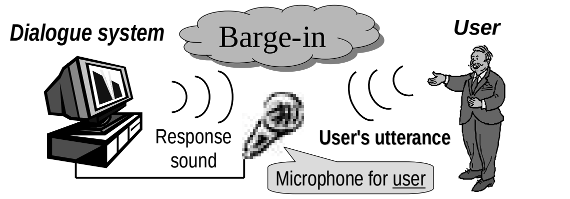
Barge-in Free Spoken Dialogue Interface Based on Sound Field Control and Microphone Array

Speech and Acoustics Lab
D1 Shigeki Miyabe

Overview

- Background
- Conventional metohd
- Proposed method
- Experimental results
  - Response sound elimination experiment
  - Speech recognition experiment
- Conclusion

Background


Dialogue system
Barge-in
User
Response sound
User's utterance
Microphone for user

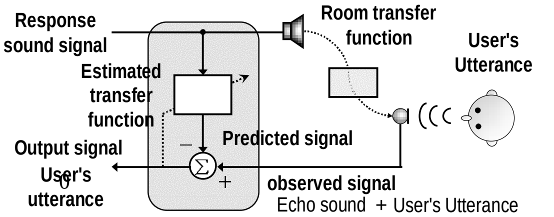
Natural
- Hands-free
- Wearing no Special equipments

Interactive
- User can input anytime even when system outputs response sound

⇨ **Degradation of speech recognition performance**

Goal: **Realize the interface for barge-in free spoken dialogue system**

Conventional Method: Acoustic Echo Canceller

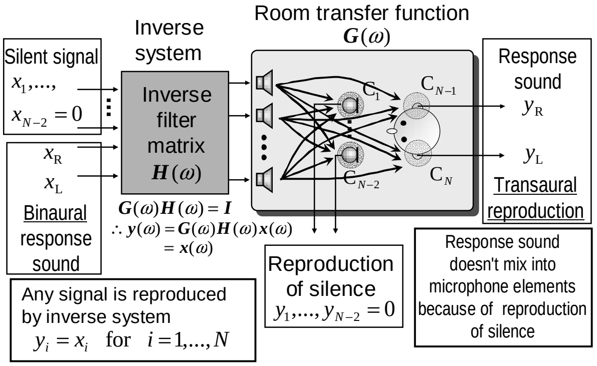

Response sound signal
Room transfer function
User's Utterance
Estimated transfer function
Predicted signal
Output signal
User's utterance
0
$-$
$+$
Σ
observed signal
Echo sound + User's Utterance

Problem

Failure estimation in barge-in situation causes lowing of performance

⇨ **double-talk detection is necessary** | Accurate detection is difficult

Proposed Method: Multiple-Output and Multiple-No-Input (MOMNI) Method


Room transfer function $\boldsymbol{G}(\omega)$

Inverse system

Silent signal $x_1,\ldots,$ $x_{N-2}=0$

x_R x_L Binaural response sound

Inverse filter matrix $\boldsymbol{H}(\omega)$

C_1 C_{N-1} C_{N-2} C_N

Response sound y_R y_L Transaural reproduction

$$\boldsymbol{G}(\omega)\boldsymbol{H}(\omega)=\boldsymbol{I}$$
$$\therefore \boldsymbol{y}(\omega)=\boldsymbol{G}(\omega)\boldsymbol{H}(\omega)\boldsymbol{x}(\omega)=\boldsymbol{x}(\omega)$$

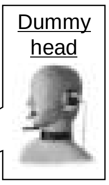
Any signal is reproduced by inverse system
$y_i = x_i$ for $i=1,\ldots,N$

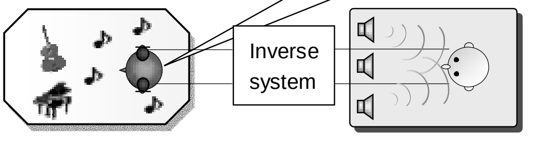
Reproduction of silence
$y_1,\ldots,y_{N-2}=0$

Response sound doesn't mix into microphone elements because of reproduction of silence

Transaural sound reproduction

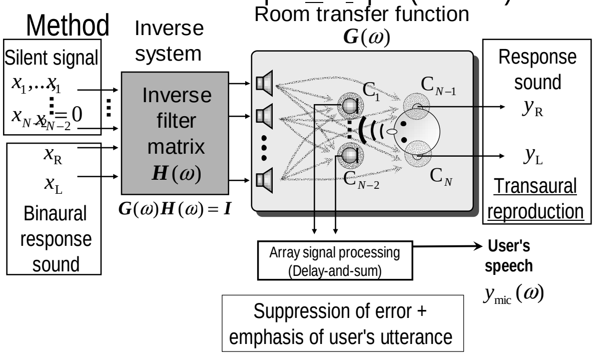
- Binaural recording
  - Observe sound includes property of humane head by using dummy head (binaural signal)
- Transaural reproduction
  - Reproduce binaural signal at user's ears
  - User can feel as if she/he is at the place where binaural signal is recorded (virtual reality of sound)


Dummy head


Inverse system

Binaural recording | Transaural reproduction

Proposed Method: Multiple-Output and Multiple-No-Input (MOMNI)



Advantages of MOMNI Method

- Robust against fluctuation of room transfer function

Error of response sound elimination

$$(\text{Error}) \propto 1/\sqrt{(\text{number of microphons}) \times (\text{number of loudspeakeers})}$$

Stable with many loudspeakers and microphone elements

Works stably with fixed filter (adaptation is unnecessary)

- Reduction of computational complexity
- Detection of double-talk is unnecessary

- Transaural reproduction of response sound
  - High quality and much presence
  - Presentation of virtual reality

Simulation

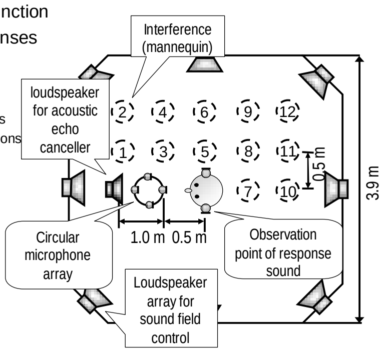
- Contents of experiment
  - Response sound elimination experiment
  - Speech recognition experiment
- Plan of experiment
  - Simulation of spoken dialogue system using impulse response measured in real-world
  - Comparison of robustness of control
- Comparing with
  - Acoustic echo canceller

Condition of Measuring Impulse Responses

Evaluate robustness against fluctuation of transfer function

Measure impulse responses caused fluctuation.

- Placing interference to cause fluctuation
  - a mannequin at 12 positions
  - imitating 12 kind of fluctuations
- Reverberation time
  - 160 ms
- Sampling frequency
  - 16 kHz

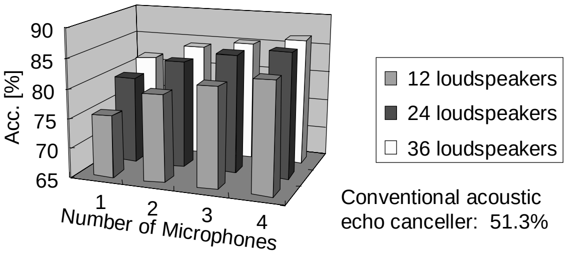

Interference (mannequin)
loudspeaker for acoustic echo canceller
2
4
6
9
12
1
3
5
8
11
7
10
0.5 m
3.9 m
1.0 m
0.5 m
Circular microphone array
Observation point of response sound
Loudspeaker array for sound field control

Response Sound Elimination

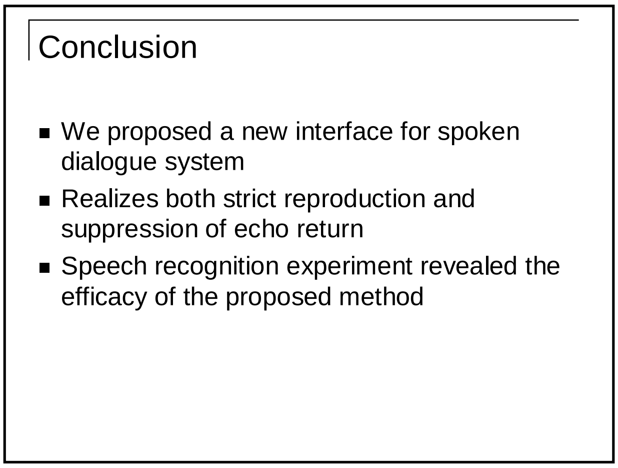
- Contents
  - Both performance of elimination at microphone and presenting response sound to user
- Evaluation score
  - Word Accuracy (Acc)

$$\text{Acc}[\%] = \{(\text{Number of words}) - (\text{Substitution Errors}) - (\text{Deletion Errors}) - (\text{Insertion Error})\} \,/\, (\text{Number of words})$$

Result

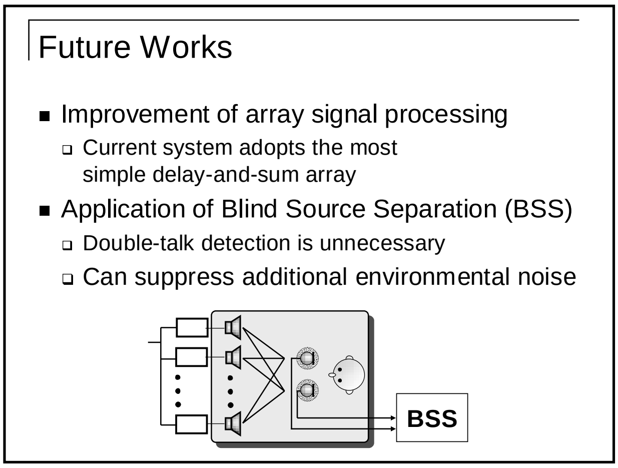


Conventional acoustic echo canceller: 51.3%



Conclusion

- We proposed a new interface for spoken dialogue system
- Realizes both strict reproduction and suppression of echo return
- Speech recognition experiment revealed the efficacy of the proposed method



Future Works

- Improvement of array signal processing
  - Current system adopts the most simple delay-and-sum array
- Application of Blind Source Separation (BSS)
  - Double-talk detection is unnecessary
  - Can suppress additional environmental noise