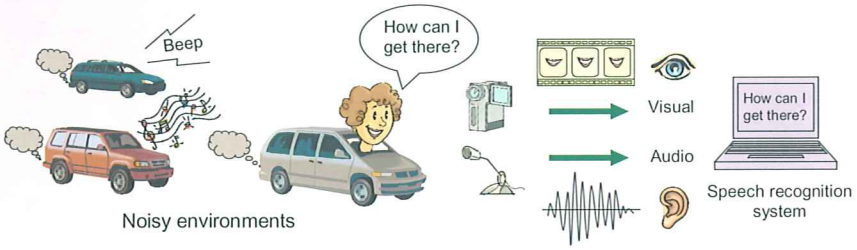
Audio-Visual Speech Database for Bimodal Speech Recognition

C. Miyajima, D. Negi, Y. Ninomiya, M. Sano, K. Mori, K. Itou, K. Takeda, and Y. Suenaga
Dept. of Media Science, Graduate School of Information Science, Nagoya University

Audio-visual speech recognition in noisy conditions

Speech recognition performance can be improved by the use of visual information from lip movements in addition to acoustic speech information.

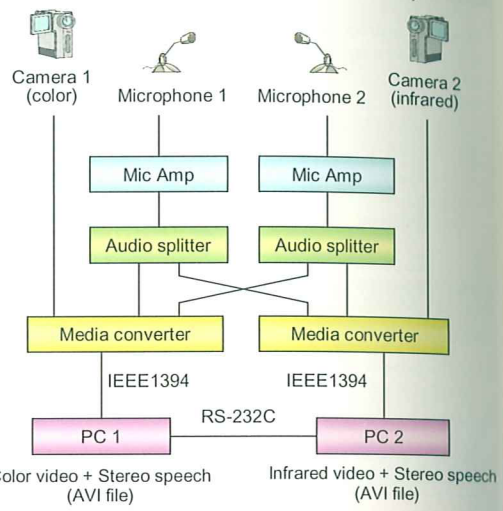


Audio-visual speech data were recorded for evaluating bimodal speech recognition performance. (Recorded in a room: AURORA-2J-AV , in a car: AURORA-3J-AV)

AURORA-project speech databases

AURORA2 (English) AURORA3 (5 languages)	Widely used speech databases for evaluating speech recognition performance (AURORA2: simulated noisy conditions, AURORA3: in-car noisy conditions)
AURORA-2J (Japanese) AURORA-3J (Japanese)	Japanese versions of AUORA2 and AURORA3J
AURORA-2J-AV (Japanese) AURORA-3J-AV (Japanese)	Audio-visual versions of AURORA-2J and AURORA-3J

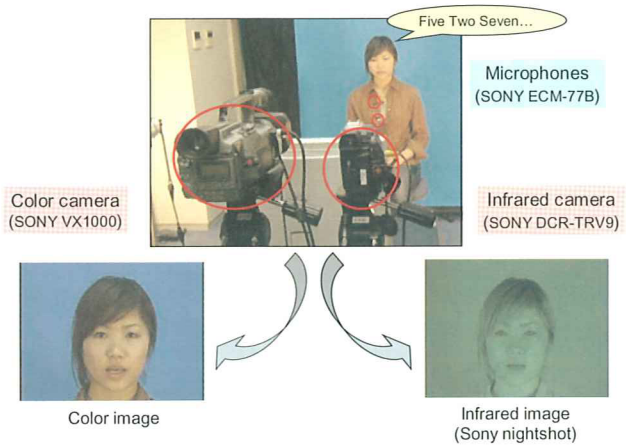
Overview of recording system
(AURORA-2J-AV/ AURORA-3J-AV)



Color and infrared videos and 2-ch speech were recorded.

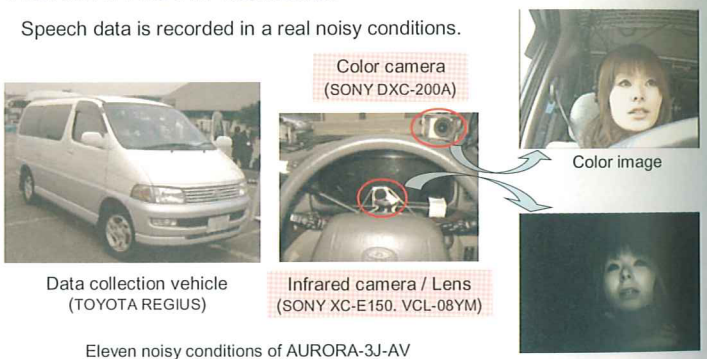
AURORA-2J-AV database

Noisy conditions are simulated by adding noise to clean speech.



AURORA-3J-AV database

Speech data is recorded in a real noisy conditions.



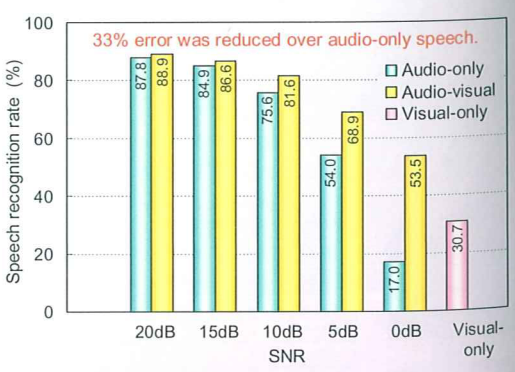
Eleven noisy conditions of AURORA-3J-AV

Recording conditions	Idling	City road	Expressway
Normal	(1)	(2)	(3)
Air-conditioner on	(4)	(5)	(6)
Audio CD player on	(7)	(8)	(9)
Window open	(10)	(11)	-

Audio-visual speech databases available for research use

Database	Audio speech	Video	Language	Number of subjects	Utterances
Tulips1 (Movellean, 1995)	Clean	Only lip region (Grayscale)	English	12	Four digits
DAVID (Chibelushi et al., 1996)	Clean	Plain/complex backgrounds (Color)	English	123	A 10-digit number Alphabet, Sentences
M2VTS (Messer et al., 1998)	Clean	Gray back (Color)	French	37	A 10-digit number
XM2VTS (Messer et al., 1999)	Clean	Blue back (Color)	English	295	Two 10-digit numbers A sentence
M2TINIT (Sako et al., 2001)	Clean	Blue back (Color)	Japanese	1	503 phonetically-balanced sentences
CUAVE (Patterson et al., 2002)	Clean	Green back (Color)	English	36	100 digits 60 numbers
AVOZES (Roland et al., 2004)	Clean	Wall back (Color)	English	20	A 10-digit number Three sentences
AVICAR (Lee et al., 2004)	Noisy (In a car)	Passenger seat (Color)	English	100	Digits, Alphabet Tel-numbers, Sentences
AURORA-2J-AV	Simulated 0-20dB noisy conditions	Blue back (Color, Infrared)	Japanese	97	70 1-7-digit numbers
AURORA-3J-AV	Noisy (In a car)	Driver's seat (Color, Infrared)	Japanese	58	110 1-7-digit numbers

Result of bimodal speech recognition



Experimental condition (AURORA-2J database)

Speech noise: Training data: Subway, babble, car, and exhibition noise (0-20dB)
Test data: Restaurant, street, airport, and station noise (0-20dB)

Image noise: Lighting conditions were changed by controlling gamma-values.