Multi-Channel Multi-Modal Speech Corpus for In -Car Communication Research

Kazuya Takeda

Graduate School of Information Science, Nagoya University

takeda@is.nagoya-u.ac.jp

Providing a human-machine interface in a car is one of the most important applications of speech signal processing, where the conventional input/output methods are unsafe and inconvenient. To develop an advanced in-car speech interface, however, not only one but many real-world problems, such as noise robustness, distortion due to distant talking and disfluency while driving, must be overcome.

In particular, the difficulty of in-car speech processing is characterized by its variety. The road and traffic conditions, the car condition and the driving movement of the driver change continuously and affect the driver's speech. Therefore, a large corpus is indispensable in the study of in-car speech, not only for training acoustic models under various background noise conditions, but also to build a new model of the combined distortions of speech.

In order to keep pace with the ever-changing environment, it may be helpful to make use of various observed signals rather than to use the speech input signal alone. Therefore, to develop advanced speech processing for in-car application, we need a corpus 1) that covers a large variety of driving condition, and 2) from which we can extract the conditions surrounding the driver. Constructing such an advanced in-car speech corpus is the goal of this project.

For the data collection, a specially built data collection vehicle (DCV) has been used for synchronous recording of seven-channel audio signals, three-channel video signals and vehicle-related signals. About 1 terabyte of data has been collected by recording the spontaneous utterances of the driver in about 60 minutes of driving for each of 800 drivers.

The drivers' utterances are recorded through dialogues between three different information systems, i.e., human operator, Wizard-of-OZ system and Automatic Speech Recognizer (ASR). The task domain of the dialogues is the restaurant guidance around the Nagoya University campus. In dialogues with a human operator and the WOZ system, we have prompted the driver to issue natural and varied utterances related to the task domain by displaying a 'prompt panel'. On the panel, a keyword, such as *fast food*, *bank*, *Japanese food* or *parking*, or a situation sentence, such as 'Today is an anniversary. Let's have a party.', 'I am so hungry. I need to eat!' or 'I am thirsty. I want a drink!', are displayed. In these modes, therefore, the driver takes the initiative in the dialogue. The operator also navigates the driver to a predetermined destination while they are having a dialogue, in order to simulate the common function of a car-navigation system. All responses of the operator are given by synthetic speech in the WOZ mode.

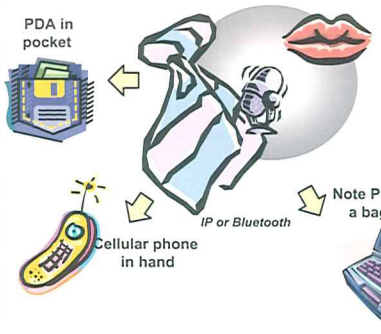
By analyzing the collected data, the dialogues in the different modes are characterized in terms of the linguistic complexity of the used grammar and speech quality. The most clear and important result of the analyses is that the driver talks to an ASR system in a louder voice with less complex sentences so that the system can recognize the utterances. The average loudness of the voice is 2 dB higher than in human navigator and WOZ systems, and the complexity of the sentences is about half of the utterances to a human operator. The regression analysis of the speech recognition accuracy also gives important suggestions on how to balance the SNR improvement and task restriction for improving the speech recognition performance.

Multi-channel multi-modal speech corpus for in-car communication research

Kazuya Takeda
Graduate School of Information Science
Nagoya University

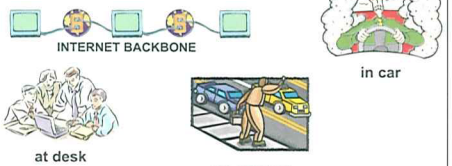
REAL WORLD SPEECH RECOGNITION

The last one feet for the ubiquitous computing



(1) Speech recognition in Everyday Life
- **Noise robustness**
- **Spontaneous speech**
- **Speaker unawareness**

(2) Distant Access to Speech Services
- **IP platforms**
- **Distributed speech recognition**
- **Voice agent**



Overview of the Presentation

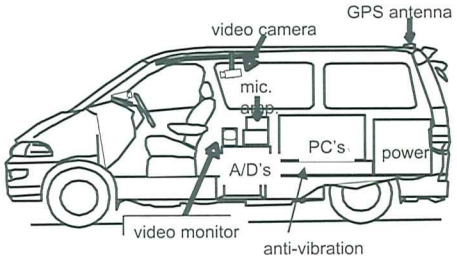
- Data collection activities and status
  - Data collection facilities and measured signals
- Basic analysis on speech communication over driving
- Multiple microphone approach for in-car robust speech recognition
  - A generalized spectral subtraction approach for multiple microphone
- Characterizing in-car speech dialogues in different communication modes
  - Comparative study among human operator, Wizard-of-OZ and ASR system

CIAIR In-car Speech Corpus

- **The importance of in-car speech processing**
  - It is acoustically highly distorted.
  - The speaker's attention to the system is low.
  - It is the only interface available in the car.
- **The features of CIAIR in-car speech corpus**
  - Large amount
    - More than 800 speakers are involved
  - Real driving condition
    - Subjects are driving on a public street while making dialogues
  - Multi-mode dialogues
    - Dialogues with 1) a human navigator, 2) a WOZ system and 3) an ASR system are recorded.
  - Multi-media recordings
    - Recorded data include multi-channel audio, multi-channel video, vehicle-related information (speed, pedals, steering handle etc..), location.

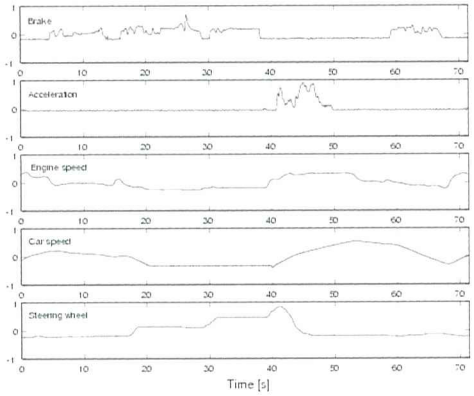
Data Collection Vehicle (DCV)

- **Simultaneous recording**
  - Speech (sound) 16 ch. (16 kHz, 16 bits)
  - Video 3 ch. (Mpeg1)
  - Engine-speed, Car-speed, Break pedal, Steering Handle (1kHz, 16bits)
  - Car location (1Hz, text)



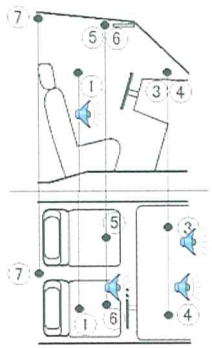
Data Collection Vehicle

Car Driving Information

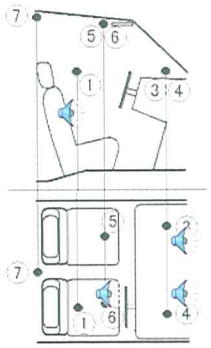


brake, acceleration, engine speed, car speed, steering wheel

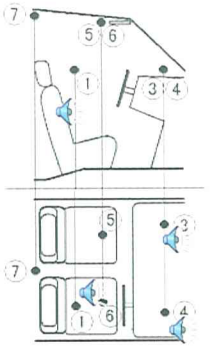
Examples of Recorded Speech through Distant Microphones



Idling



City area

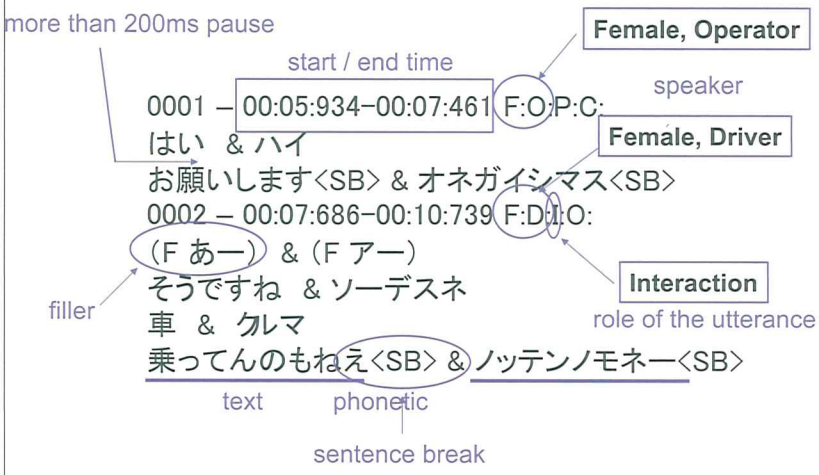


Express Way

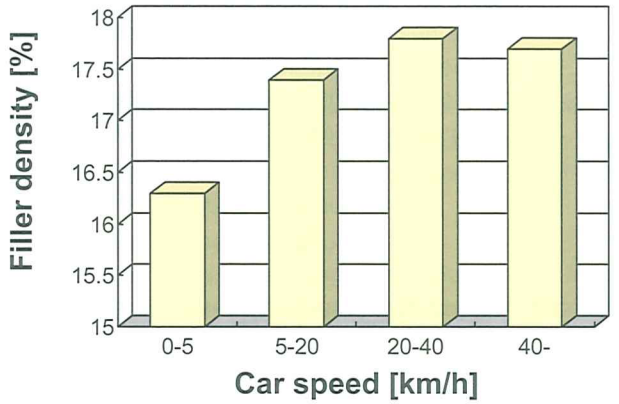
An Example of In-car Dialogue
(with a human operator)

An Example of In-car Dialogue
(with an ASR system)

Transcriptions

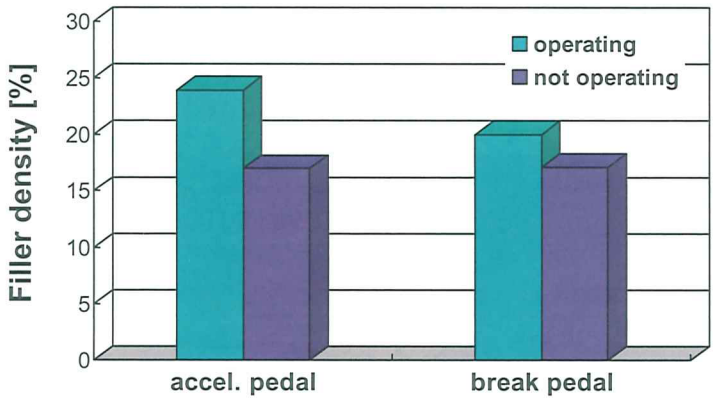


Basic Studies on In-car Dialogues
(How the disfluency affected by the car speed)



Basic Studies on In-car Dialogues

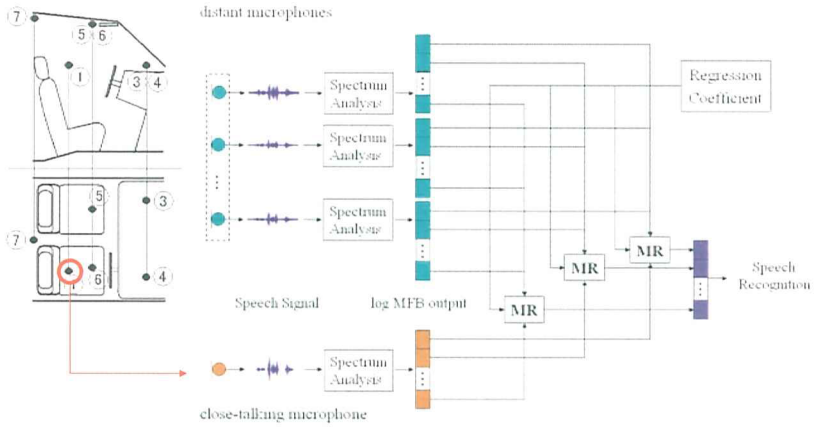
(How the disfluency affected by the driving operation)



MRLS for In-car Speech Recognition

- **Difficulties in applying (adaptive) microphone array technologies in a car**
  - Location of the speaker changes.
  - Noise source is not a point source.
- **Multiple Regression of Log Spectra (MRLS)**
  - Approximate the close-talking microphone speech from distant microphones through multiple regression
  - $\log|X_0(k)| = \Sigma\, w_i \log|X_i(k)|$
  - Minimize regression error for a given set of utterances

MRLS for Distributed Microphones

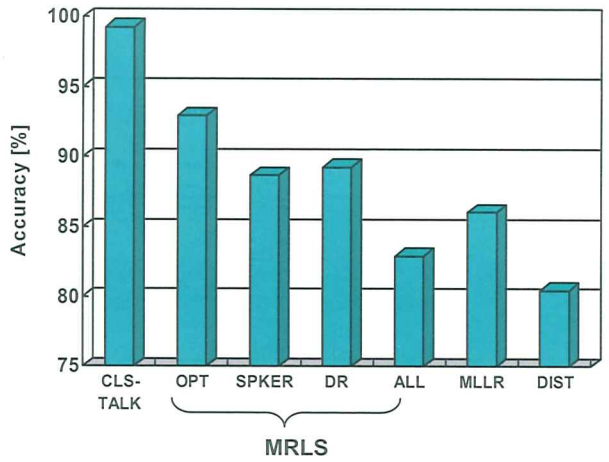


MRLS Evaluation

(experimental setup)

- **Training Set:**
  - 8,000 phonetically balanced sentences uttered while idling (6,000 sentences) and driving (2,000 sentences).
- **Test Set:**
  - 50 isolated word utterances
- **HMM:**
  - triphones share 1,000 states.
  - 16 mixture
- **Feature Parameters:**
  - MFCC, d-MFCC, d-log P
  - 250 – 8000 Hz band
  - 24 ch. Mel filter bank analysis for spectral analysis using 25 ms long frame with 10 ms shift

Performance Comparison (average over 15 different conditions)



Clustering in-car sound environment

- Clustering in-car sound environment using a spectrum feature concatenating distributed microphone signals

$$\mathbf{P}=[\mathbf{R}_{36},\mathbf{R}_{46},\mathbf{R}_{56},\mathbf{R}_{76}] \qquad \mathbf{R}_{i6}=[R_{i6}(4),\cdots,R_{i6}(24)],$$

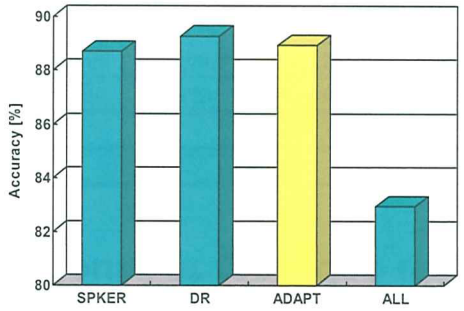
$$R_{i6}(k)=X_i(k)/X_6(k),$$

Clustering Results

	normal	CD	fan lo	fan hi	window open
Class 1	2224	190	329	8	372
Class 2	440	2477	13	4	4
Class 3	25	20	2354	2684	35
Class 4	11	13	5	0	2289

Adapting weights to sound environment

- Vary regression weights in accordance with the classification results.
- Same performance with speaker/condition dependent weights.



Dialogues in Different modes

- **HN (Human Navigator) Human vs. Human**
  - The driver makes questions on the restaurant query to human operator who sits the back seat
- **WOZ (Wizard-of-OZ) Human vs. simulated Machine**
  - The driver makes questions on the restaurant query. The operator operates touch panel and generates synthetic speech for answering.
- **ASR (Automatic Speech Recognition) Human vs. Machine**
  - A spoken dialogue system with ASR function guide Q&A dialogues about restaurant query.

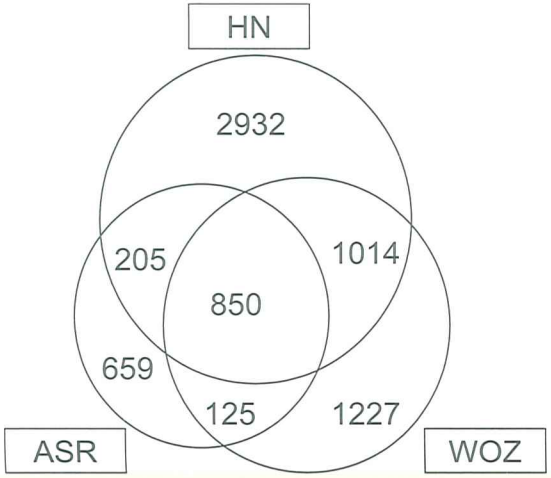
Total amount of recorded utterances

	HN		WOZ		ASR	
# of drivers	435					
duration	101430	28.2h	73116	20.3h	80978	22.5h
of driver	40186	40%	28178	39%	16979	21%
of operator	61244	60%	44938	61%	63999	79%
# of sentences		40560		32883		40149
of driver	17820	44%	13770	42%	15892	40%
average duration [sec]		2.26		2.05		1.07
of operator	22740	56%	19113	58%	24257	60%

The size of the corpus

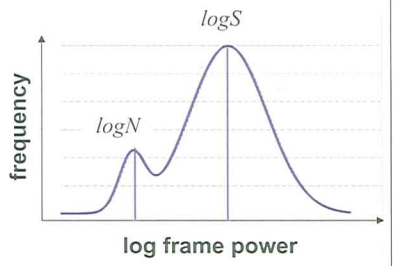
	HN		WOZ		ASR	
# of morph.	353875		195513		262354	
of driver	121919	34%	86330	44%	50374	19%
of operator	231956	66%	109183	56%	211980	81%
morph./sent.]	8.72		5.95		6.53	
of driver	6.84		6.27		3.17	
of operator	10.20		5.71		8.74	
vocabulary	7469		4064		2258	
of driver	5001		3216		1839	
of operator	4367		1524		640	
common morph.	1899	25.4%	676	16.6%	221	9.8%
driver unique	3102	41.5%	2540	62.5%	1618	71.7%
operator unique	2468	33.0%	848	20.9%	419	18.6%

Vocabulary of sessions



SNR calculation

- Background noise varies across utterances and is not easy to estimate from short time segment.
- Fitting a mixture Gaussian distribution to the log-power over an utterance using EM algorithm.
- Estimate SNR from the difference between two means.



$$SNR = 10\log_{10}\frac{S}{N}[dB]$$

Acoustic quality of speech

Dialogue mode	HN		WOZ		ASR	
microphone	CLT	DST	CLT	DST	CLT	DST
SNR [dB]	23.0	10.6	24.0	11.3	26.0	12.9

CLT: close-talking microphone/ DST: distant microphone

- Drivers talk to an ASR system, in louder voice (in 2dB) than to the human operator or Wizard-of-OZ system.

Complexity of utterance

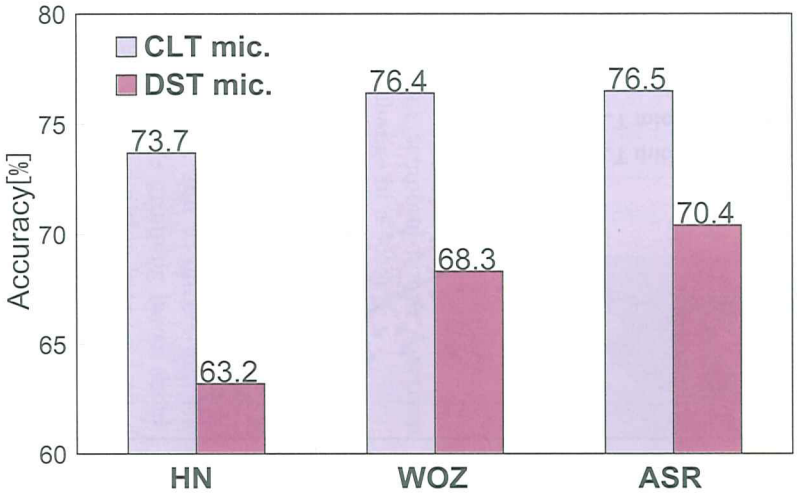
- Perplexities for bigram and trigram language models are evaluated using closed text set.

D. mode	HN	WOZ	ASR
# of drivers	535	586	575
# of sentences	22240	19044	21289
# of morph.	149213	117250	66612
size of vocabulary	5532	3694	2083
# of bigrams	35095	22277	9850
# of trigrams	67972	44322	18403
PP(bigram)	18.1	14.1	9.1
PP(trigram)	7.7	7.1	6.6

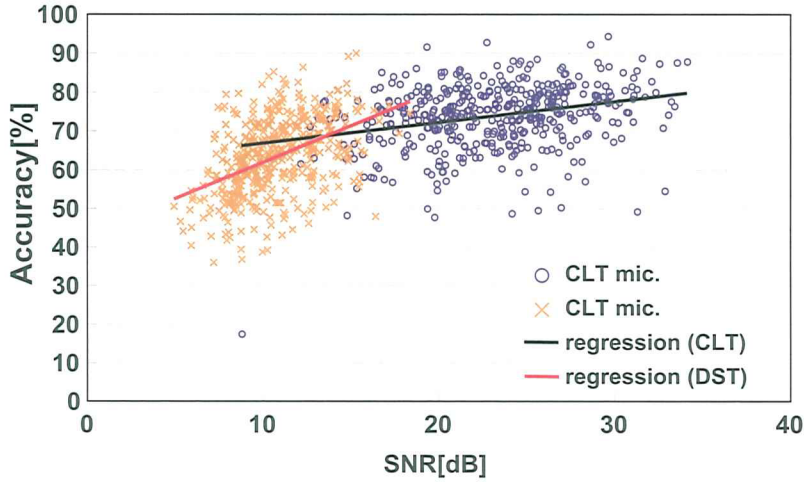
Acoustic modeling

- Two acoustic models for close-talking and distant microphones
- Training corpus
  - 25174 sentences (41.64 hours)
- Frequency band 250—8000Hz
- 2000 state triphone
- 32 mixture
- MFCC+ΔMFCC+ΔlogP
- 44 phonemes (N:)
- Skip topology for short pause model

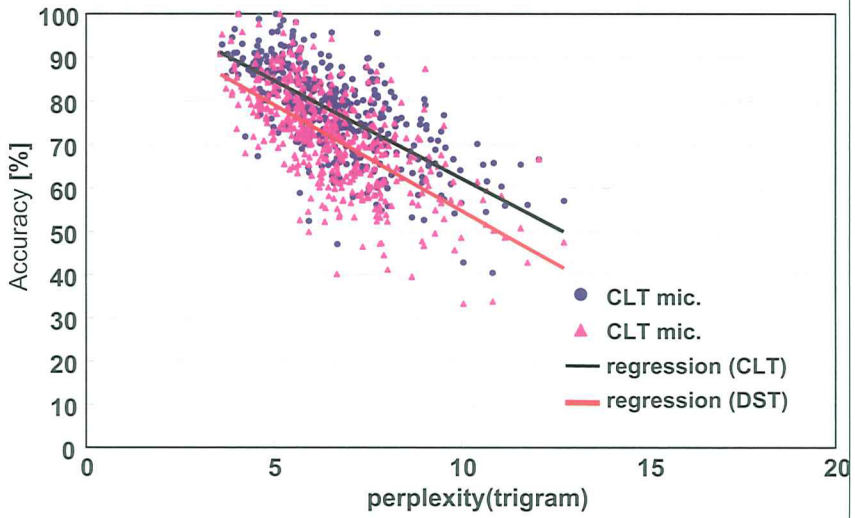
Speech recognition accuracy



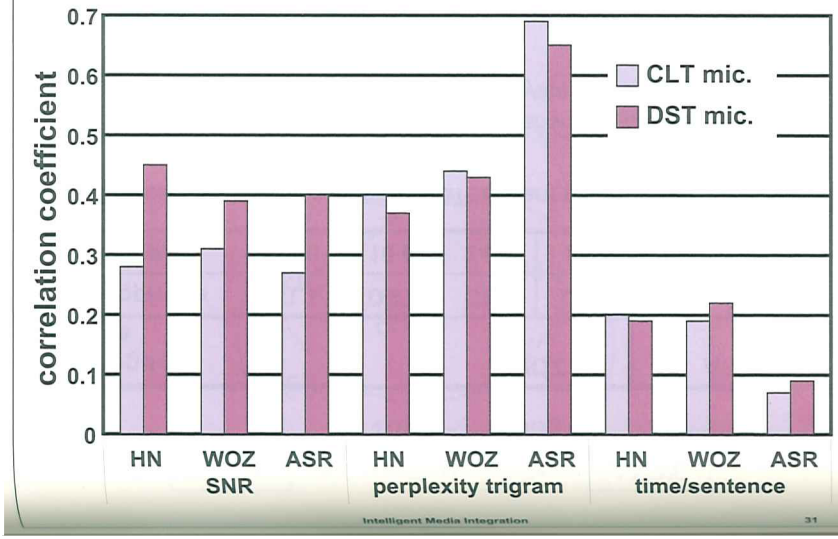
Average accuracy vs. SNR (HN)



Accuracy at different perplexity (ASR)



Correlation coefficients



Summary and future works

- Data collection activities and status
- Multiple microphone approach for in-car robust speech recognition
- Characterizing in-car speech dialogues in different communication modes

- Integration of audio-visual signals in an adverse condition.
- Modeling human behavior while driving
- An integrated framework for understanding in-car human behavior from multiple sensory signals