autonomous Speech Recognition with Noise robust Speech Features

Yoshikazu Miyanaga

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

Part 1

BASIC AUTONOMOUS SPEECH RECOGNITION SYSTEM

Conditions for Speech Recognition

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

Continuous Speech:
sentences (>2sec)

Attached Microphone
(several cm – 10cm)

Remote Microphone
(10cm – 5m)

Long Distance Microphone
(>5m)

Silent Room
（>20dB SNR)

Living Room（20～10dB SNR)

Noisy Room & Outside
（<10dB SNR)

Cloud ASR

Continuous Speech Recognition over Internet

Continuous Sentence Speech	Attached Mic (<10cm)	Silent Room (>20dB)

↓ Language Model with small Ontology

Short Sentence Speech	Attached Mic (<10cm)	Living Room (20～10dB)

↓ Array Microphone

Short Isolated Speech: (<2sec)	Remote Mic: (<5m)	Living Room (20～10dB)

Our Target

Autonomous ASR

Isolated Speech Recognition using own SW/HW

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

Long Distance Mic: (>5m)

Remote Mic: (10cm – 5m)

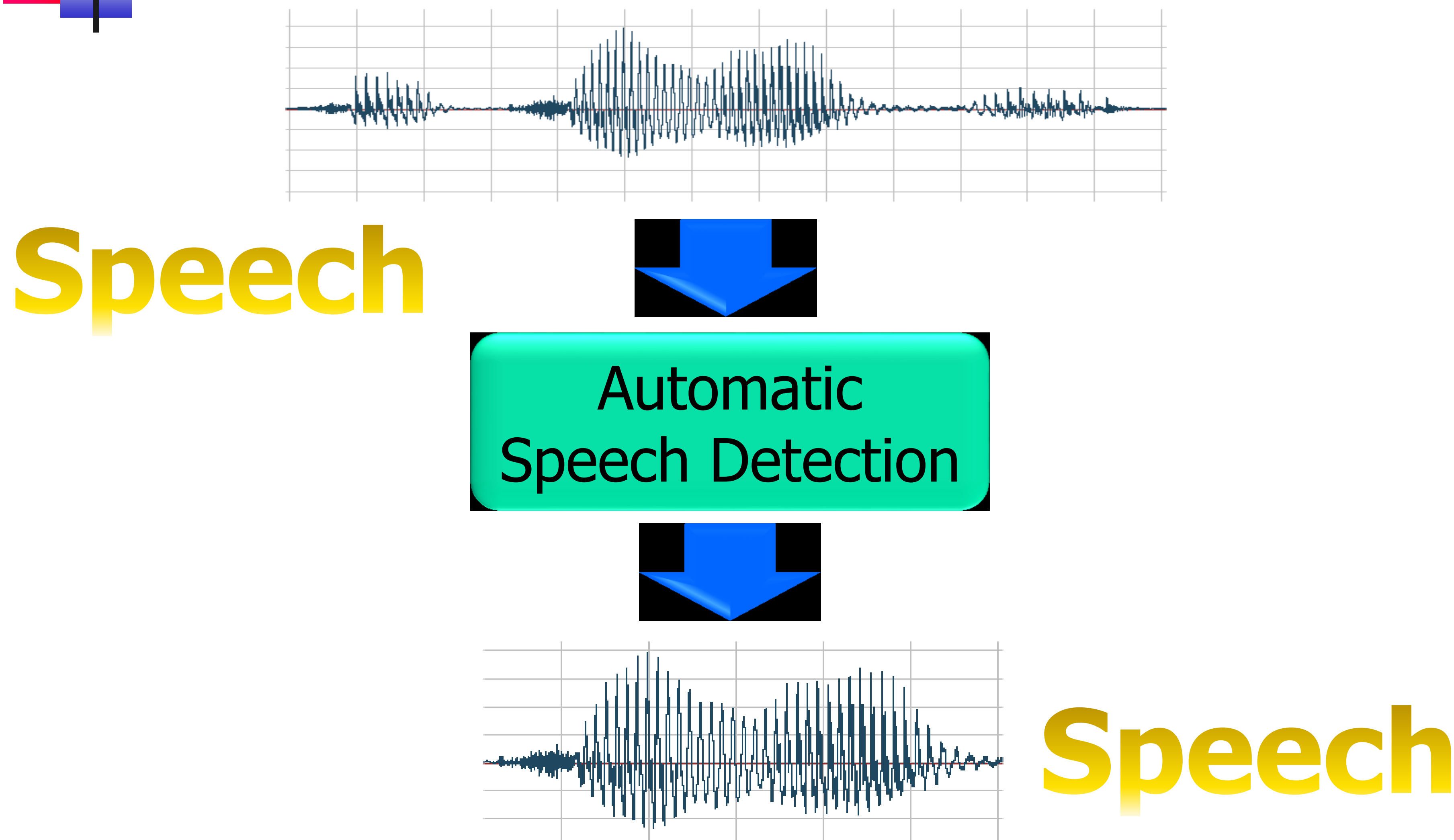
Silent Room (>20dB)

Attached Mic (several cm – 10cm)

Living Room (20～10dB)

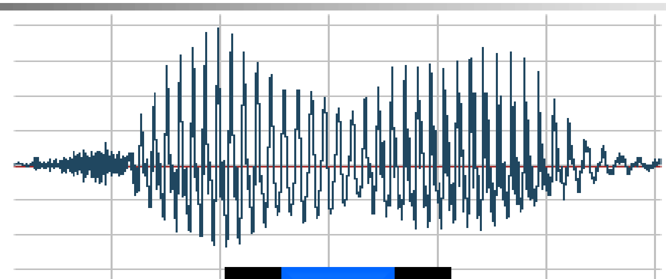
Noisy Room: exhibition (<10dB)

Voice Activity Detection



Autonomous Speech Recognition

Speech



Automatic Speech Recognition

Candidates of Recognition Results

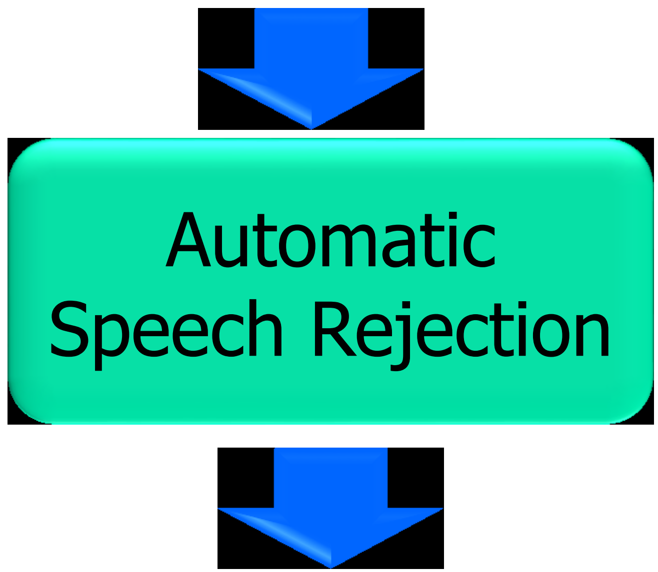
(1) Good Morning
(2) See you
(3) How are you ?

Phase

Automatic Speech Selection

Phase

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



Recognition Result: **Good Morning**

Confidential Phase

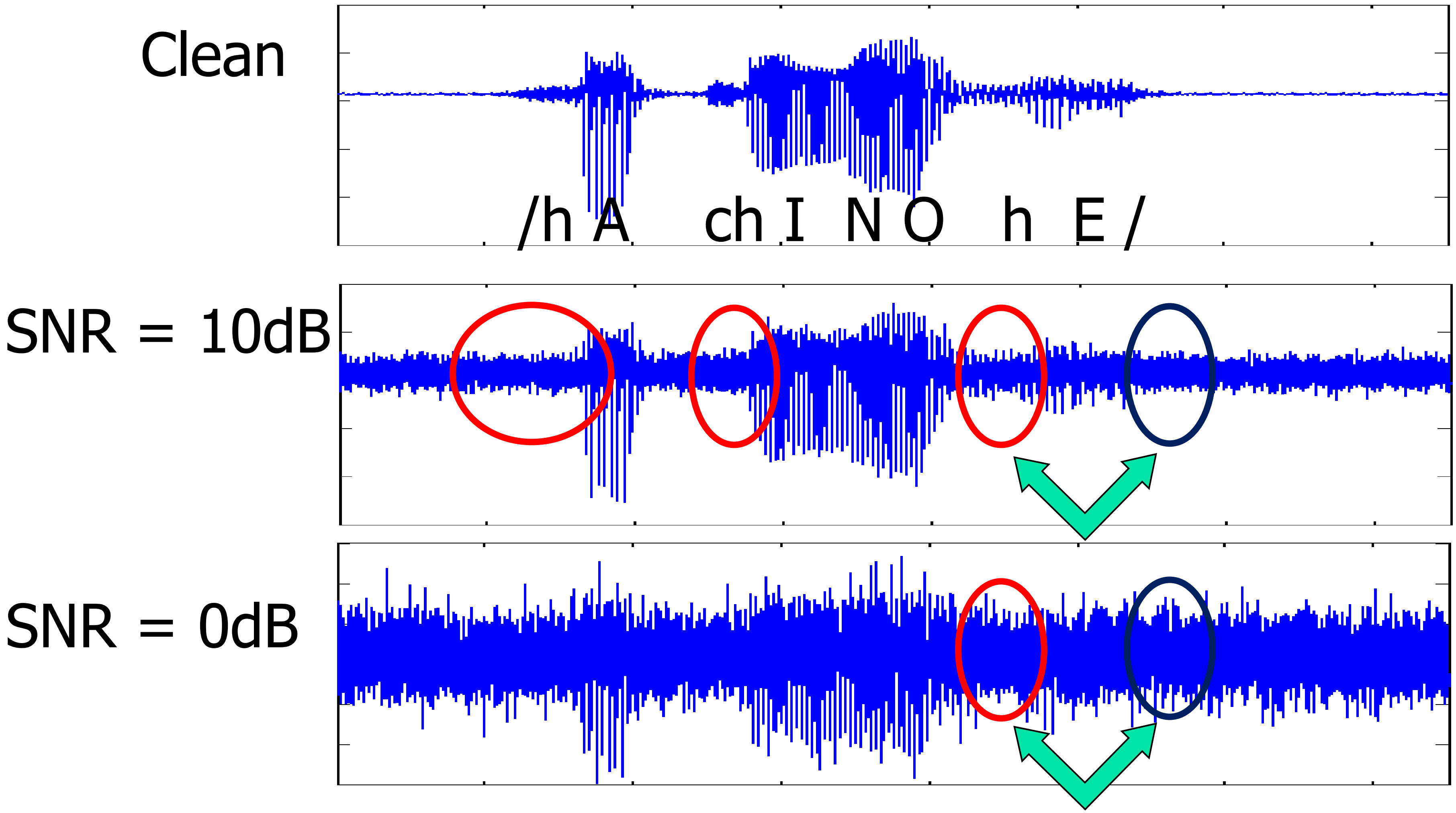
ROBI -2014-

- Producer & Sales Company
  by Deagostini Japan, and Raytron Inc, JP
- Design & Robot Controller
  by T.Takahashi, Robo-Garage Ltd
- Autonomous ASR
  by Miyanaga Lab, HU

Part 2

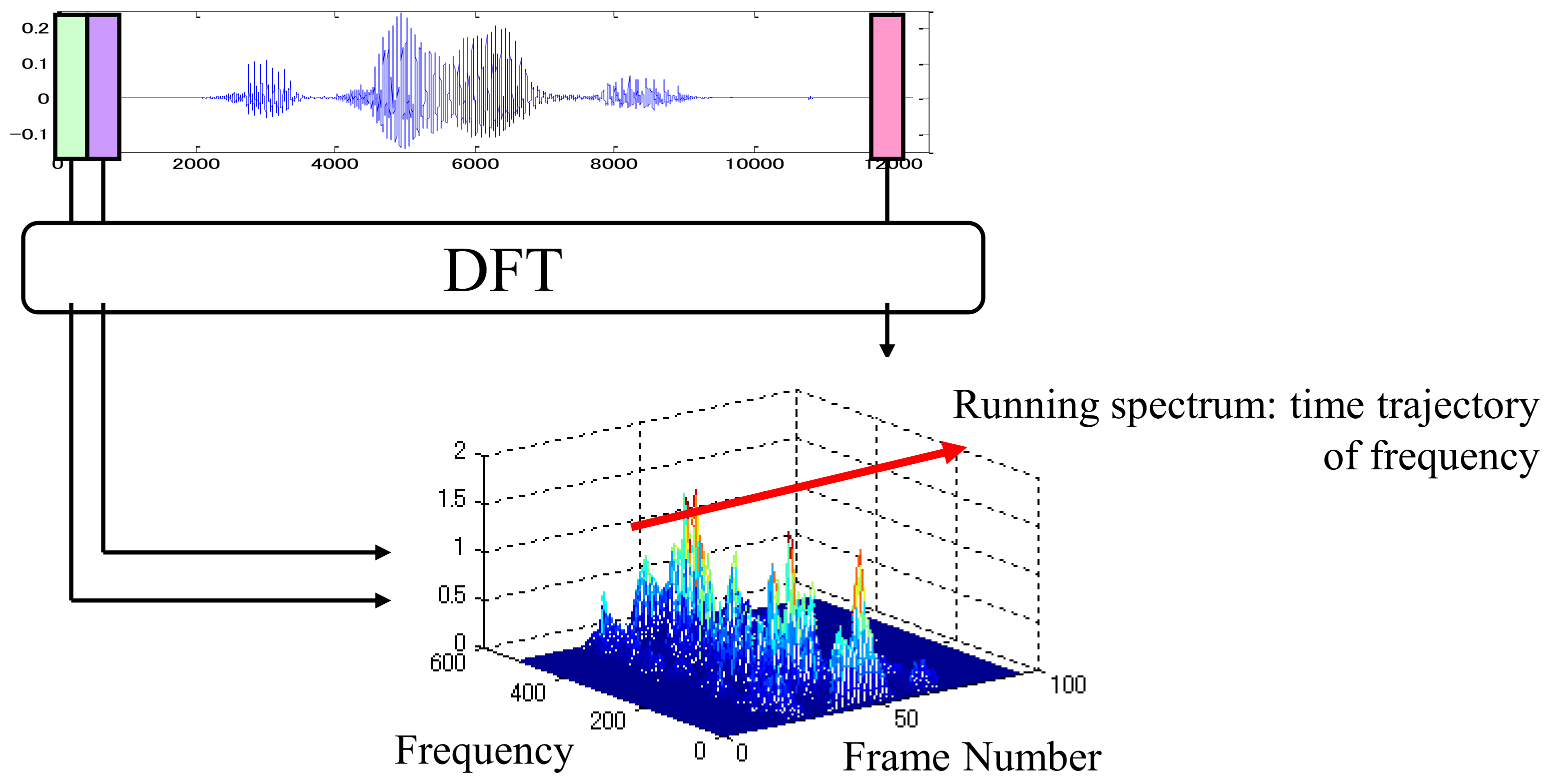
NOISE ROBUST SYSTEMS

Noise !



Running Spectrum

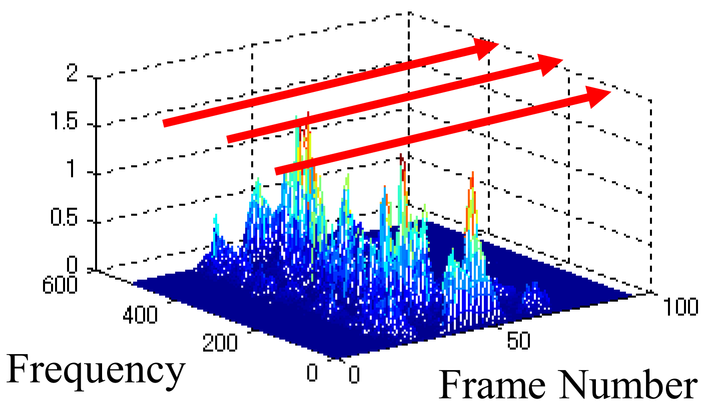
Running spectra are obtained by accumulating short-time spectrum



Modulation Spectrum

CMS, RASTA and RSF focuses on modulation spectra.

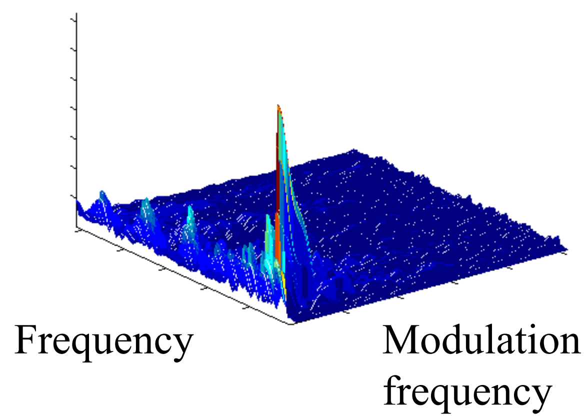
Running Spectrum



Modulation spectrum: spectrum versus time trajectory of frequency.

DFT on each frequency

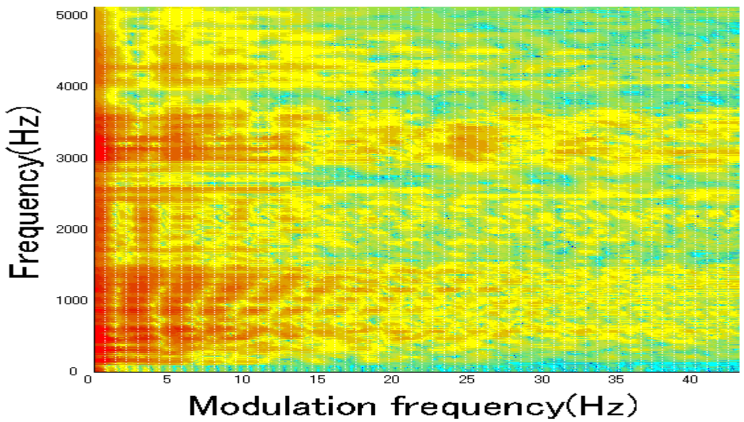
Modulation Spectrum



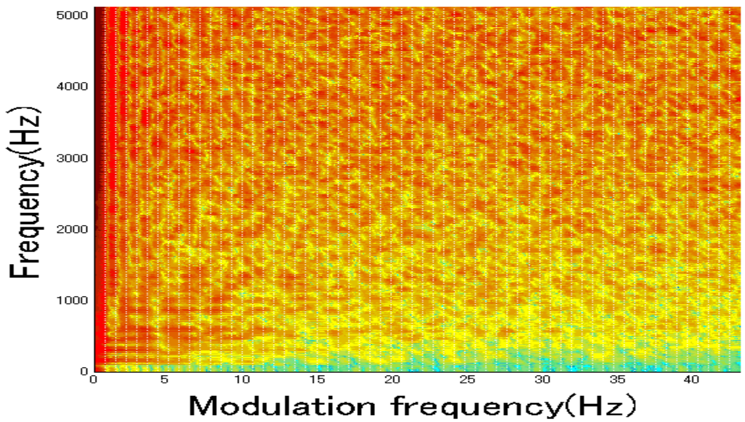
Mod-F of Clean and Noisy Speech

Speech components are dominant around 4 Hz in modulation spectrum.

Clean



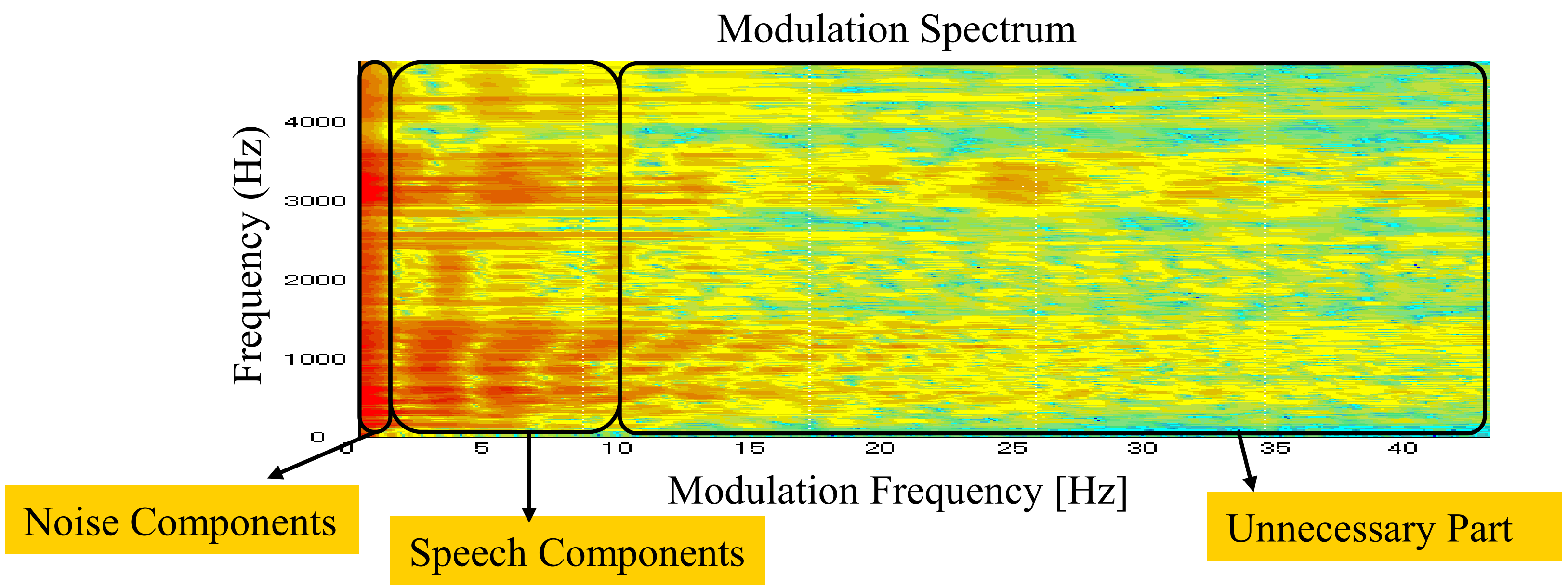
Noisy (white noise at 5 dB SNR)



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

Filtering over Running Spectrum

Speech components are dominant around 4 Hz in modulation spectrum.



RASTA (1991) and RSF (2002)

RSF (Running Spectrum Filtering)

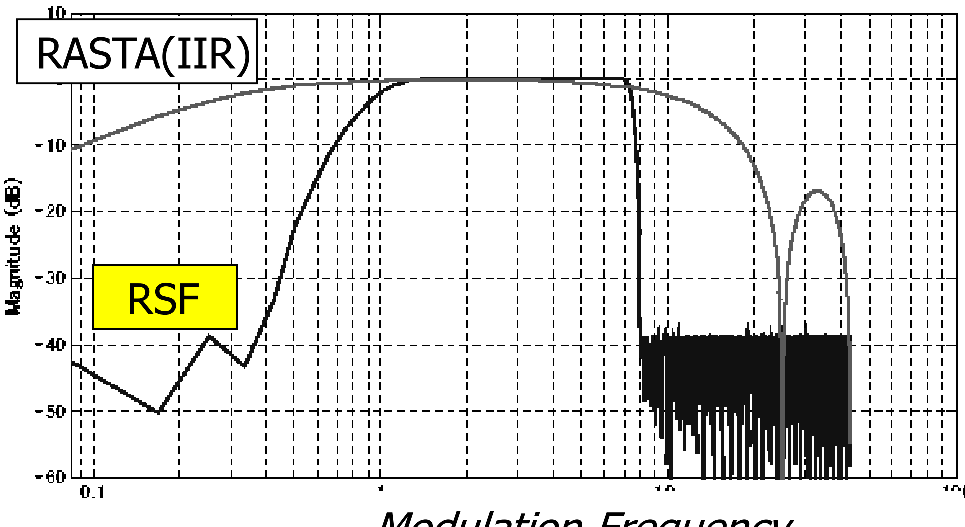
- enhances perceptual auditory components.
- decreases noise components relatively by band-pass filtering in cepstral sequences.

$$\widetilde{C}(n,k) = \sum_{i=0}^{Q} h(i) \cdot C(n-i,k)$$

Coefficients in FIR Filter

H. Hermansky, et. al., "Compensation for the effect of communication channel in auditory-like analysis of speech (RASTA-PLP)," Proceedings of European Conference on Speech Technology, 1991, pp. 1367–1370.

N.Wada, Y.Miyanaga, et. al., "A Study about the Extract of Robust Speech Characteristics on Speech Recognition System", IEICE Technical Report, DSP2002-33, pp.19-22, May 2002.



DRA

DRA (Dynamic Range Adjustment)

- normalizes amplitude of cepstral vectors in time domain (use of maximum value during utterance).
- suppresses dynamic range distortions caused by additive noise.

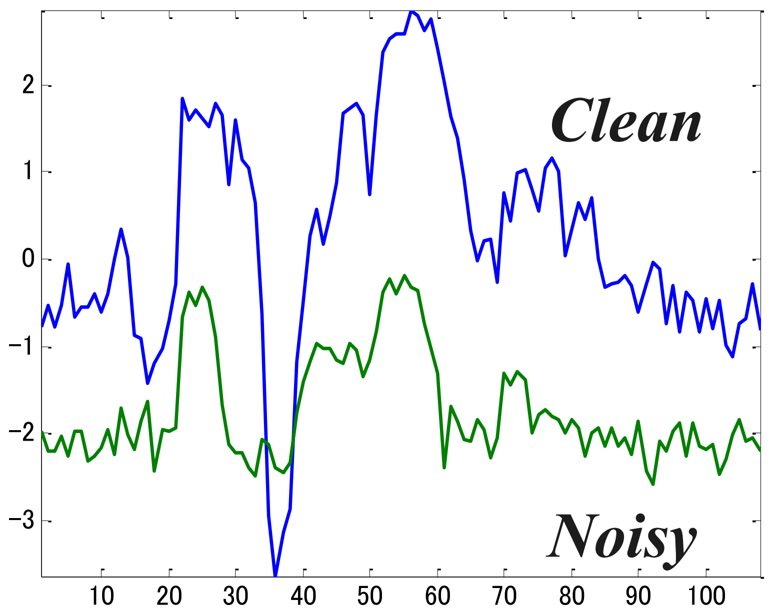
$$\overline{C}(n,k) = \frac{\widetilde{C}(n,k)}{\lambda_k}$$

$$\lambda_k = \max_{1 \le k \le T} |\widetilde{C}(n,k)|$$

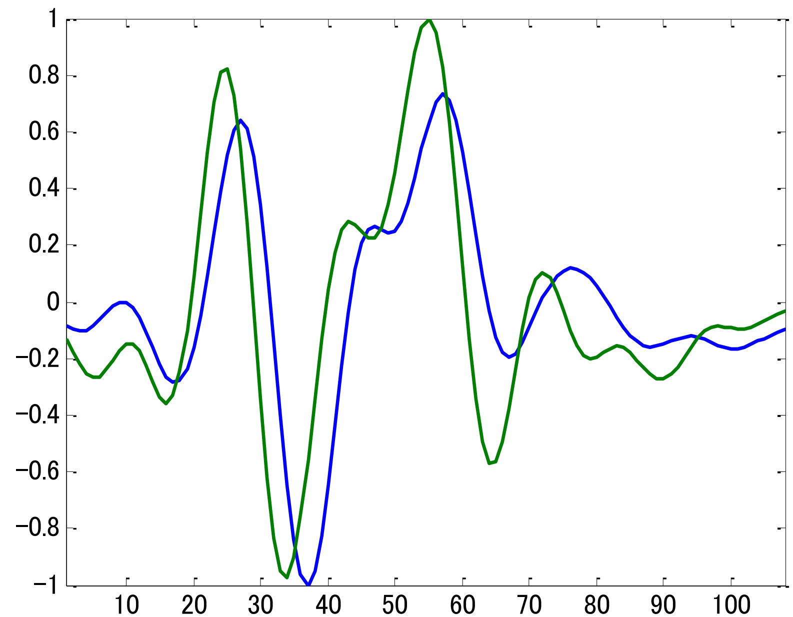
RSF/DRA (2002)

S.Yoshizawa, Y.Miyanaga et. al., "Hardware Implementation of a Noise Robust Speech Recognition System Using RSF/DRA Technique", IEICE Technical Report, CAS2003-42, VLD2003-52, DSP2003-72, pp.127-132, June 2003.

Comparison in cepstral time-trajectories at 4th order



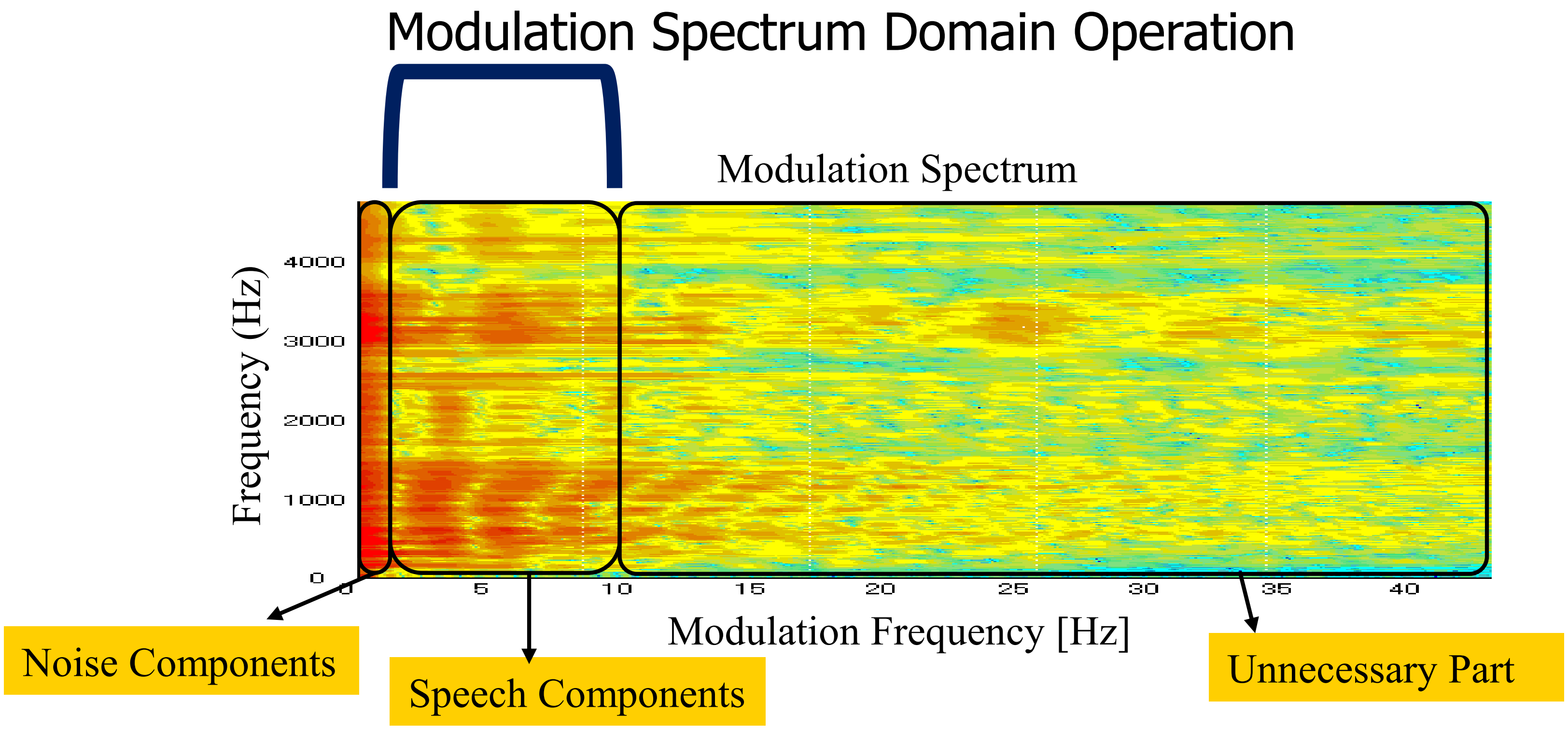
Baseline



RSF/DRA processing

RSA (Running Spectrum Analysis)

Speech components in 0.5 – 7 Hz of the Modulation Spectrum Domain are directly selected by DFT/FFT operation.



Conditions of Robust-ASR

ASR

for **Similar Japanese Pronunciation Phrases**

under Low SNR (10dB, 15dB)

Table 1. RSA passband specifications

RSA Type	LCF	HCF
(a)	1	7
(b)	1	15
(c)	1	35
(d)	1	40
(e)	0.5	7
(f)	0.5	35
(g)	0.1	7
(h)	0.1	35

Table 2. The condition of speech recognition experiments

Parameter name	Parameter value/type
Sampling	11.025 kHz (16-bit)
Frame length	23.2 ms (256 samples)
Shift length	11.6 ms (128 samples)
Pre emphasis	$1-0.97z^{-1}$
Windowing	Hanning window
Speech Feature vectors	$b_i (i = 1, \ldots, 12)$ $\triangle b_i (i = 0, \ldots, 12)$, $\triangle^2 b_i (i = 0, \ldots, 12)$,
Training Set	30 male , 30 female 3 utterances each
Testing Set	10 male, 10 female, 3 utterances each
Acoustic Model	32-states isolated phrase HMMs
Noise varieties	4 types from NOISEX-92 (white,pink, HF radio channel, babble)
SNR	10 dB, 15 dB, 20 dB
Filtering methods	RSF, RSA,

ASR Results using RSA

Table 3. Avg. recog. accur(%) for 100 common male speech

	10 dB	15 dB	20 dB
RSF	72.5	87.6	92.8
RSA:Type(a)	69.3	83.5	88.5
RSA:Type(b)	74.0	87.0	91.3
RSA:Type(c)	76.6	90.1	94.9
RSA:Type(d)	76.5	89.9	94.8
RSA:Type(e)	66.4	81.2	86.5
RSA:Type(f)	72.6	87.2	92.7
RSA:Type(g)	66.9	81.2	86.4
RSA:Type(h)	72.6	87.2	92.7

Table 4. Avg. recog. accur(%) for 6 similar pronunciation male speech

	10 dB	15 dB	20 dB
RSF	58	60	66
RSA:Type(a)	57	61	61
RSA:Type(b)	63	65	71
RSA:Type(c)	65	66	68
RSA:Type(d)	65	66	70
RSA:Type(e)	62	63	67
RSA:Type(f)	69	67	73
RSA:Type(g)	55	56	61
RSA:Type(h)	68	67	73

Table 5. Avg. recog. accur(%) for 100 common female speech

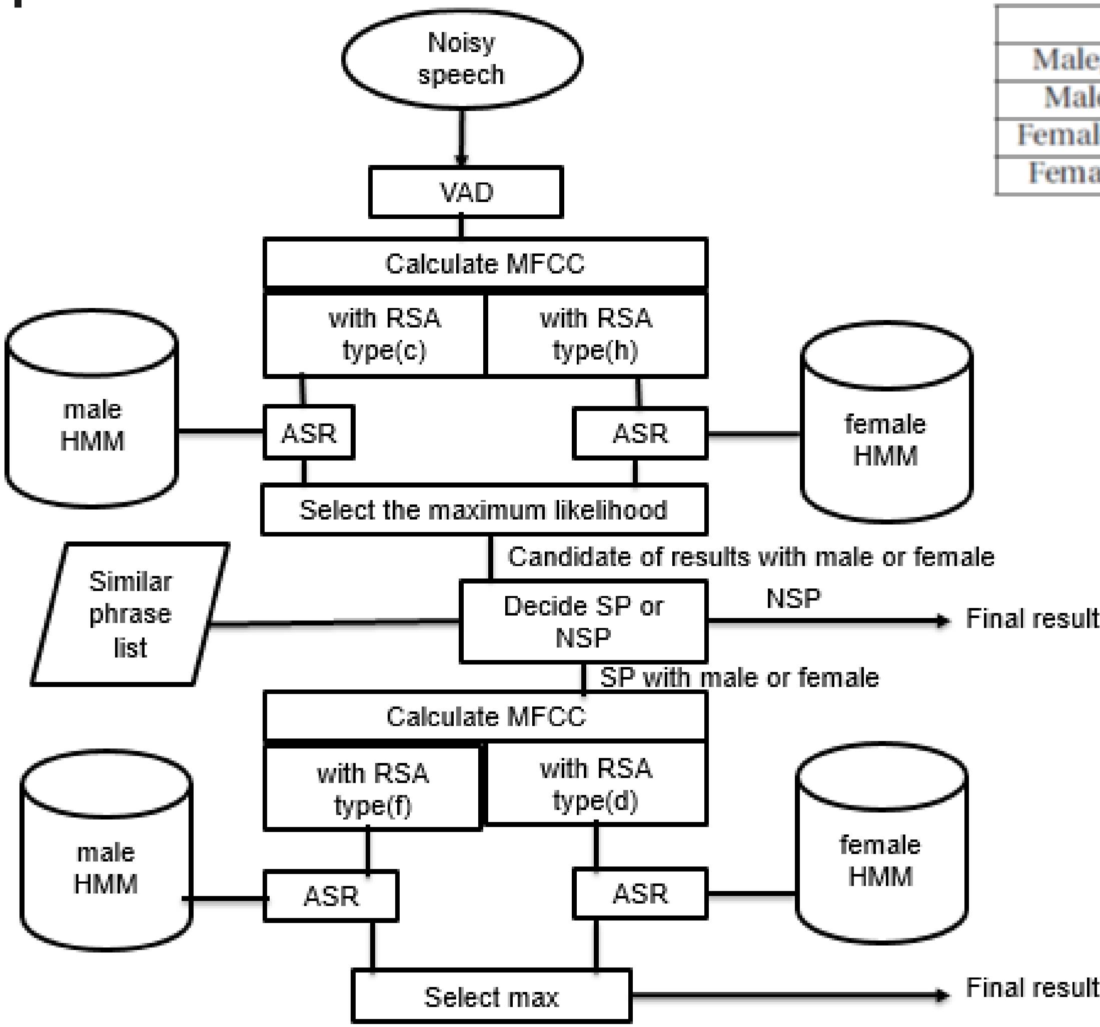
	10 dB	15 dB	20 dB
RSF	56.3	79.9	89.1
RSA:Type(a)	51.5	75.9	84.4
RSA:Type(b)	56.3	80.3	89.4
RSA:Type(c)	55.8	80.8	91.1
RSA:Type(d)	55.3	80.5	91.1
RSA:Type(e)	55.0	80.2	88.2
RSA:Type(f)	57.6	82.3	90.5
RSA:Type(g)	55.5	80.3	88.2
RSA:Type(h)	58.7	82.7	90.5

Table 6. Avg. recog. accur(%) for 6 similar pronunciation female speech

	10 dB	15 dB	20 dB
RSF	55	62	71
RSA:Type(a)	60	67	70
RSA:Type(b)	60	67	70
RSA:Type(c)	62	63	73
RSA:Type(d)	58	66	75
RSA:Type(e)	60	62	69
RSA:Type(f)	57	64	69
RSA:Type(g)	62	62	69
RSA:Type(h)	59	64	68

G. Mufungulwa, Y.Miyanaga, et. al., "Robust Speech Recognition for Similar Japanese Pronunciation Phrases Under Noisy Conditions", ISSCS2017 IEEE, to be presented, Iasi, Romania, July 2017.

Robust ASR



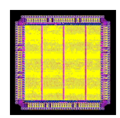
	10 dB	15 dB	20 dB
Male, NSP	4.1	2.5	2.1
Male, SP	11	7	7
Female, NSP	2.4	2.8	2.0
Female, SP	7	4	2

Improvement (%) on ASR Accuracy on NSP and SP

High Speed Eco ASR HW System

Design of Green LSI
lower clock, sub-threshold, parallel/pipeline, dynamic architecture

Definition of Real-time
180ms for speech processing



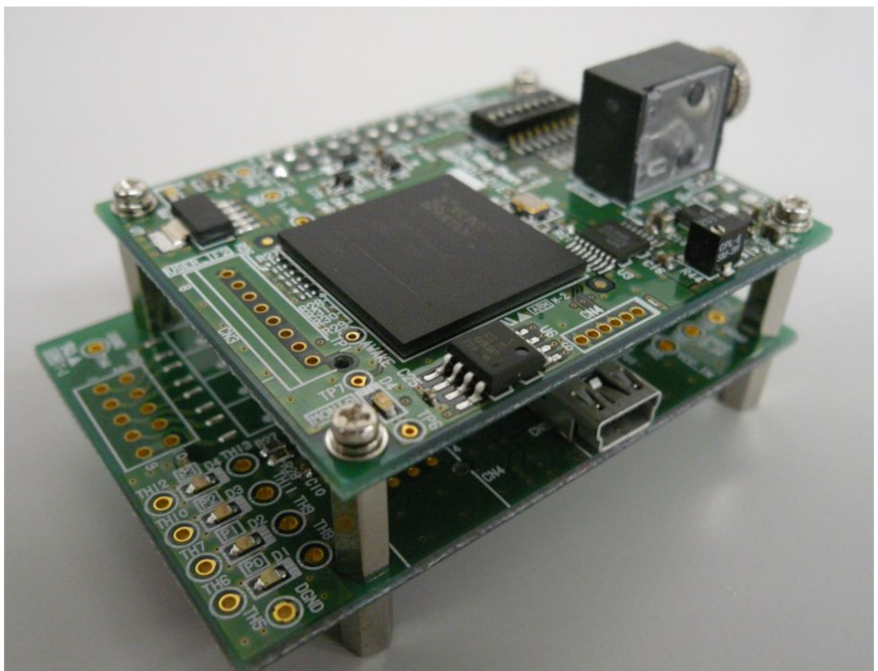
Selection of LSI Design Technology
90nm, 65nm

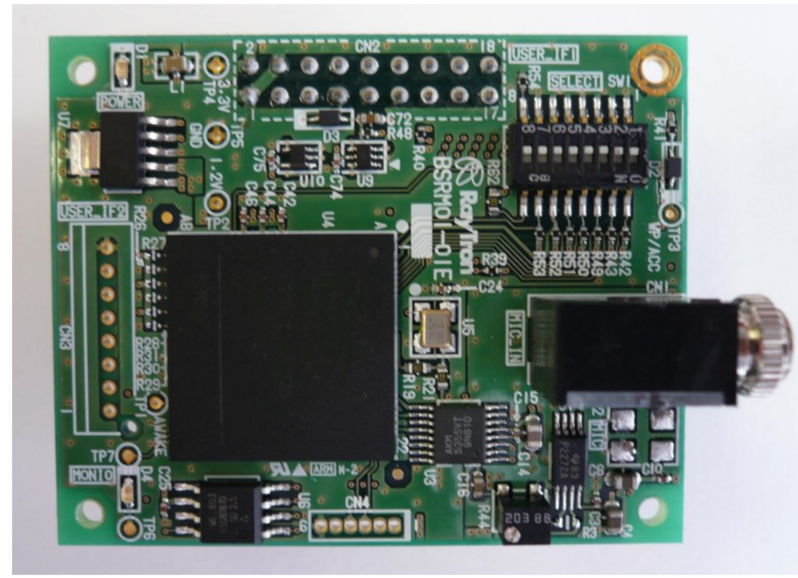
HU Robust ASR v1

S.Yoshizawa, Y.Miyanaga et. al., "A VLSI Implementation of a Word Recognition System for Low-Power Design", IEICE Technical Report, CAS2002-28, VLD2002-42, DSP2002-68, pp.13-18, June 2002.

Current HU Robust ASR v4 (2014)

PC Interface with
HU-ASR Board



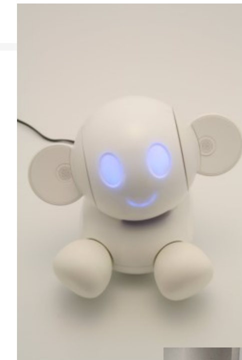


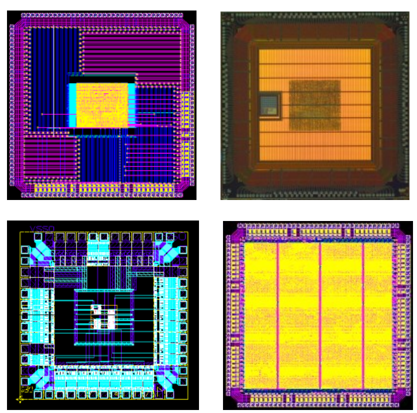
HU-ASR Board

55mm×44 mm

Robot Implementation

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





welfare

speech therapy

Summary

Autonomous ASR

Integrated Architecture of Speech Detection, Robust Speech Analysis, Speech Recognition, Speech Selection
Higher Speed Processing than DSP and Software
Superior in Energy Saving than DSP Solutions