Data Collection and Evaluation of Speech Recognition for Motorbike Riders

H. Tanaka, H. Fujimura, C. Miyajima, T. Nishino, K. Itou, and K. Takeda
Graduate School of Information Science, Nagoya University, Japan

Introduction

There has been growing demand for the ability to use a cell-phone and a navigation system while riding a motorbike.

- An eyes-free & hands-free interface is required for operating information appliances.
- We investigate the feasibility of speech recognition for motorbike riders.

Types of noise

multiplicative noise

Additional noise: Engine noise, Breath noise, Acoustics in helmet, Wind noise, Road noise

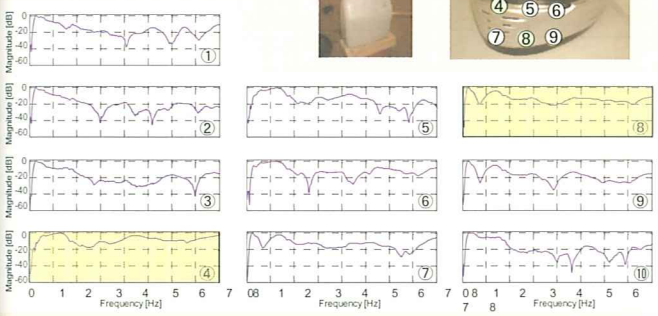
Measuring acoustic transfer function → Construction of database → Investigating SNR of speech data → Exploring the feasibility of speech recognition

Measuring Acoustic Transfer Functions

Measuring the impulse response

ATFs between artificial mouth of a head-and-torso simulator and each microphone measured using Swept Sine signals.

Room-size	7 m × 7 m × 4.5 m
Background noise	12.7 dB
Reverberation time	150 ms

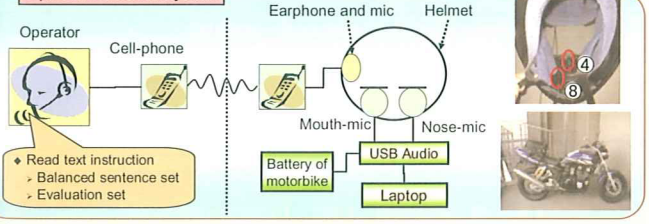


➢ Nose-mic(#4) and Mouth-mic(#8) were chosen as the appropriate microphone positions for riders' speech recognition.
✓ These were flatter and had no sharp dips in the speech band.

Construction of a Motorbike Riders' Speech Database

Collected speech data of motorbike riders

Speech collection system



Recoding equipment

Motorbike (YAMAHA XJR 400)
Helmet (ARAI RAPIDE-OR)
Microphone (RAMSA WM-S2)
USB-Audio (EDIROL UA-5)
Laptop (Panasonic Let's Note R3)
Cell-phone & earphone and mic

Statistics of the Speech Corpus of Motorbike Riders

Riding Condition	City road		Expressway	
	Number of files	Duration(h)	Number of files	Duration(h)
Balanced sentences	35,274	20.4	22,878	12.9
Words	6,152	2.8	5,942	2.6
Total	41,426	23.2	28,820	15.5

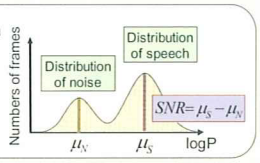
Examples of speech utterances

Device-controlling commands : "cancel"
Information retrieving commands : "I'd like to listen to Billy Joel's 'Just the Way You Are' "

Noise in the Speech Data

Estimating the SNR of an utterance

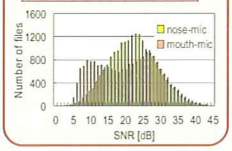
We assume the log-power distribution of the utterance as a two-mixture Gaussian; lower is noise and higher is speech. Then, the SNR of the utterance is calculated as the difference of their averages: $\mu_S - \mu_N$
Gaussians are estimated using an EM (Expectation Maximization) algorithm.



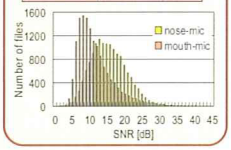
Average SNR [dB]

	City road	Expressway
Nose-mic (#4)	21.8	15.1
Mouth-mic (#8)	20.0	11.0

Distribution of SNR for city road



Distribution of SNR for expressway



➢ Average SNRs of the data captured by the mouth-mic were lower than those of the nose-mic.
✓ The noise was caused by turbulence at the gap between the helmet and the neck.
➢ The range of all the distributions is broad.
✓ Both the power of the riders' voice and the speed of the motorbike varied widely.
➢ There are two peaks in the distribution of the mouth-mic.
✓ The mouth-mic was more strongly influenced by wind noise than the nose-mic.
✓ The lower distribution mainly consists of the data uttered during moving, while the higher one mainly consists of the data uttered during stopping.

Speech recognition on motorbike

Experimental conditions

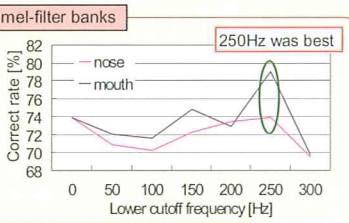
Sampling frequency	16 kHz
Analysis window	Hamming
Frame length	25 ms
Frame shift	10 ms
Number of mel-filter banks	24
Number of cepstral coefficients	12

Feature vector	MFCC (12) ΔMFCC (12) ΔLog-energy
Hidden Markov models (HMM)	Three-state triphone 1,000-stats 32-mixture
Size of vocabulary	1,000

Investigation of frequency range in mel-filter banks

Searching for the most effective cut-off frequency to avoid the effect of noise.

Frequency range	0-8000Hz
Low-frequency cutoffs	0, 50, 100, 150 200, 250, 300



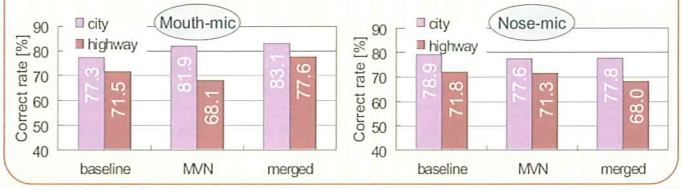
➡ Frequency range of from 250Hz to 8000Hz was used

Acoustic modeling and evaluation

Baseline : models were trained using the matched condition
e.g., the model for the mouth-mic and city roads was trained from the data captured from the mouth-mic under city road condition.

MVN : baseline + mean and variance normalization

Merged : MVN + models were trained from the data of both conditions (city road and expressway)



Conclusion

- We analyzed the acoustics in a helmet, and then confirmed microphone positions.
  - ➢ nose-mic (#4) and mouth-mic (#8)
- Explored the feasibility of speech recognition for motorbike riders.
  - ➢ Correct rate (city road condition 83.1% ; expressway condition 77.6%)