Chapter 9
In-Vehicle Speech and Noise Corpora

Nitish Krishnamurthy, Rosarita Lubag, and John H.L. Hansen

Abstract As in-vehicle speech systems become prevalent, there is a need for specific compilation of data in vehicle scenarios to develop/benchmark algorithms for speech systems. This paper describes the collection efforts and analysis of two corpora: (1) the UT-Dallas Vehicle Noise (UTD-VN) corpora and (2) the CU-Move in-car speech and noise corpora. The UTD-VN corpus is focused on addressing the variability of in-car noise environments. This corpus includes compilation of unique noise scenarios within the car (Engine idling, AC windows closed, etc.) as well as variability of these scenarios across different makes and models. Another aspect that in-car speech systems need to address along with noise is the emotional and task stress of the driver while performing the driving task. The CU-Move corpus focuses on collection of data to describe the variability of conversational speech in an in-car environment. A sample study is carried out where it is shown that these environments are unique across different vehicles using the UT-Dallas Vehicle Noise corpora. This shows that a detailed analysis of variability across vehicle platforms is necessary for successful deployment of speech systems. In our opinion, these corpora are the first to describe the environment variability along with conversational speech in an in-car environment.

Keywords Car noise • command and control • Enhancement • Environment variability • Environmental noise • Navigation • Speech • Speech recognition • Speech systems • Stress

N. Krishnamurthy (✉)
University of Texas at Dallas, Richardson, USA

Texas Instruments, Dallas, USA
e-mail: nitish@ti.com

R. Lubag • J.H.L. Hansen
University of Texas at Dallas, Richardson, USA
e-mail: john.hansen@utdallas.edu

J.H.L. Hansen et al. (eds.), *Digital Signal Processing for In-Vehicle Systems and Safety*,
DOI 10.1007/978-1-4419-9607-7_9,

9.1 Introduction

Car environments are becoming a standard/core location for conducting voice-based commerce in dialog systems, message or information exchange, and other business or entertainment exchanges. However, one of the main challenges faced by speech and audio systems today is maintaining performance under varying acoustic environmental conditions caused by different driving conditions, car make and models, along with speech variability due to task-induced and emotional stress caused during driving. Efficient use of speech systems in cars require technology to be robust across variations in vehicle environments encountered. In fact, the diversity and rich structure of noise and speech in car acoustic environments create the need for application-specific speech solutions. This is a challenging task since effective communications systems in the car need to address the issue of diversity in transportation platforms and operating conditions. Another aspect along with the environment is the emotional and task stress caused due to task variability and distraction within typical car environments. These factors lead to significant acoustical variations in speech and noise encountered within vehicle environments. The focus here is not the social or legal issues associated with speech system deployment in car environments, but the description of corpora development to address the variability encountered in car environments.

The environment in transportation platforms varies due to the different makes and models of transportation platforms along with the changing operating environments encountered. Examples of the changes in acoustic variations include road characteristics, weather, and the state of the car. Road characteristics are a significant source of noise variation in cars, and the surface properties of the road can change the properties of noise encountered (e.g., asphalt versus concrete, and smooth vs. cracks or potholes). Also, noise changes are dictated by weather conditions such as rain, snow, winds, etc. Depending on the severity, these conditions can sometimes mask other noise events/types in cars. The focus here is to study the variability in normal weather conditions.

Even though significant efforts have been made in the past to study the impact of car noise on speech, there remains a need for a corpus to enable the study of noise events across vehicles and their impact on speech systems. The UT-Dallas Vehicle Noise (UTD-VN) corpus aims to compile the variability observed across vehicles and driving conditions for a fixed set of environmental conditions. This collection is unique as it contains a comprehensive collection of noise events across different vehicle platforms. A sample analysis here formulates noise in a car environment and shows that the noise types are distinguishable across different vehicle makes and models demonstrating the necessity for noise-specific corpora. This corpus opens up new research opportunities where the knowledge gathered from studying car noise events can be exploited by in-vehicle speech systems.

Another aspect of variability in car environments is the speech variability due to task and emotional stress. The CU-Move corpus is a compilation of speech

data collected during natural conversational interaction between the user and an in-vehicle system. In the past, studies have analyzed the impact of in-vehicle noise on speech systems including use of fixed noise and speech collection in lab environments without the variability induced in either speech or noise. Recently, some studies like [1] by Kawaguchi et al. have incorporated these variations. Their corpus focuses on spontaneous conversational Japanese where the speech data was collected under car idling and driving conditions. This study does not include the environment variability of the speech due to the task-induced stress. CU-Move focuses on compiling these variations in speech under diverse acoustic conditions in the car environment along with various environments encountered in realistic driving task. This data was collected from six different vehicles. The core of this corpus includes over 300 speakers from six US cities, with five speech style scenarios including route navigation dialogs. The noise collected during this corpus identified over 14 different unique noise scenarios observed in the car environment.

The goal of CU-Move is to enable the development of algorithms and technology for robust access and transmission of information via spoken dialog systems in mobile, hands free environments. The novel aspects of CU-Move include corpora collection using microphone arrays on corpus development on speech and acoustic vehicle conditions. This setup enables research utilizing environmental classification for changing in-vehicle noise conditions and back-end dialog navigation information retrieval subsystem connected to the WWW. While previous attempts at in-vehicle speech systems have generally focused on isolated command words to set radio frequencies, temperature control, etc., the CU-Move system is focused on natural conversational interaction between the user and in-vehicle system. Since previous studies in speech recognition have shown significant losses in performance when speakers are under task or emotional stress, it is important to develop conversational systems that minimize operator stress for the driver. System advances using CU-Move include intelligent microphone arrays, auditory- and speaker-based constrained speech enhancement methods, environmental noise characterization, and speech recognizer model adaptation methods for changing acoustic conditions in the car.

Here, the focus will be on the UTD-VN corpus with mention of relevant aspects in the CU-Move corpus. In conjunction, these corpora address most of the variations in the environment and speech encountered for holistic development of in-car speech and communication systems.

9.2 The UT-Dallas Vehicle Noise Corpora

In the UTD-VN corpus, noise data samples were collected from 20 cars, five trucks, and five SUVs across 10 different noise events. To enable portable recording across vehicles, a portable, lightweight, high-fidelity data collection setup was used

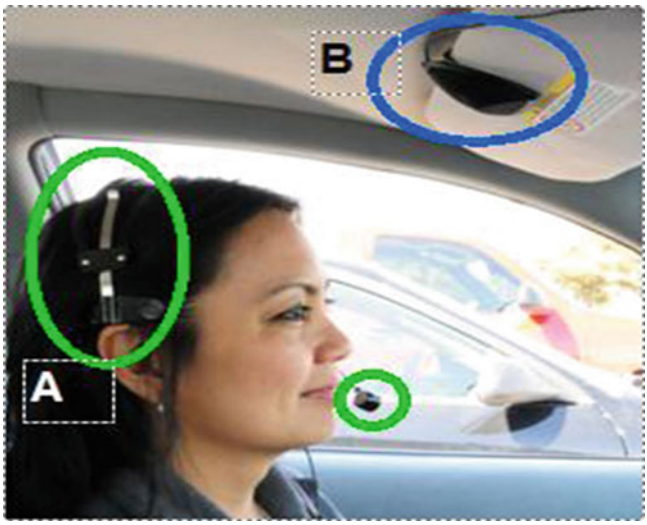

Fig. 9.1 In-vehicle portable recording setup. (**a**) Shure SM10A close talk microphone and (**b**) Shure MX 391S omni-directional microphone.

to obtain accurate recording of the noise data. The equipments used for in-vehicle data collection were:

(a) Shure SM10A close talk microphone
(b) Shure MX 391S omni-directional microphone
(c) Edirol R-09 recorder

Figure 9.1 shows the recording setup. The close-talk microphone (marked as (a)) is worn by the driver during the data collection to record the noise data observed in the closed-talking microphone under different conditions. The far field microphone (b) has been secured onto the sun visor located above the driver's seat. Meanwhile, the data collector (not in the picture) managed the recording setup.

Data is collected under the following events within the vehicle acoustic environment:

(a) NAWC: No air-conditioning with windows closed
(b) ACWC: Air-conditioner engaged with windows closed
(c) NAWO: No air-conditioning with windows open
(d) HNK: Windows closed with car horn
(e) TRN: Turn signal engaged
(f) IDL: Engine idling
(g) REV: Engine revving
(h) LDR/RDR: Left/right door opening and closing

For these fixed events, the noise varies due to weather and road conditions. To minimize the number of independent variables, such as external noise and road characteristics, the driving routes were fixed for all recordings. The average speed

of the car during the recordings was 40 mph, and data was collected on a 4-mile route of concrete roads. The route was selected so as to consist of a combination of 6-lane city roads with higher traffic density and 2-lane concrete community roads with lower traffic density. The car noise data recording was timed so as to minimize external traffic noise due to peak hours.

The data collection consisted of two parts. The first set of in-vehicle noise events were recorded in the University of Texas at Dallas parking lot. For these recordings, the vehicle was stationary, the windows were closed, and the AC was turned off. Under these vehicle conditions, the following sound events were collected:

(a) Turn signals (TRN)
(b) Horn (HNK)
(c) Front doors opening and closing (LDR/RDR)
(d) Engine idling (IDL)
(e) Revving (REV)

The average total recording time for these conditions was about 6 min.

The second set of recordings took place on roads around the University of Texas at Dallas campus. The data was collected only in dry weather conditions, where the vehicles completed the route twice. The route was 2 mi. long, with two-to-three lane roads and speed limits ranging from 30 to 40 mph. For this corpus, the route was divided into seven sections, and a particular noise condition was assigned to each section.

Figures 9.2 and 9.3 show the designated route. The seven sections of the route are also shown in the figures. As shown in Fig. 9.2, two noise conditions (ACWC, NAWC) were collected in the first loop. During sections A to D of the route, the windows and AC remained closed. In sections E to G of the route, the windows were closed and the AC was turned on with blower at full capacity. Meanwhile, Fig. 9.3 shows the four noise conditions recorded during the second loop. Section K of the route included a speech exercise. Here, the driver was asked to count out aloud from 0 to 9, three times, with the windows closed and AC turned off. Data for NAWC condition was recorded again in sections L and N. The final recording condition was ACWC in section H. The average on-the-road recording time was about 21:25 min. The priorities of this exercise were the NAWC and ACWC conditions as speech systems encounter these conditions frequently in car environments. Collection setup was designed to allow for data collection in multiple sessions to ensure that the audio recording of the car events contained variability due to different road/traffic conditions. The corpus contains a total of 8 h of car noise data.

9.3 CU-Move

The UTD-VN corpus deals with variability in fixed environments across car makes and model. Another aspect of in-car acoustics as mentioned in sect. 1 includes speech variability due to stress along with noise. These are the major causes of acoustic mismatch in a car environment. The CU-Move corpus focuses on

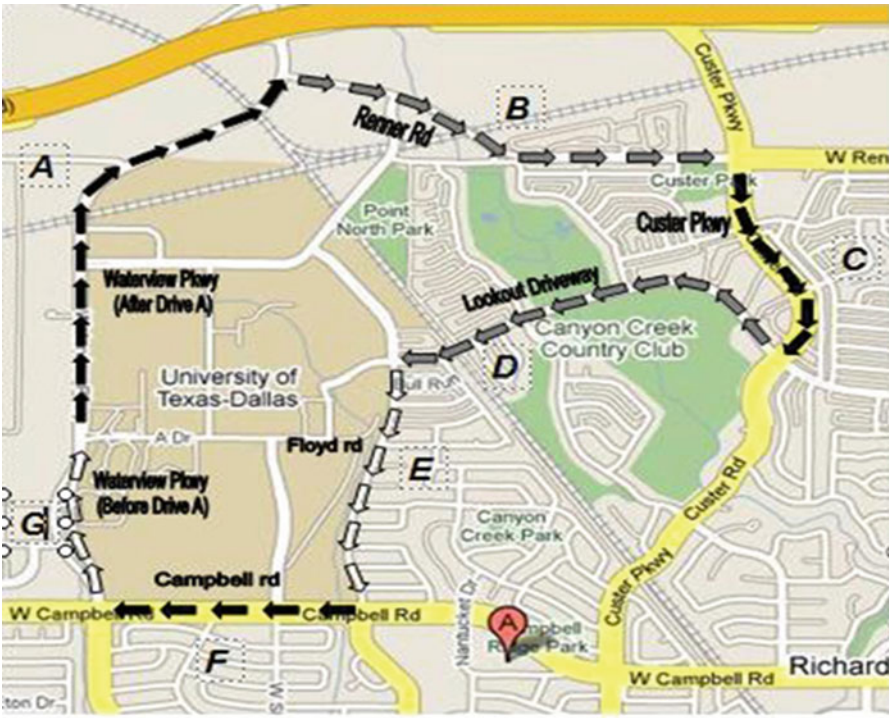

Fig. 9.2 Loops for data collection. The *dotted lines* indicate unique recording conditions

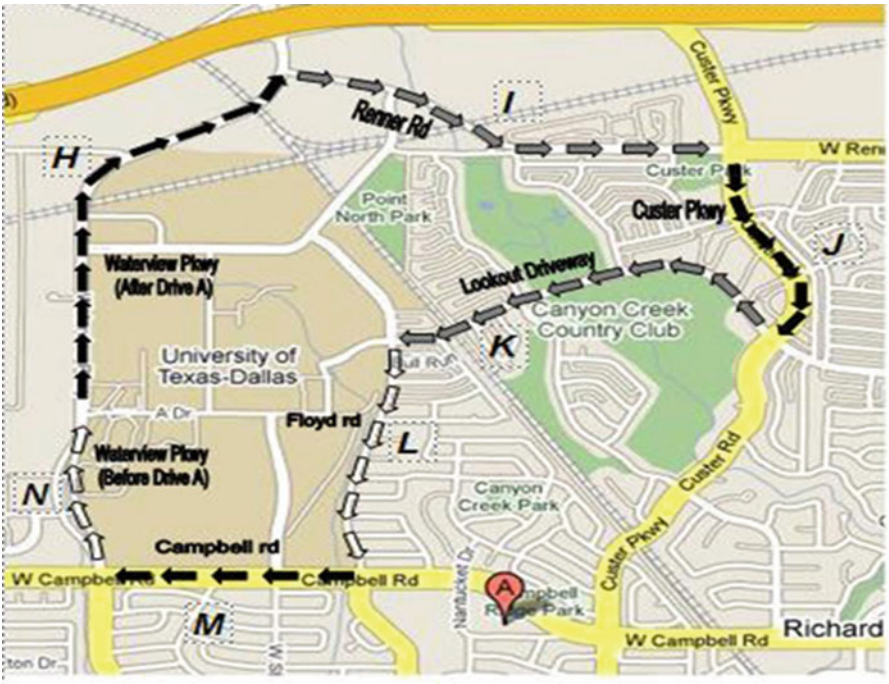

Fig. 9.3 Loops for data collection. The *dotted lines* indicate unique recording conditions

compiling speech variability for an in-car task along with the environment variability encountered for different task scenarios. This corpus consists of 2 phases:

- Phase I: Speech & speaker data collection
- Phase II: Acoustic noise data collection (CU-Move Noise)

9.3.1 Phase I: Speech and Speaker Data Collection

The speech and speaker data collection is divided in two sections. First (part 1) is structured text where the user is prompted to utter text and numbers similar to what is observed in a command and control application. The second section (part 2) is a dialog system scenario with a real person on the other end.

9.3.1.1 Part 1: Structured Text Prompts

The driver performs a fixed route that includes a combination of several driving conditions (city, highway, traffic noise, etc.). For each speaker, prompts were given for specific tasks listed below from a laptop display situated around the glove compartment of the vehicle. This portion is 30 min long. There are four subsections that include:

- Navigation direction phrases section: a collection of phrases which are determined to be useful for in-vehicle navigation interaction (prompts fixed for all speakers)
- Digits prompts section: strings of digits for the speaker to say (prompts randomized)
- Streets/address/route locations section: street names or locations within the city; some street names will be spelled, some just spoken (prompts randomized)
- Sentences – general phonetically balanced sentences section: collection of phonetically balanced sentences for the speaker to produce (prompts randomized)

9.3.1.2 Part 2: Dialog Wizard of Oz Collection

Here, the user calls a human "wizard" (WOZ) who guides the subject through various routes determined for that city. More than 100 route scenarios particular to each city were generated so that users would be traveling to locations of interest for that city. The human WOZ had access to a list of establishments for that city where subjects would request route information (e.g., "How do I get to the closest police station?", "How do I get to the Hello Deli?"). The user would call in with a modified cell phone in the car, which allows for data collection using one of the digital channels from the recorder.

9.3.2 Phase II: Acoustic Noise Data Collection (CU-Move Noise)

One of the primary goals of the CU-Move corpus is to collect speech data within realistic automobile driving conditions for route navigation and planning. Prior to selection of the vehicle used for phase II data collection across the United States, an in-depth acoustic noise data was collected on six vehicles in Boulder, Colorado. This section briefly summarizes the noise data collection scenarios.

9.3.2.1 Vehicles

A set of six vehicles were selected for in-vehicle noise analysis. These vehicles were from model years of 2000 or 2001 (all had odometer mileage readings which ranged between 11 and 8,000 mi.). The six vehicles were:

- [Cav] Chevy Cavalier compact car
- [Ven] Chevy Ventura minivan
- [SUV] Chevy SUV Blazer
- [S10] Chevy S10 extended pickup truck
- [Sil] Chevy Silverado pickup truck
- [Exp] Chevy Express cargo van

All acoustic noise conditions are collected across six vehicles: Blazer, Cavalier, Venture, Express, S10, and Silverado. The noises were labeled into 14 categories which include:

1. Idle noise: the sound of the engine after starting and not moving, windows closed
2. Noise at 45 mph, window opened 1″
3. Noise at 45 mph, window closed
4. Noise at 45 mph, window opened half way down
5. Noise at 65 mph, window opened 1″
6. Noise at 65 mph, window closed
7. Acceleration noise, window closed
8. Acceleration noise, window opened half way down
9. AC (high) noise, window closed
10. Deceleration noise, window opened 1″
11. Turn signal noise at 65 mph, window closed
12. Turn signal noise, window opened 1″
13. Turn signal noise, window closed
14. Wiper blade noise, window closed

A total of 14 noise conditions were extracted from the same environment and locations for each of the 6 GM vehicles. This noise corpus focused on describing the unique variations in noise scenarios encountered in a car environment as opposed to focusing on variations across cars. This is described in Fig. 9.4. The CU-Move

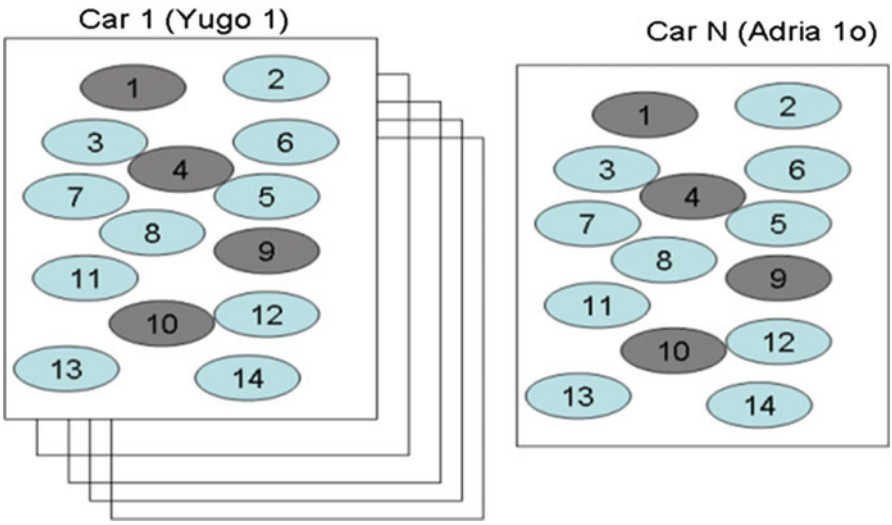

Fig. 9.4 The Scope of UTD-VN vs. CU-Move

corpus is a compilation of events encountered in a car-driving scenario as described earlier whereas the UTD-VN is a compilation of a few events across various cars and conditions.

9.4 Noise Analysis and Modeling

The car noise–environment noise samples in both the corpora can be described as a combination of noise sources active in the car, as well as the acoustic environment of the vehicle itself. In other words, the resultant car noise is a function of car-independent noise (n_e) and car-dependent noise (n_{ce}). Here, an additive model for $(\hat{n}_{ce})$ is assumed. This is illustrated in Fig. 9.5. Depending on the relative dominance of the constituent noises, the overall resultant noise observed can be of three primary types.

- *Car Internal Dominant Noise*: If car-dependent sounds such as air conditioning, horn, and engine sounds dominate, then the resulting noise $n_e$is unique to the specific car producing the sound (i.e., if $n_e << n_{ce}$ then $(\hat{n}_{ce}) \approx (n_{ce})$). For purposes of car verification/platform identification, this forms the most conducive scenario. For speech systems, it means that car specific models might be optimal for the best performance in specific car environments.
- *Car–Environment Dominant Noise*: If the observed sound is the sound of the car interacting with its environment, such as the sound of wheels on the road or wind

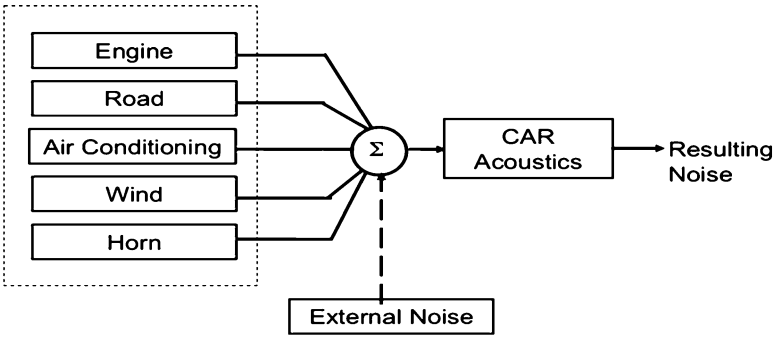

Fig. 9.5 Model of acoustic environment in the car

noise, then the resulting car noise is less car specific/dominant (i.e., $n_e > n_{ce}$). This scenario is less favorable for car verification than the previous case.

- *Environment Dominant Noise*: Finally, noise sources external to the car such as horn from a nearby car or engine sounds from a passing truck are considered outside the scope of this study. This is because these sounds are least car specific (if $n_e << n_{ce}$ then$(\hat{n}_{ce}) \approx (n_{ce})$). This would cause increased confusability in acoustic vehicle platform identification. This case would require the most generic noise models for speech systems.

In practice, it is very difficult to obtain these noise types in isolation since all noise sources cannot be controlled simultaneously in naturalistic driving. However, in the process of car noise data collection, we have minimized external noise by carefully choosing the recording conditions.

For analysis, three noise conditions in the same vehicle are analyzed for their spectral content and variability. These conditions consist of NAWC, ACWC, and NAWO, as shown in Fig. 9.6. These environments were chosen because of their high probability of occurrence. Furthermore, these noise scenarios represent unique environments because the dominant sounds in each case are different (e.g., in ACWC, AC noise is dominant).

The spectral content of the vehicle acoustic environments under ACWC, NAWC, and NAWO conditions are shown in Fig. 9.6. As seen in Fig. 9.6b when the AC is on and the windows are closed, the car noise is least time varying. The main noise sources in this environment are AC, car engine, and road noise, but the AC is the dominant source of noise. The spectral slopes indicate that the ACWC scenario has the most high-frequency content compared to the other two noise types. Also, this condition is the most conducive for car verification since the AC and the fan/air blower are the most dominant noise sources. In the other two cases, wind noise and road noise are the main noise sources. When AC is turned off, as seen in Fig. 9.6a, the car noise is a mixture of road and engine noises. The only car-dependent noise type when the AC is off and windows closed is the car engine noise which is masked by the road noise. Finally, the last plot shows NAWO condition, where the main noise sources are wind noise, road noise, and engine noise. NAWO

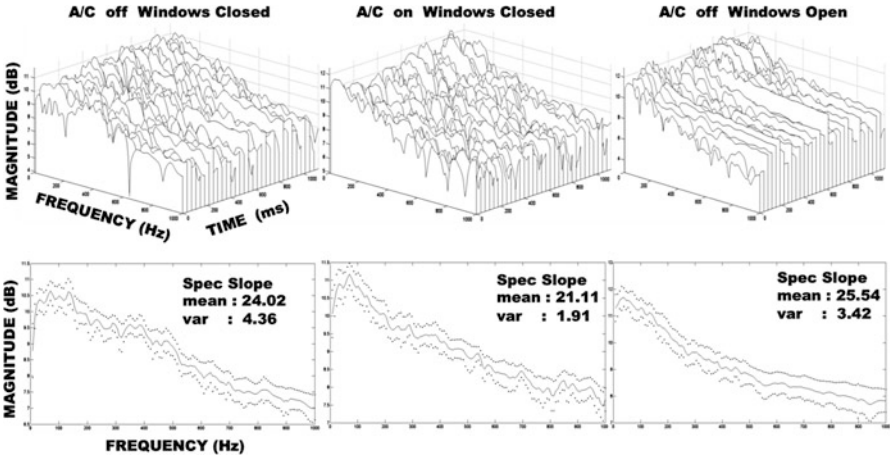

Fig. 9.6 Vehicle acoustic environments: (**a**) road and engine noise is predominantly low frequency, (**b**) road, engine, and air-conditioning shows structure in higher frequencies, and (**c**) wind noise wipes out all structure and only the aggregate remains

has the least car-dependent information as compared to the other two environments since the wind noise is external to the car and masks all car-dependent information. As seen here, car-dependent noise types are the best indicators of car types and the car-dependent AC noise, which can be viewed as a potential excitation source for the interior vehicle compartment, enables the noise to carry more car-dependent information. To study the uniqueness and the variability in different acoustic conditions across cars, the acoustic data was modeled using 13 dimensional Gaussians and the Kullback–Leibler distance was employed to analyze the in-class and across-car differences. This is illustrated in Fig. 9.7, where solid areas represent the acoustic space for a single car in a particular environment, and the smaller shaded areas represent models of the session-to-session variability in the same acoustic event.

To estimate the separation across different vehicles, the in-class and across-class KL distances are measured. If the vehicle sound events are separable within this framework, the average in-class distances will be much lower than the out-of-class distances. These distances are evaluated for three vehicles, and box plots of these distances are presented in Fig. 9.8. As seen for each of these vehicle conditions, the in-class (IS) distances are clearly separable from the out-of-class (OS) distances, indicating that under the ACWC are spectrally unique and differentiable from each other.

As evident from this discussion, car noise environment is a unique environment with a mix of car-dependent and independent noise sources. Depending on the driving conditions and road scenarios, the environments may rapidly change from one condition to the next. The analysis also reinforces the need to collect data under different scenarios as the intervehicle variations might be a significant factor to normalize for generic speech systems.

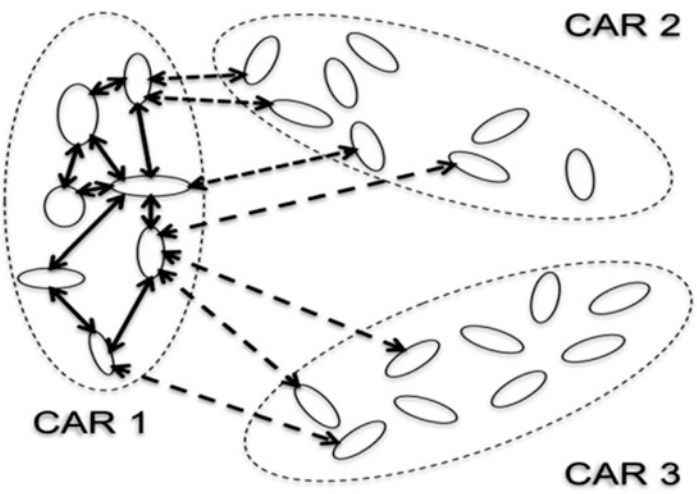

Fig. 9.7 Illustration of in-class vs. out-of-class distances for each noise event in a car. Each *dotted region* denotes a car and it encloses solid regions that denote session instances

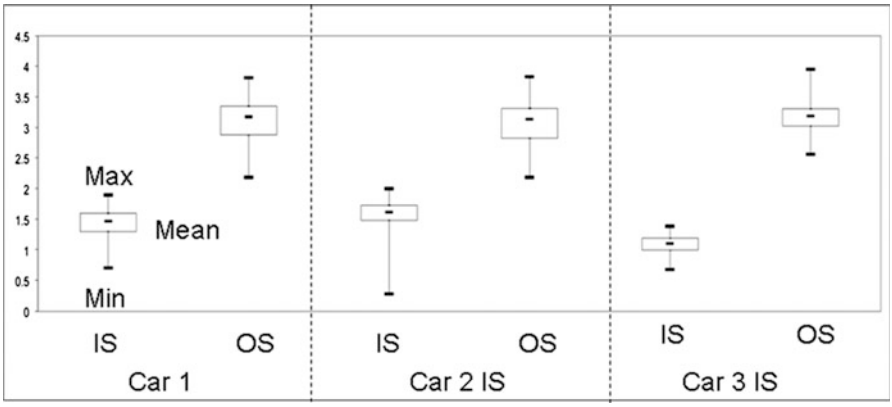

Fig. 9.8 Inset and out-of-set distances for ACWC in three cars

The CU-Move corpus has been used extensively to understand the noise properties in car environments and leveraging these properties for speech systems. Examples of these studies include [3, 4], and [5] by Akbacak and Hansen to "environmental sniffing" of variations in environment to use most appropriate models using the Rover scheme. In [6], the authors used the CU-Move corpus for advancing voice-activated route navigation in car systems. Hansen [7] includes a detailed description of the corpus along with the usage scenarios of CU-Move.

9.5 Conclusion

This paper summarizes collection efforts of the UTD-VN corpus and the CU-Move corpus. The UTD-VN corpus includes a rich variety of noise types that are frequently encountered in car environments. The UTD-VN corpus contains noise data that reflects the variability in vehicle noise events across different makes and models of cars, whereas the CU-Move corpus includes the diversity in the car environments with variations in speech due to task and driving stress. Using the UTD-VN corpus, a model for car noise was formulated and used to demonstrate the uniqueness of noise types across different vehicles. The volume, diversity, and real-world nature of these corpora make it very valuable for researchers exploring in-vehicle speech technology. The next stage of data collection would be ubiquitous data collection for in-car environments that would use multiple sensors for aiding development of integrated multi-input systems that are most suited for in-car environments.

References

1. Kawaguchi N, Matsubara S, Iwa H, Kajita H, Takeda K, Itakura F, Inagaki F (2000) Construction of speech corpus in moving car environment. In: Proceedings of the Interspeech-2000, vol 3, Beijing, pp 362–365
2. Hansen JHL, Plucienkowski J, Gallant S, Pellom B, Ward W (2000) CU-move: robust speech processing for in-vehicle speech systems. In: Proceedings of the Interspeech-2000, vol 1, Beijing, pp 524–527
3. Akbacak M, Hansen JHL (2003) Environmental sniffing: robust digit recognition for an in-vehicle environment. In: interspeech-2003/Eurospeech-2003, Geneva, pp 2177–2180
4. Akbacak M, Hansen JHL (2003) Environmental sniffing: noise knowledge estimation for robust speech systems. In: IEEE ICASSP-2003: international conference on acoustics, speech, and signal processing, vol 2, Hong Kong, pp 113–116
5. Akbacak M, Hansen JHL (2007) Environmental sniffing: noise knowledge estimation for robust speech systems, IEEE Trans Audio Speech Lang Process 15(2):465–477
6. Hansen JHL, Zhang X, Akbacak M, Yapanel U, Pellom B, Ward W (2003) CU-Move: advances in in-vehicle speech systems for route navigation. In: IEEE workshop in DSP in mobile and vehicular systems, paper 6.5, Nagoya, 4–5 April 2003, pp 1–6
7. Hansen JHL (2002) Getting started with the CU-Move corpus. CU-Move documentation