CAVEMOVE

Introduction

CAVEMOVE is a research project dedicated to the collection of audio data for the study of voice enabled technologies inside moving vehicles. The recording process involves (i) recordings of acoustic impulse responses, which are acquired at static conditions and provide the means for modeling the speech and car-audio components and (ii) recordings of acoustic noise at a wide range of both static and in-motion conditions. Data are recorded with two different microphone configurations and particularly (i) a compact circular microphone array or (ii) a distributed microphone setup.

This document provides a description of the audio recordings and acoustic impulse responses that comprise the open access dataset of CAVEMOVE. The open access dataset is available at 16 kHz sampling rate and 24 bits bit depth and is approximately 4 GB in size. It can be downloaded from zenodo through https://zenodo.org/records/13594243
Note that the open access dataset is only a subset of the full CAVEMOVE dataset.

It is recommended to use this dataset along with the corresponding python API that can be downloaded from CAVEMOVE github. It is important to note that the CAVEMOVE API relies on specific naming conventions for obtaining the audio mixtures corresponding to specific scenarios, and this document lists all the naming conventions required for correct use of the API.

For any questions with respect to CAVEMOVE dataset or API, feel free to send an email to
Andreas Symiakakis at andrysmi@ics.forth.gr
or
Nikos Stefanakis at nstefana@ics.forth.gr

CAVEMOVE project is funded by the Institute of Computer Science of the Foundation for Research and Technology-Hellas (FORTH).

Data Acquisition Process

Microphone setups

All recordings are acquired using M-audio M-Track Eight usb audio interface, at 48kHz sampling rate and at 24-bit sample size. Eight omnidirectional microphones (Shure SM93) placed at different locations are used for recording acoustic noise and IRs. These are lightweight lavalier microphones with a flat frequency response from 80 to 20000 Hz. Two different microphone configurations have been used until now, as described below.

1. Microphone array: the eight microphones are mounted on a plastic circular case, forming a circular microphone array of radius equal to 5 cm. The circular array is placed on top of the dashboard, between the driver and the front passenger (see Picture 3 and Picture 11). Microphones were mounted on the array without a windshield.
2. Distributed setup: While there are slight variations in the setup between different cars, the microphones are distributed in a way so that six of the microphones are at the front part of the car, while the other two are placed on the headliner at the back of the car (see Picture 1 and Picture 5). This setup ensures that there is at least one microphone close to each passenger. All microphones were stabilized with blue tack and were covered with a windshield.

Measurement of IRs

Acoustic Impulse Responses (IRs) are measured for different passenger locations in each car. Basic locations covered include that of the driver, the front passenger, the rear-left, rear-middle and rear-right passenger, assuming that the passengers are at normal height and that they are facing in-front. In several cases, IRs are also measured for slight displacements, e.g. by assuming that the passenger’s head is shifted a few centimeters from the nominal location, or assuming that the head is rotated.

All IRs are captured using Room Eq Wizard (https://www.roomeqwizard.com/) with a sweep tone as the excitation signal. The loudspeaker used for these measurements is the Talkbox from NTi which is especially designed for resembling the frequency response and radiation characteristics of the human voice. The loudspeaker is calibrated and it has built-in excitation signals which allow us to excite the car interior with acoustic powers that correspond to specific speech effort, particularly normal speech effort (referenced as 60 dBA at 1 m) and high speech effort (referenced as 70 dBA at 1 m). This process is very important as it will allow us to calculate the signal levels that correspond to specific speech efforts, consequently allowing us to scale the speech components so that when mixed with the noise components, a realistic balance is achieved. Additionally to the acoustic paths from the passengers to the microphones, IRs are also measured for the built-in audio system, whenever available. This process involves recording the response from all car loudspeakers simultaneously. It was accomplished by directly feeding the excitation signal required for IR estimation into the auxiliary input of the audio system, or by reproducing it through the CD player. Attention was paid so that any adjustments related to equalization settings or panning were neutralized.

Recorded conditions

As basic factors that affect the composition of noise in a car we considered the following: car speed, window aperture and ventilation/air condition level. With respect to these factors, it has been attempted to have them at fixed values for continuous temporal segments so as to achieve stationary noise conditions. A basic set of so-called "controlled" conditions is recorded in each car, spanning a specific range of driving speeds and window apertures. Particularly regarding window aperture, we consider four states designated as 0, 1, 2 and 3. Regarding the driving speeds, in each car we cover at least the range from 50 km/h to 110 km/h with steps of 10 km/h. All combinations of window apertures and driving speeds are covered in each case, excluding those that it was not possible. Additionally to these in-motion recordings, static recordings were obtained for capturing the noise produced by the built-in ventilation or air-conditioning system. Two or three different levels of ventilation power were considered in each car and microphone setup.

Synthesis Model

The ultimate goal of the data and python API delivered in the context of CAVEMOVE is to allow the engineers and researchers to easily synthesize the microphone signals that correspond to a particular scenario. Assuming that a target car and microphone setup has been chosen, the synthesis process can be compactly described as:

$$\mathbf{Y} = \mathbf{S}(p,\ Ls,\ w,\ \mathbf{x}) + \mathbf{A}(La,\ w,\ \mathbf{z}) + \mathbf{N}(s,\ w) + \mathbf{V}(l,\ w) \tag{1}$$

Briefly, each bold capital symbol is a N×M signal matrix, where M is the desired number of microphone channels and N is the duration of the synthesized sound excerpt in samples. $\mathbf{S}$ represents the filtered speech components, $\mathbf{A}$ represents the interference components produced by the built-in audio system (when available), $\mathbf{N}$ are the noise components corresponding to the particular driving condition and $\mathbf{V}$ are the noise components associated to the ventilation/air-condition functionality. Assuming that all these sound components are independent from one another, the final microphone signal can be synthesized as in Eq. (1), by means of simple superposition. As it can be seen, each component is produced as a function of user defined parameters. A brief explanation of these parameters is as follows;

- w: is an integer taking values in the range [0, 1, 2, 3], with each integer value representing a different condition with respect to the windows' apertures, as described in more detail below.
- $\mathbf{x}$: is a one-dimensional vector with the dry (ideally anechoic) speech recording that is used for the particular scenario. The user is responsible for providing an appropriate speech recording in PCM format. The audio recording can be of any length and sampling rate (it will be automatically converted to the target sampling rate).
- p: it can be selected from a given set of string values, each value corresponding to a particular location of the passenger inside the car cabin. For each combination of *p* and

w the corresponding impulse response is automatically loaded and used to produce the filtered speech components by means of convolution with $\mathbf{x}$.

- Ls: it is a user defined value, in dBA, corresponding to the speech effort. While this can be any non-negative value, we recommend values in the range between 60 and 70, with 60 corresponding to normal speech effort and 70 to a high speech effort.
- La: is a user defined value, in dBA, corresponding to the mean A-weighted acoustic level of the audio program reproduced from the built-in audio system. The corresponding interference components are automatically scaled through this value so that the signal level at a reference microphone matches the desired sound level. Again, any non-negative value is applicable and it is up to the user to set it to a realistic value.
- $\mathbf{z}$: is a one-dimensional vector representing the audio program. The user is responsible for providing an appropriate audio file in PCM format. The audio file can be of any length and sampling rate. Audio files with more than one channels are accepted but will be automatically downmixed to a monophonic audio signal.
- s: is an integer value within a given range, where each value corresponds to a different speed in km/h. The noise components corresponding to particular speed, window aperture, car and microphone setup are automatically loaded.
- l: it is an integer with values in the range [1, 2, 3], each value corresponding to a different level of the ventilation/air-conditioning system, so that higher levels produce more noise.

The synthesis approach is designed in such way that the resulting length N of the output signal **Y**, as well as of all matrices in Eq. (1) matches the length of the dry speech signal **x** when converted to the target sampling rate. To achieve this, we have incorporated a mechanism that recycles the noise sequences inherent to the construction of **A**, **N** and **V** as many times necessary in order to match the length of **x**. Note that in most cases this will not be necessary if the dry speech signal does not exceed 25 s of duration.

Additional parameters and functions

An additional parameter that affects the returned microphone signals is the use of correction gains (or not), which will apply a scaling to the produced signals so as to compensate for slight deviations in the input channel sensitivities. The sensitivity values were measured with each microphone connected to a specific input channel in the sound card, subjecting it to a pink noise sound field produced inside a semi-anechoic chamber that we have in our facilities. Although the differences between minimum and maximum sensitivity do not exceed 3.5 dB, we recommend use of correction gains, which is also the default option in all relevant function calls.

Our API provides functions that are specific to each one of the aforementioned components (i.e. speech, noise, ventilation and built-in audio system) but there is also a function with which the user can synthesize all four different components with a single function call. Additional auxiliary functions are provided for example:

- to automatically crop or extend the duration of different components, so that they can be all added together to produce a mix,
- to derive the IRs associated to a particular passenger location and microphone setup,

- to derive the steering vector associated to a particular frequency and look direction with respect to the center of the circular array, which is handy for designing beamformers in the frequency domain.

Naming conventions

Speaker Locations

Acoustic Impulse Responses are provided for at least the following locations inside the car:
d: driver
pf: front passenger
prl: rear-left passenger
prr: rear-right passenger
prm: rear-middle passenger

Windows Conditions

Regarding window aperture, we consider four states designated as 0, 1, 2 and 3:
w0: corresponds to completely closed windows.
w1: corresponds to windows of the driver and front passenger being slightly open (approx.10 cm).
w2: corresponds to completely open front windows.
w3: corresponds to all four windows being completely open.

Speed Condition

Regarding the driving speeds, in each car we cover at least the range from 50 km/h to 110 km/h with steps of 10 km/h. Speed condition is symbolized by the letter *s* followed by the speed e.g. **s70** or **s30**.

Different Versions

Some noise recordings come with more than one versions. To discriminate between different versions we append the file name with “*_ver1”,* “*_ver2*” etc. Also, some versions are obtained on coarse road surface, in which case we append the name with “*_coarse*”.

Ventilation Level

Two or three different levels of ventilation power were considered in each car:
v1
v2
v3

Car Recordings

The open access dataset includes recordings from three different cars:

1. Volkswagen Golf (2011)
2. Alfa Romeo, 146 TS (2000)
3. Smart, forfour (2019)

1. Volkswagen Golf

1.1 Distributed microphones:

The locations and numbering of the distributed microphones inside the car's cabin can be seen in Picture 1.

(a) (b)

Picture 1: Locations and numbering of the distributed microphones inside the cabin. Six microphones are located in the front (a) and two in the back (b).

1.1.1 Impulse Response Recordings

1. d50_w0[1]
2. d50_w1
3. d50_w2
4. d50_w3
5. d60_w0[2]
6. d60_w1
7. d60_w2
8. d60_w3
13. prl_w0
14. prl_w1
15. prl_w2
16. prl_w3
17. prm_w0
18. prm_w1
19. prm_w2
20. prm_w3

---

[1] d50: Driver position. 50cm distance from wheel (Picture 2).
[2] d60: Driver position. 60cm distance from wheel.

9. pf_w0
10. pf_w1
11. pf_w2
12. pf_w3
21. prr_w0
22. prr_w1
23. prr_w2
24. prr_w3

Picture 2: d50

1.1.2 Noise Recordings

1. s30_w0_coarse[3]
2. s30_w1_coarse
3. s30_w2_coarse
4. s30_w3_coarse
5. s100_w0
6. s100_w1
7. s100_w2_ver1
8. s100_w2_ver2
9. s100_w2_ver3
10. s100_w3_ver1
11. s100_w3_ver2
12. s110_w0
13. s110_w1_ver1
14. s110_w1_ver2
15. s110_w2_ver1
16. s110_w2_ver2
17. s110_w2_ver3
18. s110_w3_ver1
19. s110_w3_ver2
20. s120_w0
21. s120_w2
22. s40_w0_ver1
23. s40_w0_ver2
24. s40_w1_ver1
25. s40_w1_ver2
26. s40_w2
27. s40_w3
28. s50_w0_ver1
29. s50_w0_ver2
30. s50_w0_ver3
31. s50_w1
32. s50_w2_ver1
33. s50_w2_ver2
34. s50_w2_ver3
35. s50_w2_ver4
36. s50_w2_ver5
37. s50_w2_ver6
38. s50_w3_ver1
39. s50_w3_ver2
40. s50_w3_ver3

[3] coarse: Coarse road surface.

41. s60_w0
42. s60_w1_ver1
43. s60_w1_ver2
44. s60_w1_ver3
45. s60_w1_ver4
46. s60_w2_ver1
47. s60_w2_ver2
48. s60_w2_ver3
49. s60_w2_ver4
50. s60_w3_ver1
51. s60_w3_ver2
52. s70_w0_ver1
53. s70_w0_ver2
54. s70_w1_ver1
55. s70_w1_ver2
56. s70_w1_ver3
57. s70_w1_ver4
58. s70_w2_ver1
59. s70_w2_ver2
60. s70_w3
61. s80_w0_ver1
62. s80_w0_ver2
63. s80_w1_ver1
64. s80_w1_ver2
65. s80_w1_ver3
66. s80_w2_ver1
67. s80_w2_ver2
68. s80_w3_ver1
69. s80_w3_ver2
70. s80_w3_ver3
71. s90_w0
72. s90_w1_ver1
73. s90_w1_ver2
74. s90_w2
75. s90_w3

1.1.3 Ventilation recordings

1. v1_w0
2. v1_w1
3. v1_w2
4. v1_w3
5. v2_w0
6. v2_w1
7. v2_w2
8. v2_w3
9. v3_w0
10. v3_w1
11. v3_w2
12. v3_w3

1.1.4 Car audio

Impulse responses from the car audio were obtained for all four windows apertures in Volkswagen/distributed.

1. w0
2. w1
3. w2
4. w3

1.2 Microphone array:

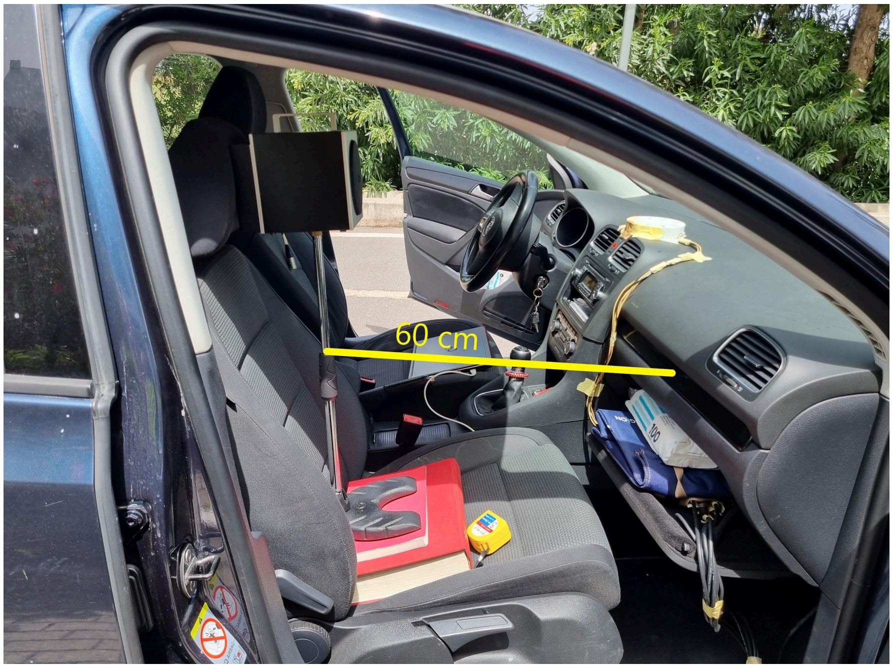
The location of the microphone array inside the car and numbering of the microphones can be seen in Picture 3.

Picture 3: Position of the microphone array inside the car's cabin and numbering of the microphones.

1.2.1 Impulse Responses

1. d50_w0[4]
2. d50_w1
3. d50_w2
4. d50_w3
5. pf60_w0[5]
6. pf60_w1
7. pf60_w2
8. pf60_w3
9. pf80_w0[6]
10. pf80_w1
11. pf80_w2
12. pf80_w3
13. prl_w0
14. prl_w1
15. prl_w2
16. prl_w3
17. prm_w0
18. prm_w1
19. prm_w2
20. prm_w3
21. prr_w0
22. prr_w1
23. prr_w2
24. prr_w3

[4] d50: Driver position. 50cm distance from wheel.
[5] pf60: Front passenger position. 60cm distance from dashboard (Picture 4).
[6] pf80: Front passenger position. 80cm distance from dashboard.



Picture 4: pf60

1.2.2 Noise Recordings

1. s30_w0_coarse[7]
2. s30_w1_coarse
3. s30_w2_coarse
4. s30_w3_coarse
5. s0_w0[8]
6. s0_w1
7. s0_w2
8. s0_w3
9. s100_w0_ver1
10. s100_w0_ver2
11. s100_w0_ver3
12. s100_w0_ver4
13. s100_w1_ver1
14. s100_w1_ver2
15. s100_w2_ver1
16. s100_w2_ver2
17. s100_w2_ver3
18. s100_w3_ver1
19. s100_w3_ver2
20. s110_w0
21. s110_w1_ver1
22. s110_w1_ver2
23. s110_w2_ver1
24. s110_w2_ver2
25. s110_w3
26. s120_w0
27. s120_w1
28. s120_w2
29. s40_w0
30. s40_w1_ver1
31. s40_w1_ver2
32. s40_w2
33. s40_w3
34. s50_w0
35. s50_w1_ver1
36. s50_w1_ver2
37. s50_w1_ver3
38. s50_w2

[7] coarse: Coarse road surface.
[8] s0: Idle conditions, stopped car.

39. s50_w3_ver1
40. s50_w3_ver2
41. s60_w0
42. s60_w1_ver1
43. s60_w1_ver2
44. s60_w2
45. s60_w3
46. s70_w0_ver1
47. s70_w0_ver2
48. s70_w0_ver3
49. s70_w1_ver1
50. s70_w1_ver2
51. s70_w2_ver1
52. s70_w2_ver2
53. s70_w3_ver1
54. s70_w3_ver2
55. s80_w0
56. s80_w1
57. s80_w2_ver1
58. s80_w2_ver2
59. s80_w2_ver3
60. s80_w2_ver4
61. s80_w2_ver5
62. s80_w3_ver1
63. s80_w3_ver2
64. s80_w3_ver3
65. s90_w0_ver1
66. s90_w0_ver2
67. s90_w1_ver1
68. s90_w1_ver2
69. s90_w2_ver1
70. s90_w2_ver2
71. s90_w2_ver3
72. s90_w3_ver1
73. s90_w3_ver2

1.2.3 Ventilation Recordings

1. v1_w0
2. v1_w1
3. v1_w2
4. v1_w3
5. v2_w0
6. v2_w1
7. v2_w2
8. v2_w3
9. v3_w0
10. v3_w1
11. v3_w2
12. v3_w3

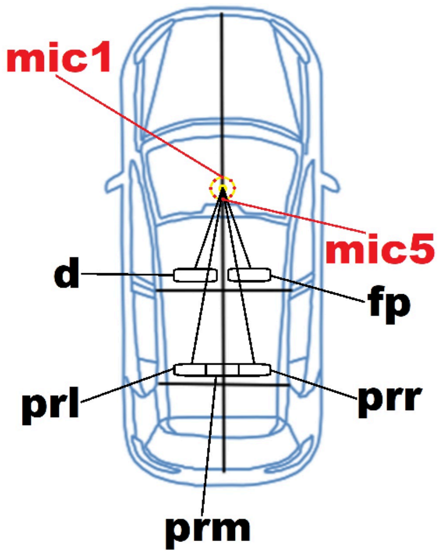
1.2.4 Car audio

Impulse responses from the car audio were obtained for all four windows apertures.

1. w0
2. w1
3. w2
4. w3

1.2.5 Steering vector

Angles for the different passenger locations required for constructing the steering vector are provided in the following picture



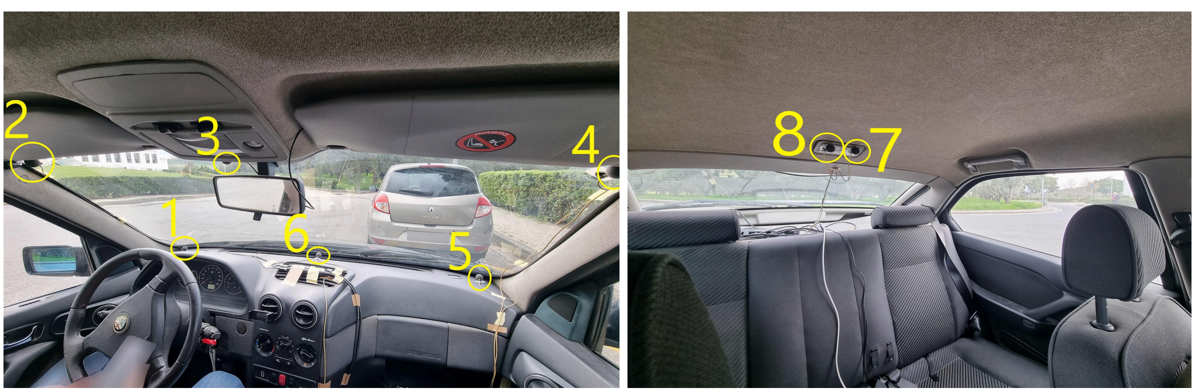
VW Golf	
location	angle (degrees)
d	-28
fp	26
prl	-12
prm	0
prr	12

2. AlfaRomeo 146

Recordings with microphone array are not provided.

2.1 Distributed microphones:

The locations and numbering of the distributed microphones inside the car can be seen in Picture 5.



(a) (b)

Picture 5: Positions and numbering of distributed microphones inside the cabin.

2.1.1 Impulse Response Recordings

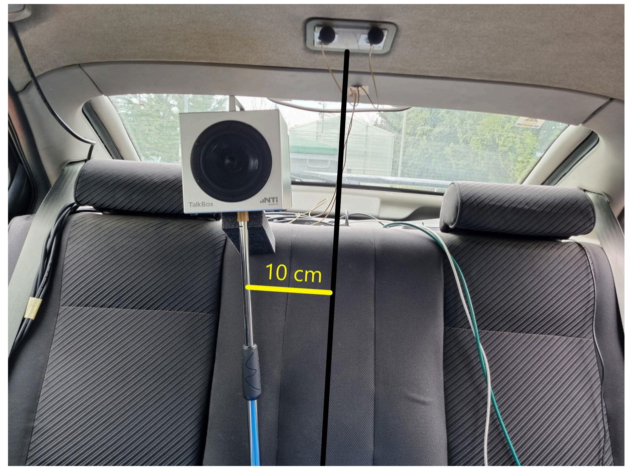
1. d45_w0[9]
2. d45_w1
3. d45_w2
4. d55_w0[10]
5. d55_w1
6. d55_w2
7. pf_w0
8. pf_w1
9. pf_w2
10. pfr90_w0[11]
11. pfr90_w1
12. pfr90_w2
13. prl66_w0[12]
14. prl66_w1
15. prl66_w2
16. prl_w0
17. Prl_w1
18. prl_w2
19. prm10l_w0[13]
20. prm10l_w1
21. prm10l_w2
22. prm10r_w0[14]
23. prm10r_w1
24. prm10r_w2
25. prm_w0
26. prm_w1
27. prm_w2
28. prr67_w0[15]
29. prr67_w1
30. prr67_w2
31. prr_w0
32. prr_w1
33. prr_w2

Picture 6: pfr90

Picture 7: prl66

[9] d45: Driver position. 45cm distance from wheel.
[10] d55: Driver position. 55cm distance from wheel.
[11] pfr90: Front passenger position.Rotated 90 degrees to the left towards the driver (Picture 6).
[12] prl66: Rear left passenger. Height: 66cm from seat base (Picture 7).
[13] prm10l: Rear middle passenger. Moved towards the left (Picture 8).
[14] prm10r: Rear middle passenger. Moved towards the right (Picture 9).
[15] prr67: Rear right passenger. Height: 67cm from seat base(Picture 10).

Picture 8: prm10l



Picture 9: prm10r

Picture 10: prr67

2.1.2 Noise Recordings

1. s30_w0_coarse[16]
2. s30_w1_coarse
3. s30_w2_coarse
4. s0_w0[17]
5. s0_w1
6. s0_w2
7. s100_w0_ver1
8. s100_w0_ver2
9. s100_w1_ver1
10. s100_w1_ver2
11. s100_w2
20. s50_w1_ver2
21. s50_w2
22. s60_w0
23. s60_w1
24. s60_w2
25. s70_w0
26. s70_w1
27. s70_w2_ver1
28. s70_w2_ver2
29. s70_w2_ver3
30. s80_w0_ver1

[16] coarse: Coarse road surface.
[17] s0: Idle conditions, stopped car.

12. s110_w0
13. s110_w1
14. s110_w2
15. s40_w0
16. s50_w0_ver1
17. s50_w0_ver2
18. s50_w0_ver3
19. s50_w1_ver1
31. s80_w0_ver2
32. s80_w1
33. s80_w2
34. s83_w2
35. s90_w0_ver1
36. s90_w0_ver2
37. s90_w1
38. s90_w2_ver1
39. s90_w2_ver2

2.1.3 Ventilation Recordings

1. v1_w0
2. v1_w1
3. v1_w2
4. v2_w0
5. v2_w1
6. v2_w2

2.1.4 Car audio

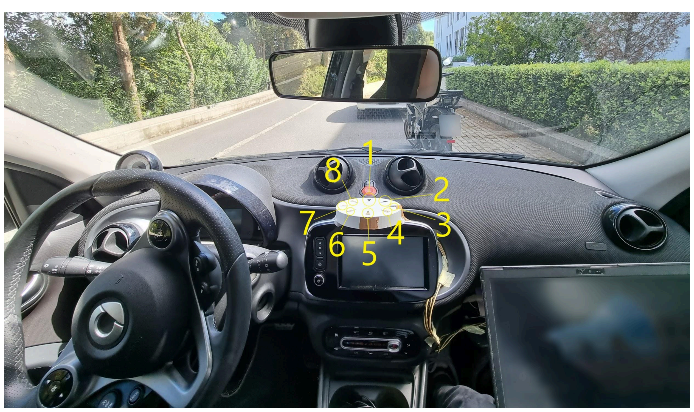
Impulse responses from the car audio were not obtained for Alfa Romeo.

3. Smart forfour

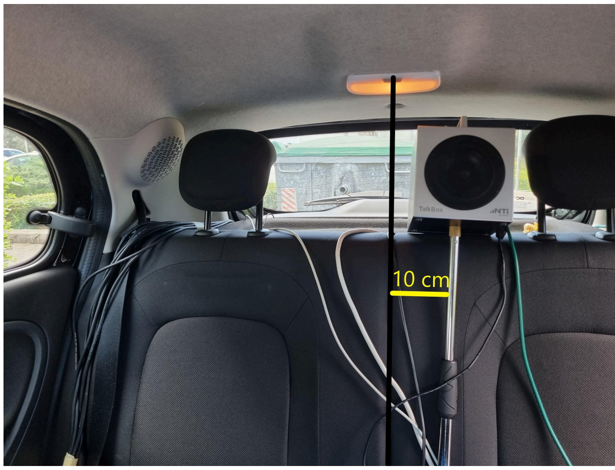
Recordings with distributed microphones are not provided.

3.1 Microphone array:

The locations of the microphone array inside the car and numbering of the microphones can be seen in Picture 11.



Picture 11: Location of the microphone array inside the car and numbering of the microphones.



Picture 12: prm10l

Picture 13: prm10r

3.1.1 Impulse Response Recordings

1. d_w0
2. d_w1
3. d_w2
4. pf_w0
5. pf_w1
6. pf_w2
7. pfr_w0
8. pfr_w1
9. pfr_w2
10. prl_w0
11. prl_w1
12. prl_w2
13. prm10l_w0[18]
14. prm10l_w1
15. prm10l_w2
16. prm10r_w0[19]
17. prm10r_w1
18. prm10r_w2
19. prm_w0
20. prm_w1
21. prm_w2
22. prr_w0
23. prr_w1
24. prr_w2

3.1.2 Noise Recordings

1. s30_w0_coarse[20]
2. s30_w1_coarse

23. s50_w2_ver2
24. s60_w0

[18] prm10l: Rear middle passenger. Moved towards the left (Picture 12).
[19] prm10r: Rear middle passenger. Moved towards the right (Picture 13).
[20] coarse: Coarse road surface.

3. s30_w2_coarse
4. s0_w0[21]
5. s0_w1
6. s0_w2
7. s100_w0
8. s100_w1_ver1
9. s100_w1_ver2
10. s100_w2
11. s110_w0
12. s110_w1
13. s110_w2
14. s120_w1
15. s40_w0
16. s40_w1
17. s40_w2
18. s50_w0_ver1
19. s50_w0_ver2
20. s50_w0_ver3
21. s50_w1
22. s50_w2_ver1

25. s60_w1_ver1
26. s60_w1_ver2
27. s60_w2
28. s70_w0_ver1
29. s70_w0_ver2
30. s70_w1_ver1
31. s70_w1_ver2
32. s70_w2
33. s80_w0_ver1
34. s80_w0_ver2
35. s80_w1
36. s80_w2_ver1
37. s80_w2_ver2
38. s90_w0_ver1
39. s90_w0_ver2
40. s90_w1_ver1
41. s90_w1_ver2
42. s90_w2_ver1
43. s90_w2_ver2

3.1.3 Ventilation Recordings

1. v1_w0
2. v1_w1
3. v1_w2
4. v2_w0
5. v2_w1
6. v2_w2
7. v3_w0
8. v3_w1
9. v3_w2

3.1.4 Car audio

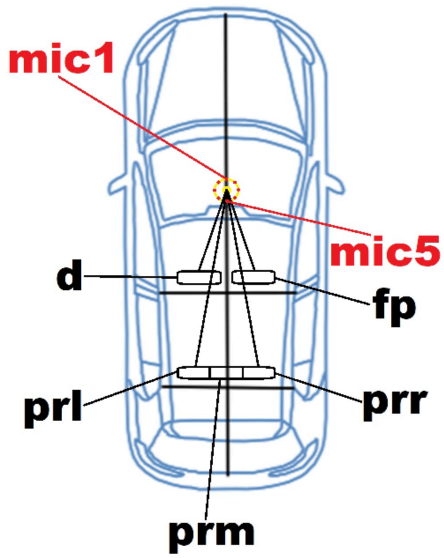
Impulse responses from the car audio were obtained for all three window apertures, state 0, 1 and 2.

1. w0
2. w1
3. w2

[21] s0: Idle conditions, stopped car.

3.1.5 Steering vector

Angles for the different passenger locations required for constructing the steering vector are provided in the following picture



smart forfour	
location	angle (degrees)
d	-25
fp	25
prl	-13
prm	0
prr	13
prm10l	-4
prm10r	4

Additional data

As already mentioned, the open-access dataset is only a subset of full CAVEMOVE dataset that is available at 48kHz sampling rate and that currently includes:

- recordings within additional cars,
- recordings with a hybrid microphone setup comprised of a four-channel microphone array and four distributed microphones
- impulse responses measured for sound sources located outside the car.

Moreover, each recording process results in approximately 2 hours of raw 8-channel audio data. This data is annotated in Praat, in an attempt to mark down various common types of noises that occur while driving. We will be happy to negotiate access to this data with interested parties and we will also be happy to hear suggestions for improving our API or for additional cases worth recording. The CAVEMOVE dataset will continue to grow until March of 2025.