Robust Speech Recognition and its ROBOT implementation

Yoshikazu Miyanaga

Hokkaido University
Laboratory of Information Communication Networks
Graduate School of Information Science and Technology
Sapporo 060-0814, Hokkaido Japan

All right reserved. Copyright ©2015- Yoshikazu Miyanaga

Part 1

NOISE ROBUST TECHNOLOGY

All right reserved. Copyright ©2015- Yoshikazu Miyanaga

Conditions for Speech Recognition

Short Isolated Speech:
words, phrase (<2sec)

Continuous Speech:
sentences (>2sec)

Attached Mic
(several cm – 10cm)

Remote Mic:
(10cm – 5m)

Long Distance Mic:
(>5m)

Silent Room
(>20dB)

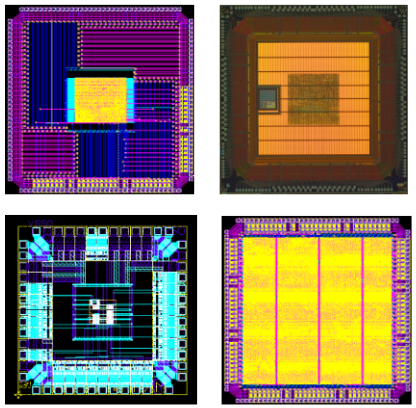
Living Room(20
～10dB)

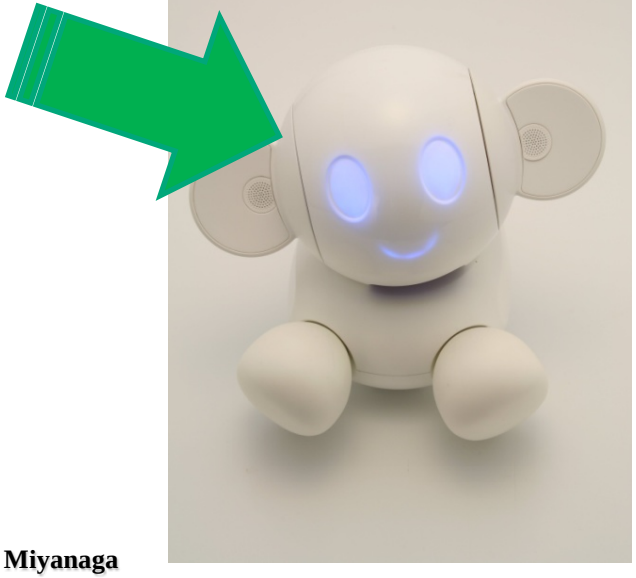
Noisy Room:
exhibition(<10dB)

All right reserved. Copyright ©2015- Yoshikazu Miyanaga

Robot Implementation

- **Autonomous Speech Recognition**
- Speech Synthesis
- Quick Response
- Control to Consumer Electronics and Machines





All right reserved. Copyright ©2015- Yoshikazu Miyanaga

Hokkaido University Speech Communication System (HU-SCS)

Short Isolated Speech: words, phrase (<2sec)

Long Distance Mic: (>5m)

Remote Mic: (10cm – 5m)

Silent Room (>20dB)

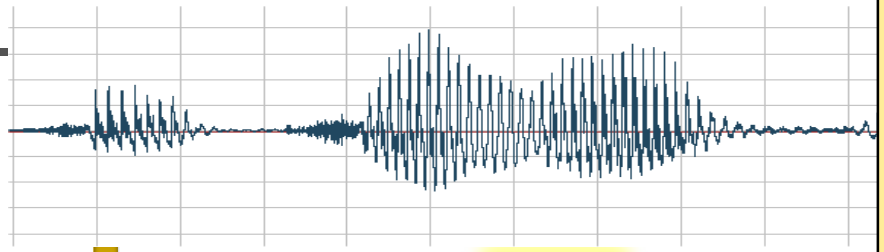
Attached Mic (several cm – 10cm)

Living Room（20～10dB)

Noisy Room: exhibition（<10dB)

All right reserved. Copyright ©2015- Yoshikazu Miyanaga

HU-SCS



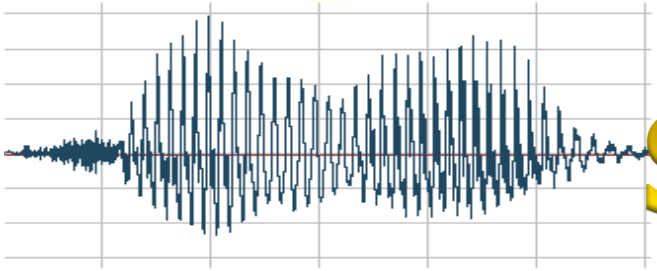
Speech

HU-SCS v4
- 99% over SNR 10dB
- BP＋Threshold Opt
- F_0 Detection

No good detection under SNR < 5dB

Automatic Speech Detection

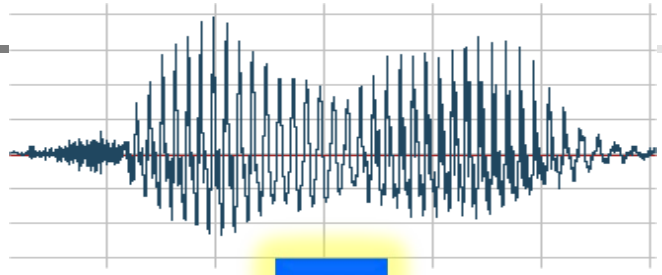
VAD
(Voice Activity Detection)



Speech

All right reserved. Copyright ©2015- Yoshikazu Miyanaga

HU-SCS



Speech

Automatic Speech Recognition

Candidates of Recognition Results
(1) Good Morning
(2) See you
(3) How are you ?

Phase

All right reserved. Copyright ©2015- Yoshikazu Miyanaga

Noise modeling

- Spectrum including noise can be modeled as,

$$X(n,\omega) = S(n,\omega)H(\omega) + A(\omega)$$

Clean spectrum	Multiplicative noise	Additive noise

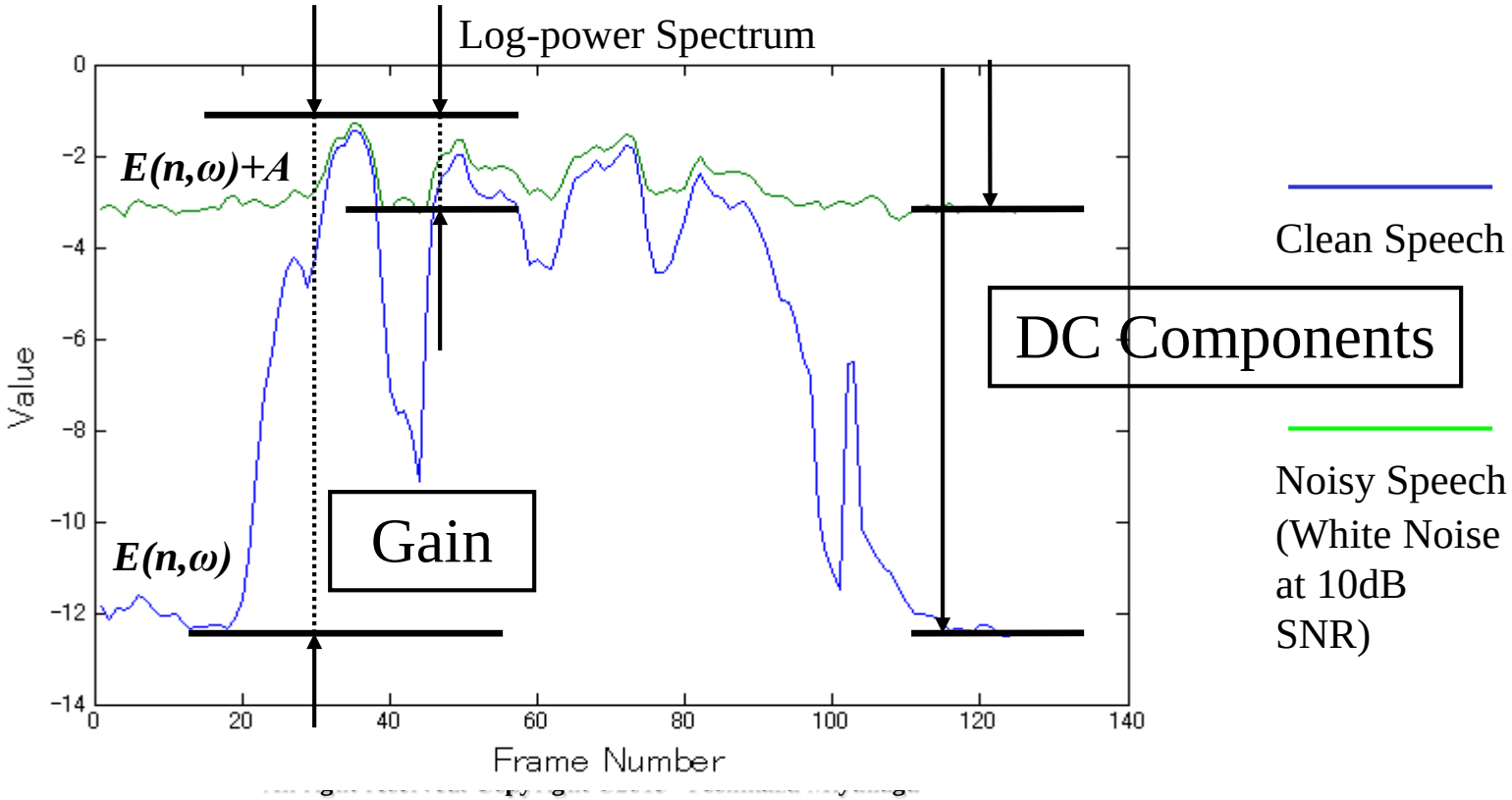
$$\log \quad E(n,\omega) = S(n,\omega)H(\omega)$$

$$\log X(n,\omega) = \log(E(n,\omega) + A(\omega))$$

All right reserved. Copyright ©2015- Yoshikazu Miyanaga

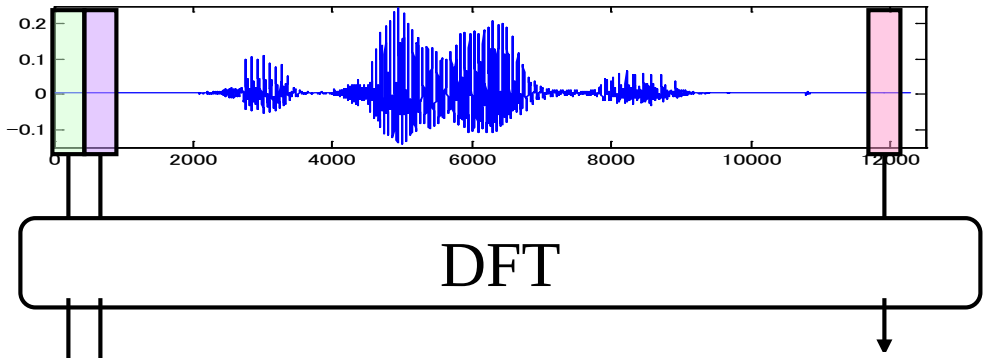
Noise Corruption in Log Power Spectrum

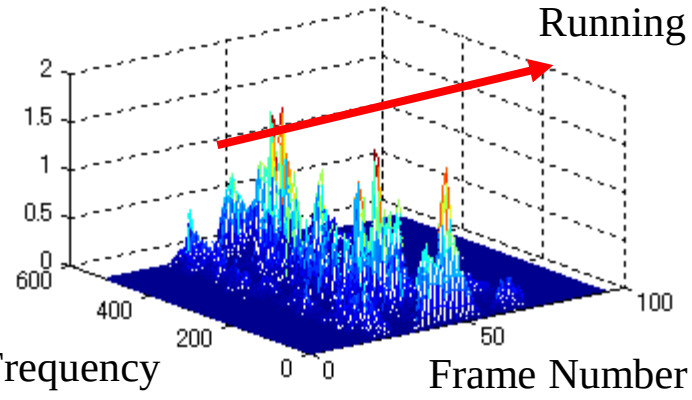
- Noise corruptions make differences on gains and DC components.



Running Spectrum

- Running spectrum is obtained by accumulating short-time spectrum





Running spectrum: time trajectory of frequency

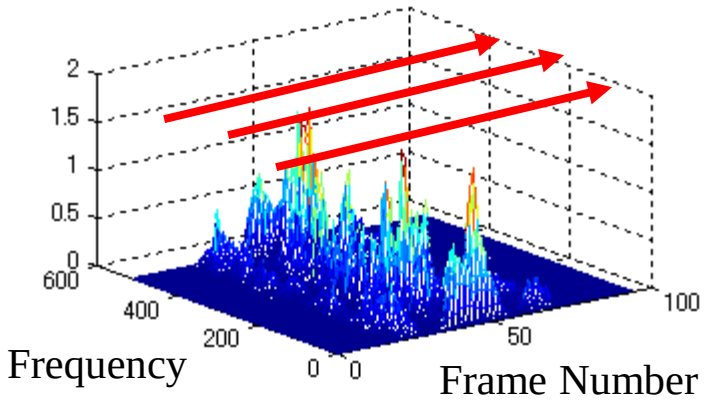
All right r

Modulation Spectrum

- RSF focuses on modulation spectrum

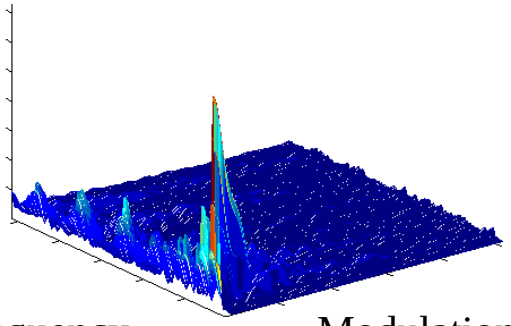
Modulation spectrum: spectrum versus time trajectory of frequency.

Running Spectrum



DFT on each frequency

Modulation Spectrum

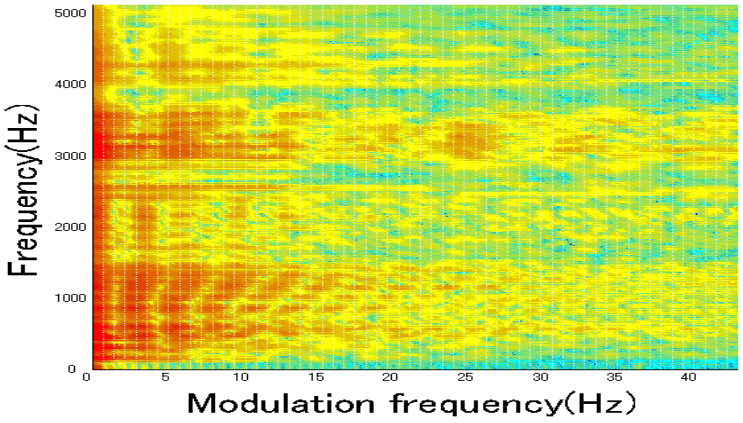


All right reserved. Copyright ©2015- Yoshikazu Miyanaga

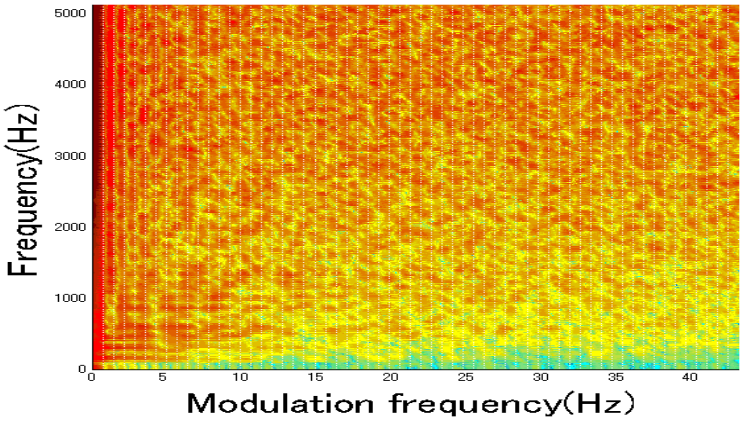
Mod-F of Clean and Noisy Speech

- Speech components are dominant around 4 Hz in modulation spectrum.

Clean



Noisy (white noise at 5 dB SNR)

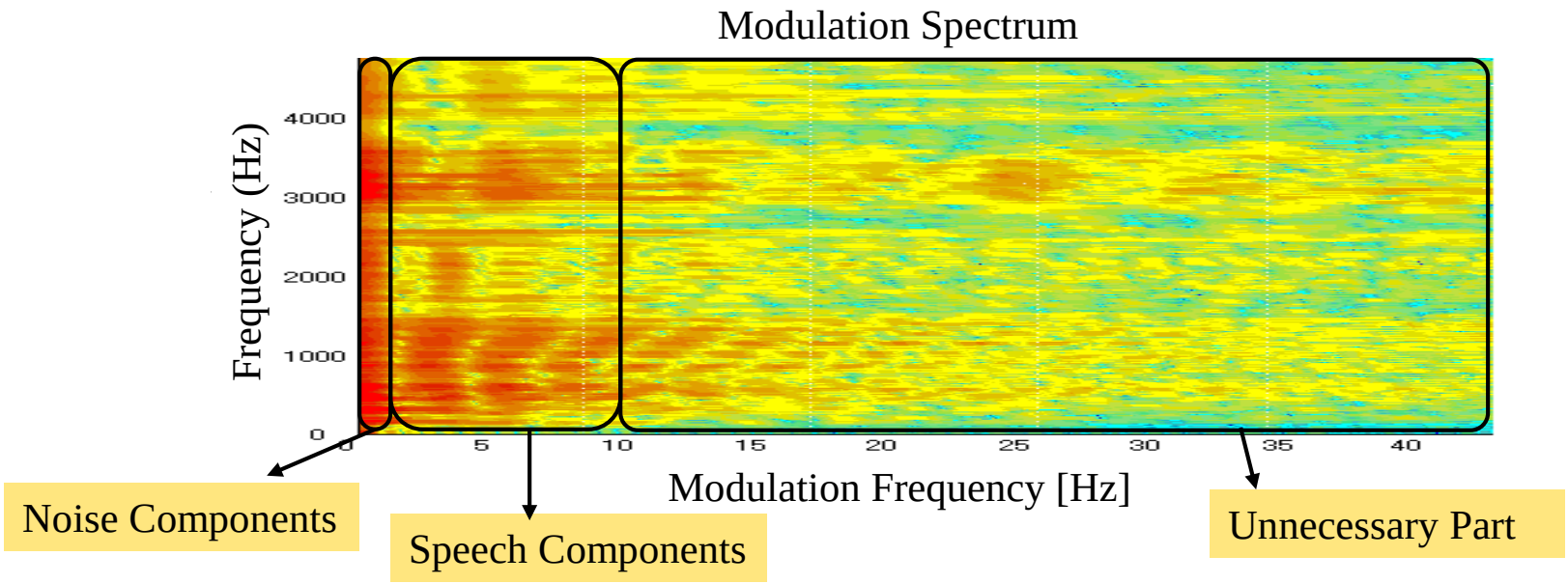


Lower modulation frequency components can be assumed as noise because of little changes in noise components.

All right reserved. Copyright ©2015- Yoshikazu Miyanaga

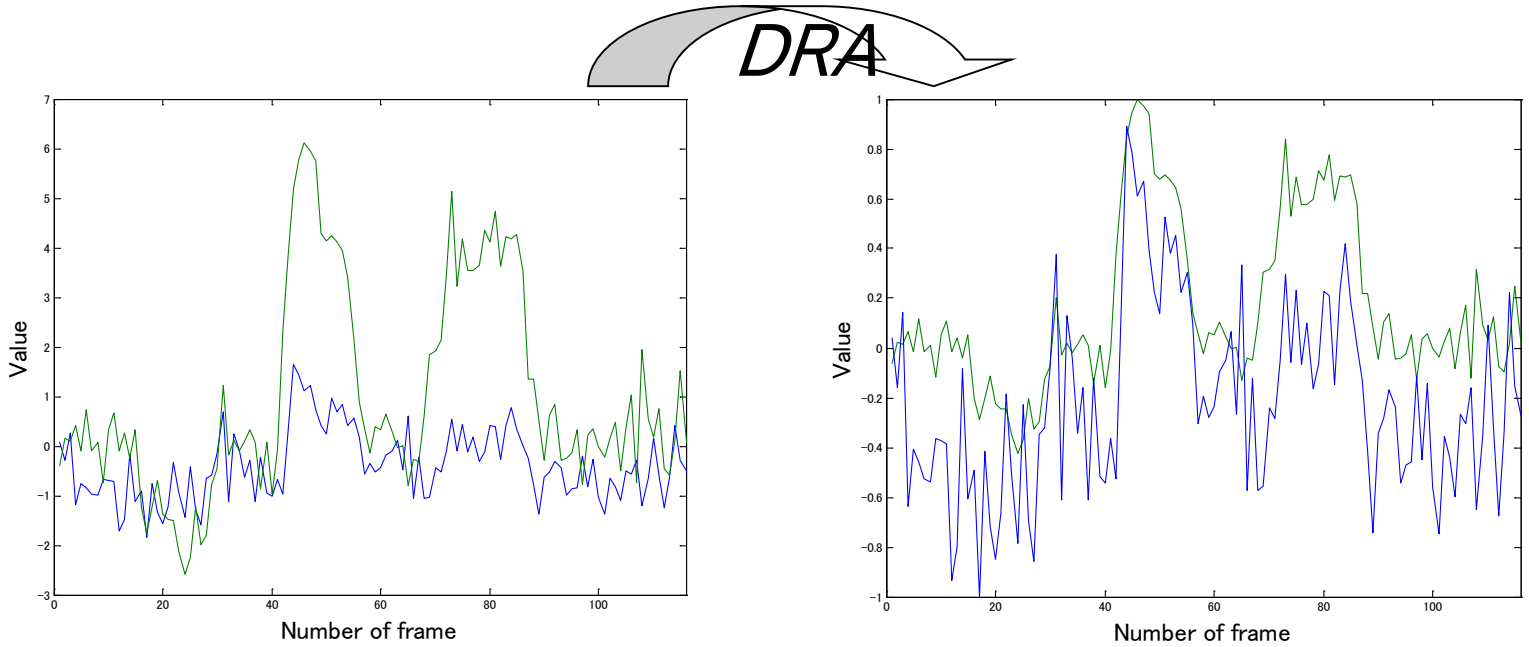
RSF (Running Spectrum Filtering)

- Speech components are dominant around 4 Hz in modulation spectrum.



All right reserved. Copyright ©2015- Yoshikazu Miyanaga

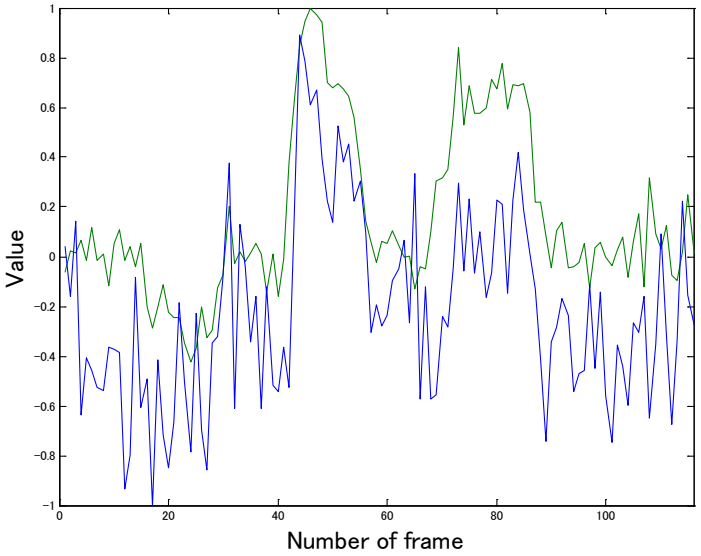
Dynamic Range Adjustment

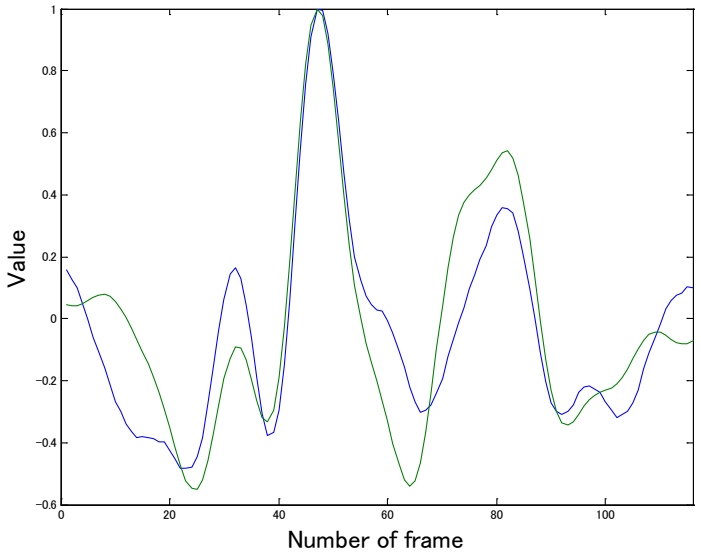


Green shows the 4th MFCC before DRA and Blue shows after DRA

All right reserved. Copyright ©2015- Yoshikazu Miyanaga

Running Spectrum Filtering

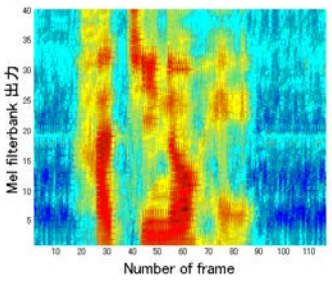




The 4th MFCC by only DRA in the left hand side, and by both DRA and RSF in the right hand side

All right reserved. Copyright ©2015- Yoshikazu Miyanaga

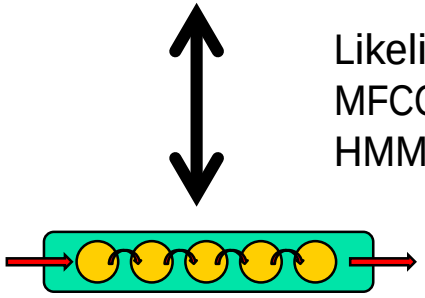
Evaluation on Likelihoods



RSF/DRA reduce noise factors.

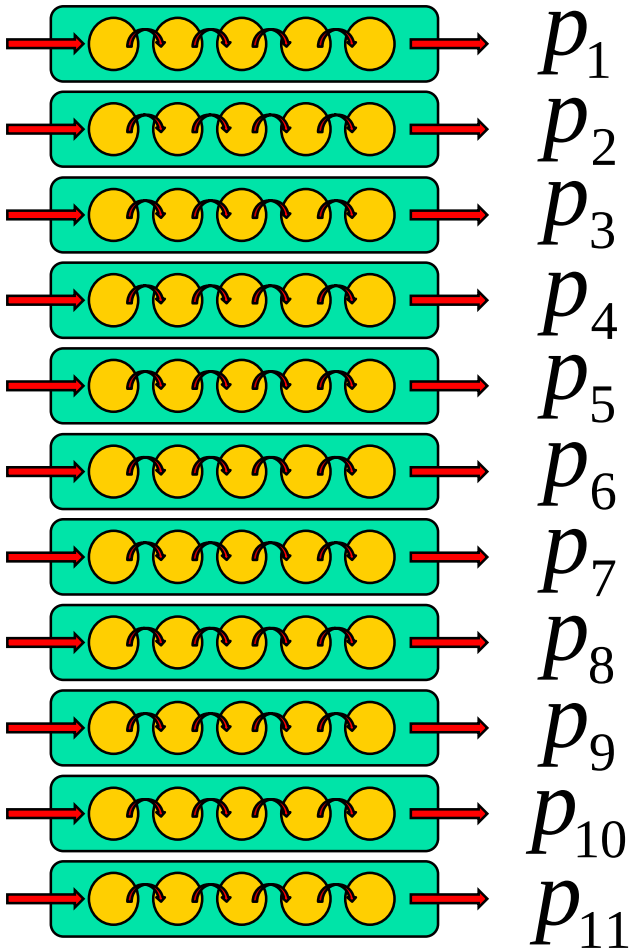
MFCC

Likelihood of MFCC into HMMs



HMM

Training HMM



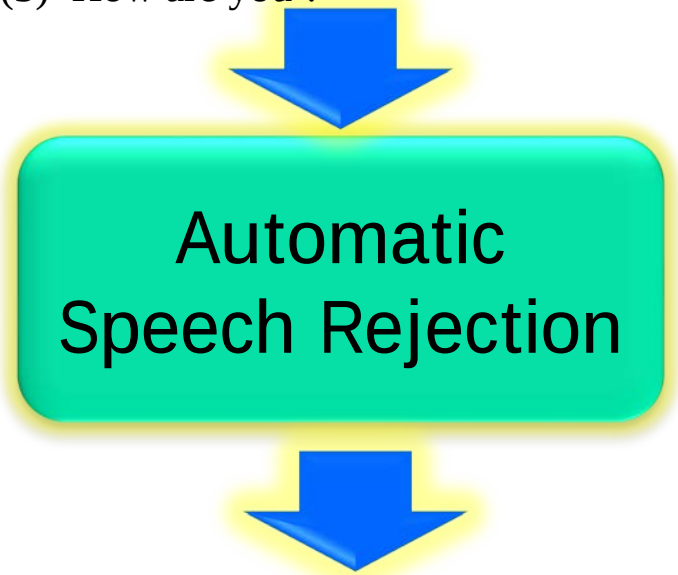
The **maximum likelihood** is selected and its label is recognized as the result.

All right reserved. Copyright ©2015- Yoshikazu Miyanaga

HU-SCS

Candidates of Recognition Results
(1) Good Morning
(2) See you
(3) How are you ?

Phase



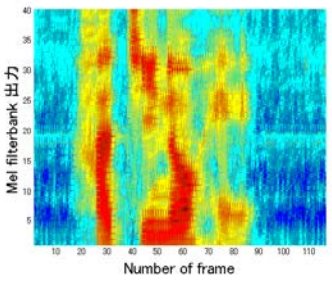
Recognition Result
Good Morning

Confidential Phase

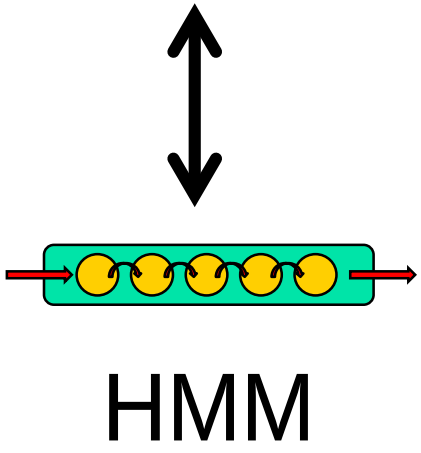
All right reserved. Copyright ©2015- Yoshikazu Miyanaga

Evaluation on Likelihoods

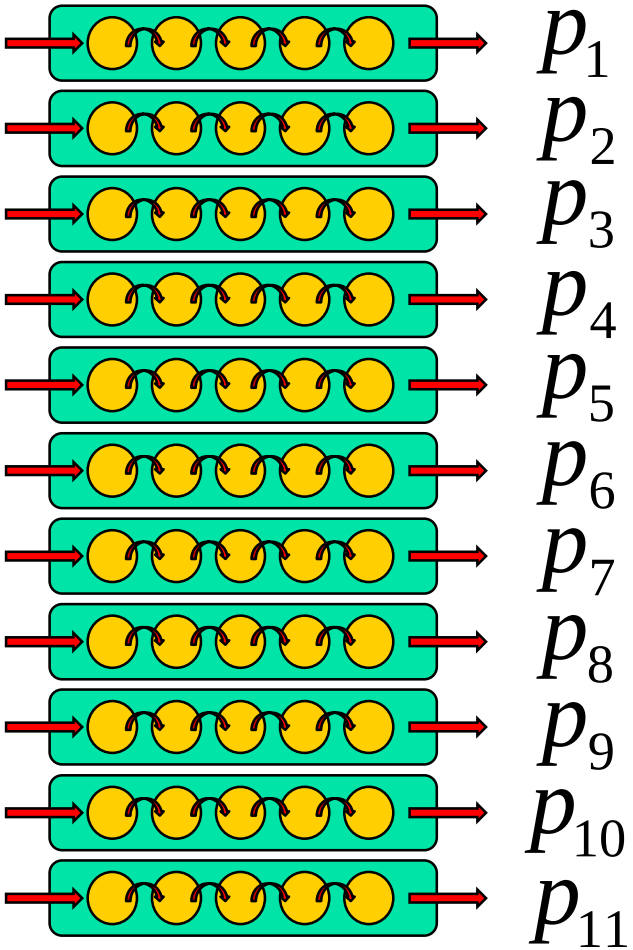
Training HMM



MFCC

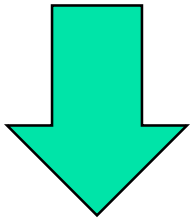


HMM



p_1
p_2
p_3
p_4
p_5
p_6
p_7
p_8
p_9
p_{10}
p_{11}

The maximum likelihood is selected and its label is recognized as the result.

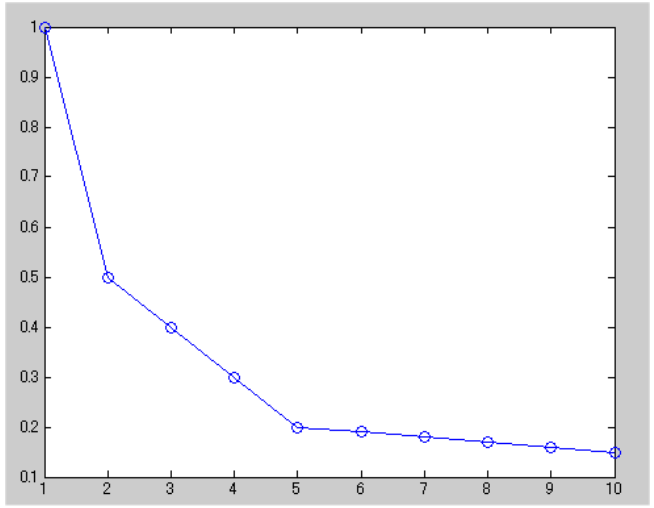


The result is correct, isn't it ?

All right reserved. Copyright ©2015- Yoshikazu Miyanaga

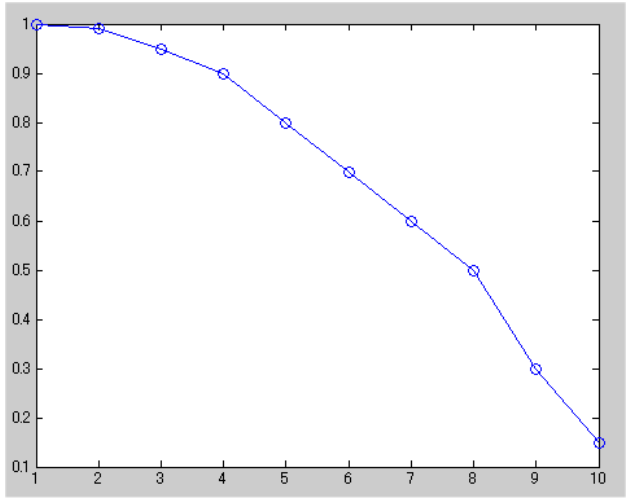
Evaluation of Reliability

Likelihood



The result of the top score is trusted.

Likelihood



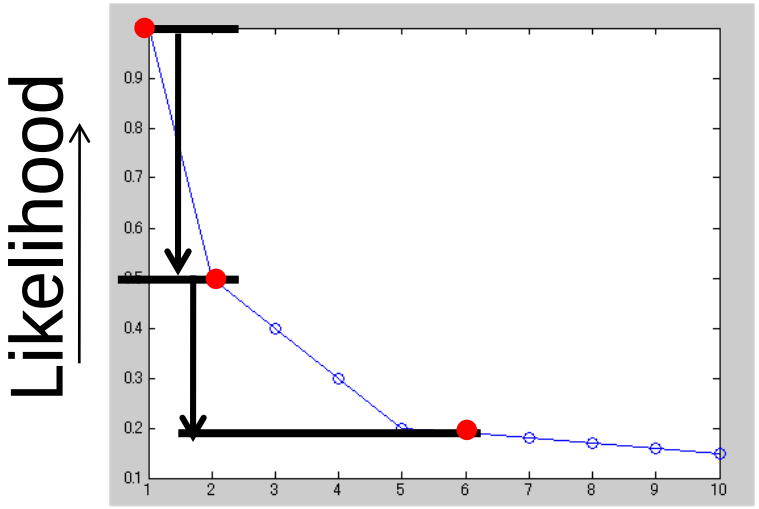
The result of the top score is **NOT** trusted.

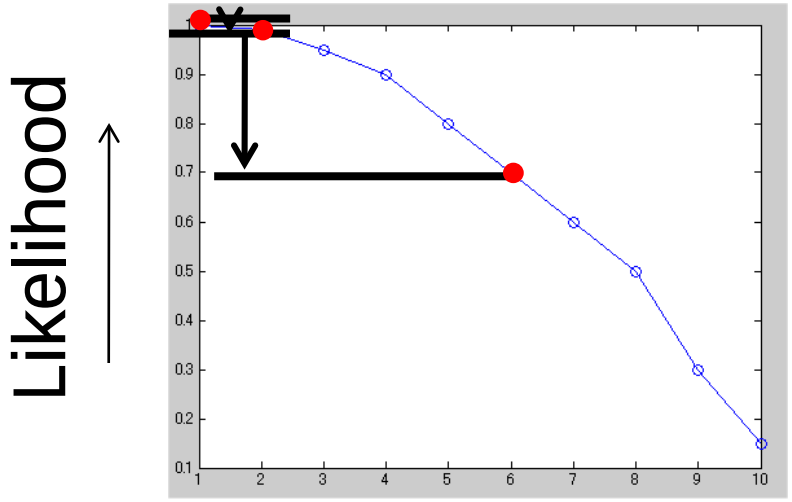
All right reserved. Copyright ©2015- Yoshikazu Miyanaga

Rejection Method

1. $\frac{1}{5}\sum_{a=k-2}^{k+2}\left[\{L(p_{a1})-L(p_{a2})\}/\{L(p_{a2})-L(p_{a,T})\}\right]>\alpha$
2. $\frac{1}{5}\sum_{a=k-2}^{k+2}\{L(p_{a1})-L(p_{a2})\}>\beta$
3. $\frac{1}{5}\sum_{a=k-2}^{k+2}\{L(p_{a1})-L(p_{a,T})\}>\Delta$

$L(p_{mi})$: Likelihood





All right reserved. Copyright ©2015- Yoshikazu Miyanaga

First SCS HW

- LSI IP
- Mobile
- Intelligent Consumer Electronics etc

Fine Advantage

HW with
Low Power

(1) Mobile Appli
- Small
- Low Power

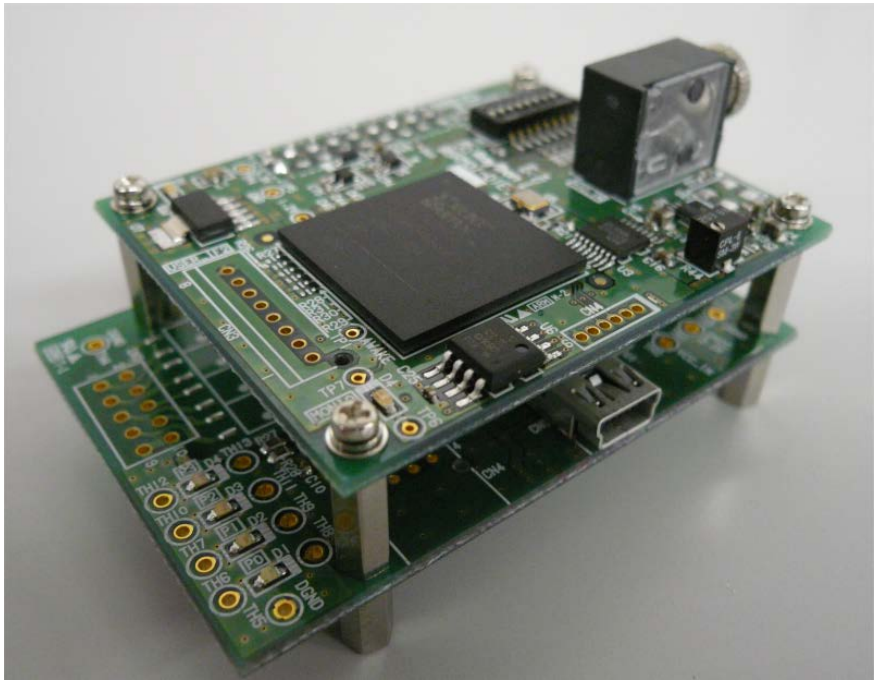
(2) PC free

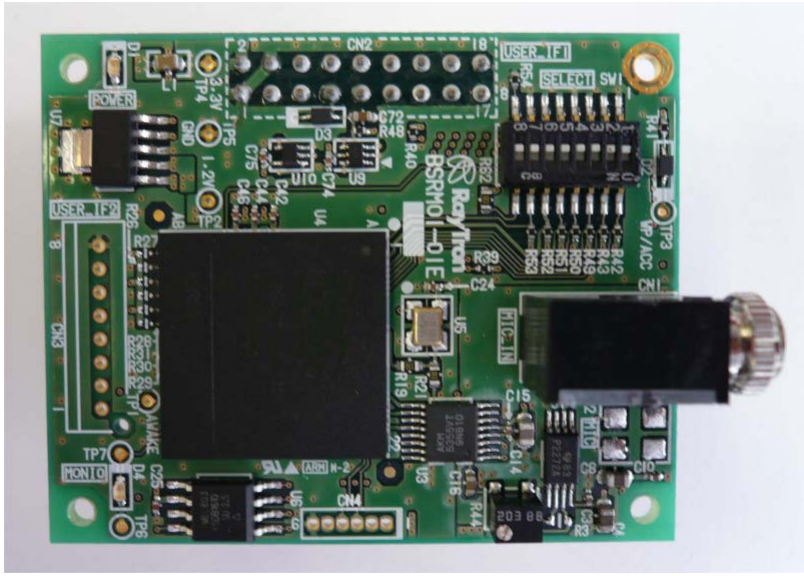
- Super Low-Power Consumption Design
- Real-Time SCS
  →180nsec/word (10MHz clk) Recognition Time
- Small Scale Design with Special Designed LSI
- Noise Reduction by Array Microphone

All right reserved. Copyright ©2015- Yoshikazu Miyanaga

Current HU-SCS v4

PC Interface with
HU-SCS Board





HU-SCS Board

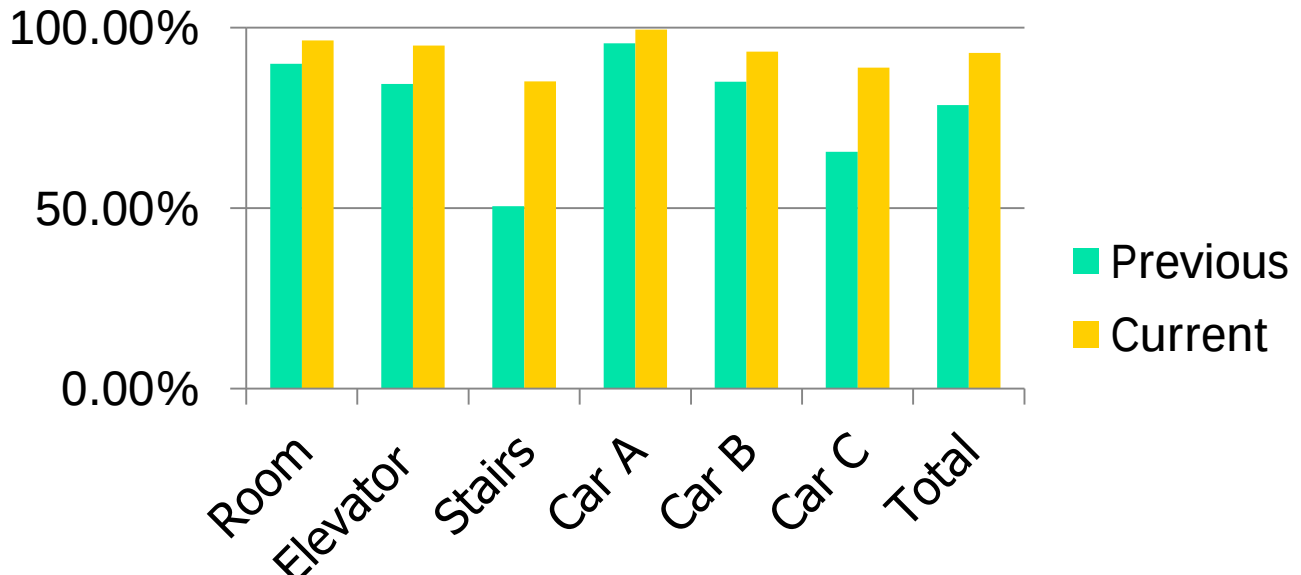
55mm × 44 mm

All right reserved. Copyright ©2015- Yoshikazu Miyanaga

Comparison on Performance

Environment	Noise Level	Correctness	
		Current	Previous
Meeting Room	50dB	96.4%	90.0%
Elevator	50dB	95.0%	84.4%
Stairs	45dB	85.1%	50.5%
Car A（Idling, No-Moving）	50dB	99.4%	95.6%
Car B（High Speed, Open Window）	75dB	93.3%	85.0%
Car C（High Speed, Audio ON（FM））	75dB	88.9%	65.6%
Total		93.0%	78.5%
※Cruiser Board（Outside, high speed）	80dB	82.7%	-

Comparisons between HU-SCS v4 and v3



All right reserved. Copyright ©2015- Yoshikazu Miyanaga

Part 2

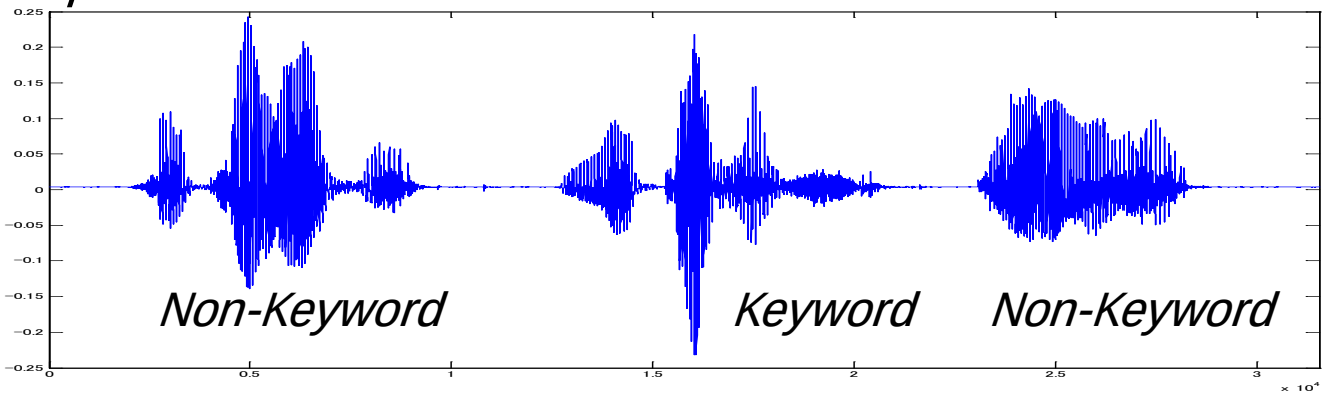
ADVANCED NOISE ROBUST TECHNOLOGY

All right reserved. Copyright ©2015- Yoshikazu Miyanaga

Sequential Word Stream

- Only keywords are recognized in a continuous speech under noisy environments.
  - Any words except keywords should not be recognized.
- The speech is given as a sequential word stream like voice commands.

Sequential word without noise



All right reserved. Copyright ©2015- Yoshikazu Miyanaga

Issues in This System

- Intelligent Keyword Rejection
  - Many of non-Keywords are recorded.

- High Precise Speech Detection
  - Keywords and non-Keywords are sequentially recorded into a system like continuous speech waveform.

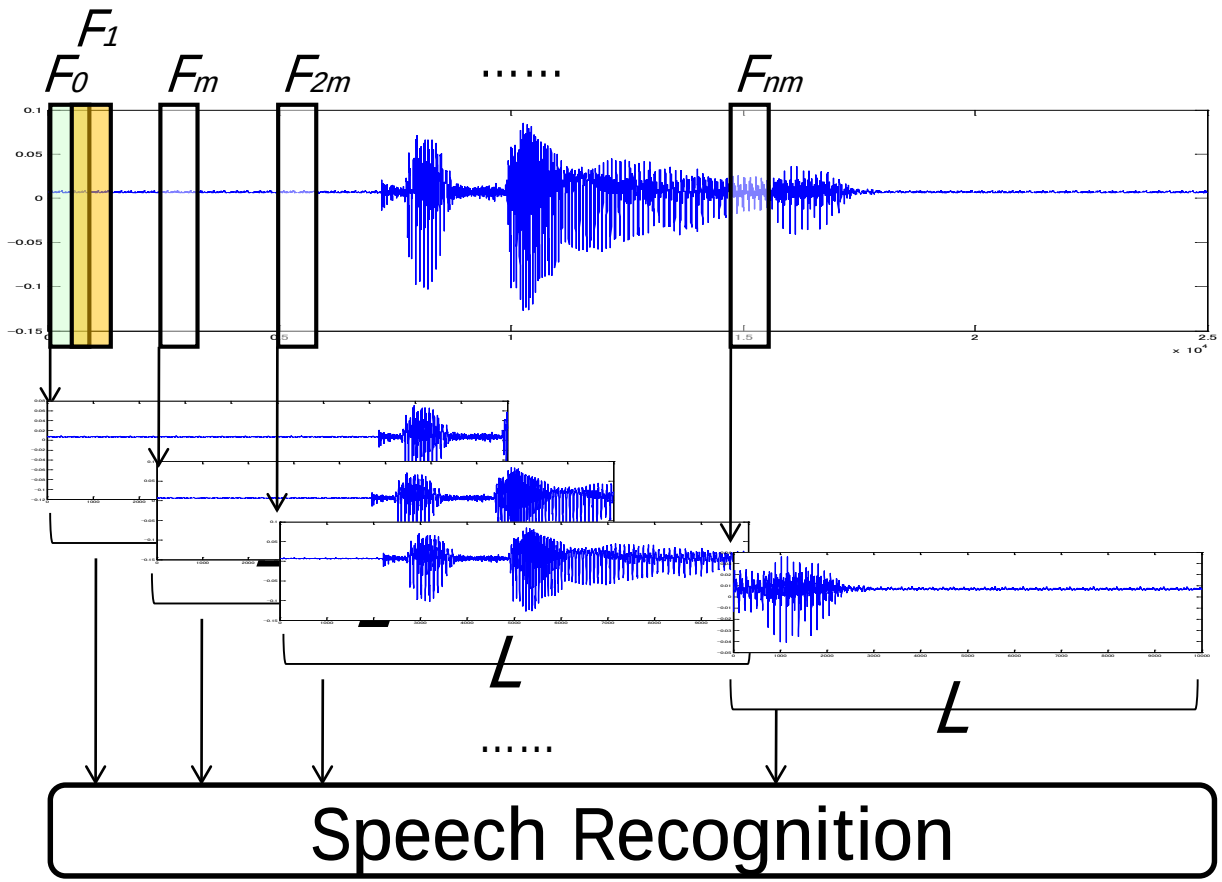
All right reserved. Copyright ©2015- Yoshikazu Miyanaga

SP-ASR Recognition Processing

- Set a segment after dividing the short-time frame

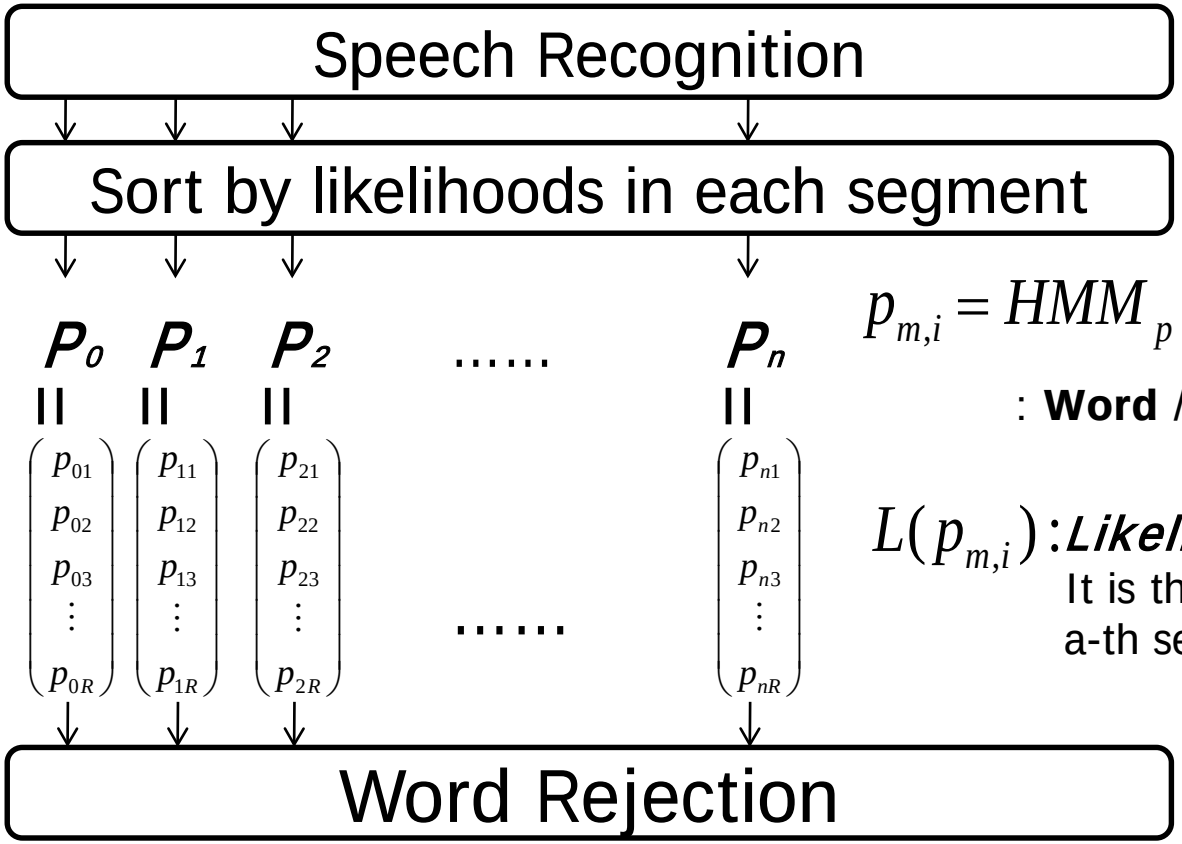
Short-time Frame

Segment



All right reserved. Copyright ©2015- Yoshikazu Miyanaga

2nd Rejection Methods



$$p_{m,i} = HMM_p(m)$$

: **Word** /*p*/ at the *m*-th segment

$L(p_{m,i})$: ***Likelihood Value***

It is the i-th largest ML in the a-th segment.

m : segment index

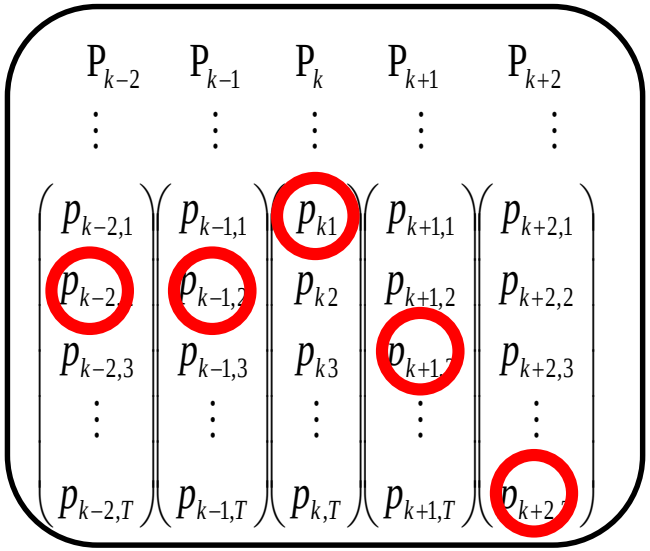
i : ranking index

Note that the vector P_i might be similar to the vector P_j

All right reserved. Copyright ©2015- Yoshikazu Miyanaga

Rejection Method

Segment Stream



k: Center segment number

1. $p_{k,1} \in \mathrm{P}_r \ (r = k-2, \cdots, k+2)$

2. $L(p_{r,i}) > L(p_{r,j})$

$$\{i \mid p_{k,1} = p_{r,i}\}, \{j \mid p_{k,2} = p_{r,j}\}$$

$$r = k-2, \cdots, k+2$$

$p_{m,i} = HMM_p(m)$: **Word** */p/* at the *m*-th segment. The index *i* shows the *i*-th largest likelihood.

$L(p_{m,i})$: ***Likelihood Value*** It is the i-th largest ML in the a-th segment.

***m** : segment index,* ***i** : ranking index*

All right reserved. Copyright ©2015- Yoshikazu Miyanaga

Experiment - Conditions

- Short-time frame
  - Width: 23.2ms(256point)　Shift: 11.6ms(128point)
- Segment
  - Length(L): 1.0s(11025point)　　Shift: 8[frame]
- Speech data:
  - Connect two words.
  - Add a non-speech section of 1.4 sec to both ends and 0.5 sec between the words.
  - 100 data by random two words combination.
- White Noise：SNR=∞,20,10dB

All right reserved. Copyright ©2015- Yoshikazu Miyanaga

Results of recognition

Recognizer／SNR	∞	20dB	10dB
Isolated Word Recognition (1 word)	99.4 %	97.6 %	84.6 %
SP-ASR with 2nd Rejection (2 sequential words)	98.5 %	96.5 %	88.5 %

- The SP-ASR recognizes keywords with the same level accuracy as the past word recognition.

All right reserved. Copyright ©2015- Yoshikazu Miyanaga

Summary

Small, Fast and Low Power

Hokkaido University
Speech Communication System

Integrated Architecture of Speech Detection, Robust Speech Analysis, Speech Recognition, Speech Rejection
Higher Speed Processing than DSP and Software
Superior in Energy Saving than DSP Solutions
Improving Noise Robustness by RSF/DRA Technique

All right reserved. Copyright ©2015- Yoshikazu Miyanaga