These include the cabin impulse response (CIR), the desired background noise, and a mathematical representation of the vehicle microphone.

Clean Speech

Clean speech utterances without background noise represent the starting point in the signal mixing. These are recorded in a studio with a high quality microphone and typically are spoken by multiple speakers from both genders with varying dialects. The microphone should be close to the mouth. Thus, these recordings are often referred to as "close talk" utterances.

Cabin Impulse Response

The CIR from the speaker's mouth to the vehicle microphone location is convolved with the close talk utterances to produce hands free speech signals. Because the vehicle microphone is some distance away and the cabin of a vehicle is bounded by both absorptive and reflective surfaces, this convolution has a significant effect on the signal.

The CIR is measured using a Head And Torso Simulator (HATS) and a single instrumentation grade microphone (see [Figure 1](#)). A pseudorandom signal known as a maximum length sequence (MLS) is played from the mouth of the HATS while being simultaneously recorded by the microphone placed at the hands-free microphone location. The CIR is computed using both the recorded signal and the original MLS. Each vehicle has a potentially different impulse response depending on the cabin interior size and shape, the surface trim, and most importantly, the vehicle microphone location.

Additionally, by adjusting the seat location, the relative position of the mouth to the microphone changes. This creates the need to measure multiple CIR's for each vehicle.



Figure 1. Example of the head and torso simulator in a vehicle to measure the cabin impulse response from the speaker's mouth to the cabin microphone location.

Background Noise

While there are many sources of noise in a vehicle, such as road, wind, wipers, blinkers, air conditioning, etc., this work focused on two of these, road noise and blower noise from the heating ventilation and air conditioning (HVAC). Road noise originates from contact between the tires and the road surface and is controlled by both the suspension design (how well it isolates the vehicle from the road), and the sound package (how insulated the cabin is acoustically from the outside). HVAC noise is controlled by the blower fan, the ducting, and the grill opening designs both at the panel vents and at the defrost outlets.

Speech recognition technology in current generation automotive vehicles uses a bandwidth from about 250 Hz up to 8 kHz. Background noise that is dominated by a booming sound (50-80 Hz) will likely have less of an effect on recognition performance than HVAC blower noise that is dominated by an air rush sound (500-3000 Hz) even though the former may sound louder to a person in the vehicle. Figure 2 shows a comparison of the frequency content for a typical set of background noise data.

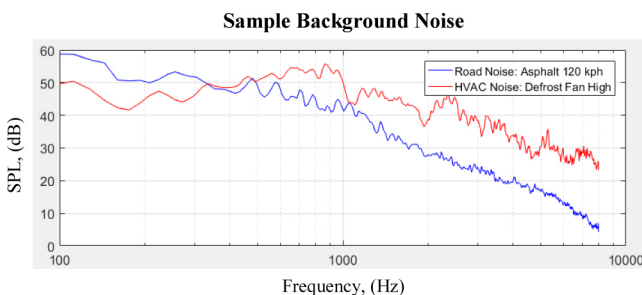


Figure 2. Sample background noise comparison of Road Noise and HVAC Noise for a typical vehicle.

For reproducibility purposes, background noises are collected on controlled test track surfaces, each of a specific roughness. A given road surface and vehicle speed combination affects different parts of the frequency spectrum. Lower speeds on coarse road surfaces, for example, will tend to result in lower frequency acoustic energy. While high vehicle speeds on smooth surfaces will result in higher frequency wind noise energy.

Microphone Characterization

It is important to recognize that microphones installed in vehicles are meant for high volume production. They typically do not have a flat frequency response like a high quality instrumentation grade microphone does. Their design is generally optimized for speech recognition and phone call quality and they tend to band limit the signal. In addition, microphone sensitivity can vary from vehicle to vehicle in the production environment. Therefore, proper representation of the vehicle microphone plays an important role in the synthetic speech recognition process.

The approach taken here was to measure a large batch of production microphones for their individual frequency response characteristics. An average in the frequency domain is determined and this spectral curve is developed into a finite impulse response (FIR) filter. This filter is then applied to the noisy speech signals to represent the effects of the vehicle microphone in the signal chain.

SYNTHETIC NOISY SPEECH CREATION

The process for generating noisy speech that includes vehicle cabin acoustics, background noise, vehicle microphone effects, and Lombard speech will now be discussed.

Lombard Speech

Lombard speech is produced by the natural tendency of a speaker to raise their voice when speaking in loud environments in order to increase audibility. This effect causes the speaker to change such characteristics as pitch, intensity, duration, and spectral slope [1]. The proposed synthetic speech process outlined in this paper utilizes a simple and convenient technique to simply scale non-Lombard speech in order to simulate the effect of a louder voice. For typical vehicle background noise levels, increasing the non-Lombard speech by 6 dB would be a first approximation, but other approaches that scale the speech based on noise level could be used as well.

Mixing Procedure

The process of mixing noise and speech in combination with the above effects is shown in Figure 3. Clean speech is filtered to account for the microphone frequency response effects, scaled to approximate the Lombard reflex and finally convolved with the cabin impulse response. The resulting signal is added to the background noise which has also been filtered by the same microphone FIR filter.